Emerging ethical considerations for the use of artificial intelligence in ophthalmology

Nicholas G. Evans PhD,[1] Danielle M. Wenner PhD,[2] I. Glenn Cohen JD,[3] Duncan Purves PhD,[4] Michael F. Chiang MD,[5] Daniel SW Ting MD PhD,[6] Aaron Y. Lee MD MSCI[7]

[1] Department of Philosophy, University of Massachusetts Lowell, MA
[2] Department of Philosophy and Center for Ethics & Policy, Carnegie Mellon University, PA
[3] Harvard Law School, MA
[4] Department of Philosophy, University of Florida, FL
[5] National Eye Institute, National Institutes of Health, Bethesda, MD
[6] Singapore National Eye Center, Duke-NUS Medical School
[7] Department of Ophthalmology, University of Washington

Running Head: Ethical Considerations for AI in Medicine

Financial Support: NGE is supported by NSF 1734521, the Greenwall Foundation Faculty Scholars Program, the Davis Educational Foundation, and the US Department of Defense Minerva Research Initiative. AYL is supported by NIH/NEI K23EY029246 and an unrestricted grant from Research to Prevent Blindness. DMW is supported by the Greenwall Foundation Faculty Scholars Program. I.G.C. is supported by a grant from the Collaborative Research Program for Biomedical Innovation Law, a scientifically independent collaborative research program supported by a Novo Nordisk Foundation grant (NNF17SA0027784) and by Diagnosing in the Home: The Ethical, Legal, and Regulatory Challenges and Opportunities of Digital Home Health, a grant from the Gordon and Betty Moore Foundation (grant agreement number 9974). DP is supported by NSF 1917712.

The sponsors / funding organizations had no role in the design or conduct of this research.

Competing Interests
A.L. reports support from the US Food and Drug Administration, grants from Santen, Regeneron, Carl Zeiss Meditec, and Novartis, personal fees from Genentech, Topcon, and Verana Health, outside of the submitted work. This article does not reflect the opinions of the Food and Drug Administration. I.G.C. serves as a bioethics consultant for Otsuka on their Abilify MyCite product and for Dawnlight. D. Ting is the patent holder of a deep learning system for retinal diseases, and the equity owner of EyRIS Ptd Ltd. M.C. reports previously receiving personal fees from Novartis, previously receiving grant support from Genentech, the National Institues of Health, and the National Science Foundation, and previously having been an equity owner of InTeleretina, LLC.

Acronyms: Machine Learning (ML); Deep Learning (DL); Artificial Intelligence (AI); Retinopathy of Prematurity (ROP).

Keywords: ethics, artificial intelligence, machine learning, explainability, patient outcomes

Corresponding Author:
Nicholas G. Evans
883 Broadway Street
Dugan 200F
Lowell MA 01853
Ph: 978-934-4996
Email: Nicholas_evans@uml.edu

Emerging ethical considerations for the use of artificial intelligence in medicine

Rapid developments in artificial intelligence (AI) promise improved diagnosis and care for patients, but raise ethical issues.[1–5] Over six months, in consultation with the American Academy of Ophthalmology (AAO) Committee on AI, we analyzed potential ethical concerns, with a focus on applications of AI in ophthalmology that are deployed or will be deployed in the near future.[6] We identified three pressing issues: 1) *transparency*, paradigmatically through the explanation or interpretation of AI models; 2) attribution of *responsibility* issues for particular harms arising from the use or misuse of AI; and 3) *scalability* of use cases and screening infrastructure.

1) Transparency. The ability to understand why a machine learning model has produced a particular result is an oft-cited ethical principle for AI.[4,5,7–10] We distinguish between AI that are *interpretable,* or governed by models that are directly understandable by humans, and AI that are too complex for any human to comprehend (sometimes called "black box" models), requiring *post hoc explainability* for how results are produced.[4] Recent work has shown that lack of transparency is associated with decreased accuracy of AI algorithms.[11,12] Issues of transparency may arise, for example, in diagnosing diabetic retinopathy, glaucoma, age-related macular degeneration and retinopathy of prematurity (ROP).[1]

Transparency may also be important when an AI does not perform as expected or gives a false answer. Given a novel image to analyze, for example, AI may misdiagnose a patient based on an incomplete or inadequate "training set." Machine learning (ML) and especially deep learning (DL) platforms need to be "trained" on large amounts of historical data (e.g. fundus photography) to learn which features of an image are associated with a particular

condition. When a novel image is presented that is atypical, such as if a diabetic retinopathy AI is given a central retinal vein occlusion, the AI may provide false or even nonsense answers. Without transparency it may be impossible to explain why a particular failure occurred: even if the general explanation is that "the training set is insufficiently broad," what data are missing or needed may be opaque.

Transparency is arguably secondary to the capacity for AI to improve patient outcomes and public health. ML systems in ophthalmology have been tested, but to date only one trial has demonstrated improved patient outcomes.[13] Experiences in other specialties, such as a 2017 trial of using automated interpretation of cardiotocographs in labor, have found no improvement in clinical outcomes as a result of AI.[14] Thus, transparency may be insufficient to justify the use of AI if it fails to improve patient outcomes.

The degree to which transparency is obligatory may also depend on the medical specialty. In some cases, accurate, empirically verified results may be sufficient. In infectious disease, for example, broad-spectrum antibiotics may be tried in the absence of detailed information of a pathogen.[8] Ophthalmology, however, is highly explainable in diagnostic terms with strict definitions for most diseases. Deferring to AI may present a significant decrease in confidence in the diagnostic process, especially when there are only modest increases in verifiability. The degree to which this arises, and how this trade-off between transparency and confidence varies by specialty, needs further investigation.

Lack of sufficient transparency may exacerbate other issues in the use of medical AI. While human physicians can reflect upon and justify their actions to colleagues, an AI's mistakes are predetermined through training. Errors may propagate from a single point of failure if they become the diagnostic standard across, say, an entire hospital network. Patients may seek a second opinion, but if an algorithm is widely distributed, they may be diagnosed by the same system at separate clinics. Future AI may be able to revise their predictions in response to new data gained through operation in the real world, but this presents its own challenges, especially if these revisions lack transparency. Excessive trust in AI may be worse for patient outcomes than if AI were approached more skeptically.[15]

Sometimes, the benefits of AI may outweigh transparency concerns. Consider retinopathy of prematurity (ROP), a leading cause of childhood blindness worldwide. The clinical benefit of screening is well-established but hampered globally by cost and human labor requirements.[1] AI may provide a low-cost screening option in resource-scarce settings, where even modest improvements in testing and treatment could have a significant impact given the steep long-term costs of ROP.[16] While challenges translating diagnosis to treatment in low-income settings remain,[17] the large potential benefits and low cost justify the use of AI.

Explainable AI may obviate some of these transparency concerns. Cynthia Rudin has noted, however, that explainability may be a misnomer. Instead, the focus should be on creating models that are inherently interpretable, rather than attempting to generate solutions for unexplainable AI.[8] For the foreseeable future, then, a tension exists between deploying

black-boxed AI immediately or waiting for explainable AI, where delays might come at the cost of improvements to patient outcomes.

2) Responsibility. Ethical frameworks may distinguish between the responsibility for ensuring AI performs in a certain way and the moral or legal liability when harms occur. Here, we only deal with *ethical* responsibilities and not, e.g. legal liability, though these are related issues. In healthcare, a "responsibility gap" arises when responsibility cannot be easily attributed to one or more actor, including hospitals, health and malpractice insurers, individual physicians and nurses, and so on. In ophthalmology, one private company, IDx, has accepted responsibility for errors in their AI, effectively attempting to close the responsibility gap through *claiming responsibility* for AI outcomes, enshrining this in legal terms by purchasing liability and malpractice insurance on behalf of the platform.[5]

Companies are responsible for ensuring AI algorithms function appropriately and safely when used as indicated, but may not be for off-label uses. In their consideration of the legal aspects of AI, for example, IDx claims their principles require creators "assume liability for harm caused by the diagnostic output of the device when used properly and on-label."[5] Responsibility for ensuring appropriate off-label use may thus seem to fall to the provider, but the fragile nature of these models means even strong associations between patient outcomes and off-label AI use post-market may be undermined if subtle changes in patient characteristics cause the algorithm to produce flawed results.[13,21] Whether providers can responsibly determine appropriate use based on these unknown variations is unclear.

Responsibility issues may become more acute in future adaptive AI that update their weightings of factors associated with a diagnosis in response to new data. Here, responsibility for appropriate use might include managing which data is retained by the system. For these adaptive regimes, evaluating performance for on-label and off-label conditions will require continuous post-market monitoring, rather than the current pre-market approval approach for pharmaceuticals or other devices.

Allocating responsibility at the level of governance and regulation is an additional challenge. Others have argued that regulation of AI should focus on continuous monitoring[18] with a "system" view that sees new AI as part of a larger network of actors and institutions and evaluates its performance in the context of that network.[15] The obligation to promote benefits and reduce harms is jointly held by, and distributed between, the creators and users of an AI. Implementing this in practice, however, would require overhauling the institutions that govern medical innovation and practice.

One preliminary approach would require large, adaptive clinical trials of human adjudication versus AI diagnosis. This approach could validate AI performance in a variety of contexts to improve outcomes, adapt to other potential uses, and develop trust in the system. In 2018, engineers at Google demonstrated that image adjudication images by retinal specialists improved algorithmic outcomes for the diagnosis of diabetic retinopathy.[19] In the same year, IDx reported that their autonomous AI-based diagnostic exceeds human reference standards.[13] Last year, two AI-assisted ROP diagnosis packages were approved for use[20] as part of China's developing medical AI landscape.[21] When

specialist opinion can be linked to correct surrogate outcomes or risk of poor outcomes, these trials become an intermediate step towards demonstrating the efficacy of AI, improving patient outcomes, enhancing trust, and providing a broader context for AI use.

3) Scalability and implementation. One promise of AI is to automate high volume screening. Consider a near-future hypothetical. In the United Kingdom the English National Health Service Diabetic Eye Screening Program screened more than 2 million patients in 2015-16 for diabetic retinopathy.[22] We could imagine a case in which this service incorporates AI diagnosis, an implementation that could place most diabetic retinopathy cases in the country under a single algorithm.

Two failure modes exist for mass AI-driven diagnostics. First, standard errors in diagnostics matter at scale: a sensitivity of 99.9% for a test that applies to a condition affecting hundreds of millions of patients still entails hundreds of thousands of false negatives.[2] Importantly, transitioning to AI could redistribute false positives or false negatives in a population. This raises concerns of justice if, for example, AI misdiagnoses disproportionately impact disadvantaged groups, as has occurred with pulse oximeters[23] and x-ray datasets,[24] resulting in a form of "health poverty" where individuals, groups, or populations are unable to benefit from AI due to a scarcity of representative data, and may even be harmed by it at the population level.[25] The degree to which this may occur with ophthalmological AI applications is an empirical question. We do, however, know that racial bias in ophthalmological clinical trials is an ongoing concern,[26] and this trend could continue into AI development if it goes unchecked.

However, the distribution of harms using AI might be traded against the distribution of services through the deployment of AI, such that:

1) Some patients have worse outcomes than others because of the distribution of risk by AI; yet
2) Those patients have better outcomes than they would otherwise have had because
    a. the AI is ultimately less biased than physician treatment alone; or
    b. the benefits of access to services outweigh the potential harms of bias; or
    c. both.

Consider the proliferation of telemedicine during the COVID-19 pandemic, particularly for individuals who may have otherwise delayed diagnosis or treatment.[27,28] AI-assisted diagnostics could make it easier to diagnose patients remotely and at local points of care using e.g. new innovations such as slit-lamp biomicroscopes used with smartphones[29] and AI-based interpretation of results. A potential tradeoff arises between errors caused by AI when a physician cannot directly access the patient, and benefits of receiving early diagnosis. In rare or emergent cases (such as pandemic) where the risk of travel to a medical facility presents additional risk, AI may provide preliminary guidance on whether or not to seek care inside a clinical setting.[30] Moreover, even if AI does produce worse outcomes than physician diagnosis, AI might be justified to the extent delayed or missed diagnosis is worse.

The social benefit of AI to telemedicine relies in part, however, on the extent to which inequalities of access to information technology can be remedied. Telemedicine is unevenly adopted by providers, may not be supported by insurers, and depends on reliable internet access. Smartphone penetration, however, may be higher than access to specialist medical care in some if not many areas, and thus there may be favorable tradeoffs through local AI-driven solutions. Like other emerging technologies, the setting in which medical AI will be implemented is a major determinant of the risks and benefits.

A second failure mode is a systemic failure that affects all or most users simultaneously. These very low probability, very high consequence events could arise, for example, in the case of a continual learning AI system intended to improve with additional data[31] but which through sustained machine error ultimately diverges radically from its original parameters and begins assigning false results. Depending on how submissions to the AI are structured, "adversarial uses" could arise in which intentionally doctored images are submitted to achieve the same effect.

Protection from systemic failures is unlikely to be achieved through self-governance, and will require regulatory action to guard against. Adding ongoing cybersecurity and fault tree testing to the approval requirements is one solution, but two challenges arise. First pre-market regulation does typically entail continuous monitoring of the system; study of results by human analysts; and quality control tests against the algorithm to prevent system failures may become dysfunctional on a large-scale level. Second, the FDA only regulates medical devices, of which IDx is one, but some AI (such the Apple Watch pulse

oximeter) may constitute a "general wellness product" designed to be sold directly to consumers.[32] Addressing both challenges might reduce the possibility of low probability/high consequence events, but represent tradeoffs in system efficiency and resource use around AI in medicine.

In response, the FDA and similar agencies in other countries might require reform to accommodate the challenges presented by AI. Alternatively, the mismatch between the current regulatory structure and the potential impacts of AI in medicine might mean that the FDA is ultimately not well-suited for regulating AI. In the latter case, a new agency may be required, or governance could occur through a different mechanism entirely, e.g. through government payment choices in national health insurance schemes.

AI presents a range of novel opportunities to improve medical care and to make healthcare more widely accessible to patients. However, the use of AI raises many ethical concerns, even in cases where it augments the capabilities of human physicians and technicians. These issues are partly endogenous to AI, and partly a function of the regulatory, social, and political circumstances in which it is developed and implemented. Realizing the full benefits of AI will require reaching a consensus on which tradeoffs are acceptable as this technology is implemented at scale.

Acknowledgements

The authors would like to thank the American Academy of Ophthalmology (AAO) Task Force on AI, Dr. Ron Pelton, and the AAO Ethics Committee for their review and comments on this manuscript.

References

1. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*. 2019;103(2):167-175. doi:10.1136/bjophthalmol-2018-313173

2. Lee A. Machine diagnosis. *Nature*. Published online April 10, 2019. doi:10.1038/d41586-019-01112-x

3. Lin D, Lin H. Translating artificial intelligence into clinical practice. *Ann Transl Med*. 2020;8(11). doi:10.21037/atm.2019.11.110

4. Emanuel EJ, Wachter RM. Artificial Intelligence in Health Care: Will the Value Match the Hype? *JAMA*. 2019;321(23):2281-2282. doi:10.1001/jama.2019.4914

5. Abràmoff MD, Tobey D, Char DS. Lessons Learned About Autonomous AI: Finding a Safe, Efficacious, and Ethical Path Through the Development Process. *American Journal of Ophthalmology*. 2020;214:134-142. doi:10.1016/j.ajo.2020.02.022

6. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*. 2019;1(9):389-399. doi:10.1038/s42256-019-0088-2

7. London AJ. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*. 2019;49(1):15-21. doi:10.1002/hast.973

8. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-215. doi:10.1038/s42256-019-0048-x

9. Gunning D. Explainable Artificial Intelligence (XAI). Presented at: DARPA/I20 Program Update; November 2017; Washington DC.

10. Hoffman RR, Klein G, Mueller ST. Explaining Explanation For "Explainable Ai." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 2018;62(1):197-201. doi:10.1177/1541931218621047

11. Floridi L, Cowls J, Monica Beltrametti, et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds & Machines*. 2018;28(4):689-707. doi:10.1007/s11023-018-9482-5

12. Shah MP, Merchant SN, Awate SP. Abnormality detection using deep neural networks with robust quasi-norm autoencoding and semi-supervised learning. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. ; 2018:568-572. doi:10.1109/ISBI.2018.8363640

13. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Medicine*. 2018;1(1):1-8. doi:10.1038/s41746-018-0040-6

14. Brocklehurst P, Field D, Greene K, et al. Computerised interpretation of fetal heart rate during labour (INFANT): a randomised controlled trial. *The Lancet*. 2017;389(10080):1719-1729. doi:10.1016/S0140-6736(17)30568-8

15. Gerke S, Babic B, Evgeniou T, Cohen IG. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *npj Digit Med*. 2020;3(1):1-4. doi:10.1038/s41746-020-0262-2

16. Frick KD, Mashfeghi DM. A Case for Universal Eye Screening. Retina Today. Published 2015. Accessed June 2, 2020. http://retinatoday.com/2015/06/a-case-for-universal-eye-screening/

17. Brown R, Evans NG. The social value of candidate HIV cures: actualism versus possibilism. *Journal of Medical Ethics*. Published online July 2016:medethics-2015-103125. doi:10.1136/medethics-2015-103125

18. Babic B, Gerke S, Evgeniou T, Cohen IG. Algorithms on regulatory lockdown in medicine. *Science*. 2019;366(6470):1202-1204. doi:10.1126/science.aay9547

19. Beede E, Baylor E, Hersch F, et al. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Association for Computing Machinery; 2020:1-12. doi:10.1145/3313831.3376718

20. New progress in AI research and development! Two Tangwang Fundus Image Aided Diagnosis Software Approved for Market. Accessed August 23, 2020. http://www.cnpharm.com/c/2020-08-11/748522.shtml

21. Li R, Yang Y, Wu S, et al. Using artificial intelligence to improve medical services in China. *Ann Transl Med*. 2020;8(11):711. doi:10.21037/atm.2019.11.108

22. Scanlon PH. The English National Screening Programme for diabetic retinopathy 2003–2016. *Acta Diabetol*. 2017;54(6):515-525. doi:10.1007/s00592-017-0974-1

23. Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS. Racial Bias in Pulse Oximetry Measurement. *New England Journal of Medicine*. 2020;383(25):2477-2478. doi:10.1056/NEJMc2029240

24. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *PNAS*. 2020;117(23):12592-12594. doi:10.1073/pnas.1919012117

25. Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. *The Lancet Digital Health*. 2021;3(4):e260-e265. doi:10.1016/S2589-7500(20)30317-4

26. Coney JM. Racial Bias in Clinical Trials: What You Need to Know. Retina Today. Published March 2021. Accessed July 5, 2021. https://retinatoday.com/articles/2021-mar/racial-bias-in-clinical-trials-what-you-need-to-know

27. Evans NG. The Ethics of Social Distancing. *The Philosopher's Magazine*. 2020;89(2):96-103.

28. Saleem SM, Pasquale LR, Sidoti PA, Tsai JC. Virtual Ophthalmology: Telemedicine in a COVID-19 Era. *Am J Ophthalmol*. Published online April 30, 2020. doi:10.1016/j.ajo.2020.04.029

29. Spaide T, Wu Y, Yanagihara RT, et al. Using Deep Learning to Automate Goldmann Applanation Tonometry Readings. *Ophthalmology*. 2020;127(11):1498-1506. doi:10.1016/j.ophtha.2020.04.033

30. Evans NG, Sekkarie MA. Allocating scarce medical resources during armed conflict: ethical issues. *Disaster and Military Medicine*. 2017;3(1):5. doi:10.1186/s40696-017-0033-z

31. Lee CS, Lee AY. Clinical applications of continual learning machine learning. *The Lancet Digital Health*. 2020;2(6):e279-e281. doi:10.1016/S2589-7500(20)30102-3

32. Minssen T, Gerke S, Aboy M, Price N, Cohen G. Regulatory responses to medical machine learning. *Journal of Law and the Biosciences*. 2020;7(1). doi:10.1093/jlb/lsaa002