The Artificial Intelligence Doctor: Considerations for the Clinical Implementation of Ethical AI

29

Julius M. Kernbach, Karlijn Hakvoort, Jonas Ort, Hans Clusmann, Georg Neuloh, and Daniel Delev

29.1 Introduction

The field of medicine has become increasingly data-driven, with artificial intelligence (AI) and machine learning (ML) attracting much interest across disciplines [1–4]. While the implementation in patient care still lags behind, almost every type of clinician is predicted to use some form of AI technology in the foreseeable future [3]. Evolving with the industrialization of AI, where the academic and industrial boundaries of AI-associated research are increasingly blurred, the number of ML-based algorithms developed for clinical and commercial application within health care is continuously increasing. Realizing the accompanying rising ethical concerns, many institutions, governments, and companies alike have since formulated sets of rules and principles to inform research and guide the implementation into clinical care [5]. More than 80 policies on "Ethical AI" have since been proposed [6], including popular examples such as the European Commission's AI strategy [7], the UK's Royal College of Physicians' Task Force Report [8], the AI Now Institute's Report [9], as well as statements from major influences form the industry (e.g., Google, Amazon, IBM) [10]. Collectively, there appears to be a widespread agreement between the distinct proposals regarding meta-level aims, including the use of AI for the common good, preventing harm while upholding people's rights, and following widely-respected values of privacy, fairness, and autonomy. Demonstrating considerable overlap, the suggested pillars building Ethical AI converge to the principles of autonomy, beneficence, non-maleficence, justice and fairness, privacy, responsibility, and transparency [6]. While certain principles, generally describing the four bioethical principles of autonomy, beneficence, non-maleficence, and justice, are well-known in healthcare, AI-specific concerns arise regarding the autonomy, accountability, and need of explicability of AI-based systems.

Until now, there are relatively few neurosurgical papers implementing AI. However, the recent trend demonstrates the growing interest in ML and AI in neurosurgery [11, 12]. From a clinician's point of view, AI can be untransparent, and without methodological foundations, pose a severe risk to patients' care. How can we make AI transparent for clinicians and patients? How do we choose which clinical decisions are going to be delegated to AI? How do we prevent adverse events caused by AI algorithms? When the AI agent makes wrong decisions—who can be held responsible? There is a clear increase of directives and papers on AI ethics [6, 10] offering guidelines to these critical questions. This article non-exhaustively covers basic practical guidelines regarding AI-specific ethical aspects that will be useful for every ML or AI researcher, author, and reviewer aiming to ensure ethical innovation in AI-based medical research.

29.2 Transparency and Explicability

Research in AI systems rapidly advances across medical disciplines; however, the trust placed in developed applications lags behind [13]. Many proposals on ethical AI guidelines acknowledge the lack of algorithmic transparency and accountability as the most prevalent problems to address [6]. As humans and responsible clinicians, we must understand and interpret the outcome of an AI or ML model. With the European Union being at the forefront of shaping the

Julius M. Kernbach and Karlijn Hakvoort have contributed equally to this work.

J. M. Kernbach (✉) · K. Hakvoort · J. Ort · D. Delev
Neurosurgical Artificial Intelligence Laboratory Aachen (NAILA), RWTH Aachen University Hospital, Aachen, Germany

Department of Neurosurgery, Faculty of Medicine, RWTH Aachen University, Aachen, Germany
e-mail: jkernbach@ukaachen.de

H. Clusmann · G. Neuloh
Department of Neurosurgery, Faculty of Medicine, RWTH Aachen University, Aachen, Germany

V. E. Staartjes et al. (eds.), *Machine Learning in Clinical Neuroscience*, Acta Neurochirurgica Supplement 134,
https://doi.org/10.1007/978-3-030-85292-4_29

international debate on Ethical AI, the General Data Protection Regulation (GDPR) was introduced in 2018. Herein, articles 13–14 mandates "meaningful information about the logic involved" for all decisions made by artificially intelligent systems [14]. This *right to an explanation* of the directive implies that any clinician using AI-based decision-making is legally bound to convey patients with explanations to the applied ML and AI models' inner workings. Suppose the AI-based decision cannot be explained. In that case, the clinician ends up in the uncomfortable position of vouching for the application's trustworthiness without being able to interpret its methodology and outcome. Unfortunately, many ML and AI models are considered "black boxes" that do not explain their predictions in a comprehensible way. The consequent lack of transparency and explicability of predictive models in medicine can have severe consequences [15, 16].

The precise lack of interpretability has been exacerbated with the rise and popularity of deep learning (DL) models. As a form of representation learning with multiple layers of abstraction, DL methods are extremely good at discovering intricate patterns in high-dimensional data [17, 18] that are beyond the human scope of perception. DL methods have produced promising results in speech recognition, visual object recognition, object detection, and many other domains such as drug discovery and genomics. They frequently outperformed different ML algorithms in image recognition and computer vision [19–21], speech recognition [22, 23] and more. DL methods, including deep neural networks, are increasingly complex and challenging—if not impossible—to interpret because the function relating the input data through multiple complex layers of neurons to the final outcome vector is far too complex to comprehend. Fortunately, in the spirit of "Explainable AI" [24–26], approaches have been developed to address the black box problem. Broadly, Explainable AI involves creating a second (post hoc) model to explain the first black box model [26]. Successful analytical approaches to "open the black box" have since been proposed. One example are local interpretable model-agnostic explanations (LIME), which can explain the predictions of a classifier in a comprehensible manner by learning an interpretable model locally around the prediction [27]. Other implementations primarily rely on assessing variable importance, such as RISE (Randomized Input Sampling for Explanation), which probes deep image classification modes with a randomly masked version of the input image [28]. However, particularly in the clinical context, evidence to whether post hoc approximations can adequately explain deep models remains very limited [27, 29, 30].

With the increasing success of AI and, in particular, DL, a "myth of accuracy-interpretability trade-off" arise, meaning that complicated deep models are necessary for excellent predictive performance [26]. However, more complex models are often not more accurate, particularly when the data are structured with a good representation in terms of naturally meaningful features. In DL, the inherent complexity scales to large datasets [17, 31]. Particularly successful examples of employed DL include studies on electronic health records, as demonstrated by Rajkomar and colleagues in >200,000 adult patients cumulating a total of >46.8 billion data points [32], and large prospective population cohort studies of >500,000 participants from the UK Biobank [33]. But even in the big-data omics fields, such as imaging or genomics, investigations in part question the superiority of DL compared to simple models based on available data. Schulz and colleagues showed that the increase in performance of linear models in brain imaging does not saturate at the limit of current data availability, and DL is not beneficial at the currently exploitable sample sizes such as those based on the UK Biobank (>10,000 3D multimodal brain images [34]. In the prediction of genomic phenotypes, DL performance was competitive to linear models but did not outperform linear models by a sizable margin (>100,000 participants with >500,000 features) [35]. Historically, linear models have long dominated data analysis, as complex transformations into rich high-dimensional spaces were computationally infeasible. In small sample sizes particularly, complex methods with high variance such as many DL methods tend to overfit: the algorithm performs "too well" on training data to the extent that it negatively impacts the interpretation of new data. Less complex models such as general linear models are generally less prone to overfitting—especially with regularization strategies applied [36, 37].

The best practice recommendations on predictive modeling hence include considerations of the given structure on the input data, the choice of feature engineering, sample size and model complexity, and more [38–40] and should always be considered when selecting the appropriate models for a given predictive modeling task.

29.3 Fairness and Bias

There is global agreement that AI should be fair and just [6]. Herein, unfairness relates explicitly to the effect of unwanted *bias* and *discrimination*. While biased decision-making is hardly unique to AI and ML, research demonstrated that ML models tend to amplify societal bias in the available training data [41, 42]. Skewed training data is a major influence on bias amplification and can lead to severe adverse events arising from the lack of inclusion of ethical minorities. Esteva and colleagues used DL to identify skin cancer from photographs using 129,450 images (with only 5% of dark-skinned participants). While the classification works en par with expert knowledge on light skin, it fails to diagnose melanoma in people with dark skin colors [3, 43]. This highlights

the importance of deliberate data acquisition that is representable and diverse (e.g., regarding race, gender), focusing on including minorities. Many of the ML applications available today can be considered "narrow AI," that is, they help with specific tasks on specific types of data. An AI system trained on a certain patient cohort cannot unconsciously be used on an entirely different population. Therefore, the limits of generalizability should always be kept in mind. However, even in balanced data sets, bias may be amplified due to spurious (mostly unlabeled) correlations. For example, in a balanced picture data set of 50% men cooking and 50% women cooking, unlabeled influences, e.g., children, which co-occur more often with women, can be labeled cooking as well. Hence, more women will be associated with cooking [30]. To counteract unwanted bias in balanced data sets, adversarial debiasing was proposed [30, 44, 45]. Models are trained adversarially to preserve task-specific information while eliminating, e.g., gender-specific cues in images. The removal of features associated with the protected variable (gender, ethnicity, age, or others) within the intermediate representation leads to less biased predictions in balanced data sets. Protected variables include gender, race, and socioeconomic status. Failure to address the societal bias could ultimately widen the present gap in health outcome [3, 46].

We welcome increasing diversity within a research group itself, which increases detection of possible (unconsciousness) biases. Nowadays, diversity is an important factor in obtaining European and national research funding [47]. For every AI application, it should clearly be outlined which patient characteristics within training were available. An extensive table with patient characteristics, including sex, age, ethical background, length, weight, and BMI, as well as detailed disease information should be included. Major sources of bias should be described within the limitation section as well. It is important to realize that most biases are unintended and do not arise deliberately. Despite attempts to reduce biases, these can occur when not expected at all.

29.4 Liability and Legal Implications

While the important ethical issues mentioned above are still a matter of intensive and critical debate, the first steps toward structured and transparent software legalization using ML have been successfully made. The Medical Device Regulation (MDR, EU Regulation 2017/745) is an essential step toward better software use regulation, aiming at improved safety and transparency. MDR and the Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745, which was endorsed by the Medical Device Coordination Group (MDCG), accurately address the definition of software. Herein, software is regarded as a medical device, meaning that medical device software (MDSW) is any software that is intended to be used alone or in combination for any purpose mentioned by the definition of medical device, i.e., used for diagnostic, prevention, prediction, prognosis or treatment of a disease (for a full report, c.f. to the EU 2017/745). MDSW can be independent and still qualifies as such regardless of its physical localization (i.e., cloud).

Furthermore, the MDR defines software as a set of instructions that processes input data and creates output data. Thus, MDR encompasses to a full extend any use of AI technology. One needs to look more precisely at the decision steps assisting the qualification as MDSW. Here, one will unmistakably find that if the software is not acting for the individual patient's benefit, it is not covered by the MDR. A more critical interpretation of this part could suggest that software or AI technology, which is not used in a clinical setup, is not considered by the MDR. This is indeed the usual case when AI technology is used in an experimental and scientific setting. However, in this setting, any discoveries or assistance by the AI technology should not be directly used to influence patients' diagnostics or treatment. In the case of IBM Watson's AI for Oncology program [15], the developed algorithm for the recommendation of treatment choices for patients with cancer frequently suggested harmful and erroneous treatment regimes. If the harmful algorithm were to be integrated into the actual clinical routine, many patients would have suffered preventable harm. Compared to errors on the single doctor-patient level, the faulty AI recommender would have inflicted harm on an exponentially higher level. Following this line of thought and embracing the ethical axiom of "primam non nocere," one can argue that any software, AI technology, or ML algorithm, which is intended to be used for clinical decision-making of any kind, needs to be CE or FDA approved. Although this is inevitably associated with considerable effort, it will guarantee that every software life cycle will include all the steps of paramount importance, such as hazard management and quality management. Although the software does not directly harm a patient, it still can create harmful situations by providing incorrect information. This gap has been successfully addressed by the Rule 11 of the MDR. Consequently, many software applications (including AI, ML, and statistical tools like risk calculators) will fall into Class IIa or Class IIb. Indeed, all these regulating measures may seem less progressive. Still, they try to solve the legal question of liability by introducing terms as the *intended purpose* and the use outside of it.

One further problem in AI liability is that the law, including tort law, "is built on legal doctrines that are focused on human conduct, which when applied to AI, may not function" [48]. Moreover, until now, there is no clear legal definition of AI that can be used as a foundation for new laws regarding its use since existing definitions were created to understand AI instead of regulating it. The legal definitions are, therefore, often circular and/or subjective [49].

Additionally, *adopting* AI applications that might influence clinical decision-making may "evolve dynamically in ways that are at times unforeseen by system designers" [50]. With adaptation, the AI system gains *autonomy*. But our definition of what is considered autonomous or intelligent is still ill-defined and will likely change over time due to rapid developments within the field of AI [49].

Until AI definitions and regulations are clearly defined, care should be warranted to use AI-assisted tools. Clinical decision-making algorithms could be allocated to research purposes only, which demands the approval of an ethical commission, patient insurance, and patients' consent before its use. AI has already been proven very helpful—especially in making diagnoses and predicting prognosis and outcome—also within the field of neurosurgery [11]. In the end, every outcome from an AI algorithm should be checked against the current medical gold-standard and clinical guidelines. For future considerations, the development of concise AI definitions and regulations is relevant to deflect potential harm.

29.5 Conclusion

With the continuously advancing field of AI, fostering trust in the clinical implementation of AI applications becomes imperative. Almost every type of clinician is predicted to use some form of AI technology in the foreseeable future, hence, shaping the ethical and regulatory use of AI becomes increasingly important. In the article, we reviewed *transparency and algorithmic explicability* as the trade-off between complexity and available data, the *mitigation of unwanted biases* that even affect balanced data sets, and the *legal considerations* when advancing AI in health care. We introduce approaches, including post hoc models and adversarial attacks, to combat the above problems and foster Ethical AI.

Acknowledgments and Disclosure *Funding:* J. M. K. and D. D. are supported by the Bundesministerium für Bildung und Forschung (BMBF COMPLS3–022).

Conflicts of Interest/Competing Interests None of the authors has any conflict of interest to disclose.

References

1. Deo RC. Machine learning in medicine. Circulation. 2015;132(20):1920–30. https://doi.org/10.1161/CIRCULATIONAHA.115.001593.
2. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science. 2015;349:255–60. https://doi.org/10.1126/science.aaa8415.
3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56.
4. Vellido A. Societal issues concerning the application of artificial intelligence in medicine. Kidney Dis. 2019;5(1):11–7.
5. Whittlestone J, Alexandrova A, Nyrup R, Cave S. The role and limits of principles in AI ethics: towards a focus on tensions. In: AIES 2019—Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society; 2019. p. 195–200.
6. Jobin A, Ienca M, Vayena E. Artificial intelligence: the global landscape of ethics guidelines. arXiv; 2019.
7. European Commision. Artificial intelligence: commission takes forward its work on ethics guidelines; 2019.
8. Reznick RK, Harris K, Horsley T. Artificial intelligence (AI) and emerging digital technologies; 2020.
9. Crawford K, Dobbe R, Dryer T, et al. AI now 2019 report. New York: AI Now Institute; 2019.
10. Floridi L. Establishing the rules for building trustworthy AI. Nat Mach Intell. 2019;1(6):261–2.
11. Bonsanto MM, Tronnier VM. Artificial intelligence in neurosurgery. Chirurg. 2020;91(3):229–34.
12. Senders JT, Arnaout O, Karhade AV, Dasenbrock HH, Gormley WB, Broekman ML, Smith TR. Natural and artificial intelligence in neurosurgery: a systematic review. Clin Neurosurg. 2018;83(2):181–92.
13. Dreiseitl S, Binder M. Do physicians value decision support? A look at the effect of decision support systems on physician opinion. Artif Intell Med. 2005;33(1):25–30. https://doi.org/10.1016/j.artmed.2004.07.007.
14. Goodman B, Flaxman S. European Union regulations on algorithmic decision making and a "right to explanation". AI Mag. 2017;38(3):50–7. https://doi.org/10.1609/aimag.v38i3.2741.
15. Ross C, Swetlitz I. IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. Stat+; 2018.
16. Varshney KR, Alemzadeh H. On the safety of machine learning: cyber-physical systems, decision sciences, and data products. Big Data. 2016;5:246–55.
17. Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning. Cambridge, MA: The MIT Press; 2016.
18. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436.
19. Djuric U, Zadeh G, Aldape K, Diamandis P. Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. NPJ Precis Oncol. 2017;1:22. https://doi.org/10.1038/s41698-017-0022-1.
20. Farabet C, Couprie C, Najman L, Lecun Y. Learning hierarchical features for scene labeling. IEEE Trans Pattern Anal Mach Intell. 2013;35(8):1915–29. https://doi.org/10.1109/TPAMI.2012.231.
21. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Adv Neural Inf Proces Syst. 2012. https://doi.org/10.1061/(ASCE)GT.1943-5606.0001284.
22. Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Process Mag. 2012;29(6):82–97. https://doi.org/10.1109/MSP.2012.2205597.
23. Mikolov T, Deoras A, Povey D, Burget L, Černocký J. Strategies for training large scale neural network language models. In: 2011 IEEE workshop on automatic speech recognition and understanding, ASRU 2011, PRO; 2011. https://doi.org/10.1109/ASRU.2011.6163930.
24. Albers DJ, Levine ME, Stuart A, Mamykina L, Gluckman B, Hripcsak G. Mechanistic machine learning: how data assimilation leverages physiologic knowledge using Bayesian inference to forecast the future, infer the present, and phenotype. J Am Med Inform Assoc. 2018;25(10):1392–401. https://doi.org/10.1093/jamia/ocy106.

25. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng. 2018;2(10):749–60. https://doi.org/10.1038/s41551-018-0304-0.
26. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206–15. https://doi.org/10.1038/s42256-019-0048-x.
27. Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. https://doi.org/10.1145/2939672.2939778.
28. Petsiuk V, Das A, Saenko K. RISE: randomized input sampling for explanation of black-box models. arXiv; 2018.
29. Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell. 2019;1(11):501–7.
30. Wang T, Zhao J, Yatskar M, Chang KW, Ordonez V. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In: Proceedings of the IEEE International Conference on Computer Vision 2019-October; 2019. p. 5309–18.
31. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24–9. https://doi.org/10.1038/s41591-018-0316-z.
32. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning for EHR—supplement. NPJ Digit Med. 2018;1:18.
33. Weng SF, Vaz L, Qureshi N, Kai J. Prediction of premature all-cause mortality: a prospective general population cohort study comparing machine-learning and standard epidemiological approaches. PLoS One. 2019;14:e0214365. https://doi.org/10.1371/journal.pone.0214365.
34. Schulz MA, Thomas Yeo BT, Vogelstein JT, Mourao-Miranada J, Kather JN, Kording K, Richards B, Bzdok D. Deep learning for brains?: different linear and nonlinear scaling in UK biobank brain images vs. machine-learning datasets. In: bioRxiv; 2019. https://doi.org/10.1101/757054.
35. Bellot P, de los Campos G, Pérez-Enciso M. Can deep learning improve genomic prediction of complex human traits? Genetics. 2018;210(3):809–19. https://doi.org/10.1534/genetics.118.301298.
36. Hastie T, Tibshirani R, Friedman J. Springer series in statistics the elements of statistical learning—data mining, inference, and prediction. Berlin: Springer; 2009.
37. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Curr Med Chem. 2000. https://doi.org/10.1007/978-1-4614-7138-7.
38. Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2013. https://doi.org/10.1007/978-1-4614-6849-3.
39. Neeman T. Clinical prediction models: a practical approach to development, validation, and updating by Ewout W. Steyerberg. Int Stat Rev. 2009;77(2):320–1. https://doi.org/10.1111/j.1751-5823.2009.00085_22.x.
40. Poldrack RA, Huckins G, Varoquaux G. Establishment of best practices for evidence for prediction: a review. JAMA Psychiat. 2020;77(5):534–40. https://doi.org/10.1001/jamapsychiatry.2019.3671.
41. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: Advances in neural information processing systems; 2016.
42. Yao S, Huang B. Beyond parity: fairness objectives for collaborative filtering. In: Advances in neural information processing systems; 2017.
43. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8.
44. Xie Q, Dai Z, Du Y, Hovy E, Neubig G. Controllable invariance through adversarial feature learning. In: Advances in neural information processing systems; 2017.
45. Zhang BH, Lemoine B, Mitchell M. Mitigating unwanted biases with adversarial learning. In: AIES 2018—Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society; 2018. https://doi.org/10.1145/3278721.3278779.
46. Stringhini S, Carmeli C, Jokela M, et al. Socioeconomic status and the 25 × 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1.7 million men and women. Lancet. 2017;389(10075):1229–37.
47. European Commission E. Work Programme 2018–2020: Science with and for society; 2018.
48. Bathaee Y. The artificial intelligence black box and the failure of intent and causation. Harv J Law Technol. 2018;31(2):889–936.
49. Buiten MC. Towards intelligent regulation of artificial intelligence. Eur J Risk Regul. 2019;10(1):41–59.
50. Gasser U, Almeida VAF. A layered model for AI governance. IEEE Internet Comput. 2017;21(6):58–62.