Duke Law

White Paper:

Addressing Bias in Artificial Intelligence

The Current Regulatory Landscape

THOMSON REUTERS®

Executive summary

As artificial intelligence (AI) grows in importance, regulation will increasingly affect its development. Although most targeted algorithmic regulation is still theoretical, certain themes exist that can provide keen insight into future regulation. Specifically, regulators and legislators in Europe and North America are identifying high-risk uses of AI and requiring the performance and publication of impact assessments and reports. In the meantime, regulators intend to rely on existing antidiscrimination laws to combat the kind of algorithmic bias that AI can exacerbate. These laws emphasize the concept of AI *explainability*, which can enable people to better understand the logic used by AI systems to make predictions, recommendations, or decisions.

Through this concept of explainability, those affected by AI systems can both understand the outcome of AI decisions and challenge these outcomes where relevant. This report reviews the regulatory landscape regarding bias requirements and also analyzes several obstacles to bias identification, such as the imprecise definition of fairness and the lack of regard for proxy variables.

Regulators and legislators in Europe and North America are identifying high-risk uses of AI and requiring the performance and publication of impact assessments and reports.

Beyond technical considerations, this report also shares key ways that companies working in this area could align their AI principles with global standards established by the United Nations Educational, Scientific and Cultural Organization (UNESCO), the Organisation for Economic Co-operation and Development (OECD), and the National Institute of Standards and Technology (NIST).

Finally, this report shares opportunities for putting these principles into practice through techniques and tools that can be embedded throughout the AI development life cycle, including audits, explainable AI (XAI), checklists, and ethical matrices. Practicing these methods can embed fairness within tools that shape AI development and establish best practices that increase auditability and transparency.

Sources of potential bias in AI

NIST's 2023 AI Risk Management Framework suggests that bias in AI can be categorized into three main sources: "systemic, computational and statistical, and human-cognitive."[1] Managing bias in AI systems requires consideration of all three categories, including where bias occurs because of more than one simultaneously.

The seminal works of scholars Timnit Gebru and Joy Buolamwini in the field of AI fairness, accountability, and explainability have identified how gender and racial biases can arise in machine learning (ML) tools due to incomplete or unrepresentative data sets, a source of bias that could be categorized both as systemic and computational.[2] Their findings have broad implications for ML tools, showing that if individuals from certain intersectional demographic groups are underrepresented in the data upon which AI systems are trained, the AI might perform disproportionately worse when applied to these groups.

Biases can have devastating effects on individuals and groups when biased AI is used to make predictions, recommendations, or decisions in the real world.

There is also the issue of historic bias. For example, when AI tools are trained on large historical corpora, any biases that *already* exist in these corpora will be embedded in the AI systems which, when deployed, could perpetuate discrimination, including discrimination based on protected characteristics such as race, gender, and social status.

Biases are not always overtly perceivable in AI systems. However, biases can have devastating effects on individuals and groups when biased AI is used to make predictions, recommendations, or decisions in the real world based on data about these individuals or groups, including in critical sectors. Even if biases do not directly produce discriminatory outcomes via automated decision-making, they can also reinforce negative stereotypes when embedded within AI systems.

1 National Institute of Standards and Technology (NIST), "Artificial Intelligence Risk Management Framework (AI RMF 1.0)" January 2023; https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf.

2 Timnit Gebru. Joy Buolamwini. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81: 1-15, 2018. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

Overview of legislation and regulation on AI bias

Although there is global interest in regulating biased algorithms, there are currently few enforceable codes. As a result, regulators are looking to existing antidiscrimination laws to help combat algorithmic bias before meaningful legislation is passed.

> ***"If you don't hold yourself accountable, the FTC may do it."***

In the United States, the Federal Trade Commission (FTC) is one of the organizations at the forefront, declaring in a blog post for developers and users of algorithms that "if you don't hold yourself accountable, the FTC may do it."[3] Alongside the FTC is the Equal Employment Opportunity Commission (EEOC), which can insinuate itself in cases where algorithms may produce disparate impacts under the Civil Rights Act of 1964.

Meanwhile, current European antidiscrimination law is far broader than its American counterpart and could also be used to combat algorithmic discrimination both before and after the finalized EU AI Act is enacted.

Federal Trade Commission

In the US, many federal agencies have expressed their interest in regulating the use of biased algorithms, but the FTC is leading the charge.[4] Although the Department of Justice (DOJ) is the traditional gatekeeper in antidiscrimination enforcement, the FTC leads in areas where the law is "unsettled."[5] Congress recognized this reality by centering the proposed Algorithmic Accountability Act around the FTC. Due to the lack of direct federal legislation, agencies are relying on existing civil rights statutes and precedent in order to police bias, the most pertinent of which is Section 5 of the FTC Act.

Enacted in 1917, Section 5 prohibits "unfair or deceptive *acts or practices* in or affecting commerce" (emphasis added). Here, "commerce" is interpreted broadly to allow maximal applicability. A violation need only be unfair or deceptive, not both.

An act or practice qualifies as unfair if it: *i*) causes or is likely to cause substantial injury to consumers; *ii*) cannot be reasonably avoided by consumers; and *iii*) is not outweighed by countervailing benefits to consumers or to competition. Public policy – as established by statute, regulation, or judicial decisions – may be considered with all other evidence in

3 "FTC Outlines Approach to Discrimination in AI and Foreshadows Potential Enforcement," J.D. Supra, April 27, 2021.
4 *Id.*
5 *Id.*, 15 USC §45(a)(1).

determining whether an act or practice is unfair.[6] The FTC has indicated it is prepared to treat the sale of racially biased algorithms as unfair under Section 5.[7]

An algorithm could also be deemed deceptive if *i*) it involves a "representation, omission, or practice [that] misleads or is likely to mislead the consumer;" *ii*) it is reasonable under the circumstances for the consumer to be misled; and *iii*) the misleading representation, omission, or practice is material.[8]

When determining whether a violation has occurred, the FTC applies a "does more harm than good" balancing test weighing the discriminatory impact with the algorithm's social or commercial benefit. As a hypothetical example, if a bank employs a credit-scoring algorithm that considers the crime rate of an applicant's zip code, this could result in redlining, where the bank refuses to provide services to potential customers because they live in a low-income or crime-ridden area. The bank would be required to show a benefit *greater* than the social cost of the discriminatory algorithm to survive any potential FTC challenge.

To avoid such a potential legal action, the FTC suggests that algorithms should be well tested, transparent, and built upon inclusive data sets.[9]

Equal Employment Opportunity Commission

The Equal Employment Opportunity Commission (EEOC) enforces policies that mitigate AI bias within the context of their larger responsibility to regulate antidiscrimination. The prominence of the Commission in regulating issues of AI fairness is evident from the recent launch of their Initiative on Artificial Intelligence and Algorithmic Fairness,[10] as well as recent EEOC guidance about employers' use of algorithmic, automated, or AI decision-making tools for job applicant and employee-related decisions.[11]

Established under the Civil Rights Act of 1964, the EEOC is tasked with enforcing that statute and subsequent employment rights legislation. In the case of biased algorithms, Title VII of the Civil Rights Act prohibits algorithms from facilitating employment discrimination against a protected class either directly (disparate treatment) or indirectly (disparate impact). Currently, disparate impact actions are primarily restricted to employment claims, with the EEOC being the agency responsible for enforcement. Disparate treatment, or explicit discrimination, is covered under a broader range of civil rights statutes.

6 William Kovacic, "The Application of Section 5 of the Federal Trade Commission Act," U.S. Federal Trade Commission; Presentation Fall Forum, November 12, 2009.

7 Jillson, *Aiming for truth, fairness, and equity in your company's use of AI*, Federal Trade Commission Business Blog (2021).

8 *Id.*

9 *Id.*

10 "EEOC Launches Initiative on Artificial Intelligence and Algorithmic Fairness," U.S. Equal Employment Opportunity Commission, 28 October 2021. https://www.eeoc.gov/newsroom/eeoc-launches-initiative-artificial-intelligence-and-algorithmic-fairness.

11 "Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII of the Civil Rights Act of 1964," U.S. Equal Employment Opportunity Commission," 18 May 2023.

Although *disparate impact* is relatively restricted, it provides a useful guide for bias testing. To establish a disparate impact claim, courts will commonly apply the "four-fifths rule." After a positive employment decision is made, the lowest selection rate of a protected class must be greater than 80% of the nonprotected group's selection rate.[12] For example, if 100 men and 200 women apply for 10 open positions, and the company hires seven men and three women, a disparate impact claim exists.

Male Selection Rate = 7% (7 hires out of 100 applicants)

Female Selection Rate = 1.5% (3 hires out of 200 applicants)

Female/Male Selection Rate = 21.4% (1.5% divided by 7%)

After a disparate impact claim is established, the accused party has an opportunity to provide "substantial legitimate justification."[13] For example, the positions above may have required an advanced STEM degree and only three female applicants possessed one.

Finally, if a substantial, legitimate justification exists, the analysis asks whether there are less discriminatory alternatives. For example, is it reasonable for the position to require an advanced STEM degree? While disparate impact is unlikely to be expanded beyond its current limits, the four-fifths rule provides a useful metric to test for bias and is likely to be borrowed in non-Title VII and Fair Housing Act actions that implicate algorithmic bias. Nonetheless, the EEOC cautioned employers that the "four-fifths rule" may not always be the appropriate standard to assess disparate impact, advising employers to ask any vendors that they employ for their algorithmic decision-making tools to confirm which standard is being relied upon to evaluate these tools.[14]

The European Union

Although European nondiscrimination law has not yet been mandated by regulators as a tool to combat algorithmic bias, current European legal frameworks for assessing discrimination may be relevant for algorithms in the future. However, it remains uncertain whether EU nondiscrimination law is broad enough in scope to address these challenges.[15] Much like the US and its disparate treatment/disparate impact duality, the European Union has a similar two-tiered framework whereby biased AI may be deemed "direct discrimination" (disparate treatment) or "indirect discrimination" (disparate impact).[16] Similar to the US, actions arising from indirect discrimination are not available in all cases. However, European coverage of indirect discrimination is far greater than America's relatively limited disparate-impact theory.

12 Equal Employment Opportunity Commission, "Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures," Federal Register 44, (Mar 2, 1979).

13 *Id.*

14 "Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII of the Civil Rights Act of 1964," U.S. Equal Employment Opportunity Commission," 18 May 2023.

15 European Commission, Directorate-General for Justice and Consumers, Gerards, J., Xenidis, R. "Algorithmic discrimination in Europe: challenges and opportunities for gender equality and non-discrimination law." 2021, Publications Office. https://data.europa.eu/doi/10.2838/544956.

16 Justyna Maliszewska-Nienartowicz, *Direct and Indirect Discrimination in European Union Law – How to Draw a Dividing Line?*, International Journal of Social Sciences 3, 2014.

For example, unjustified, indirect discrimination is forbidden by governments or government-affiliated agencies under the European Convention on Human Rights (ECHR).

While the European Court of Justice did not extend the ECHR to private actors, other EU regulations provide similar coverage with EU directives that ban indirect discrimination on the basis of sex and on the basis of race or ethnicity.

To determine whether discrimination has occurred under EU law, one asks whether "an apparently neutral provision, criterion, or practice would put persons having a particular racial or ethnic origin, religion or belief, sex, disability, age or sexual orientation at a particular disadvantage compared with other people."[17]

Next, a court or regulator asks whether the practice is objectively justified and by a legitimate aim; and whether the means of achieving that aim are appropriate and necessary.[18] Under current law, it is difficult for harmed parties to bring claims against biased algorithms. However, as future legislation increases transparency and explainability, the number of discrimination claims under existing civil rights legislation will grow.

17 *Id.*
18 *Id.*

Trends in emerging legislation related to AI bias

The rapid development of AI poses a serious challenge to the slow-moving legislative process. Few meaningful statutes confronting algorithmic bias are under consideration and even fewer have been enacted into law.

However, certain common themes signaling the future of regulation can be observed that relate to AI bias, even if they do not explicitly address the issue of biased algorithms.

Administrative oversight

It is common practice for legislators to empower agencies to enforce, update, and interpret a statute.[19] An agency's duties vary by statute and directive, but common tasks involve collecting audit and transparency reports, maintaining repositories, and enforcing compliance.

Certain common themes signaling the future of regulation can be observed that relate to AI bias, even if they do not explicitly address the issue of biased algorithms.

This kind of administrative oversight is recommended by current and proposed algorithmic bias legislation. For example, the proposed Algorithmic Accountability Act would elevate the FTC's role in policing algorithmic bias and discrimination.[20] Additionally, the recently introduced nonbinding principles issued by the White House called the *Blueprint for an AI Bill of Rights* recommend performing technical "disparity assessments" using demographic data and algorithmic impact assessments, and highlight the need for "disparity mitigation" strategies.[21]

In the EU a regulation originally proposed in 2021, *Laying Down Harmonised Rules on Artificial Intelligence* (otherwise known as the AI Act), while not yet enacted, is likely to set international precedent in AI law for the comprehensive risk-based framework it proposes.[22] It is also possible that antidiscrimination laws will be given new life in the wake of the EU AI Act's transparency requirements. These transparency requirements, though not explicitly framed as "bias requirements," are positioned to minimize the risk of biased AI-assisted decisions

19 For example, the FTC, EEOC, and DOJ as described above.
20 Proposed Algorithmic Accountability Act of 2022. H.R.6580 — 117th Congress (2021-2022).
21 "Blueprint for an AI Bill of Rights," The White House (The United States Government, October 4, 2022), https://www.whitehouse.gov/ostp/ai-bill-of-rights.
22 Feedback from: University of Cambridge (Leverhulme Centre for the Future of Intelligence and Centre for the Study of Existential Risk," European Commission, 6 August 2021,

and discriminatory effects in critical areas such as education and training, employment, important services, law enforcement, and the judiciary.[23]

Notably, the proposed EU AI Act would mandate tiered levels of administrative oversight based on a comprehensive framework dividing algorithms into one of four risk categories: unacceptable, high, limited, and minimal. While certain AI practices deemed to pose an unacceptable risk would be entirely banned and those bringing minimal risk would have no requirements, the oversight demanded for high-risk systems includes transparency or conformity assessments and technical documentation. Such oversight would be enforced through the governance systems of each European Member state, although a European AI Board would be established to facilitate cooperation across states.

Transparency, impact reports and repositories

Transparency or impact reporting requirements are another common theme in emerging and enacted legislation. While the complexity required varies, most legislation asks for some degree of impact assessment.

> ***While the complexity required varies, most legislation asks for some degree of impact assessment.***

Among the most demanding is New York City's 2023 requirement that all employers conduct third-party bias audits of any algorithm involved in a hiring decision.[24] Similarly, the EU AI Act's proposed eight-part "conformity assessment" will likely also require outside assistance due to its complexity,[25] although that assessment is not yet mandatory.

Federal legislation, as well as proposed actions by other states in the United States, have included less rigorous "impact assessments." For example, the Algorithmic Accountability Act would require that developers and users of high-risk algorithms *i*) describe the system in detail; *ii*) assess the relative costs and benefits of the system; *iii*) determine the risks to the privacy and security of personal information; and *iv*) explain the steps taken to minimize those risks, if discovered.[26]

Although it is unclear what the final form would look like, Canada has adopted a questionnaire format.[27] The United Kingdom's Algorithmic Transparency Standard also uses a questionnaire-style assessment. Currently, the Ada Lovelace Institute's impact assessment is recommended by the UK government but is not yet mandated.[28]

23 Proposal for a Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts, §40. EUR-Lex-52021PC0206-EN-EUR-Lex. eur-lex.europa.eu. §3.5.

24 New York City Department of Consumer and Worker Protection. "Notice of Adoption of Final Rule." Local Law 144 of 2021. April 5 2023. https://rules.cityofnewyork.us/wp-content/uploads/2023/04/DCWP-NOA-for-Use-of-Automated-Employment-Decisionmaking-Tools-2.pdf.

25 Proposal for a Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts, §40. EUR-Lex-52021PC0206-EN-EUR-Lex. eur-lex.europa.eu.

26 Proposed Algorithmic Accountability Act of 2022. H.R.6580 — 117th Congress (2021-2022).

27 "Algorithmic Impact Assessment Tool," Government of Canada, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html.

28 "UK to pilot world-leading approach to improve ethical adoption of AI in healthcare," UK Department of Health and Social Care Press Release (Feb 8, 2022). https://www.gov.uk/government/news/uk-to-pilot-world-leading-approach-to-improve-ethical-adoption-of-ai-in-healthcare.

In addition to compelling transparency, most emerging legislation requires that administrative agencies maintain repositories. For example, under the Algorithmic Accountability Act, the FTC would host a publicly accessible repository for all impact assessments. Similarly, the EU AI Act would also require conformity assessments to be collected by national regulators.[29]

The near uniformity of emerging legislation in requiring that impact assessment reports be delivered to enforcing agencies or be made publicly available is a sign that officials are emphasizing accountability and transparency in automated decisions.

Proactive policymaking by jurisdictions like the EU is likely to inform the more unregulated, patchwork approach to AI legislation in the US. In this light, the standards set by the proposed EU AI Act offer a window into obligations that organizations might eventually face within the US when developing AI-powered algorithms. In particular, the EU AI Act specifies a number of high-risk domains that are heavily restricted or outright banned. As drafted, the EU AI Act will classify certain systems related to the following areas as high-risk AI, subjecting them to requirements including rigorous testing, proper data documentation procedures, and the implementation of an accountability framework:

1. critical infrastructures
2. educational or vocational training
3. employment, worker management and access to self-employment
4. essential private and public services
5. law enforcement (within specific bounds)
6. migration, asylum and border control management
7. administration of justice and democratic processes[30]

Additionally, in the latest draft of the EU AI Act, the European Parliament's Internal Market Committee and Civil Liberties Committee jointly adopted amendments to the draft that, among other modifications, expanded the list of banned AI to include remote biometric identification systems that evaluate biometric data in real time and in publicly accessible places; certain "post" remote biometric identification systems; predictive policing systems; biometric categorization systems that use sensitive characteristics; emotion recognition systems in the areas of law enforcement, border management, workplaces, and educational institutions; and systems that create or expand facial recognition databases through the indiscriminate scraping of biometric data from social media or CCTV footage.[31]

29 Proposal for a Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts, §40. EUR-Lex-52021PC0206-EN-EUR-Lex. eur-lex.europa.eu.

30 Proposal for a Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts, §40. EUR-Lex-52021PC0206-EN-EUR-Lex. eur-lex.europa.eu.

31 Amendments adopted by the European Parliament on 14 June 2023 on the Proposal for a Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts. eur-lex.europa.eu.

Compared to these explicitly banned use-cases, the amendments adopted by the European Parliament also expanded the list of high-risk systems, systems which are permitted, provided that they undergo careful monitoring. Key among the requirements that the proposed EU AI Act would impose on providers of high-risk AI systems is transparency.[32]

Given the pace at which generative foundation models, like GPT, have recently evolved, it is perhaps unsurprising that the European Parliament specifically focused on additional transparency requirements for these models in the most recent June 2023 amendments to the proposed EU AI Act.[33] Due to the complexity and unexpected impact of foundation models, the Parliament adopted additional transparency obligations that were tailored for these models, including: unique obligations to disclose content as being generated by AI outputs; restrictions on models to prevent the generation of illegal outputs; and requirements to publish copyrighted data used for training. If passed, these imposed obligations for the EU AI Act would come after China has already passed legislation on foundation models and synthetically generated content, though China does not yet have a comprehensive national AI law.[34]

32 Proposal for a Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts, §40. EUR-Lex-52021PC0206-EN-EUR-Lex. eur-lex.europa.eu. §43.

33 Amendments adopted by the European Parliament on 14 June 2023 on the Proposal for a Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts. eur-lex.europa.eu.

34 China's AI Regulations and How They Get Made," Carnegie Endowment For International Peace, July 10 2023, https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117.

Obstacles to bias identification

Despite the range of global legislation that explicitly or implicitly addresses algorithmic bias, several obstacles to identifying biases in AI pose challenges for providers when interpreting and complying with laws. One impediment to identifying bias is how difficult it can be to define what constitutes bias and, conversely, fairness.

Conversations surrounding bias in AI often focus on age, gender, and ethnicity. However, there is a wider sphere of protected characteristics in reference to which algorithms can be systematically biased, including national origin, sexual identity, and disability status. What constitutes bias may vary among jurisdictions. For example, different countries may have different minority populations, which will inform which groups tend to be underrepresented. Thus, defining *fairness* in both a legal and a technical context poses substantial difficulties.

Legislators, judges, and regulators struggle to agree on a universal definition beyond "not unfair."

Fairness is a core component of the legal system in Western countries; however, legislators, judges, and regulators struggle to agree on a universal definition beyond "not *unfair*." The US Supreme Court recognized this dilemma in 1931 when it admitted that fairness "belongs to that class of phrases which do not admit of precise definition."[35]

Recent algorithmic bias legislation also fails to provide a useful definition. However, a common emphasis on "equality of opportunity" is clear in Western legal systems. For example, the various equal protection laws in the US share a common desire to promote *fairness* by limiting the impact of protected characteristics on significant life opportunities.

This theme also exists beyond the civil rights context. For example, the Federal Communications Commission's (FCC) now-defunct Fairness Doctrine rested upon equality of opportunity. Between 1949 to 1987, the FCC required news organizations to cover controversial issues of public importance in a manner that "fairly reflected differing viewpoints."[36] The FCC abolished this doctrine due to First Amendment concerns, not because of an evolving understanding of *fairness*. A universal definition of *fairness* continues to elude transatlantic legal systems; however, equality of opportunity will likely be central to a future legal definition of algorithmic fairness.

35 Federal Trade Commission v. Raladam Co., 283 U.S. 643, 648 (1931).
36 *FTC v. Raladam Co.*, 283 U.S. 643, 648 (1931).

In the technical context, there is increasing recognition that defining fairness in AI and ML applications is challenging, and that not all fairness metrics can be mathematically satisfied at the same time.[37]

One final impediment to bias identification to be considered is the tension between technical methods of identifying biases and legal limitations related to data privacy and antidiscrimination laws. Attempts to identify biases by auditing training data often require assessing different demographic groups to which individuals in the data belong. Collecting this data could run counter to antidiscrimination law protections that impose restrictions on retaining and using demographic data.[38]

It could also run counter to privacy laws in certain jurisdictions like the EU's General Data Protection Regulation (GDPR), which restricts the over-collection of "special category data" like race, or state privacy laws, including the California Consumer Privacy Act of 2018 (CCPA), the Virginia Consumer Data Protection Act (VCDPA), and Colorado Privacy Act (ColoPA), which govern the collection of personal data.[39]

37 Reuben Binns. "Fairness in Machine Learning: Lessons from Political Philosophy," *Proceedings of Machine Learning Research* 81:149-159, 2018. https://arxiv.org/abs/1712.03586.

38 Alice Xiang, "Reconciling Legal and Technical Approaches to Algorithmic Bias" (January 4, 2021). 88 Tennessee Law Review 649 (2021), https://ssrn.com/abstract=3650635.

39 *Ibid.* See article 9 of *EU General Data Protection Regulation (GDPR)*: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016, on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.

Minimizing AI-based risk in the future

Guidelines

There are several AI ethics guidelines that are emerging as authoritative standards across global organizations. Scholars have noted a growing proliferation and even consensus of AI ethics principles, although they recognize that there is little agreement on uniform methods of implementation.[40]

Among this sea of principles includes work by organizations like UNESCO and the OECD that has been widely hailed and adopted. In 2021, UNESCO produced their Recommendation on the *Ethics of Artificial Intelligence*, which focused on integrating ethics throughout the AI life cycle and may be particularly useful when considering how to ethically identify biases in AI systems throughout the development cycle. UNESCO's document includes a focus on fairness and nondiscrimination in the AI product itself and additionally compels organizations to "ensure inclusive access to and participation in the development of AI."[41] Following this principle could reinforce ethical practices as an application is being developed, including a focus on stakeholder engagement and participatory design that invites diverse perspectives during the design stages of the AI product life cycle.

OECD's 2019 Principles on Artificial Intelligence are another widely adopted set of guidelines.[42] These principles are aligned with best practices for ethical AI regarding inclusivity, fairness, explainability, security, and accountability. Given this harmony, future augmentations to OECD's principles can serve as a marker for how businesses might alter or develop their own principles.

Techniques

An area that will increase in regulatory relevance is the auditing of AI risk, bias, and fairness. Taking inspiration from the audit tradition standard in cybersecurity, AI auditing will be a key method by which to enforce regulation and establish expanding norms.

Both internal and external audits are increasingly used in the AI arena. Internal audits are to be deployed as part of the development process by the company creating the tool, and give an organization the chance to see everything from the beginning of a project through to examining the model design, training data, and stakeholders involved. This allows an

40 Jessica Fjeld *et al.*, "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI," SSRN Electronic Journal, 2020, https://doi.org/10.2139/ssrn.3518482.

41 UNESCO Recommendations on the Ethics of Artificial Intelligence, https://unesdoc.unesco.org/ark:/48223/pf0000380455 (2021).

42 Recommendation of the Council on Artificial Intelligence," OECD, May 24, 2019, https://one.oecd.org/document/C/MIN(2019)3/FINAL/en/pdf.

organization to change processes if biases are uncovered internally. As has been widely recognized, however, potential conflicts of interest could arise during an internal audit. Future regulation is likely to stress internal audits as recommended best practices, but regulators will lack the ability to easily verify and enforce compliance of such internal audits.

Thus, the practice of external auditing of AI is likely to increase in popularity. The *Blueprint for an AI Bill of Rights* focuses on ongoing evaluation and reporting, including by independent third-party evaluators.[43] External audits have the benefit of avoiding potential conflicts of interest, given that they are conducted by independent observers. Businesses are likely to face challenges when adopting this approach, however, because proprietary data, AI models, and other internal information cannot be easily shared to third parties without adequate safeguards being put in place.

As much as possible, businesses should try to anticipate auditability requirements and frame the data and models that they might be required to provide to third parties.

One additional issue with external auditing is that it is conducted after AI or models have already been built and deployed. Regulation might try to increase regular procedural audits as necessary endeavors for any AI products – a development that will require a response from organizations investing in AI capabilities.

Some auditability techniques have been proposed that focus on interweaving auditing with the AI development process. Google published an auditability framework that includes the collection of artifacts and documentation. These documents could, in turn, be used as evidence that materials for AI transparency have been prepared.[44] Whether a business uses this internal audit approach or one developed in-house, it would plausibly benefit from including audits throughout its AI development life cycle, particularly as material for an external audit to review.

Combining auditing with steps in the AI development cadence is an approach reflected in the NIST AI Risk Management Framework, which also includes steps on how to map, measure, manage, and govern AI risks as part of the AI development process from design through deployment.[45]

Tools

Tools for constructing fair AI range on a continuum of technical complexity. On one end of this continuum are technical features of AI models that increase their transparency and auditability. There is increasing consensus on the need for explainable AI (XAI) that focuses

43 "Blueprint for an AI Bill of Rights," The White House (The United States Government, October 4, 2022), https://www.whitehouse.gov/ostp/ai-bill-of-rights.

44 Timnit Gebru, *et al.* "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing," https://arxiv.org/abs/2001.00973.

45 "AI Risk Management Framework (AI RMF 1.0)," NIST, January 26, 2023, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

on technical avenues to enable humans to more easily interpret decisions made by models. There are also weighting criteria and other technical interventions to select for desirable fairness metrics in AI models. On the other end of the continuum are tools that guide the development process through conceptual steps and considerations that attack AI bias through the ethical reasoning of the developers themselves.

By enabling good ethical reasoning habits from the individuals building AI, tools can have positive outcomes on AI through the actions of their developers. These kinds of tools do not necessarily require the individuals using them to possess formal training in ethical or moral theories.

Expanding outside of technical approaches are checklists such as Microsoft's fairly comprehensive AI Fairness Checklist.[46] Checklists are a long-standing method for reinforcing procedural behaviors across industries, and organizations have started to independently create their own checklists for development.

While checklists are an incredibly popular tool for actioning AI ethics, it is worth noting their limitations. In many cases, checklists essentially constitute rote box-checking exercises, which can lead teams not to take them seriously. Technology ethics scholar Thilo Hagendorff calls for moving beyond checklists to more "situation sensitive ethical approach[es] based on virtues and personality dispositions, knowledge expansions, responsible autonomy, and freedom of action."[47] This suggestion speaks to more organizational- and cultural-related changes that businesses might implement in the product development process, including by examining how deadlines and other pressures of innovation interact with the time needed to consider ethics as part of development.

Ethical matrices that have been used in other fields have been adopted for the purpose of evaluating AI. For example, scholar Cathy O'Neil borrows the Mepham Ethical Matrix to evaluate how components of an AI application reflect core values.[48] Microsoft has also published its own matrix approach, Harms Modeling, which attempts to address harms that could arise from a given application across stakeholder groups.[49]

These standalone tools have clear positions in the design process into which they are plugged. Most broadly, tools such as the Markkula Center's Ethical Toolkit for Engineering/ Design Practice can function as companions to existing or new products that are being developed.[50] These tools encourage participation in ethical reasoning processes among development teams and the leadership who use them in daily project management activities, which might be a good starting place for systematization of AI ethics principles.

46 Madaio et al., "AI Fairness Checklist," Microsoft, accessed March 25, 2022, https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4t6dA.

47 Thilo Hagendorff, "The Ethics of AI Ethics: An Evaluation of Guidelines," Minds and Machines 30, no. 1 (January 2020): pp. 99-120, https://doi.org/10.1007/s11023-020-09517-8.

48 S. Matthew Liao, Hanna Gunn, and Cathy O'Neil, "Near-Term Artificial Intelligence and the Ethical Matrix," in *Ethics of Artificial Intelligence* (New York, NY: Oxford University Press, 2020), pp. 237-270.

49 Harms Modeling - Azure Application Architecture Guide," Harms Modeling - Azure Application Architecture Guide | Microsoft Docs, November 8, 2021, https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling.

50 Shannon Vallor, "An Ethical Toolkit for Engineering/Design Practice," Markkula Center for Applied Ethics (Santa Clara University), accessed March 25, 2022, https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit.

Conclusion

Efforts are underway to codify interpretations of bias pertaining to algorithms into law, and practitioners developing AI tools will need to comply with any such developments. The proposed law on AI in Europe, though still subject to review, stands to have a significant impact on the development of most AI tools if passed, given its horizontal nature. However, its proportional and risk-based approach to regulating AI systems also serves as a clue for future possible regulatory moves in global jurisdictions.

Teams should aim to implement rigorous processes to assess how AI is being developed, as well as its possible impacts on various populations.

As more investment is directed towards sector-specific regulatory bodies to target algorithmic bias, we are likely to see the growing influence of governmental agencies like the EEOC and NIST in setting standards for regulation. NIST's 2023 Risk Management Framework will likely become a widely referenced playbook for organizations looking to adopt flexible and thorough AI governance and standards in the coming years. The *Blueprint for an AI Bill of Rights* in the US suggests that guidelines, techniques, and tools will continue to play an important role in the ensuring of safe and effective AI systems free of algorithmic discrimination.[51]

AI practitioners can strive to apprehend regulatory developments by developing XAI applications and incorporating technical audits in the development life cycle of their tools. Project teams and management can also strategically evaluate AI using risk matrices and checklists. These teams should aim to implement rigorous processes to assess how AI is being developed, as well as its possible impacts on various populations.

Emerging legislation has not settled on unified requirements for identifying and mitigating biases in AI. However, the growing body of guidelines, techniques, and tools recommended in this report offer a pathway for organizations looking to build fairer AI tools that are evergreen in the face of changing legal landscapes.

51 "Blueprint for an AI Bill of Rights," The White House (The United States Government, October 4, 2022), https://www.whitehouse.gov/ostp/ai-bill-of-rights.

Authors

Victoria Matthews is an artificial intelligence policy specialist who researches and advises on topics including global AI regulation; responsible data governance; and fairness, accountability, and transparency in AI/ML applications. She works with AI researchers across North America, Asia, and Europe on AI ethics problems. Matthews has experience leading AI ethics advocacy and education projects, as well as developing tools and processes for the ethical review of AI systems.

Matthews received her master's degree in Technology Ethics and Policy from Duke University and received her bachelor's degree from the University of Oxford, where she founded Moral IT, an organization that explores legal, ethical, and policy solutions for more responsible technologies.

Matt Murphy is a technology strategy and trustworthy AI consultant. He helps companies implement frameworks and governance for responsible and ethical AI and other emerging technologies. Murphy has advised executive clients at a myriad of private companies and government agencies, including national security organizations in the US intelligence community, on implementing robotics, AI/ML, and other capabilities. His other work has focused on technology scouting, futures analysis, and software development.

Murphy received his master's degree in Technology Ethics and Policy from Duke University. His research focuses on methods and tools that translate ethical principles into practice for developers and technologists. He does work on fairness and explainability in AI/ML applications, surveillance and privacy, data ownership, and online platform accountability.

Credits

This paper was published in conjunction with Thomson Reuters and the Thomson Reuters Institute.

Thomson Reuters

Thomson Reuters is a leading provider of business information services. Our products include highly specialized information-enabled software and tools for legal, tax, accounting and compliance professionals combined with the world's most global news service — Reuters.

For more information on Thomson Reuters, visit **tr.com** and for the latest world news, **reuters.com**.

Thomson Reuters Institute

The Thomson Reuters Institute brings together people from across the legal, corporate, tax & accounting and government communities to ignite conversation and debate, make sense of the latest events and trends and provide essential guidance on the opportunities and challenges facing their world today. As the dedicated thought leadership arm of Thomson Reuters, our content spans blog commentaries, industry-leading data sets, informed analyses, interviews with industry leaders, videos, podcasts and world-class events that deliver keen insight into a dynamic business landscape.

Visit **thomsonreuters.com/institute** for more details.

THOMSON REUTERS®