VIEWPOINT



Advancing health equity with artificial intelligence

Nicole M. Thomasian^{1,2} · Carsten Eickhoff^{3,4} · Eli Y. Adashi¹

Accepted: 14 November 2021 © The Author(s), under exclusive licence to Springer Nature Limited 2021

Abstract

Population and public health are in the midst of an artificial intelligence revolution capable of radically altering existing models of care delivery and practice. Just as AI seeks to mirror human cognition through its data-driven analytics, it can also reflect the biases present in our collective conscience. In this Viewpoint, we use past and counterfactual examples to illustrate the sequelae of unmitigated bias in health-care artificial intelligence. Past examples indicate that if the benefits of emerging AI technologies are to be realized, consensus around the regulation of algorithmic bias at the policy level is needed to ensure their ethical integration into the health system. This paper puts forth regulatory strategies for uprooting bias in healthcare AI that can inform ongoing efforts to establish a framework for federal oversight. We highlight three overarching oversight principles in bias mitigation that maps to each phase of the algorithm life cycle.

Keywords Health equity \cdot Artificial intelligence \cdot Machine learning \cdot Health policy \cdot Algorithmic bias

Introduction

Artificial intelligence (AI) continues to feature prominently in the health sector by way of its ever-growing contribution to population and public health practice, management, and surveillance [1, 2]. However, just as AI seeks to mirror human cognition through its data-driven analytics, it can also reflect the biases present in our collective conscience. Indeed, a number of algorithms guiding population medicine

Nicole M. Thomasian nicole_thomasian@brown.edu

¹ Warren Alpert Medical School of Brown University, Brown University, 222 Richmond Street, Providence, RI 02906, USA

² The Harvard Kennedy School of Government, Harvard University, Cambridge, MA, USA

³ Center for Biomedical Informatics, Brown University, Providence, RI, USA

⁴ Department of Computer Science, Brown University, Providence, RI, USA

initiatives, federal healthcare reimbursements, and clinical management have been found to discriminate against protected social groups [3–6]. In this *Viewpoint*, we argue that federal authorities should implement a regulatory framework for healthcare AI that incorporates standards to ensure health equity. To support this argument, we present past and counterfactual examples of racially biased algorithms alongside their respective mitigation strategies. It should be noted, however that many of the principles described herein with respect to race will also apply to other social identifiers such as sex, income bracket, etc. The overarching objective of this paper is, therefore, to raise awareness around the important topic of bias in AI among health experts and policymakers to guide future oversight directions.

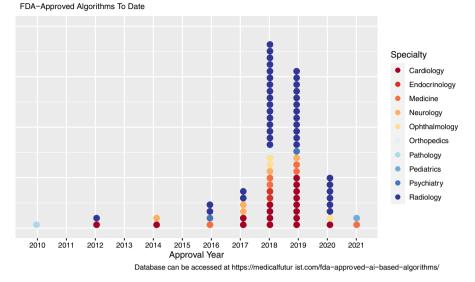
Background

Highly visible examples of biased computation in healthcare, such as race corrections for glomerular filtration rate and pulmonary function, continue to endure despite decades of efforts to eradicate them [5]. It could be even more difficult to undo algorithmic bias in AI that, due to model complexity and varying degrees of artificial intelligence literacy, may not be readily apparent to the user. This idea of concealed discrimination is not new to the health sector and can occur regardless of intent. Implicit bias contributes to delays in melanoma diagnoses among dark-skinned patients that translate to worse outcomes [7]. Another example would be well-documented disparities in the administration of pain medication on the basis of race [8].

In terms of public health at large, this pattern of obfuscation as a means of reinventing racism can be illustrated using the concept of the 'submerged state.' The submerged state makes reference to the use of concealed welfare in the form of federal disbursements or tax credit programs as an under-recognized means of reinforcing the racial wealth gap [9]. A recent example impacting public health systems in the US would be the appropriations of the Coronavirus Aid, Relief, and Economic Security (CARES) Act funds using a formula based on lost healthcare facility revenue, a miscalibration of need such that hospitals in communities of color remain inadequately reimbursed [4]. Our goal is to avoid the case where unfettered applications of artificial intelligence in healthcare become a nidus for bias 'submerged' within the complexity of AI models.

Federal governance

In January 2021, the US Food and Drug Administration (FDA) indicated its intent to develop a regulatory guidance for AI in their 'Artificial Intelligence/Machine Learning-Based Action Plan,' an update that follows an earlier FDA discussion paper in April 2019 [10]. As of September 2021, however, the FDA has already approved a total of 77 artificial intelligence applications under the 'Software as Medical Devices' (SaMDs) classification for use in clinical settings (see Fig. 1). Pursuant



Artifical Intelligence-based Software as Medical Devices

Fig. 1 FDA-Approved Artificial Intelligence-based Algorithms as of September 2021

to section 520(o) of the 21st Century Cures Act, SaMD refers to software that is intended to primarily drive clinical decision making or to analyze patient health data or medical images [11]. It should be noted that a large proportion of artificial intelligence algorithms are exempted by this definition and are already in widespread use throughout the health sector. Therefore, while the establishment of a federal guidance for clinical AI that incorporates health equity considerations would set a much-needed precedent for fairness in medical informatics, there is no guarantee that such a framework would trickle down to the FDA-exempt majority of healthcare AI that still influence resource allocation, access to public health services, and medical care. Limitations in federal authority make it incumbent on developers, public health practitioners, and policymakers to also assist in shifting the status quo of bias mitigation as a desirable addition to an unequivocal inclusion.

A lifecycle approach to regulation

'Artificial intelligence' refers to computational models that automate tasks typically performed by humans, and this umbrella term encompasses machine learning algorithms (See Box 1) [12]. From a regulatory perspective, it is critical for policymakers to understand that bias mitigation should not end with AI model development but, rather, extend across the product lifecycle. The AI lifecycle moves from model development to validation to implementation, with maintenance and updates also possible in the post-implementation period. Bias can enter at any point during the AI lifecycle. Below, we illustrate key topics in algorithmic bias with strategies to address ongoing barriers that should be incorporated into a federal legal guidance

Box 1 Glossary of key terms	
Term	Definition
Artificial intelligence (AI)	An umbrella term referring to computational technologies that automate tasks typically performed by humans
Machine learning	A subset of AI that refers to models that can learn from examples without the explicit programming of rules
Healthcare AI	An umbrella term referring to AI for use in the health sector (i.e., disease surveillance, diagnostics and treatment, resource allocation, delivery of health services, workflow, etc.)
Protected group	Groups that face discrimination due to a shared social characteristic that are protected under the federal legal code (i.e., race, gender, age, ability, etc.)
Algorithmic bias	An algorithm's performance, allocation, or outcome for a protected social group puts them at a (dis-)advantage with respect to the unprotected social group
Health equity	The ability of all patients to attain their full health potential is the same across all groups [36]
Development	Creation of the model: a process that encompasses data pre-processing, model training/validation/testing efforts
Validation (regulatory)	Assessment of model performance prior to its formal implementation
Implementation	Integration of the AI model into the healthcare setting for real-world use
Maintenance	Updates made to the AI model after it is in real-world use to assure a contin- ued high-quality performance
Training	A process where the model learns trends or categories from data
Validation (model)	A process that confirms the generality of the trained model and explores different hyperparameter choices
Testing	A process that evaluates model performance on an unseen dataset
Pre-training	A process that trains a model on a large, non-specific dataset prior to subse- quent fine-tuning on the actual dataset to improve overall performance
Federated learning	Each institution trains a model using their home data and the model weights are communicated to a centralized server to develop an aggregate model; there is no sharing of protected health information
Cyclic weight transfer	An institution trains a model using their home data and passes the updated model weights to the next institution, the process repeats until all institu- tions have participated; there is no sharing of protected health information
Bias accounting	The process of measuring bias, when applicable to the algorithm's intended use case
Bias mitigation	The process of correcting for bias, when applicable to the algorithm's intended use case
Positive predictive value	The likelihood that if you screen positive that you actually have the disease
Negative predictive value	The likelihood that if you screen negative that you actually do not have the disease
Equalized odds	No difference in sensitivity and specificity across all groups
Predictive parity	No difference in positive predictive value rates across all groups
Demographic parity	No difference in positive outcome rates across all groups
Validation (AI lifecycle)	Evaluation of model performance prior to formal implementation
Interpretability	The degree to which the decision process of AI is understandable to humans
Continuously learning AI	AI that can update in real-time to learn from incoming data

Box 1 Glossary of key terms

for advancing health equity in AI. We also highlight a key tension: how can we account for racial disparities in care access and quality while avoiding the use of the social indicator of race to draw inferences about intrinsic human biology? The bias mitigation principles described herein are illustrated using broad strokes, as they are intended as a general guide for public health practitioners and policymakers. A glossary of key terms can be found in Box 1.

Development

Approaches to bias mitigation in the development phase address data- or model-specific factors. Below, we describe strategies on both sides of this dichotomy, providing real-world examples for context.

Data factors

Developers should consider how limitations in data quality can threaten model performance with respect to race. At a minimum, data should include patients from various racial backgrounds in cases where failure to account for race is linked to known disparities in care. This principle can be challenging in practice, however, particularly when working with rare diseases. Take acromegaly, for example, an endocrine pathology that results from excess growth hormone production and classically presents with characteristic changes in facial bone structure. Facial recognition software with AI is currently being explored for early detection of acromegaly, but studies to-date have been performed in exclusively white or asian populations [13–15]. The incidence of acromegaly does not change with respect to race, and early diagnosis is critical to enhance patient outcomes and quality of life [16]. This situation begs the question: how can we improve the availability of racially diverse data in order to promote health equity?

Creation of curated, open databases with deidentified or aggregate patient information is one excellent solution to combat racially imbalanced data [2]. However, this option might not always be feasible for data sharing in facial and skin images, geographically linked, or environmental exposure data due to privacy concerns [17]. Another option would be to pretrain models with large, non-specific datasets or with synthetic data generation to boost the model's ability to recognize a diversity of cases prior to training with the original data (See Box 1) [15]. Finally, collaborative training techniques such as federated learning and cyclic weight transfer are just two examples of options to boost data availability, as they can allow for model training across multiple institutions without transfer of patient data (See Box 1) [18, 19]. To promote transparency, developers should document the distributions of patient characteristics for race in aggregate, and they should explain their reasoning for implementing or not implementing techniques to improve the model's performance with respect to race. Finally, developers should think about and convey how any subjective or missing health data, such as entries in the electronic medical record, for example, may be coded with preexisting human biases, and how this may affect their model predictions with respect to race. For example, consider a population medicine algorithm that identifies patients who might benefit from a specialized community chronic pain clinic and associated physical rehabilitation program. If this model uses subjective criteria such as provider-reported pain scores, this could bias against people of color who are generally perceived as having less pain than their white counterparts [20]. This impact of data reliability on model performance should also be subsequently assessed both during model evaluation and implementation using bias-auditing techniques.

Model factors

Next, we will discuss how bias can be engineered into models by their design. Failure to account for race in contexts where known disparities are relevant to the algorithm's intended use case has been repeatedly shown to entrench bias [3, 21, 22]. Consider, for example, an algorithm that enrolls patients into high-risk care management programs using a risk score [3]. One such widely used, race-blind model was found to systematically under enroll black patients, who were sicker than their white counterparts for a given risk score [3]. The increased enrollment threshold for black patients to gain access to these beneficial programs was due to the effect of proxy variables, namely healthcare costs, which were lower in black participants. This relationship is believed to be owing to a long-standing lack of trust in the healthcare system and disparities in health access among black Americans. Both of these observations are well described in the literature and the latter finding is also consistent with the itemized distribution of healthcare expenses across groups in the study [3, 23]. By accounting for race through adjustment of the data label choice, the authors were able to correct the algorithm so that placement was adjudicated fairly [3].

The above example illustrates how failing to assess how race and other aspects of social identity are treated by the model can compromise health equity, and it brings us to our discussion of bias accounting. It is important to note that there are many metrics that can be used for bias accounting during development (i.e., equalized odds, predictive parity, demographic parity, etc.) (See Box 1). Bias metric(s) should be carefully selected based on the intended use of a specific algorithm as they are not universally compatible [21]. Take, for example, an algorithm that is in compliance with predictive parity such that all groups are selected at an identical positive predictive value rate. That same model could dually defy the principle of equalized odds such that sensitivity and specificity differ across groups. We can revisit our earlier example of implicit bias in the diagnosis of skin cancer to illustrate this principle. Let us consider a hypothetical health surveillance image recognition algorithm where patients can take a picture of their mole(s) using a Smartphone App to see if they should seek further evaluation to rule out malignancy. We ensure that this model is in compliance with predictive parity but is later found to disproportionately underpredict dark-skinned patients with malignant melanomas as having benign lesions of low concern. How is this possible? One interpretation lies in our failure to account for false-negative rates



across groups. False negatives represent instances where patients with a malignant lesion are falsely reassured, disparities in which implicit bias is thought to contribute to and that we want to avoid replicating in our model [7]. Therefore, a metric that limits discrepancies in false-negative rates across groups such as equalized odds might be preferred for this algorithm's intended use case. As this process will look differently for each algorithm, it is critical for developers to provide well-documented evidence and reasoning to justify tradeoffs in chosen bias mitigation techniques [3, 21]. Ideally this information would be made publicly available in a deidentified format whenever possible, but should, at a minimum, be reviewed by those overseeing implementation of the algorithm prior to its use. Bias adjustments are to be made in cases where the social impacts of race are clearly linked to access, diagnosis, or treatment of your healthcare question, **not to adjust for or to explain biological variation.**

Validation

Model validation on retrospective data alone is not rigorous enough to account for biases that can emerge in real-world conditions. Prospective studies in healthcare settings, ideally in the form of a randomized trial, should be federally mandated for AI applications with the potential to cause a patient bodily harm or death. To navigate the tension between optimizing model performance and safety without diminishing incentives to develop new AI, it may be beneficial to develop a tiered-risk system linked to specific validation criteria. Regulators can look towards AI-reporting recommendations in academia as a reference when formulating validation criteria, as the research community has recently developed a number of checklists designed to quantify the robustness of artificial intelligence studies [24, 25]. Again, as many AI applications in population and public health will be exempt from federal review, we would encourage those overseeing implementation at their institution or organization to play an active role in upholding rigorous validation criteria for algorithms prior to any large-scale rollouts.

Model interpretability is an important step in validation that also doubles as a bias mitigation strategy by providing explanations about the 'inner-workings' of an AI model in a way that is understandable to humans. Briefly, interpretable AI can be achieved using 'model-specific' methods that are constrained to a certain model type or broadly adaptable 'model-agnostic' techniques [26] (see Box 1). Interpretability pipelines can highlight important model logic or features at two levels: global methods assess population-level performance, whereas local explanations reveal the reasoning underlying a specific model prediction instance. Both approaches to interpretability can be useful depending on the intended use case and both are becoming increasingly available and computationally affordable [26–29]. Developers can use interpretability to detect bias during the validation stage because it can highlight when the model is (1) using race when it should not be or (2) not using race when it should be in its evaluation of a given dataset or input.

Implementation

Interpretability is also critical in controlling for human factors that can manifest during the implementation phase. Users of healthcare AI may, quite reasonably, find it hard to trust a model without understanding how it works. This resistance can have the untoward effect of preventing the realization of the equitable outcomes that a well-built algorithm was designed to achieve. We feel that the healthcare field must move towards interpretability as a standard prerequisite for all models slated for real-world implementation, with explanations from developers for cases in which their implementation is not feasible or reliable. In the future, it may be possible to leverage artificial intelligence to automate interpretability pipelines to streamline and promote their use by developers [30].

Interpretability is only part of the story when it comes to the discussion of human factors during algorithm implementation [31, 32] Developers should attempt to understand why population and public health practitioners choose to use or to ignore model recommendations and develop controls for these factors whenever possible. For example, cumbersome user interface, user fatigue, or organizational constraints such as ties to financial reimbursement or liability could cause humans to ignore the advice of an unbiased algorithm. Therefore, it is important for those overseeing implementation to measure user uptake and qualitative user experience (the latter should also take place during development) in addition to outcomes data during post-market re-appraisals.

Next, continual bias auditing and surveillance are a key arbiter of equity in the implementation phase. Performance and bias checks should be a shared responsibility among healthcare AI developers and end organizations, and they should be incorporated into algorithm tuning and maintenance at regular intervals. Checks may need to occur more or less frequently depending on the automatic update permissions of the algorithm in the real-world setting and the risk to health and safety. Regulators should work with involved stakeholders to develop a system that does not use overly burdensome reporting requirements that would discourage the use of continuously learning AI that can update in real time [33] (see Box 1). A solution may also be baked into this problem, as there is a potential to leverage AI for future automation of this bias screening process to promote the uptake and frequency of surveillance. We would also advocate for the creation of AI task forces at the federal, state, and local public health levels to spearhead the continuous review of models for bias and broader quality control in the post-implementation period.

In a related point, the transparency and stewardship of AI algorithms are currently stymied by a general lack of inventory, and this should be made a priority moving forward. For example, as of September 1st 2021, FDA-approved artificial intelligence algorithms are currently indexed as approved medical devices without a distinct search filter or database, which has prompted the creation of thirdparty databases for tracking (see Fig. 1) [34, 35]. Since many AI applications are in use in an auxiliary decision support, allocation, or workflow capacity not requiring FDA approval, it is critical for population and public health organizations to maintain their own dedicated inventories for bias surveillance purposes. While the examples outlined in this paper discuss race consciousness and racism in healthcare AI, many of these principles will also translate to other aspects of patient social identity such as sex, religion, and income brackets. Further research is needed to determine how best to pick up on bias related to intersectionality and to non-discrete identifiers (such as gender, age, lifestyle, ability, and employment sector that are harder to classify, stratify over, and rectify. Still, we envision AI as a powerful force for advancing health equity that, if applied with the proper controls, can mitigate persistent inequalities that plague our healthcare through fair and unbiased evaluation. Looking ahead, we must develop harmonized standards for health equity in AI and make every effort to uphold them at the federal, industry, academic, and community levels.

Funding Professor Adashi is serving as Co-Chair of the Safety Advisory Board of Ohana Biosciences, Inc. Professor Eickhoff is a co-founder and shareholder of codiag AG and x-cardiac GmbH.

References

- 1. Ting DSW, Carin L, Dzau V, Wong TY. Digital technology and COVID-19. Nat Med. 2020;26(4):459-61.
- Bai HX, Thomasian NM. RICORD: a precedent for open AI in COVID-19 image analytics. Radiology. 2021;299:E219.
- Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–53.
- Kakani P, Chandra A, Mullainathan S, Obermeyer Z. Allocation of COVID-19 relief funding to disproportionately Black counties. JAMA. 2020;324(10):1000–3.
- Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight—Reconsidering the use of race correction in clinical algorithms. N Engl J Med. 2020;383(9):874–82.
- Grobman WA, Lai Y, Landon MB, Spong CY, Leveno KJ, Rouse DJ, et al. Development of a nomogram for prediction of vaginal birth after cesarean delivery. Obstetr Gynecol. 2007;109(4):806–12.
- Adamson AS, Smith A. Machine learning and health care disparities in dermatology. JAMA Dermatol. 2018;154(11):1247–8.
- Tamayo-Sarver JH, Hinze SW, Cydulka RK, Baker DW. Racial and ethnic disparities in emergency department analgesic prescription. Am J Public Health. 2003;93(12):2067–73.
- Mettler S. The submerged state: how invisible government policies undermine American democracy. Chicago: University of Chicago Press; 2011.
- US Food and Drug Administration. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan January 2021. https://www.fda.gov/media/145022/ download.
- 11. 21 U.S.C. § 520(o).
- 12. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24–9.
- Schneider HJ, Kosilek RP, Günther M, Roemmler J, Stalla GK, Sievers C, et al. A novel approach to the detection of acromegaly: accuracy of diagnosis by automatic face classification. J Clin Endocrinol Metab. 2011;96(7):2074–80.
- Kong X, Gong S, Su L, Howard N, Kong Y. Automatic detection of acromegaly from facial photographs using machine learning methods. EBioMedicine. 2018;27:94–102.
- 15. Wei R, Jiang C, Gao J, Xu P, Zhang D, Sun Z, et al. Deep-learning approach to automatic identification of facial anomalies in endocrine disorders. Neuroendocrinology. 2020;110(5):328–37.
- 16. AlDallal S. Acromegaly: a challenging condition to diagnose. Int J Gen Med. 2018;11:337–43.
- 17. Mhasawade V, Zhao Y, Chunara R. Machine learning and algorithmic fairness in public and population health. Nat Mach Intell. 2021;3(8):659–66.

- Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep. 2020;10(1):12598.
- Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. J Healthc Inform Res. 2020;5:1–19.
- Sabin JA, Greenwald AG. The influence of implicit bias on treatment recommendations for 4 common pediatric conditions: pain, urinary tract infection, attention deficit hyperactivity disorder, and asthma. Am J Public Health. 2012;102(5):988–95.
- 21. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med. 2018;169(12):866–72.
- 22. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R, editors. Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference; 2012: ACM.
- 23. Muhammad KG. The condemnation of blackness: race, crime, and the making of modern urban America. Cambridge: Harvard University Press; 2010.
- Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Chan A-W, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26(9):1364–74.
- Mongan J, Moy L, Kahn CE Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol: Artif Intell. 2020;2(2):e200029.
- Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. Wiley Interdiscip Rev: Data Min Knowl Discov. 2020;10(5):e1379.
- Ribeiro MT, Singh S, Guestrin C, editors. 'Why should i trust you?' Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016.
- Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D, editors. Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision; 2017.
- 29. Reyes M, Meier R, Pereira S, Silva CA, Dahlweid F-M, Tengg-Kobligk HV, et al. On the interpretability of artificial intelligence in radiology: challenges and opportunities. Radiol: Artif Intell. 2020;2(3):e190043.
- Kovalerchuk B, Neuhaus N, editors. Toward Efficient Automation of Interpretable Machine Learning. 2018 IEEE International Conference on Big Data (Big Data); 2018 10-13 Dec. 2018.
- Friedman B, Nissenbaum H. Bias in computer systems: association for computing machinery; 1996. 330–47 p.
- 32. Sujan M, Furniss D, Grundy K, Grundy H, Nelson D, Elliott M, et al. Human factors challenges for the safe use of artificial intelligence in patient care. BMJ Health Care Inform. 2019;26(1):e100081.
- 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):53.
- 34. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med. 2020;3(1):118.
- American College of Radiology Data Science Institute. FDA Cleared AI Algorithms. https://models.acrdsi.org/.
- 36. Center for Disease Control. https://www.cdc.gov/chronicdisease/healthequity/index.htm. Accessed 11 Mar 2020.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.