How US law will evaluate artificial intelligence for covid-19

Daniel E Ho and colleagues explore the legal implications of using artificial intelligence in the response to covid-19 and call for more robust evaluation frameworks

Numerous proposals, prototypes, and models have emerged for using artificial intelligence (AI) and machine learning to predict individual risk related to covid-19. In the United States, for instance, the Department of Veterans Affairs uses individualised risk scores to allocate medical resources to people with covid-19,[1] and prisons have sought to detect symptoms by processing inmates' phone calls.[2] Further tools, such as vulnerability predictions for individuals[3] and voice based detection of infection,[4] are on the horizon. But use of AI for such purposes has given rise to questions about legality.

When a state or federal government seeks to use AI models to predict an individual's risk of covid-19, the key legal questions will ultimately turn on how effective the models are and how much they burden legal interests. We focus on two of the most salient legal concerns under US law: privacy and discrimination. Challenges on privacy or discrimination grounds might appear in a variety of contexts, including challenges to regulatory decisions, tort actions, or lawsuits under health privacy laws. We argue that the basic need to balance benefits against burdens runs through all of these legal regimes. Governments implementing risk scoring tools must show that their tools produce valid, reliable predictions and burden individuals' civil liberties no more than necessary. In evaluating the legality of public health use of algorithms, courts will likely also probe how the output of these tools is used to shape policies and programs. But showing that a model performs well and does not exceedingly burden privacy and other interests are essential preconditions for lawful deployment.

KEY MESSAGES

- A proliferation of models using AI and machine learning are in use or under development to predict individuals' covid-19 related risk
- The use of personally identifiable information, including race, raises legal concerns over privacy and anti-discrimination, which we illustrate in the context of US law
- The underlying legal principles are essentially an assessment of effectiveness and burdens of AI and machine learning tools
- More robust evaluation of AI and machine learning tools will be necessary to support the adoption and legality of rapidly proliferating tools

Governing legal principles

Privacy law

Government intrudes on privacy when it forces people to reveal what they reasonably expect will be shielded from public view.[5] Health information is an archetypical example. The health information privacy provisions of the federal Health Insurance Portability and Accountability Act (HIPAA) restrict disclosures and uses of identifiable health information on the basis that such disclosures necessarily harm patients' dignity. However, across the many legal regimes regulating privacy, the fact that the government harms privacy is not enough to establish that it has violated a person's privacy rights. Rather, governments may violate privacy only when the volume of personal information intruded on is disproportionate to achieving the government's purpose.[5] The risk of a privacy violation is especially great when the government intrudes into a space so intricately connected to a person's identity that the intrusion "depersonalize[s] and dehumanize[s]."[6] Health information privacy is commonly held to safeguard "personal dignity" and "[protect] patients from embarrassment, stigma, and discrimination."[7]

Antidiscrimination law

US antidiscrimination law consists of two basic doctrines. Firstly, disparate treatment (or intentional discrimination) occurs when an actor treats individuals differently because they are members of a protected class, such as a racial minority group.[8] When the actor engages in disparate treatment it must offer a justification, with the strength of the rationale calibrated to the protected class.

Secondly, disparate impact occurs when an actor takes a facially neutral action that differentially burdens a protected class. Disparate impact doctrine applies only in circumstances outlined by statute, such as employment, healthcare institutions with federal funding, and housing.[8] In these domains, regulated parties may not use any tool that, even unintentionally, results in disparate outcomes, unless justified by "business necessity."[8]

The use of AI to combat covid-19 potentially raises both antidiscrimination and privacy concerns. Data hungry algorithms can pose privacy challenges by incorporating voluminous and intimate personal information. Models commonly use gender and race, potentially running afoul of disparate treatment,[3 8] and risk scores may vary systematically across such demographic groups.[9]

Evaluating harms and trade-offs

Effectiveness

What legal standards must a government meet for deploying a machine learning application? The stringency of a court's inquiry will vary depending on the legal claim. In a constitutional claim of disparate racial treatment, for example, the government's use of an AI tool would receive strict scrutiny. Policymakers would first have to show that an important government interest is at stake—an easy argument in the covid-19 context. Next, the government's action would have to show that the policy is sufficiently well tailored to serving that interest. In practice, this distils to two questions: Is the policy likely to advance the government's objective (effectiveness)? And are there alternative ways of achieving that objective that are less burdensome on individual interests (burden)?

Other claims might be analysed under more deferential legal regimes. In nearly all cases, however, accurately quantifying

the effectiveness and burdens of AI models is central. For instance, federal courts can strike down the actions of administrative agencies if those actions are deemed "arbitrary and capricious." A health agency would need to provide a reasoned explanation for a model and provide the evidence considered in its appraisal.

Courts and policymakers are often poorly equipped to make such assessments; table 1 summarises key aspects. The problem is not merely a lack of technical competence. It is that courts and policymakers seeking to assess model performance will have to wade into the AI/machine learning field's replication crisis.[10] The lack of incentives for robust evaluation is exacerbated in machine learning by model complexity, data volume, and computational demands. This has resulted in influential models performing worse in practice than originally reported. For instance, a machine learning based risk score for whether a covid-19 patient requires rapid response was found to have "limited value to guide clinical decision-making" for most patients—but only after it was deployed to over 100 US hospitals.[11] Similarly, epidemiological models—some of which are based on machine learning[12]—have been shown to produce unreliable forecasts of actual covid-19 infection rates—but only after their adoption.[13] Given the challenges facing even experts, evaluating risk scoring models poses a daunting task for policymakers and courts.

As governments turn to increasingly powerful tools focused on individualised predictions, the potential to harm privacy and antidiscrimination rights will grow. Proper evaluation will be at the heart of whether new tools appropriately trade off effectiveness and burdens.

Privacy

How would the government sustain the legality of an AI tool despite its privacy risks? While it is more problematic for the government to possess information closer to the core of a person's identity, the government's interest may be so weighty that even the most personal information can be seized and potentially used. For example, protected health information is deemed highly sensitive under the Health Insurance Portability and Accountability Act and is thus normatively bound up with dignitary concerns. But the act allows the release of protected health information to prevent imminent risks to public health or safety.

Weighing these considerations is a complex endeavour. Firstly, protecting privacy may degrade the accuracy of AI or machine learning models.[14] Secondly, in circumstances as severe as a pandemic, failure to deploy effective AI or machine learning tools can itself lead to dignitary harm by neglecting tools that could help control a pandemic more effectively. Privacy protections may then pose not just an accuracy-burden trade-off, but a trade-off between burdens: sheltering for extended periods of time because of an ineffective tool may cause greater indignity than having protected health information revealed. Thirdly, privacy protections can themselves have disparate impact, degrading accuracy more for minority groups than majority groups.[15] Essentially, the government faces a highly complex trade-off between effectiveness, dignity, and equality.

The leading technical framework for navigating these competing concerns is differential privacy, which works by adding random noise to aggregate statistics to prevent inferences about private individual attributes. In this framework, policymakers can directly select the extent of privacy protection by setting how much any individual's data influence aggregate statistics. The technical fix, however, should not obscure the need for trade-offs. Policymakers may need to offer as much justification for sacrificing privacy as for prioritising it.

Bias

Assessment of bias for covid-19 risk scoring may seem more straightforward. Disparate treatment may well prohibit the use of a protected attribute (eg, race) to generate risk scores. Disparate impact may occur when the use of the risk score leads to decisions (eg, preventing someone from going to work) that affect racial groups differentially.

But implementing these divergent metrics simultaneously is another matter entirely, raising profound questions of structural sources of bias. For example, evidence suggests differences in the risk of contracting and succumbing to covid-19 between African-American and Caucasian patients.[9] This creates a catch-22 situation: a model that is blind to protected attributes, like race, may be more likely to produce risk scores correlated with those protected attributes. Deploying such a model to, say, determine which employees could return to work could produce disparate impact. Technical solutions suggest adjusting the machine learning process to conform to fairness constraints, such as the requirement that outcomes be independent of group status, conditional on the model's risk score.[16] But calibrating risk scores by race raises important constitutional concerns, as government classifications based on race are deemed particularly noxious.

Collecting a wider range of socioeconomic predictors may eliminate the need for race variables in models, but increasing the volume of data collected may be infeasible or aggravate concerns about privacy. Just as with differential privacy then, the pure engineering solution of imposing one fairness definition, given conflicting effects, cannot solve the underlying value trade-offs.

The high likelihood of the differential impact of risk models makes it all the more critical for policymakers to insist on reliable evidence as to the efficacy. Knowing that machine learning tools may well engage in disparate treatment or cause disparate impact means that policymakers must be prepared to show that such tools are necessary to achieve public health goals, and to establish quantitatively the difference in efficacy between models that impose potential discrimination harms.

Table 1 | Main dimensions of effectiveness and burdens of artificial intelligence and machine learning systems

Concern	Example of failure
Effectiveness	
Accuracy	Hospitalisation risk model fails to predict actual hospitalisations
Replicability	Feature selection for risk model cannot be replicated
Generalisability	Risk models works in one hospital, but not another
Explainability	Outputs of the risk model cannot be explained easily to a human user, limiting take-up rate by decision makers
Burden	
Bias	Risk model performs well only on one demographic group on which it is trained
Privacy	Risk model uses and/or discloses sensitive information about individual
Due care/process	Individualised patient assessment is compromised owing to risk score

Towards an evaluation framework

The deployment of AI in the fight against covid-19 is an important moment for algorithmic governance. There is an abundance of models and a shortage of coordinated and consistent standards and evaluation. To give government uses of AI or machine learning the strongest prospects of pass-

ing legal muster, we spell out elements of a robust evaluation framework that deals with effectiveness and burdens.

Firstly, the framework must be transparent to provide a basis for evaluation: source code, learned parameters, and base data should be released to the extent allowed by privacy concerns. Secondly, evaluation should be independent of model development, ideally conducted at arm's length. Thirdly, evaluation methods and metrics must be thorough, robust, and interoperable, tackling performance across demographic groups and privacy fairness trade-offs. Lastly, interoperability permits evaluation results to be compiled in a single location, enabling decision makers to assess models efficiently, while not impeding multiple, decentralised innovation efforts.

One model framework is the National Institute of Standards and Technology (NIST) evaluation of facial recognition technology.[17] NIST enables any algorithm to be submitted and tested for accuracy, bias, and a range of other criteria using standardised tests and benchmark data. Specific methods for validating covid-19 models have been examined by researchers,[11 12] but have fallen short on assessment of burdens. Certainly, the NIST framework is not perfect, particularly as facial recognition is deployed to a much wider range of domains not represented in the NIST data. However, NIST provides a good example of the kind of robust evaluative framework that the law may ultimately demand. After development and evaluation through this framework, deployments can then also be evaluated through adaptive trials to assess operational performance in the human context of deployment.[18 19]

To illustrate this, consider the application to the Veterans Affairs' adoption of a risk scoring model for patients in hospital with covid-19.[1 11] Under the framework laid out above, competing vendors would submit their models in a standardised format to an independent party, such as NIST or an academic clearing house. The third party would run each vendor's tool on a hold-out set of data, providing an authoritative audit of the benefits and burdens they offer. For example, the agency might audit the degree to which a given tool's performance depends on access to invasive data or the extent to which it scores protected subgroups differently. Ideally, the evaluator's process would be a stable, interoperable pipeline—much like NIST's facial recognition evaluator—such that the assessment process is not resource intensive.

Such an evaluation protocol will not only help AI applications survive legal challenges, but also cultivate public trust in a highly contentious time for AI governance and public health.

Contributors and sources: DEH teaches administrative and antidiscrimination law and is associate director of the Stanford's Institute for Human-Centered Artificial Intelligence. MK and PH are graduate students focusing on privacy, administrative law, and machine learning. DMS and DEH have published extensively on American health law and policy. MK, DEH, and PH conceived of this project. All authors discussed the results and contributed to the final manuscript. DEH acts as guarantor.

Competing interests: We have read and understood BMJ policy on declaration of interests and have no relevant interests to declare.

Provenance and peer review: Commissioned; externally peer reviewed.

This collection of articles was proposed by the WHO Department of Digital Health and Innovation and commissioned by *The BMJ*. *The BMJ* retained full editorial control over external peer review, editing, and publication of these articles. Open access fees were funded by WHO.

Mark Krass, doctoral candidate[1,2]

Peter Henderson, doctoral student[1,3]

Michelle M Mello, professor[1,4]

David M Studdert, professor[1,4]

Daniel E Ho, professor[1,2,5,6]

[1]Stanford Law School, Stanford University, Stanford, CA, USA

[2]Department of Political Science, Stanford University School of Humanities and Sciences, Stanford, CA, USA

[3]Department of Computer Science, Stanford University School of Engineering, Stanford, CA, USA

[4]Stanford Health Policy and Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA

[5]Stanford Institute for Human-Centered Artificial Intelligence, Stanford, CA, USA

[6]Stanford Institute for Economic Policy Research, Stanford, CA, USA

Correspondence to: D E Ho
dho@law.stanford.edu

1 Veterans Affairs experiments with AI "to-go." *FedScoop* 2020. https://www.fedscoop.com/va-ai-to-go/
2 Lacy A, Speri A, Smith J, Biddle S. Prisons attempt to track coronavirus-related keywords in inmate phone calls. *Intercept* 2020 Apr 21. https://theintercept.com/2020/04/21/prisons-inmates-coronavirus-monitoring-surveillance-verus/
3 DeCaprio D, Gartner J, Burgess T, et al. Building a COVID-19 vulnerability index. ArXiv 2020. [Preprint.] https://arxiv.org/abs/2003.07347
4 Carnegie Mellon University. COVID voice detector. https://cvd.lti.cmu.edu/
5 Carpenter v United States [2018] 585 US.
6 Sklansky DA. Too much information: how not to think about privacy and the Fourth Amendment. *Calif Law Rev* 2014;102:1069-122.
7 Gostin LO, Nass S. Reforming the HIPAA privacy rule: safeguarding privacy and promoting research. *JAMA* 2009;301:1373-5. doi:10.1001/jama.2009.424
8 Ricci v DeStefano [2009] 557 US 557.
9 Gross CP, Essien UR, Pasha S, Gross JR, Wang SY, Nunez-Smith M. Racial and ethnic disparities in population-level covid-19 mortality. *J Gen Intern Med* 2020;35:3097-9. doi:10.1007/s11606-020-06081-w
10 Hutson M, et al. Eye-catching advances in some AI fields are not real. *Science* | 2020 May 27. https://www.sciencemag.org/news/2020/05/eye-catching-advances-some-ai-fields-are-not-real
11 Singh K, Valley TS, Tang S. Validating a widely implemented deterioration index model among hospitalized COVID-19 patients. *medRxiv*2020. [Preprint.] http://www.medrxiv.org/content/10.1101/2020.04.24.20079012v2
12 Dandekar R, Barbastathis G. Quantifying the effect of quarantine control in covid-19 infectious spread using machine learning. *medRxiv* 2020:2020.04.03.20052084. [Preprint.] 2020.04.03.20052084. doi:10.1101/2020.04.03.20052084.
13 Marchant R, Samia NI, Rosen O, Tanner MA, Cripps S. Learning as we go: an examination of the statistical accuracy of COVID 19 daily death count predictions. ArXiv 2020. May 3. [Preprint]. https://arxiv.org/abs/2004.04734
14 Bagdasaryan E, Poursaeed O, Shmatikov V. Differential privacy has disparate impact on model accuracy. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, eds. *Advances in neural information processing systems 32*. Curran, 2019:15479-88.
15 Ruggles S, Fitch C, Magnuson D, Schroeder J. Differential privacy and census data: implications for social and economic research. *AEA Pap Proc* 2019;109:403-8. doi:10.1257/pandp.20191107.
16 Barocas S, Hardt M, Narayanan A. *Fairness in machine learning*. 2019. https://fairmlbook.org/
17 Grother P, Ngan M, Hanaoka K. Face recognition vendor test (FRVT). Part 2: Identification. NIST interagency/internal report (NISTIR)). Report no 8271. doi:10.6028/NIST.IR.8271
18 Horwitz LI, Kuznetsova M, Jones SA. Creating a learning health system through rapid-cycle, randomized testing. *N Engl J Med* 2019;381:1175-9. doi:10.1056/NEJMsb1900856
19 Stanford HAI. COVID + AI: the road ahead. https://hai.stanford.edu/watch-covid-ai-road-ahead

Cite this as: ***BMJ*** **2021;372:n234**
http://dx.doi.org/10.1136/bmj.n234