Volume 6, No. 4 | July–August 2023

RAIL

The Journal of Robotics, Artificial Intelligence & Law

fastcase FULL COURT PRESS

RAIL

The Journal of Robotics, Artificial Intelligence & Law

Volume 6, No. 4 | July–August 2023

THE JOURNAL OF ROBOTICS, ARTIFICIAL INTELLIGENCE & LAW (ISSN 2575-5633 (print) /ISSN 2575-5617 (online) at $495.00 annually is published six times per year by Full Court Press, a Fastcase, Inc., imprint. For customer support, please contact Fastcase, Inc., 729 15th Street, NW, Suite 500, Washington, D.C. 20005, 202.999.4777 (phone), or email customer service at support@fastcase.com.

Publishing Staff
Publisher: Morgan Morrissette Wright
Production Editor: Sharon D. Ray
Cover Art Design: Juan Bustamante

Cite this publication as:

The Journal of Robotics, Artificial Intelligence & Law (Fastcase)

A Full Court Press, Fastcase, Inc., Publication

Editorial Office

729 15th Street, NW, Suite 500, Washington, D.C. 20005
https://www.fastcase.com/

POSTMASTER: Send address changes to THE JOURNAL OF ROBOTICS, ARTIFICIAL INTELLIGENCE & LAW, 729 15th Street, NW, Suite 500, Washington, D.C. 20005.

Articles and Submissions

Direct editorial inquiries and send material for publication to:

Steven A. Meyerowitz, Editor-in-Chief, Meyerowitz Communications Inc., 26910 Grand Central Parkway, #18R, Floral Park, NY 11005, smeyerowitz@meyerowitzcommunications.com, 631.291.5541.

Material for publication is welcomed—articles, decisions, or other items of interest to attorneys and law firms, in-house counsel, corporate compliance officers, government agencies and their counsel, senior business executives, scientists, engineers, and anyone interested in the law governing artificial intelligence and robotics. This publication is designed to be accurate and authoritative, but neither the publisher nor the authors are rendering legal, accounting, or other professional services in this publication. If legal or other expert advice is desired, retain the services of an appropriate professional. The articles and columns reflect only the present considerations and views of the authors and do not necessarily reflect those of the firms or organizations with which they are affiliated, any of the former or present clients of the authors or their firms or organizations, or the editors or publisher.

QUESTIONS ABOUT THIS PUBLICATION?

For questions about the Editorial Content appearing in these volumes or reprint permission, please contact:

Morgan Morrissette Wright, Publisher, Full Court Press at mwright@fastcase.com or at 202.999.4878

For questions or Sales and Customer Service:

Customer Service
Available 8 a.m.–8 p.m. Eastern Time
866.773.2782 (phone)
support@fastcase.com (email)

Sales
202.999.4777 (phone)
sales@fastcase.com (email)

ISSN 2575-5633 (print)
ISSN 2575-5617 (online)

Why a Data Scientist Needs a Lawyer to Correct Algorithmic Discrimination

Bradley Merrill Thompson*

In this article, the author explains why a collaboration between data scientists and attorneys is the key to developing algorithms that are compliant with antidiscrimination laws.

We all have heard the stories. A recruiting app that prefers men. Facial recognition software that cannot recognize Black women. Clinical decision support tools for evaluating kidney disease that often give doctors the wrong advice when the patient is Black. Triage software that puts White people ahead of Black and Brown people. The list is long and growing.

These are real harms to our friends and neighbors, costing them economic opportunities or causing physical injury in the case of health care. For organizations seeking to use artificial intelligence (AI)–powered tools, they also create potentially expensive legal liabilities and damage in the court of public opinion.

Local, state, and federal agencies are racing to implement regulations to address these issues. Regardless of the domain, a common thread in the current and proposed regulations is bias. Soon, the Department of Health and Human Services may require that users of algorithms in health care evaluate those algorithms for bias.[1] New York City presently requires all AI-powered selection and hiring tools to be audited for bias, and several municipalities are currently considering similar regulations. In the European Union, big tech companies will have to conduct annual audits of their AI systems from 2024, and the upcoming AI Act will require audits of "high-risk" AI systems.[2]

But what exactly do the antidiscrimination laws require? While the law is often unclear, just picking age discrimination as one example, some attorneys would argue that:

- Age may be considered by an airline as "bona fide occupational qualification" for safety reasons.

Robotics, Artificial Intelligence & Law / July–August 2023, Vol. 6, No. 4, pp. 277–285.

ISSN 2575-5633 (print)/ISSN 2575-5617 (online)

- Age may not generally be considered by software companies when hiring new programmers.
- Age, in certain circumstances, may be considered affirmatively as part of a company's comprehensive recruitment strategy to help older Americans get programming jobs where the software company can establish that its own hiring patterns disadvantaged older workers historically.
- Age may not be considered in scheduling access to radiological imaging.[3]
- Age may be considered when deciding who should receive a particular transplanted organ.
- Age may not be considered when deciding who should go on a physically demanding business trip.
- Age might unconsciously be considered in promotions so long as there is no statistical evidence of disparate impact overall.
- Age may not be considered by a health insurance company in deciding whether to cover an expensive procedure.[4]
- Age may, and really must, be considered when diagnosing macular degeneration.
- Age may not be considered in targeting advertising for certain credit services in social media.[5]
- Age may be considered in targeting advertising for certain healthcare products in social media.

If your head hurts now, that is understandable.

Age, like sex, race, and dozens of other demographic factors, is a protected class in America that is supposed to be free from discrimination. But knowing that hardly provides sufficient guidance for the development of a wide range of algorithms that make, or advise on, decisions impacted by age.

Why Isn't This Simple?

Why can't we just tell data scientists to fix it? Is this not just somehow an error in their math? It is called machine learning after all. Can't experts just fix the machine so as not to discriminate?

Unfortunately, the answer is no in many cases. There are two intertwined layers to the complexity that make remediation a matter of judgment, using expert knowledge in both data science and law.

Those layers are (1) the technical challenges around finding and correcting bias, and (2) the legal complexities and changing nature around defining what bias is and when it is acceptable.

The Technical Challenges

This article is certainly not meant to be a comprehensive technical explanation of the challenges in finding and correcting bias,[6] but a brief discussion helps put the legal issues in appropriate context.[7]

Bias can creep into algorithmic decision-making in many ways, including training an algorithm on data that contains encoded prior human bias (i.e., the historical decisions on which the model is trained were biased), failing to ensure that the training data includes adequate representation from categories of interest (either through poor sampling technique or downright exclusion), assumptions or mistakes made in the design and coding of the algorithm, and measurement bias where the data collected for training differ from that collected in the real world, to name just a few.

Can we overcome those problems simply by making sure that the data analyzed by the algorithm excludes data on age, sex, race, and so forth? The problem is that sensitive information can be inferred from other information, which has value that we do not want to lose. For example, the names Emily, Shaquana, and Soo, statistically speaking, say something about race and sex. Playing volleyball in college says something about sex. The year of graduation says something about age. Zip code says something about race, as does graduating from a historically black college or a university in India. In the health care realm, a medical record might indicate that you had uterine cancer, that you had your prostate removed in 1969, or that you had gender reassignment surgery done.

Algorithms are just looking for statistical associations, so it would be easy for a machine learning model to infer information about protected classes. Further, getting rid of all such information creates new problems because the information might be important to a person's credentials in the case of an employment decision. Being on the team that won the NCAA championship in volleyball speaks to work ethic and team mentality. Graduating from a prestigious school in India is not something you want to hide. And in medicine, important medical history cannot be removed safely without risking misdiagnosis.

Nor can we just avoid the problem by always finding perfect data without any bias whatsoever to train the model. In the real world, discrimination exists. We cannot assume that any real-world data will be free from bias. Finding good data, let alone excellent data, is perhaps the biggest challenge to machine learning. Frequently, we can't even find data that exactly measures what we want to measure, so we must use surrogate variables that estimate the attribute we really care about.

In United States, there are many protected classes, including age, ancestry, color, disability, ethnicity, gender, gender identity or expression, genetic information, HIV/AIDS status, military status, national origin, pregnancy, race, religion, sex, sexual orientation, veteran status, and receipt of income from any public assistance program. Many of those categories have more than two classes, with race, for example, including White, Black or African American, American Indian or Alaska Native, Asian, and Native Hawaiian or Other Pacific Islander.

Thus, if you are developing a model, for example, to guide decision-making around health insurance coverage, you need to train the model to render an accurate assessment for an elderly, disabled, Jewish, lesbian woman from the Philippines (who is of Pacific Islander race but Hispanic in ethnicity) honorably discharged from the Marines who is HIV-positive. How easy is it to find a health insurance coverage training set that is both big enough to include sufficient data regarding each of her vulnerabilities and completely free of bias? That data set does not exist.

You would think at least finding discrimination in an algorithm would be easy, but no. There are lots of competing tools to test for bias, but none of them are the gold standard, each with their pluses and minuses. Indeed, the current tools are so weak that the National Institutes of Health has launched a competition to try to get developers to create better tools for this purpose.[8]

A key part in testing algorithms is having a metric to determine whether unlawful discrimination exists or not. But we don't even have a uniformly agreed upon measure of bias, and indeed there are literally dozens of different "fairness metrics" each with their own focus, some of which might be legally useful and others clearly not.[9]

Further, some forms of discrimination are simply invisible in the outputs of the algorithms, and only can be detected based on real-world impact in the specific market or domain in which it is being used. Testing for that requires something akin to a clinical trial. In those situations, we are typically looking for disparate impacts on any of the protected classes.[10]

When bias is found, under some basic limits of the technology, often it cannot simply be eradicated. There is no free lunch. Unless it is based on finding new, suitable data to supplement the existing training set, improving equality in the overall output typically means reducing overall accuracy. And we typically like our models to be accurate. Further, the data scientist's toolbox includes ways to make trade-offs between different subpopulations, improving equity in one area at the cost of accuracy in another area. When is that appropriate?

Sometimes a data scientist is tempted to put her thumb on the scale, helping those who are disadvantaged through a mathematical adjustment. But often these well-accepted data science techniques involve value judgments that can run contrary to the law[11] and create big problems for the overall accuracy and effectiveness of the model.

In medicine, as an example, the National Institutes of Health explained that for more than 20 years, physicians have been boosting the scores of Black Americans for the estimated glomerular filtration rate, or eGFR. That calculation is used to determine the prognosis and treatment of people with kidney disease, including when a person should go on hemodialysis or get a transplant. The problem is boosting the scores based on self-reported race, when there is great variance within the genetic ancestry of people who identify as Black, means that many people who identified as Black were misdiagnosed with respect to the risk of kidney disease.[12]

So, if (1) we cannot cleanse the data of the attributes of concern; (2) we cannot find perfect, sufficient data set free of bias; (3) in fact we cannot even be sure that we can find bias; and (4) often we cannot just use math to fix it, we need a different strategy. We must get a little creative and we also must be practical. And that is where the legal judgments described below come into play.

Fighting Algorithmic Bias Throughout the Product Life Cycle

There is general agreement that bias needs to be fought throughout the product life cycle. An emerging standard is being developed by the National Institute Standards and Technology, or NIST, that centers around best practices at each stage of the product life cycle, from pre-design, to design, to deployment and finally testing and evaluation.[13]

During that product life cycle, there are several areas where developers can fight bias, including:

- Prior to the design of the algorithm, having good governance practices in place to make sure there is appropriate responsibility assigned.
- During the design phase, making sure that the design team includes a culturally diverse group with cross functional representation that will be more likely to spot problematic design elements.
- During deployment, having policies in place that ensure the company appropriately instructs its customers on how to use the product.
- During the testing and evaluation phase, ensuring appropriate auditing of the model's output, as well as put in place appropriate monitoring to ensure the algorithm doesn't develop biased tendencies in the future.

The list of proactive steps to avoid bias goes on and on.

Where the Law Enters In

Given those technical challenges, there are judgments that need to be made, and this is where the law squarely comes into play. Some companies have left data scientists to do their own thing based on their own ethical values and the product is, well, based on the ethical values of the individuals designing the algorithms. But if the goal of the company is to achieve legal compliance and reduce legal risk, the technical decisions need to be informed by the laws.

It should come as no surprise that the laws are complicated and ever changing. Indeed, it seems likely the Supreme Court will yet again re-examine the topic of affirmative action.[14] So it should seem obvious that we can't simply ask data scientists to come up with legally compliant algorithms by themselves.

But where specifically should data scientists and lawyers collaborate to ensure that algorithms comply with antidiscrimination laws? Following are just two examples.

- Development phase

Attorneys are trained to help companies develop compliance programs under federal law, including the federal sentencing guidelines. Compliance programs in this case give the company legal shelter from claims that the company was not diligent enough in working to detect violations of antidiscrimination laws. Following federal guidelines for compliance programs, which are generally accepted now as best practices, helps companies mitigate their legal risk.

- Testing phase

Here are a few of the areas where attorneys and data scientists should collaborate to help achieve compliance in the testing phase:

1. What do the new laws that are requiring audits specifically require the company to do?
2. Which evaluation metric is legally the correct one for the testing? While courts have generally recognized certain statistical tests for employment decision-making by humans, there is disagreement about how new machine learning algorithms should be evaluated for fairness. There are a myriad and growing number of so-called fairness metrics, including:
    - Statistical parity;
    - Equalized odds;
    - Predictive parity; and
    - Overall accuracy equality.

   Should we test for fairness at the group level or at the individual level?[15] The fact is, choosing the right fairness metric for the test depends on:
    - The specific application (certain differences in treatment may be justified in a particular case);
    - The technology used (e.g., basic classification versus natural language output); and
    - The particular legal requirement the testing is trying to satisfy (e.g., FTC versus EEOC fairness).

   One size does not fit all, and the field is constantly evolving.
3. When are retrospective tests on existing data sufficient, as opposed to a prospective trial? What data, including edge cases, are necessary for the testing?

4. Exactly which demographic categories need to be evaluated? For example, does every algorithm have to be accurate for every single racial and ethnic group that is present in the United States, however small? What about other protected categories? For machine learning models used outside of employment decision-making, exactly which categories need to be considered at all?
5. To what degree do these antidiscrimination laws apply to the particular intended use? Does every product need to work well for everyone?
6. If testing reveals gaps in performance, does the law allow those deficiencies to be addressed through labeling, or does the model have to be retrained on data that might be hard to find?
7. No model is perfect. What is good enough in terms of overall performance? What is good enough in terms of the performance in each individual subpopulation?
8. When unfairness is found, how do you fix it? Which tradeoffs are acceptable? Can you put your finger on the scale and mathematically make adjustments to compensate or is that reverse discrimination?
9. What about software that works well in the hands of the developers, but then produces biased results when used in a new population? Regulators are making it clear that because there can be differences between the data on which a model is trained and ultimately used, even some users have an affirmative legal obligation to ensure that the model does not behave in a biased manner once it is in the users' hands.
10. What are the risks that bias could develop over time, such that retesting is necessary? If it is necessary, how often and what kind?

These are the types of questions that arise in the evaluation phase of the machine learning model's life cycle.

Conclusion

It seems like this should be easy but it is not. In some ways, because it is so mechanical, it is easier to find certain types of discrimination committed by an algorithm that it would be if we were auditing a purely human decision. But even though all the software code and data are right there before our eyes, evaluating the performance of these models is difficult at best. We cannot achieve perfection, but collaboration between data scientists and attorneys is the key to developing algorithms that are compliant with antidiscrimination laws.

Notes

* The author, a member of Epstein Becker & Green, P.C., may be contacted at bthompson@ebglaw.com.

1. https://jamanetwork.com/journals/jama/fullarticle/2800369.
2. https://www.technologyreview.com/2022/10/24/1062071/do-ai-systems-need-to-come-with-safety-warnings/.
3. https://www.scu.edu/ethics/focus-areas/bioethics/resources/aged-based-health-care-rationing/.
4. https://www.hhs.gov/civil-rights/for-providers/clearance-medicare-providers/age-discrimination-act-requirements/index.html and https://www.consumercomplianceoutlook.org/2019/third-issue/from-catalogs-to-clicks-the-fair-lending-implications-of-targeted-internet-marketing/#.
5. https://www.consumerfinance.gov/about-us/blog/cracking-down-on-discrimination-in-the-financial-sector/.
6. Jie Xu, et al., Algorithmic Fairness in Computational Medicine (Sept. 6, 2022), https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(22)00432-7/fulltext.
7. https://www.technologyreview.com/2021/02/11/1017955/auditors-testing-ai-hiring-algorithms-bias-big-questions-remain/.
8. https://ncats.nih.gov/funding/challenges/bias-detection-tools-in-health-care.
9. https://arxiv.org/pdf/1912.00761.pdf.
10. Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, Cass R Sunstein, Discrimination in the Age of Algorithms, Journal of Legal Analysis, Volume 10, 2018, pp. 113-74, https://doi.org/10.1093/jla/laz001 https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086.
11. https://hai.stanford.edu/news/legal-approach-affirmative-algorithms.
12. https://www.nih.gov/news-events/news-releases/nih-supported-study-suggests-alternative-race-based-kidney-function-calculations.
13. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf.
14. https://www.nytimes.com/2022/10/31/us/supreme-court-harvard-unc-affirmative-action.html.
15. https://www.law.georgetown.edu/georgetown-law-journal/wp-content/uploads/sites/26/2020/04/Is-Algorithmic-Affirmative-Action-Legal.pdf.