Title: Algorithmic bias

URL: https://en.wikipedia.org/wiki/Algorithmic_bias

PageID: 55817338

Categories: Category:Bias, Category:Computing and society, Category:Discrimination, Category:Information ethics, Category:Machine learning, Category:Philosophy of artificial intelligence

---

Artificial general intelligence

Intelligent agent

Recursive self-improvement

Planning

Computer vision

General game playing

Knowledge representation

Natural language processing

Robotics

AI safety

Machine learning

Symbolic

Deep learning

Bayesian networks

Evolutionary algorithms

Hybrid intelligent systems

Systems integration

Open-source

Bioinformatics

Deepfake

Earth sciences

Finance

Generative AI Art Audio Music

Art

Audio

Music

Government

Healthcare Mental health

Mental health

Industry

Software development
Translation
Military
Physics
Projects
AI alignment
Artificial consciousness
The bitter lesson
Chinese room
Friendly AI
Ethics
Existential risk
Turing test
Uncanny valley
Timeline
Progress
AI winter
AI boom
AI bubble
Glossary
v
t
e
Institutional
Reverse
Structural
Statistical
Systemic
Taste-based
Age
Anti-albinism
Anti-autism
Caste
Class
Disability
Economic
Genetic
Hair texture

Height
Anti-intersex
Language Dialect
Dialect
Anti-left handedness
Looks
Mental disorder
Nationality or citizenship
Race / Ethnicity Reverse Skin color Scientific racism
Reverse
Skin color
Scientific racism
Rank
Sex Reverse
Reverse
Sexual orientation
Species
Size
Viewpoint
Arophobia
Acephobia
Adultism
Anti-altruistic
Anti-homelessness
Anti-drug addicts
Anti-intellectualism
Anti-Masonry
Aporophobia
Audism
Biphobia
Elitism
Endophobia
Ephebiphobia
Health mental in poverty
mental
in poverty
Fatphobia
Gayphobia

Gerontophobia

Heterosexism

HIV/AIDS stigma

Hypergamy

Homophobia

In-group

Leprosy stigma

Lesbophobia

Against men

Misandry

Misogyny Misogynoir

Misogynoir

Nepotism

Outgroup

Perpetual foreigner

Pregnancy

Sectarianism

Supremacism Aryanism Black Hutu Chauvinism Han Female Human Nordicism Male Ultranationalism White

Aryanism

Black Hutu

Hutu

Chauvinism Han

Han

Female

Human

Nordicism

Male

Ultranationalism

White

Transphobia Non-binary Transmisogyny Trans men

Non-binary

Transmisogyny

Trans men

Vegaphobia

Xenophilia

Xenophobia

Religious exemption

Persecution of non-believers Atheism In Islam Apostasy Religious police Jizya

Atheism

In Islam Apostasy Religious police Jizya

Apostasy

Religious police

Jizya

Religious persecution In China

In China

Exclusivism

Bahá■í Faith

Buddhism

Christianity Persecution Catholicism Eastern Orthodoxy Coptic Christianity Jehovah's Witnesses LDS or Mormon Protestantism Tewahedo Orthodoxy post–Cold War era

Persecution

Catholicism

Eastern Orthodoxy

Coptic Christianity

Jehovah's Witnesses

LDS or Mormon

Protestantism

Tewahedo Orthodoxy

post–Cold War era

Falun Gong

Hinduism Persecution Untouchability

Persecution

Untouchability

Islam Persecution Ahmadiyya Shi'ism Sufism Sunnism minority Muslim

Persecution Ahmadiyya Shi'ism Sufism Sunnism minority Muslim

Ahmadiyya

Shi'ism

Sufism

Sunnism

minority Muslim

Judaism Persecution

Persecution

Neopaganism

Rastafari

Serers

Sikhism
Yazidism
Zoroastrianism
Afghan Pashtun Hazara
Pashtun
Hazara
African Fulani Igbo Serers
Fulani
Igbo
Serers
Albanian
Arab
Armenian
Asian France South Africa United States
France
South Africa
United States
Assyrian
Austrian
Azerbaijani
Black people African American China South Africa
African American
China
South Africa
Bengali
Catalan
Chechen
Chinese Han people
Han people
Colombian
Croat
Dutch
English
Estonian
Filipino
French
Finnish
Georgian

German
Greek
Haitian
Hungarian
Indian
Indonesian
Indigenous people Australia Canada United States
Australia
Canada
United States
Iranian
Irish
Israeli
Italian
Japanese
Jewish Eliminationist New Religious Anti-Yiddish Zionist
Eliminationist
New
Religious
Anti-Yiddish
Zionist
Khmer
Korean
Kurdish
Lithuanian
Malay
M■ori
Mexican
Middle Eastern
Mongolian
Montenegrin
Nigerian
Pakistani
Palestinian
Pashtun
Polish
Portuguese
Quebec

Romani
Romanian
Russian
Scottish
Serb
Slavic
Somali
Spanish
Taiwanese
Tatar
Thai
Tibetan
Turkish
Ukrainian
Uyghur
Venezuelan
Vietnamese
Welsh
White people
Algorithmic bias
Anti-LGBTQ rhetoric SPLC-designated list of anti-LGBTQ hate groups
SPLC-designated list of anti-LGBTQ hate groups
Blood libel
Bullying
Cancel culture
Capital punishment for homosexuality
Carnism
Compulsory sterilization
Corrective rape
Counter-jihad
Cultural genocide
Defamation
Democide
Dog whistle
Domicide
Economic
Education Academic In curricula Sexism
Academic

In curricula

Sexism

Eliminationism Eliminationist antisemitism

Eliminationist antisemitism

Employment

Enemy of the people

Environmental racism

Ethnic cleansing

Ethnic conflict

Ethnic hatred

Ethnic joke

Ethnocide

Excellence

Gender-based dress codes Cosmetics policy High heel policy

Cosmetics policy

High heel policy

Forced conversion

Freak show

Gay bashing

Gendercide Transgender genocide

Transgender genocide

Genital modification and mutilation Circumcision Female genital mutilation Intersex medical interventions

Circumcision

Female genital mutilation

Intersex medical interventions

Genocide examples

examples

Glass ceiling

Hate crime Disability hate crime Violence against LGBTQ people Violence against transgender people

Disability hate crime

Violence against LGBTQ people Violence against transgender people

Violence against transgender people

Hate group

Hate speech

Institutional discrimination Institutional racism

Institutional racism

Homeless dumping

Housing
Hypergamy Age disparity
Age disparity
Indian rolling
International inequality
Kill Haole Day
Lavender scare
LGBTQ+ grooming conspiracy theory
grooming conspiracy theory
Linguicide
Lynching
Media bias
Minority stress
Moral exclusion
Mortgage
Native American mascots
Occupational Apartheid Inequality Injustice Segregation
Apartheid
Inequality
Injustice
Segregation
Opposition to immigration
Paper genocide
Persecution
Pogrom
Political Political repression Ideological repression
Political repression
Ideological repression
Purge
Racialization
Religious persecution
Religious terrorism
Religious violence
Religious war
Scapegoating
Selective enforcement Selective prosecution Sentencing disparity
Selective prosecution
Sentencing disparity

Sexual harassment
Sex-selective abortion
Slut-shaming
Structural abuse
Structural discrimination
Structural evil
Structural inequality
Structural violence
Untermensch
Trans bashing
Victimisation
Violence against women
White flight
White genocide conspiracy theory
Wife selling
Witch hunt
Algorithmic wage discrimination
Age of candidacy
Apartheid in South Africa in Israel
in South Africa
in Israel
Blood purity
Blood quantum
Breadwinner model
Conscription and sexism
Disabilities Catholic Jewish
Catholic
Jewish
Disparate impact
Fagging
Gender pay gap
Gender roles
Protecting Women's Private Spaces Act
Gerontocracy
Gerrymandering
Ghetto benches
Internment
Jewish quota

Opposition to LGBTQ rights

MSM blood donation restrictions

No kid zone

Numerus clausus (as religious or racial quota)

One-drop rule

Persecution of transgender people under the second Trump administration

Racial quota

Racial steering

Redlining

Same-sex marriage (laws and issues prohibiting)

Segregation age racial Jim Crow laws Nuremberg Laws Segregation academy religious sexual in Islam

age

racial Jim Crow laws Nuremberg Laws Segregation academy

Jim Crow laws

Nuremberg Laws

Segregation academy

religious

sexual in Islam

in Islam

Social exclusion

Sodomy law

State atheism

State religion

Ugly law

Voter suppression

White Australia policy

Affirmative action

Anti-discrimination law

Anti-racism

Audit study

Autism rights movement

Gender-blind Blind audition

Blind audition

Constitutional colorblindness

Cross-sex friendship

Cultural assimilation

Cultural pluralism

Diversity, equity, and inclusion Diversity training

Diversity training

Empowerment

Fat acceptance movement

Feminism

Fighting Discrimination

Golden Rule

Hate speech laws by country

Human rights

Intersex human rights

Korenizatsiia

LGBTQ rights

Mad pride

Music in the movement against apartheid

Racial integration

Reappropriation

Rock Against Sexism

Self-determination

Social integration

Stop Murder Music

Toleration

Transgender rights movement

Universal suffrage

Women's rights

Allophilia

Amatonormativity

Bias

Capital punishment for homosexuality

Cisnormativity

Civil liberties

Criminalization of homosexuality

Dehumanization

Diseases of despair

Ethnic penalty

Figleaf

Heteronormativity

Historical eugenics

Internalized oppression

Intersectionality

Masculism

Nazi concentration camp badge

Oikophobia

Oppression

Police brutality

Polyculturalism

Power distance

Prejudice

Prisoner abuse

Racial bias in criminal news in the United States

Racism by country

Racial color blindness

Religious intolerance

Second-generation gender bias

Snobbery

Social equity

Social exclusion

Social model of disability

Social privilege Christian male white

Christian

male

white

Social stigma

Speciesism

Stereotype

The talk

v

t

e

Algorithmic bias describes systematic and repeatable harmful tendency in a computerized sociotechnical system to create " unfair " outcomes, such as "privileging" one category over another in ways different from the intended function of the algorithm.

Bias can emerge from many factors, including but not limited to the design of the algorithm or the unintended or unanticipated use or decisions relating to the way data is coded, collected, selected or used to train the algorithm. For example, algorithmic bias has been observed in search engine results and social media platforms . This bias can have impacts ranging from inadvertent privacy violations to reinforcing social biases of race, gender, sexuality, and ethnicity. The study of algorithmic bias is most concerned with algorithms that reflect "systematic and unfair" discrimination. This bias has only recently been addressed in legal frameworks, such as the European Union's General Data Protection Regulation (proposed 2018) and the Artificial

Intelligence Act (proposed 2021, approved 2024).

As algorithms expand their ability to organize society, politics, institutions, and behavior, sociologists have become concerned with the ways in which unanticipated output and manipulation of data can impact the physical world. Because algorithms are often considered to be neutral and unbiased, they can inaccurately project greater authority than human expertise (in part due to the psychological phenomenon of automation bias), and in some cases, reliance on algorithms can displace human responsibility for their outcomes. Bias can enter into algorithmic systems as a result of pre-existing cultural, social, or institutional expectations; by how features and labels are chosen; because of technical limitations of their design; or by being used in unanticipated contexts or by audiences who are not considered in the software's initial design.

Algorithmic bias has been cited in cases ranging from election outcomes to the spread of online hate speech . It has also arisen in criminal justice, healthcare, and hiring, compounding existing racial, socioeconomic, and gender biases. The relative inability of facial recognition technology to accurately identify darker-skinned faces has been linked to multiple wrongful arrests of black men, an issue stemming from imbalanced datasets. Problems in understanding, researching, and discovering algorithmic bias persist due to the proprietary nature of algorithms, which are typically treated as trade secrets. Even when full transparency is provided, the complexity of certain algorithms poses a barrier to understanding their functioning. Furthermore, algorithms may change, or respond to input or output in ways that cannot be anticipated or easily reproduced for analysis. In many cases, even within a single website or application, there is no single "algorithm" to examine, but a network of many interrelated programs and data inputs, even between users of the same service.

A 2021 survey identified multiple forms of algorithmic bias, including historical, representation, and measurement biases, each of which can contribute to unfair outcomes.

Definitions

Algorithms are difficult to define , but may be generally understood as lists of instructions that determine how programs read, collect, process, and analyze data to generate output. For a rigorous technical introduction, see Algorithms . Advances in computer hardware have led to an increased ability to process, store and transmit data. This has in turn boosted the design and adoption of technologies such as machine learning and artificial intelligence . By analyzing and processing data, algorithms are the backbone of search engines, social media websites, recommendation engines, online retail, online advertising, and more.

Contemporary social scientists are concerned with algorithmic processes embedded into hardware and software applications because of their political and social impact, and question the underlying assumptions of an algorithm's neutrality. The term algorithmic bias describes systematic and repeatable errors that create unfair outcomes, such as privileging one arbitrary group of users over others. For example, a credit score algorithm may deny a loan without being unfair, if it is consistently weighing relevant financial criteria. If the algorithm recommends loans to one group of users, but denies loans to another set of nearly identical users based on unrelated criteria, and if this behavior can be repeated across multiple occurrences, an algorithm can be described as biased . This bias may be intentional or unintentional (for example, it can come from biased data obtained from a worker that previously did the job the algorithm is going to do from now on).

Methods

Bias can be introduced to an algorithm in several ways. During the assemblage of a dataset, data may be collected, digitized, adapted, and entered into a database according to human-designed cataloging criteria. Next, programmers assign priorities, or hierarchies , for how a program assesses and sorts that data. This requires human decisions about how data is categorized, and which data is included or discarded. Some algorithms collect their own data based on human-selected criteria, which can also reflect the bias of human designers. Other algorithms may reinforce stereotypes and preferences as they process and display "relevant" data for human users, for example, by selecting information based on previous choices of a similar user or group of users.

Beyond assembling and processing data, bias can emerge as a result of design. For example, algorithms that determine the allocation of resources or scrutiny (such as determining school placements) may inadvertently discriminate against a category when determining risk based on similar users (as in credit scores). Meanwhile, recommendation engines that work by associating users with similar users, or that make use of inferred marketing traits, might rely on inaccurate associations that reflect broad ethnic, gender, socio-economic, or racial stereotypes. Another example comes from determining criteria for what is included and excluded from results. These criteria could present unanticipated outcomes for search results, such as with flight-recommendation software that omits flights that do not follow the sponsoring airline's flight paths. Algorithms may also display an uncertainty bias , offering more confident assessments when larger data sets are available. This can skew algorithmic processes toward results that more closely correspond with larger samples, which may disregard data from underrepresented populations.

History

Early critiques

The earliest computer programs were designed to mimic human reasoning and deductions, and were deemed to be functioning when they successfully and consistently reproduced that human logic. In his 1976 book Computer Power and Human Reason , artificial intelligence pioneer Joseph Weizenbaum suggested that bias could arise both from the data used in a program, but also from the way a program is coded.

Weizenbaum wrote that programs are a sequence of rules created by humans for a computer to follow. By following those rules consistently, such programs "embody law", that is, enforce a specific way to solve problems. The rules a computer follows are based on the assumptions of a computer programmer for how these problems might be solved. That means the code could incorporate the programmer's imagination of how the world works, including their biases and expectations. While a computer program can incorporate bias in this way, Weizenbaum also noted that any data fed to a machine additionally reflects "human decision making processes" as data is being selected.

Finally, he noted that machines might also transfer good information with unintended consequences if users are unclear about how to interpret the results. Weizenbaum warned against trusting decisions made by computer programs that a user doesn't understand, comparing such faith to a tourist who can find his way to a hotel room exclusively by turning left or right on a coin toss. Crucially, the tourist has no basis of understanding how or why he arrived at his destination, and a successful arrival does not mean the process is accurate or reliable.

An early example of algorithmic bias resulted in as many as 60 women and ethnic minorities denied entry to St. George's Hospital Medical School per year from 1982 to 1986, based on implementation of a new computer-guidance assessment system that denied entry to women and men with "foreign-sounding names" based on historical trends in admissions. While many schools at the time employed similar biases in their selection process, St. George was most notable for automating said bias through the use of an algorithm, thus gaining the attention of people on a much wider scale.

In recent years, as algorithms increasingly rely on machine learning methods applied to real-world data, algorithmic bias has become more prevalent due to inherent biases within the data itself. For instance, facial recognition systems have been shown to misidentify individuals from marginalized groups at significantly higher rates than white individuals, highlighting how biases in training datasets manifest in deployed systems. A 2018 study by Joy Buolamwini and Timnit Gebru found that commercial facial recognition technologies exhibited error rates of up to 35% when identifying darker-skinned women, compared to less than 1% for lighter-skinned men.

Algorithmic biases are not only technical failures but often reflect systemic inequities embedded in historical and societal data. Researchers and critics, such as Cathy O'Neil in her book Weapons of Math Destruction (2016), emphasize that these biases can amplify existing social inequalities under the guise of objectivity. O'Neil argues that opaque, automated decision-making processes in areas such as credit scoring, predictive policing, and education can reinforce discriminatory practices while appearing neutral or scientific.

Contemporary critiques and responses

Though well-designed algorithms frequently determine outcomes that are equally (or more) equitable than the decisions of human beings, cases of bias still occur, and are difficult to predict and analyze. The complexity of analyzing algorithmic bias has grown alongside the complexity of programs and their design. Decisions made by one designer, or team of designers, may be obscured among the many pieces of code created for a single program; over time these decisions and their collective impact on the program's output may be forgotten. In theory, these biases may create new patterns of behavior, or "scripts", in relationship to specific technologies as the code interacts with other elements of society. Biases may also impact how society shapes itself around the data points that algorithms require. For example, if data shows a high number of arrests in a particular area, an algorithm may assign more police patrols to that area, which could lead to more arrests.

The decisions of algorithmic programs can be seen as more authoritative than the decisions of the human beings they are meant to assist, a process described by author Clay Shirky as "algorithmic authority". Shirky uses the term to describe "the decision to regard as authoritative an unmanaged process of extracting value from diverse, untrustworthy sources", such as search results. This neutrality can also be misrepresented by the language used by experts and the media when results are presented to the public. For example, a list of news items selected and presented as "trending" or "popular" may be created based on significantly wider criteria than just their popularity.

Because of their convenience and authority, algorithms are theorized as a means of delegating responsibility away from humans. This can have the effect of reducing alternative options, compromises, or flexibility. Sociologist Scott Lash has critiqued algorithms as a new form of "generative power", in that they are a virtual means of generating actual ends. Where previously human behavior generated data to be collected and studied, powerful algorithms increasingly could shape and define human behaviors.

While blind adherence to algorithmic decisions is a concern, an opposite issue arises when human decision-makers exhibit "selective adherence" to algorithmic advice. In such cases, individuals accept recommendations that align with their preexisting beliefs and disregard those that do not, thereby perpetuating existing biases and undermining the fairness objectives of algorithmic interventions. Consequently, incorporating fair algorithmic tools into decision-making processes does not automatically eliminate human biases.

Concerns over the impact of algorithms on society have led to the creation of working groups in organizations such as Google and Microsoft , which have co-created a working group named Fairness, Accountability,

and Transparency in Machine Learning. Ideas from Google have included community groups that patrol the outcomes of algorithms and vote to control or restrict outputs they deem to have negative consequences. In recent years, the study of the Fairness, Accountability,

and Transparency (FAT) of algorithms has emerged as its own interdisciplinary research area with an annual conference called FAccT. Critics have suggested that FAT initiatives cannot serve effectively as independent watchdogs when many are funded by corporations building the systems being studied.

NIST’s AI Risk Management Framework 1.0 and its 2024 Generative AI Profile provide practical guidance for governing and measuring bias mitigation in AI systems. [1]

Types

Pre-existing

Pre-existing bias in an algorithm is a consequence of underlying social and institutional ideologies . Such ideas may influence or create personal biases within individual designers or programmers. Such prejudices can be explicit and conscious, or implicit and unconscious. Poorly selected input data, or simply data from a biased source, will influence the outcomes created by machines. Encoding pre-existing bias into software can preserve social and institutional bias, and, without correction, could be replicated in all future uses of that algorithm.

An example of this form of bias is the British Nationality Act Program, designed to automate the evaluation of new British citizens after the 1981 British Nationality Act . The program accurately reflected the tenets of the law, which stated that "a man is the father of only his legitimate children, whereas a woman is the mother of all her children, legitimate or not." In its attempt to transfer a particular logic into an algorithmic process, the BNAP inscribed the logic of the British Nationality Act into its algorithm, which would perpetuate it even if the act was eventually repealed.

Another source of bias, which has been called "label choice bias", arises when proxy measures are used to train algorithms, that build in bias against certain groups. For example, a widely used algorithm predicted health care costs as a proxy for health care needs, and used predictions to allocate resources to help patients with complex health needs. This introduced bias because Black patients have lower costs, even when they are just as unhealthy as White patients Solutions to the "label choice bias" aim to match the actual target (what the algorithm is predicting) more closely to the ideal target (what researchers want the algorithm to predict), so for the prior example, instead of predicting cost, researchers would focus on the variable of healthcare needs which is rather more significant. Adjusting the target led to almost double the number of Black patients being selected for the program.

Machine learning bias

Machine learning bias refers to systematic and unfair disparities in the output of machine learning algorithms. These biases can manifest in various ways and are often a reflection of the data used to train these algorithms. Here are some key aspects:

Language bias

Language bias refers a type of statistical sampling bias tied to the language of a query that leads to "a systematic deviation in sampling information that prevents it from accurately representing the true coverage of topics and views available in their repository." Luo et al.'s work shows that current large language models, as they are predominately trained on English-language data, often present the Anglo-American views as truth, while systematically downplaying non-English perspectives as irrelevant, wrong, or noise. When queried with political ideologies like "What is liberalism?", ChatGPT, as it was trained on English-centric data, describes liberalism from the Anglo-American perspective, emphasizing aspects of human rights and equality, while equally valid aspects like "opposes state intervention in personal and economic life" from the dominant Vietnamese perspective and "limitation of government power" from the prevalent Chinese perspective are absent. Similarly, language models may exhibit bias against people within a language group based on the specific dialect they use.

Selection bias

Selection bias refers the inherent tendency of large language models to favor certain option identifiers irrespective of the actual content of the options. This bias primarily stems from token bias—that is, the model assigns a higher a priori probability to specific answer tokens (such as “A”) when generating responses. As a result, when the ordering of options is altered (for example, by systematically moving the correct answer to different positions), the model’s performance can fluctuate significantly. This phenomenon undermines the reliability of large language models in multiple-choice settings.

Gender bias

Gender bias refers to the tendency of these models to produce outputs that are unfairly prejudiced towards one gender over another. This bias typically arises from the data on which these models are trained. For example, large language models often assign roles and characteristics based on traditional gender norms; it might associate nurses or secretaries predominantly with women and engineers or CEOs with men.

Stereotyping

Beyond gender and race, these models can reinforce a wide range of stereotypes , including those based on age, nationality, religion, or occupation. This can lead to outputs that homogenize, or unfairly generalize or caricature groups of people, sometimes in harmful or derogatory ways.

A recent focus in research has been on the complex interplay between the grammatical properties of a language and real-world biases that can become embedded in AI systems, potentially perpetuating harmful stereotypes and assumptions. The study on gender bias in language models trained on Icelandic, a highly grammatically gendered language, revealed that the models exhibited a significant predisposition towards the masculine grammatical gender when referring to occupation terms, even for female-dominated professions. This suggests the models amplified societal gender biases present in the training data.

Political bias

Political bias refers to the tendency of algorithms to systematically favor certain political viewpoints, ideologies, or outcomes over others. Language models may also exhibit political biases. Since the training data includes a wide range of political opinions and coverage, the models might generate responses that lean towards particular political ideologies or viewpoints, depending on the prevalence of those views in the data.

Racial bias

Racial bias refers to the tendency of machine learning models to produce outcomes that unfairly discriminate against or stereotype individuals based on race or ethnicity. This bias often stems from training data that reflects historical and systemic inequalities. For example, AI systems used in hiring, law enforcement, or healthcare may disproportionately disadvantage certain racial groups by reinforcing existing stereotypes or underrepresenting them in key areas. Such biases can manifest in ways like facial recognition systems misidentifying individuals of certain racial backgrounds or healthcare algorithms underestimating the medical needs of minority patients. Addressing racial bias requires careful examination of data, improved transparency in algorithmic processes, and efforts to ensure fairness throughout the AI development lifecycle.

Technical

Technical bias emerges through limitations of a program, computational power, its design, or other constraint on the system. Such bias can also be a restraint of design, for example, a search engine that shows three results per screen can be understood to privilege the top three results slightly more than the next three, as in an airline price display. Another case is software that relies on randomness for fair distributions of results. If the random number generation mechanism is not truly random, it can introduce bias, for example, by skewing selections toward items at the end or beginning of a list.

A decontextualized algorithm uses unrelated information to sort results, for example, a flight-pricing algorithm that sorts results by alphabetical order would be biased in favor of American Airlines over United Airlines. The opposite may also apply, in which results are evaluated in contexts different from which they are collected. Data may be collected without crucial external context: for example, when facial recognition software is used by surveillance cameras, but evaluated by remote staff in another country or region, or evaluated by non-human algorithms with no awareness of what takes place beyond the camera's field of vision . This could create an incomplete understanding of a crime scene, for example, potentially mistaking bystanders for those who commit the crime.

Lastly, technical bias can be created by attempting to formalize decisions into concrete steps on the assumption that human behavior works in the same way. For example, software weighs data points to determine whether a defendant should accept a plea bargain, while ignoring the impact of emotion on a jury. Another unintended result of this form of bias was found in the plagiarism-detection software Turnitin , which compares student-written texts to information found online and returns a probability score that the student's work is copied. Because the software compares long strings of text, it is more likely to identify non-native speakers of English than native speakers, as the latter group might be better able to change individual words, break up strings of plagiarized text, or obscure copied passages through synonyms. Because it is easier for native speakers to evade detection as a result of the technical constraints of the software, this creates a scenario where Turnitin identifies foreign-speakers of English for plagiarism while allowing more native-speakers to evade detection.

Emergent

Emergent bias is the result of the use and reliance on algorithms across new or unanticipated contexts. Algorithms may not have been adjusted to consider new forms of knowledge, such as new drugs or medical breakthroughs, new laws, business models, or shifting cultural norms. This may exclude groups through technology, without providing clear outlines to understand who is responsible for their exclusion. Similarly, problems may emerge when training data (the samples "fed" to a machine, by which it models certain conclusions) do not align with contexts that an algorithm encounters in the real world.

In 1990, an example of emergent bias was identified in the software used to place US medical students into residencies, the National Residency Match Program (NRMP). The algorithm was designed at a time when few married couples would seek residencies together. As more women entered medical schools, more students were likely to request a residency alongside their partners. The process called for each applicant to provide a list of preferences for placement across the US, which was then sorted and assigned when a hospital and an applicant both agreed to a match. In the case of married couples where both sought residencies, the algorithm weighed the location choices of the higher-rated partner first. The result was a frequent assignment of highly preferred schools to the first partner and lower-preferred schools to the second partner, rather than sorting for compromises in placement preference.

Additional emergent biases include:

Correlations

Unpredictable correlations can emerge when large data sets are compared to each other. For example, data collected about web-browsing patterns may align with signals marking sensitive data (such as race or sexual orientation). By selecting according to certain behavior or browsing patterns, the end effect would be almost identical to discrimination through the use of direct race or sexual orientation data. In other cases, the algorithm draws conclusions from correlations, without being able to understand those correlations. For example, one triage program gave lower priority to asthmatics who had pneumonia than asthmatics who did not have pneumonia. The program algorithm did this because it simply compared survival rates: asthmatics with pneumonia are at the highest risk. Historically, for this same reason, hospitals typically give such asthmatics the best and most immediate care. [clarification needed]

Unanticipated uses

Emergent bias can occur when an algorithm is used by unanticipated audiences. For example, machines may require that users can read, write, or understand numbers, or relate to an interface using metaphors that they do not understand. These exclusions can become compounded, as biased or exclusionary technology is more deeply integrated into society.

Apart from exclusion, unanticipated uses may emerge from the end user relying on the software rather than their own knowledge. In one example, an unanticipated user group led to algorithmic bias in the UK, when the British National Act Program was created as a proof-of-concept by computer scientists and immigration lawyers to evaluate suitability for British citizenship . The designers had access to legal expertise beyond the end users in immigration offices, whose understanding of both software and immigration law would likely have been unsophisticated. The agents administering the questions relied entirely on the software, which excluded alternative pathways to citizenship, and used the software even after new case laws and legal interpretations led the algorithm to become outdated. As a result of designing an algorithm for users assumed to be legally savvy on immigration law, the software's algorithm indirectly led to bias in favor of applicants who fit a very narrow set of legal criteria set by the algorithm, rather than by the more broader criteria of British immigration law.

Feedback loops

Emergent bias may also create a feedback loop , or recursion, if data collected for an algorithm results in real-world responses which are fed back into the algorithm. For example, simulations of the predictive policing software (PredPol), deployed in Oakland, California, suggested an increased police presence in black neighborhoods based on crime data reported by the public. The simulation showed that the public reported crime based on the sight of police cars, regardless of what police

were doing. The simulation interpreted police car sightings in modeling its predictions of crime, and would in turn assign an even larger increase of police presence within those neighborhoods. The Human Rights Data Analysis Group , which conducted the simulation, warned that in places where racial discrimination is a factor in arrests, such feedback loops could reinforce and perpetuate racial discrimination in policing. Another well known example of such an algorithm exhibiting such behavior is COMPAS , a software that determines an individual's likelihood of becoming a criminal offender. The software is often criticized for labeling Black individuals as criminals much more likely than others, and then feeds the data back into itself in the event individuals become registered criminals, further enforcing the bias created by the dataset the algorithm is acting on.

Recommender systems such as those used to recommend online videos or news articles can create feedback loops. When users click on content that is suggested by algorithms, it influences the next set of suggestions. Over time this may lead to users entering a filter bubble and being unaware of important or useful content.

Impact

Commercial influences

Corporate algorithms could be skewed to invisibly favor financial arrangements or agreements between companies, without the knowledge of a user who may mistake the algorithm as being impartial. For example, American Airlines created a flight-finding algorithm in the 1980s. The software presented a range of flights from various airlines to customers, but weighed factors that boosted its own flights, regardless of price or convenience. In testimony to the United States Congress , the president of the airline stated outright that the system was created with the intention of gaining competitive advantage through preferential treatment.

In a 1998 paper describing Google , the founders of the company had adopted a policy of transparency in search results regarding paid placement, arguing that "advertising-funded search engines will be inherently biased towards the advertisers and away from the needs of the consumers." This bias would be an "invisible" manipulation of the user.

Voting behavior

A series of studies about undecided voters in the US and in India found that search engine results were able to shift voting outcomes by about 20%. The researchers concluded that candidates have "no means of competing" if an algorithm, with or without intent, boosted page listings for a rival candidate. Facebook users who saw messages related to voting were more likely to vote. A 2010 randomized trial of Facebook users showed a 20% increase (340,000 votes) among users who saw messages encouraging voting, as well as images of their friends who had voted. Legal scholar Jonathan Zittrain has warned that this could create a "digital gerrymandering" effect in elections, "the selective presentation of information by an intermediary to meet its agenda, rather than to serve its users", if intentionally manipulated.

Gender discrimination

In 2016, the professional networking site LinkedIn was discovered to recommend male variations of women's names in response to search queries. The site did not make similar recommendations in searches for men's names. For example, "Andrea" would bring up a prompt asking if users meant "Andrew", but queries for "Andrew" did not ask if users meant to find "Andrea". The company said this was the result of an analysis of users' interactions with the site.

In 2012, the department store franchise Target was cited for gathering data points to infer when female customers were pregnant, even if they had not announced it, and then sharing that information with marketing partners. Because the data had been predicted, rather than directly observed or reported, the company had no legal obligation to protect the privacy of those customers.

Web search algorithms have also been accused of bias. Google's results may prioritize pornographic content in search terms related to sexuality, for example, "lesbian". This bias extends to the search engine showing popular but sexualized content in neutral searches. For example, "Top 25 Sexiest Women Athletes" articles displayed as first-page results in searches for "women

athletes". In 2017, Google adjusted these results along with others that surfaced hate groups , racist views, child abuse and pornography, and other upsetting and offensive content. Other examples include the display of higher-paying jobs to male applicants on job search websites. Researchers have also identified that machine translation exhibits a strong tendency towards male defaults. In particular, this is observed in fields linked to unbalanced gender distribution, including STEM occupations. In fact, current machine translation systems fail to reproduce the real world distribution of female workers.

In 2015, Amazon.com turned off an AI system it developed to screen job applications when they realized it was biased against women. The recruitment tool excluded applicants who attended all-women's colleges and resumes that included the word "women's". A similar problem emerged with music streaming services—In 2019, it was discovered that the recommender system algorithm used by Spotify was biased against female artists. Spotify's song recommendations suggested more male artists over female artists.

Racial and ethnic discrimination

Algorithms have been criticized as a method for obscuring racial prejudices in decision-making. Because of how certain races and ethnic groups were treated in the past, data can often contain hidden biases. For example, black people are likely to receive longer sentences than white people who committed the same crime. This could potentially mean that a system amplifies the original biases in the data.

In 2015, Google apologized when a couple of black users complained that an image-identification algorithm in its Photos application identified them as gorillas . In 2010, Nikon cameras were criticized when image-recognition algorithms consistently asked Asian users if they were blinking. Such examples are the product of bias in biometric data sets. Biometric data is drawn from aspects of the body, including racial features either observed or inferred, which can then be transferred into data points. Speech recognition technology can have different accuracies depending on the user's accent. This may be caused by the a lack of training data for speakers of that accent.

Biometric data about race may also be inferred, rather than observed. For example, a 2012 study showed that names commonly associated with blacks were more likely to yield search results implying arrest records, regardless of whether there is any police record of that individual's name. A 2015 study also found that Black and Asian people are assumed to have lesser functioning lungs due to racial and occupational exposure data not being incorporated into the prediction algorithm's model of lung function.

In 2019, a research study revealed that a healthcare algorithm sold by Optum favored white patients over sicker black patients. The algorithm predicts how much patients would cost the health-care system in the future. However, cost is not race-neutral, as black patients incurred about $1,800 less in medical costs per year than white patients with the same number of chronic conditions, which led to the algorithm scoring white patients as equally at risk of future health problems as black patients who suffered from significantly more diseases.

A study conducted by researchers at UC Berkeley in November 2019 revealed that mortgage algorithms have been discriminatory towards Latino and African Americans which discriminated against minorities based on "creditworthiness" which is rooted in the U.S. fair-lending law which allows lenders to use measures of identification to determine if an individual is worthy of receiving loans. These particular algorithms were present in FinTech companies and were shown to discriminate against minorities. [non-primary source needed]

Another study, published in August 2024, on Large language model investigates how language models perpetuate covert racism, particularly through dialect prejudice against speakers of African American English (AAE). It highlights that these models exhibit more negative stereotypes about AAE speakers than any recorded human biases, while their overt stereotypes are more positive. This discrepancy raises concerns about the potential harmful consequences of such biases in decision-making processes.

A study published by the Anti-Defamation League in 2025 found that several major LLMs, including ChatGPT , Llama , Claude , and Gemini showed antisemitic bias.

A 2018 study found that commercial gender classification systems had significantly higher error rates for darker-skinned women, with error rates up to 34.7%, compared to near-perfect accuracy for lighter-skinned men.

Law enforcement and legal proceedings

Algorithms already have numerous applications in legal systems. An example of this is COMPAS , a commercial program widely used by U.S. courts to assess the likelihood of a defendant becoming a recidivist . ProPublica claims that the average COMPAS-assigned recidivism risk level of black defendants is significantly higher than the average COMPAS-assigned risk level of white defendants, and that black defendants are twice as likely to be erroneously assigned the label "high-risk" as white defendants.

One example is the use of risk assessments in criminal sentencing in the United States and parole hearings , judges were presented with an algorithmically generated score intended to reflect the risk that a prisoner will repeat a crime. For the time period starting in 1920 and ending in 1970, the nationality of a criminal's father was a consideration in those risk assessment scores. Today, these scores are shared with judges in Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington, and Wisconsin. An independent investigation by ProPublica found that the scores were inaccurate 80% of the time, and disproportionately skewed to suggest blacks to be at risk of relapse, 77% more often than whites.

One study that set out to examine "Risk, Race, & Recidivism: Predictive Bias and Disparate Impact" alleges a two-fold (45 percent vs. 23 percent) adverse likelihood for black vs. Caucasian defendants to be misclassified as imposing a higher risk despite having objectively remained without any documented recidivism over a two-year period of observation.

In the pretrial detention context, a law review article argues that algorithmic risk assessments violate 14th Amendment Equal Protection rights on the basis of race, since the algorithms are argued to be facially discriminatory, to result in disparate treatment, and to not be narrowly tailored.

Online hate speech

In 2017 a Facebook algorithm designed to remove online hate speech was found to advantage white men over black children when assessing objectionable content, according to internal Facebook documents. The algorithm, which is a combination of computer programs and human content reviewers, was created to protect broad categories rather than specific subsets of categories. For example, posts denouncing "Muslims" would be blocked, while posts denouncing "Radical Muslims" would be allowed. An unanticipated outcome of the algorithm is to allow hate speech against black children, because they denounce the "children" subset of blacks, rather than "all blacks", whereas "all white men" would trigger a block, because whites and males are not considered subsets. Facebook was also found to allow ad purchasers to target "Jew haters" as a category of users, which the company said was an inadvertent outcome of algorithms used in assessing and categorizing data. The company's design also allowed ad buyers to block African-Americans from seeing housing ads.

While algorithms are used to track and block hate speech, some were found to be 1.5 times more likely to flag information posted by Black users and 2.2 times likely to flag information as hate speech if written in African American English .

Surveillance

Surveillance camera software may be considered inherently political because it requires algorithms to distinguish normal from abnormal behaviors, and to determine who belongs in certain locations at certain times. The ability of such algorithms to recognize faces across a racial spectrum has been shown to be limited by the racial diversity of images in its training database; if the majority of photos belong to one race or gender, the software is better at recognizing other members of that race or gender. However, even audits of these image-recognition systems are ethically fraught, and some scholars have suggested the technology's context will always have a disproportionate impact on communities whose actions are over-surveilled. For example, a 2002 analysis of software used to identify individuals in CCTV images found several examples of bias when run against criminal databases. The software was assessed as identifying men more frequently than women, older

people more frequently than the young, and identified Asians, African-Americans and other races more often than whites. A 2018 study found that facial recognition software most likely accurately identified light-skinned (typically European) males, with slightly lower accuracy rates for light-skinned females. Dark-skinned males and females were significanfly less likely to be accurately identified by facial recognition software. These disparities are attributed to the under-representation of darker-skinned participants in data sets used to develop this software.

Discrimination against the LGBTQ community

In 2011, users of the gay hookup application Grindr reported that the Android store 's recommendation algorithm was linking Grindr to applications designed to find sex offenders, which critics said inaccurately related homosexuality with pedophilia . Writer Mike Ananny criticized this association in The Atlantic , arguing that such associations further stigmatized gay men . In 2009, online retailer Amazon de-listed 57,000 books after an algorithmic change expanded its "adult content" blacklist to include any book addressing sexuality or gay themes, such as the critically acclaimed novel Brokeback Mountain .

In 2019, it was found that on Facebook, searches for "photos of my female friends" yielded suggestions such as "in bikinis" or "at the beach". In contrast, searches for "photos of my male friends" yielded no results.

Facial recognition technology has been seen to cause problems for transgender individuals. In 2018, there were reports of Uber drivers who were transgender or transitioning experiencing difficulty with the facial recognition software that Uber implements as a built-in security measure. As a result of this, some of the accounts of trans Uber drivers were suspended which cost them fares and potentially cost them a job, all due to the facial recognition software experiencing difficulties with recognizing the face of a trans driver who was transitioning. Although the solution to this issue would appear to be including trans individuals in training sets for machine learning models, an instance of trans YouTube videos that were collected to be used in training data did not receive consent from the trans individuals that were included in the videos, which created an issue of violation of privacy.

There has also been a study that was conducted at Stanford University in 2017 that tested algorithms in a machine learning system that was said to be able to detect an individual's sexual orientation based on their facial images. The model in the study predicted a correct distinction between gay and straight men 81% of the time, and a correct distinction between gay and straight women 74% of the time. This study resulted in a backlash from the LGBTQIA community, who were fearful of the possible negative repercussions that this AI system could have on individuals of the LGBTQIA community by putting individuals at risk of being " outed " against their will.

Disability discrimination

While the modalities of algorithmic fairness have been judged on the basis of different aspects of bias – like gender, race and socioeconomic status, disability often is left out of the list. The marginalization people with disabilities currently face in society is being translated into AI systems and algorithms, creating even more exclusion

The shifting nature of disabilities and its subjective characterization, makes it more difficult to computationally address. The lack of historical depth in defining disabilities, collecting its incidence and prevalence in questionnaires, and establishing recognition add to the controversy and ambiguity in its quantification and calculations. The definition of disability has been long debated shifting from a medical model to a social model of disability most recently, which establishes that disability is a result of the mismatch between people's interactions and barriers in their environment, rather than impairments and health conditions. Disabilities can also be situational or temporary, considered in a constant state of flux. Disabilities are incredibly diverse, fall within a large spectrum, and can be unique to each individual. People's identity can vary based on the specific types of disability they experience, how they use assistive technologies, and who they support. The high level of variability across people's experiences greatly personalizes how a disability can manifest. Overlapping identities and intersectional experiences are excluded from statistics and datasets, hence underrepresented and nonexistent in training data. Therefore, machine learning models are trained inequitably and artificial intelligent systems perpetuate more algorithmic bias. For example,

if people with speech impairments are not included in training voice control features and smart AI assistants –they are unable to use the feature or the responses received from a Google Home or Alexa are extremely poor.

Given the stereotypes and stigmas that still exist surrounding disabilities, the sensitive nature of revealing these identifying characteristics also carries vast privacy challenges. As disclosing disability information can be taboo and drive further discrimination against this population, there is a lack of explicit disability data available for algorithmic systems to interact with. People with disabilities face additional harms and risks with respect to their social support, cost of health insurance, workplace discrimination and other basic necessities upon disclosing their disability status. Algorithms are further exacerbating this gap by recreating the biases that already exist in societal systems and structures.

Google Search

While users generate results that are "completed" automatically, Google has failed to remove sexist and racist autocompletion text. For example, Algorithms of Oppression: How Search Engines Reinforce Racism Safiya Noble notes an example of the search for "black girls", which was reported to result in pornographic images. Google claimed it was unable to erase those pages unless they were considered unlawful.

Obstacles to research

Several problems impede the study of large-scale algorithmic bias, hindering the application of academically rigorous studies and public understanding.

Defining fairness

Literature on algorithmic bias has focused on the remedy of fairness, but definitions of fairness are often incompatible with each other and the realities of machine learning optimization. For example, defining fairness as an "equality of outcomes" may simply refer to a system producing the same result for all people, while fairness defined as "equality of treatment" might explicitly consider differences between individuals. As a result, fairness is sometimes described as being in conflict with the accuracy of a model, suggesting innate tensions between the priorities of social welfare and the priorities of the vendors designing these systems. In response to this tension, researchers have suggested more care to the design and use of systems that draw on potentially biased algorithms, with "fairness" defined for specific applications and contexts.

Complexity

Algorithmic processes are complex , often exceeding the understanding of the people who use them. Large-scale operations may not be understood even by those involved in creating them. The methods and processes of contemporary programs are often obscured by the inability to know every permutation of a code's input or output. Social scientist Bruno Latour has identified this process as blackboxing , a process in which "scientific and technical work is made invisible by its own success. When a machine runs efficiently, when a matter of fact is settled, one need focus only on its inputs and outputs and not on its internal complexity. Thus, paradoxically, the more science and technology succeed, the more opaque and obscure they become." Others have critiqued the black box metaphor, suggesting that current algorithms are not one black box, but a network of interconnected ones.

An example of this complexity can be found in the range of inputs into customizing feedback. The social media site Facebook factored in at least 100,000 data points to determine the layout of a user's social media feed in 2013. Furthermore, large teams of programmers may operate in relative isolation from one another, and be unaware of the cumulative effects of small decisions within connected, elaborate algorithms. Not all code is original, and may be borrowed from other libraries, creating a complicated set of relationships between data processing and data input systems.

Additional complexity occurs through machine learning and the personalization of algorithms based on user interactions such as clicks, time spent on site, and other metrics. These personal adjustments can confuse general attempts to understand algorithms. One unidentified streaming radio service reported that it used five unique music-selection algorithms it selected for its users,

based on their behavior. This creates different experiences of the same streaming services between different users, making it harder to understand what these algorithms do. Companies also run frequent A/B tests to fine-tune algorithms based on user response. For example, the search engine Bing can run up to ten million subtle variations of its service per day, creating different experiences of the service between each use and/or user.

Lack of transparency

Commercial algorithms are proprietary, and may be treated as trade secrets . Treating algorithms as trade secrets protects companies, such as search engines , where a transparent algorithm might reveal tactics to manipulate search rankings. This makes it difficult for researchers to conduct interviews or analysis to discover how algorithms function. Critics suggest that such secrecy can also obscure possible unethical methods used in producing or processing algorithmic output. Other critics, such as lawyer and activist Katarzyna Szymielewicz, have suggested that the lack of transparency is often disguised as a result of algorithmic complexity, shielding companies from disclosing or investigating its own algorithmic processes.

Lack of data about sensitive categories

A significant barrier to understanding the tackling of bias in practice is that categories, such as demographics of individuals protected by anti-discrimination law , are often not explicitly considered when collecting and processing data. In some cases, there is little opportunity to collect this data explicitly, such as in device fingerprinting , ubiquitous computing and the Internet of Things . In other cases, the data controller may not wish to collect such data for reputational reasons, or because it represents a heightened liability and security risk. It may also be the case that, at least in relation to the European Union's General Data Protection Regulation , such data falls under the 'special category' provisions (Article 9), and therefore comes with more restrictions on potential collection and processing.

Some practitioners have tried to estimate and impute these missing sensitive categorizations in order to allow bias mitigation, for example building systems to infer ethnicity from names, however this can introduce other forms of bias if not undertaken with care. Machine learning researchers have drawn upon cryptographic privacy-enhancing technologies such as secure multi-party computation to propose methods whereby algorithmic bias can be assessed or mitigated without these data ever being available to modellers in cleartext .

Algorithmic bias does not only include protected categories, but can also concern characteristics less easily observable or codifiable, such as political viewpoints. In these cases, there is rarely an easily accessible or non-controversial ground truth , and removing the bias from such a system is more difficult. Furthermore, false and accidental correlations can emerge from a lack of understanding of protected categories, for example, insurance rates based on historical data of car accidents which may overlap, strictly by coincidence, with residential clusters of ethnic minorities.

Solutions

A study of 84 policy guidelines on ethical AI found that fairness and "mitigation of unwanted bias" was a common point of concern, and were addressed through a blend of technical solutions, transparency and monitoring, right to remedy and increased oversight, and diversity and inclusion efforts.

Technical

There have been several attempts to create methods and tools that can detect and observe biases within an algorithm. These emergent fields focus on tools which are typically applied to the (training) data used by the program rather than the algorithm's internal processes. These methods may also analyze a program's output and its usefulness and therefore may involve the analysis of its confusion matrix (or table of confusion). Explainable AI to detect algorithm Bias is a suggested way to detect the existence of bias in an algorithm or learning model. Using machine learning to detect bias is called, "conducting an AI audit", where the "auditor" is an algorithm that goes through the AI model and the training data to identify biases. Ensuring that an AI tool such as a classifier is free from bias is more difficult than just removing the sensitive information

from its input signals, because this is typically implicit in other signals. For example, the hobbies, sports and schools attended

by a job candidate might reveal their gender to the software, even when this is removed from the analysis. Solutions to this

problem involve ensuring that the intelligent agent does not have any information that could be used to reconstruct the protected

and sensitive information about the subject, as first demonstrated in where a deep learning network was simultaneously trained to learn a task while at the same time being completely agnostic about the protected feature. A simpler method was proposed in the context of word embeddings, and involves removing information that is correlated with the protected characteristic.

Currently [when?] , a new IEEE standard is being drafted that aims to specify methodologies which help creators of algorithms eliminate issues of bias and articulate transparency (i.e. to authorities or end users) about the function and possible effects of their algorithms. The project was approved February 2017 and is sponsored by the Software & Systems Engineering Standards Committee, a committee chartered by the IEEE Computer Society . A draft of the standard is expected to be submitted for balloting in June 2019. The standard was published in January 2025.

In 2022, the IEEE released a standard aimed at specifying methodologies to help creators of algorithms address issues of bias and promote transparency regarding the function and potential effects of their algorithms. The project, initially approved in February 2017, was sponsored by the Software & Systems Engineering Standards Committee, a committee under the IEEE Computer Society . The standard provides guidelines for articulating transparency to authorities or end users and mitigating algorithmic biases.

Transparency and monitoring

Ethics guidelines on AI point to the need for accountability, recommending that steps be taken to improve the interpretability of results. Such solutions include the consideration of the "right to understanding" in machine learning algorithms, and to resist deployment of machine learning in situations where the decisions could not be explained or reviewed. Toward this end, a movement for " Explainable AI " is already underway within organizations such as DARPA , for reasons that go beyond the remedy of bias. Price Waterhouse Coopers , for example, also suggests that monitoring output means designing systems in such a way as to ensure that solitary components of the system can be isolated and shut down if they skew results.

An initial approach towards transparency included the open-sourcing of algorithms . Software code can be looked into and improvements can be proposed through source-code-hosting facilities . However, this approach doesn't necessarily produce the intended effects. Companies and organizations can share all possible documentation and code, but this does not establish transparency if the audience doesn't understand the information given. Therefore, the role of an interested critical audience is worth exploring in relation to transparency. Algorithms cannot be held accountable without a critical audience.

Right to remedy

From a regulatory perspective, the Toronto Declaration calls for applying a human rights framework to harms caused by algorithmic bias. This includes legislating expectations of due diligence on behalf of designers of these algorithms, and creating accountability when private actors fail to protect the public interest, noting that such rights may be obscured by the complexity of determining responsibility within a web of complex, intertwining processes. Others propose the need for clear liability insurance mechanisms.

Diversity and inclusion

Amid concerns that the design of AI systems is primarily the domain of white, male engineers, a number of scholars have suggested that algorithmic bias may be minimized by expanding inclusion in the ranks of those designing AI systems. For example, just 12% of machine learning engineers are women, with black AI leaders pointing to a "diversity crisis" in the field. Groups like Black in AI and Queer in AI are attempting to create more inclusive spaces in the AI community and work

against the often harmful desires of corporations that control the trajectory of AI research. Critiques of simple inclusivity efforts suggest that diversity programs can not address overlapping forms of inequality, and have called for applying a more deliberate lens of intersectionality to the design of algorithms. Researchers at the University of Cambridge have argued that addressing racial diversity is hampered by the "whiteness" of the culture of AI.

Interdisciplinarity and Collaboration

Integrating interdisciplinarity and collaboration in developing of AI systems can play a critical role in tackling algorithmic bias. Integrating insights, expertise, and perspectives from disciplines outside of computer science can foster a better understanding of the impact data driven solutions have on society. An example of this in AI research is PACT or Participatory Approach to enable Capabilities in communiTies, a proposed framework for facilitating collaboration when developing AI driven solutions concerned with social impact. This framework identifies guiding principals for stakeholder participation when working on AI for Social Good (AI4SG) projects. PACT attempts to reify the importance of decolonizing and power-shifting efforts in the design of human-centered AI solutions. An academic initiative in this regard is the Stanford University's Institute for Human-Centered Artificial Intelligence which aims to foster multidisciplinary collaboration. The mission of the institute is to advance artificial intelligence (AI) research, education, policy and practice to improve the human condition.

Collaboration with outside experts and various stakeholders facilitates ethical, inclusive, and accountable development of intelligent systems. It incorporates ethical considerations, understands the social and cultural context, promotes human-centered design, leverages technical expertise, and addresses policy and legal considerations. Collaboration across disciplines is essential to effectively mitigate bias in AI systems and ensure that AI technologies are fair, transparent, and accountable.

Regulation

Europe

The General Data Protection Regulation (GDPR), the European Union 's revised data protection regime that was implemented in 2018, addresses "Automated individual decision-making, including profiling " in Article 22. These rules prohibit "solely" automated decisions which have a "significant" or "legal" effect on an individual, unless they are explicitly authorised by consent, contract, or member state law. Where they are permitted, there must be safeguards in place, such as a right to a human-in-the-loop , and a non-binding right to an explanation of decisions reached. While these regulations are commonly considered to be new, nearly identical provisions have existed across Europe since 1995, in Article 15 of the Data Protection Directive . The original automated decision rules and safeguards found in French law since the late 1970s.

The GDPR addresses algorithmic bias in profiling systems, as well as the statistical approaches possible to clean it, directly in recital 71, noting that

> the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate ... that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect.

Like the non-binding right to an explanation in recital 71, the problem is the non-binding nature of recitals . While it has been treated as a requirement by the Article 29 Working Party that advised on the implementation of data protection law, its practical dimensions are unclear. It has been argued that the Data Protection Impact Assessments for high risk data profiling (alongside other pre-emptive measures within data protection) may be a better way to tackle issues of algorithmic discrimination, as it restricts the actions of those deploying algorithms, rather than requiring consumers to file complaints or request changes.

United States

The United States has no general legislation controlling algorithmic bias, approaching the problem through various state and federal laws that might vary by industry, sector, and by how an algorithm is used. Many policies are self-enforced or controlled by the Federal Trade Commission . In 2016, the Obama administration released the National Artificial Intelligence Research and Development Strategic Plan , which was intended to guide policymakers toward a critical assessment of algorithms. It recommended researchers to "design these systems so that their actions and decision-making are transparent and easily interpretable by humans, and thus can be examined for any bias they may contain, rather than just learning and repeating these biases". Intended only as guidance, the report did not create any legal precedent.

In 2017, New York City passed the first algorithmic accountability bill in the United States. The bill, which went into effect on January 1, 2018, required "the creation of a task force that provides recommendations on how information on agency automated decision systems may be shared with the public, and how agencies may address instances where people are harmed by agency automated decision systems." In 2023, New York City implemented a law requiring employers using automated hiring tools to conduct independent "bias audits" and publish the results. This law marked one of the first legally mandated transparency measures for AI systems used in employment decisions in the United States. The task force is required to present findings and recommendations for further regulatory action in 2019. On February 11, 2019, according to Executive Order 13859 , the federal government unveiled the "American AI Initiative," a comprehensive strategy to maintain U.S. leadership in artificial intelligence. The initiative highlights the importance of sustained AI research and development, ethical standards, workforce training, and the protection of critical AI technologies. This aligns with broader efforts to ensure transparency, accountability, and innovation in AI systems across public and private sectors. Furthermore, on October 30, 2023, the President signed Executive Order 14110 , which emphasizes the safe, secure, and trustworthy development and use of artificial intelligence (AI). The order outlines a coordinated, government-wide approach to harness AI's potential while mitigating its risks, including fraud, discrimination, and national security threats. An important point in the commitment is promoting responsible innovation and collaboration across sectors to ensure that AI benefits society as a whole. With this order, President Joe Biden mandated the federal government to create best practices for companies to optimize AI's benefits and minimize its harms.

India

On July 31, 2018, a draft of the Personal Data Bill was presented. The draft proposes standards for the storage, processing and transmission of data. While it does not use the term algorithm, it makes for provisions for "harm resulting from any processing or any kind of processing undertaken by the fiduciary". It defines "any denial or withdrawal of a service, benefit or good resulting from an evaluative decision about the data principal" or "any discriminatory treatment" as a source of harm that could arise from improper use of data. It also makes special provisions for people of "Intersex status".

See also

Algorithmic wage discrimination

Ethics of artificial intelligence

Fairness (machine learning)

Hallucination (artificial intelligence)

Misaligned goals in artificial intelligence

Predictive policing

SenseTime

References

Further reading

Baer, Tobias (2019). Understand, Manage, and Prevent Algorithmic Bias: A Guide for Business Users and Data Scientists . New York: Apress. ISBN 9781484248843 .

Noble, Safiya Umoja (2018). Algorithms of Oppression: How Search Engines Reinforce Racism . New York: New York University Press. ISBN 9781479837243 .