Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People

White House Office of Science and Technology Policy
https://www.whitehouse.gov/ostp/ai-bill-of-rights/

Applying the Blueprint for an AI Bill of Rights

https://www.whitehouse.gov/ostp/ai-bill-of-rights/applying-the-blueprint-for-an-ai-bill-of-rights/

This framework describes protections that should be applied with respect to all automated systems that have the potential to meaningfully impact individuals' or communities' exercise of:

- Rights, Opportunities, or Access
- Civil rights, civil liberties, and privacy, including freedom of speech, voting, and protections from discrimination, excessive punishment, unlawful surveillance, and violations of privacy and other freedoms in both public and private sector contexts;
- Equal opportunities, including equitable access to education, housing, credit, employment, and other programs; or,
- Access to critical resources or services, such as healthcare, financial services, safety, social services, non-deceptive information about goods and services, and government benefits.

Definitions

https://www.whitehouse.gov/ostp/ai-bill-of-rights/definitions/

RIGHTS, OPPORTUNITIES, OR ACCESS: "Rights, opportunities, or access" is used to indicate the scoping of this framework. It describes the set of: civil rights, civil liberties, and privacy, including freedom of speech, voting, and protections from discrimination, excessive punishment, unlawful surveillance, and violations of privacy and other freedoms in both public and private sector contexts; equal opportunities, including equitable access to education, housing, credit, employment, and other programs; or, access to critical resources or services, such as healthcare, financial services, safety, social services, non-deceptive information about goods and services, and government benefits.

SENSITIVE DATA: Data and metadata are sensitive if they pertain to an individual in a sensitive domain (defined below); are generated by technologies used in a sensitive domain; can be used to infer data from a sensitive domain or sensitive data about an individual (such as disability-related data, genomic data, biometric data, behavioral data, geolocation data, data related to interaction with the criminal justice system, relationship history and legal status such as custody and divorce information, and home, work, or school environmental data); or have the reasonable potential to be used in ways that are likely to expose individuals to meaningful harm, such as a loss of privacy or financial harm due to identity theft. Data and metadata generated by or about those who are not yet legal adults is also sensitive, even if not related to a sensitive domain. Such data includes, but is not limited to, numerical, text, image, audio, or video data.

SENSITIVE DOMAINS: “Sensitive domains” are those in which activities being conducted can cause material harms, including significant adverse effects on human rights such as autonomy and dignity, as well as civil liberties and civil rights. Domains that have historically been singled out as deserving of enhanced data protections or where such enhanced protections are reasonably expected by the public include, but are not limited to, health, family planning and care, employment, education, criminal justice, and personal finance. In the context of this framework, such domains are considered sensitive whether or not the specifics of a system context would necessitate coverage under existing law, and domains and data that are considered sensitive are understood to change over time based on societal norms and context.

Examples of Automated Systems

https://www.whitehouse.gov/ostp/ai-bill-of-rights/examples-of-automated-systems/

Examples of automated systems for which the Blueprint for an AI Bill of Rights should be considered include those that have the potential to meaningfully impact:

Civil rights, civil liberties, or privacy, including but not limited to:
…

- Systems with a potential privacy impact such as smart home systems and associated data, systems that use or collect health-related data, systems that use or collect education-related data, criminal justice system data, ad-targeting systems, and systems that perform big data analytics in order to build profiles or infer personal information about individuals; …

Equal opportunities, including but not limited to:

- Education-related systems such as algorithms that purport to detect student cheating or plagiarism, admissions algorithms, online or virtual reality student monitoring systems, projections of student progress or outcomes, algorithms that determine access to resources or programs, and surveillance of classes (whether online or in-person);...

Safe and Effective Systems: You Should Be Protected From Unsafe or Ineffective Systems

https://www.whitehouse.gov/ostp/ai-bill-of-rights/safe-and-effective-systems-3/

NA

Algorithmic Discrimination Protections: You Should Not Face Discrimination by Algorithms and Systems Should Be Used and Designed in an Equitable Way

https://www.whitehouse.gov/ostp/ai-bill-of-rights/algorithmic-discrimination-protections-2/

- An automated system using nontraditional factors such as educational attainment and employment history as part of its loan underwriting and pricing model was found to be much more likely to charge an applicant who attended a Historically Black College or University (HBCU) higher loan prices for refinancing a student loan than an applicant who did not attend an HBCU. This was found to be true even when controlling for other credit-related factors.[iii]
- A predictive model marketed as being able to predict whether students are likely to drop out of school was used by more than 500 universities across the country. The model was found to use race directly as a predictor, and also shown to have large disparities by race; Black students were as many as four times as likely as their otherwise similar white peers to be deemed at high risk of dropping out. These risk scores are used by advisors to guide students towards or away from majors, and some worry that they are being used to guide Black students away from math and science subjects.[v]
- The National Disabled Law Students Association expressed concerns that individuals with disabilities were more likely to be flagged as potentially suspicious by remote proctoring AI systems because of their disability-specific access needs such as needing longer breaks or using screen readers or dictation software.[xvi]

Data Privacy: You Should Be Protected from Abusive Data Practices Via Built-In Protections and You Should Have Agency Over How Data About You is Used

https://www.whitehouse.gov/ostp/ai-bill-of-rights/data-privacy-2/

Enhanced protections and restrictions for data and inferences related to sensitive domains, including health, work, education, criminal justice, and finance, and for data pertaining to youth should put you first.

Continuous surveillance and monitoring should not be used in education, work, housing, or in other contexts where the use of such surveillance technologies is likely to limit rights, opportunities, or access.

Extra Protections for Data Related to Sensitive Domains

Some domains, including health, employment, education, criminal justice, and personal finance, have long been singled out as sensitive domains deserving of enhanced data protections. This is due to the intimate nature of these domains as well as the inability of individuals to opt out of these domains in any meaningful way, and the historical discrimination that has often accompanied data knowledge.[x] Domains understood by the public to be sensitive also change over time, including because of technological developments. Tracking and monitoring technologies, personal tracking devices, and our extensive data footprints are used and misused more than ever before; as such, the protections afforded by current legal guidelines may be inadequate. The American public deserves assurances that data related to such sensitive domains is protected and used appropriately and only in narrowly defined contexts with clear benefits to the individual and/or society.

To this end, automated systems that collect, use, share, or store data related to these sensitive domains should meet additional expectations. Data and metadata are sensitive if they pertain to an individual in a sensitive domain (defined below); are generated by technologies used in a sensitive domain; can be used to infer data from a sensitive domain or sensitive data about an individual (such as disability-related data, genomic data, biometric data, behavioral data, geolocation data, data related to interaction with the criminal justice system, relationship history and legal status such as custody and divorce information, and home, work, or school environmental data); or have the reasonable potential to be used in ways that are likely to expose individuals to meaningful harm, such as a loss of privacy or financial harm due to identity theft.

Data and metadata generated by or about those who are not yet legal adults is also sensitive, even if not related to a sensitive domain. Such data includes, but is not limited to, numerical, text, image, audio, or video data. “Sensitive domains” are those in which activities being conducted can cause material harms, including significant adverse effects on human rights such as autonomy and dignity, as well as civil liberties and civil rights. Domains that have historically been singled out as deserving of enhanced data protections or where such enhanced protections are reasonably expected by the public include, but are not limited to, health, family planning and care, employment, education, criminal justice, and personal finance. In the context of this framework, such domains are considered sensitive whether or not the specifics of a system context would necessitate coverage under existing law, and domains and data that are considered sensitive are understood to change over time based on societal norms and context.

[Examples provided:]

…

- School audio surveillance systems monitor student conversations to detect potential “stress indicators” as a warning of potential violence.[xiii] Online proctoring systems claim to detect if a student is cheating on an exam using biometric markers.[xiv] These systems have the potential to limit student freedom to express a range of emotions at school and may inappropriately flag students with disabilities who need accommodations or use screen readers or dictation software as cheating.[xv]

…

- Companies collect student data such as demographic information, free or reduced lunch status, whether they've used drugs, or whether they've expressed interest in LGBTQI+ groups, and then use that data to forecast student success.[xvii] Parents and education experts have expressed concern about collection of such sensitive data without express parental consent, the lack of transparency in how such data is being used, and the potential for resulting discriminatory impacts.

What should be expected of automated systems?

In addition to the privacy expectations above for general non-sensitive data, any system collecting, using, sharing, or storing sensitive data should meet the expectations below. Depending on the technological use case and based on an ethical assessment, consent for sensitive data may need to be acquired from a guardian and/or child.

Provide enhanced protections for data related to sensitive domains

Necessary functions only. Sensitive data should only be used for functions strictly necessary for that domain or for functions that are required for administrative reasons (e.g., school attendance records), unless consent is acquired, if appropriate, and the additional expectations in this section are met. Consent for non-necessary functions should be optional, i.e., should not be required, incentivized, or coerced in order to receive opportunities or access to services. In cases where data is provided to an entity (e.g., health insurance company) in order to facilitate payment for such a need, that data should only be used for that purpose.

Ethical review and use prohibitions. Any use of sensitive data or decision process based in part on sensitive data that might limit rights, opportunities, or access, whether the decision is automated or not, should go through a thorough ethical review and monitoring, both in advance and by periodic review (e.g., via an independent ethics committee or similarly robust process). In some cases, this ethical review may determine that data should not be used or shared for specific uses even with consent. Some novel uses of automated systems in this context, where the algorithm is dynamically developing and where the science behind the use case is not well established, may also count as human subject experimentation, and require special review under organizational compliance bodies applying medical, scientific, and academic human subject experimentation ethics rules and governance procedures.

Data quality. In sensitive domains, entities should be especially careful to maintain the quality of data to avoid adverse consequences arising from decision-making based on flawed or inaccurate data. Such care is necessary in a fragmented, complex data ecosystem and for datasets that have limited access such as for fraud prevention and law enforcement. It should be not left solely to individuals to carry the burden of reviewing and correcting data. Entities should conduct regular, independent audits and take prompt corrective measures to maintain accurate, timely, and complete data.

Limit access to sensitive data and derived data. Sensitive data and derived data should not be sold, shared, or made public as part of data brokerage or other agreements. Sensitive data includes data that can be used to infer sensitive information; even systems that are not directly marketed as sensitive domain technologies are expected to keep sensitive data private. Access to such data should be limited based on necessity and based on a principle of local control, such that those individuals closest to the data subject have more access while those who are less proximate do not (e.g., a teacher has access to their students' daily progress data while a superintendent does not).

Reporting. In addition to the reporting on data privacy (as listed above for non-sensitive data), entities developing technologies related to a sensitive domain and those collecting, using, storing, or sharing sensitive data should, whenever appropriate, regularly provide public reports describing: any data security lapses or breaches that resulted in sensitive data leaks; the number, type, and outcomes of ethical pre-reviews undertaken; a description of any data sold, shared, or made public, and how that data was assessed to determine it did not present a sensitive data risk; and ongoing risk identification and management procedures, and any mitigation added based on these procedures. Reporting should be provided in a clear and machine-readable manner.

How these principles can move into practice

A school board's attempt to surveil public school students—undertaken without adequate community input—sparked a state-wide biometrics moratorium.[xx] Reacting to a plan in the city of Lockport, New York, the state's legislature banned the use of facial recognition systems and other "biometric identifying technology" in schools until July 1, 2022.[xxi] The law additionally requires that a report on the privacy, civil rights, and civil liberties implications of the use of such technologies be issued before biometric identification technologies can be used in New York schools.

[Relevant Footnotes:]

[ii] See, e.g., Nir Kshetri. School surveillance of students via laptops may do more harm than good. The Conversation. Jan. 21, 2022. https://theconversation.com/school-surveillance-of-students-via-laptops-may-do-more-harm-than-good-170983; …

[xiii] Jack Gillum and Jeff Kao. Aggression Detectors: The Unproven, Invasive Surveillance Technology Schools are Using to Monitor Students. ProPublica. Jun. 25, 2019. https://features.propublica.org/aggression-detector/the-unproven-invasive-surveillance-technology-schools-are-using-to-monitor-students/

[xiv] Drew Harwell. Cheating-detection companies made millions during the pandemic. Now students are fighting back. Washington Post. Nov. 12, 2020. https://www.washingtonpost.com/technology/2020/11/12/test-monitoring-student-revolt/

[xv] See, e.g., Heather Morrison. Virtual Testing Puts Disabled Students at a Disadvantage. Government Technology. May 24, 2022. https://www.govtech.com/education/k-12/virtual-testing-puts-disabled-students-at-a-disadvantage; Lydia X. Z. Brown, Ridhi Shetty, Matt Scherer, and Andrew Crawford. Ableism And Disability Discrimination In New Surveillance Technologies: How new surveillance technologies in education, policing, health care, and the workplace disproportionately harm disabled people. Center for Democracy and Technology Report. May 24, 2022. https://cdt.org/insights/ableism-and-disability-discrimination-in-new-surveillance-technologies-how-new-surveillance-technologies-in-education-policing-health-care-and-the-workplace-disproportionately-harm-disabled-people/

[xx] ACLU of New York. What You Need to Know About New York’s Temporary Ban on Facial Recognition in Schools. Accessed May 2, 2022. https://www.nyclu.org/en/publications/what-you-need-know-about-new-yorks-temporary-ban-facial-recognition-schools

[xxi] New York State Assembly. Amendment to Education Law. Enacted Dec. 22, 2020. https://nyassembly.gov/leg/?default_fld=&leg_video=&bn=S05140&term=2019&Summary=Y&Text=Y

Notice and Explanation: You Should Know that an Automated System is Being Used and Understand How and Why It Contributes to Outcomes That Impact You

https://www.whitehouse.gov/ostp/ai-bill-of-rights/notice-and-explanation/

- A formal child welfare investigation is opened against a parent based on an algorithm and without the parent ever being notified that data was being collected and used as part of an algorithmic child maltreatment risk assessment.[ii] The lack of notice or an explanation makes it harder for those performing child maltreatment assessments to validate the risk assessment and denies parents knowledge that could help them contest a decision.

Human Alternatives, Consideration, and Fallback: You Should Be Able to Opt Out, Where Appropriate, and Have Access to a Person Who Can Quickly Consider and Remedy Problems You Encounter

https://www.whitehouse.gov/ostp/ai-bill-of-rights/human-alternatives-consideration-and-fallback/

Why this principle is important

Automated systems with an intended use within sensitive domains, including, but not limited to, criminal justice, employment, education, and health, should additionally be tailored to the purpose, provide meaningful access for oversight, include training for any people interacting with the system, and incorporate human consideration for adverse or high-risk decisions.

In the criminal justice system, employment, education, healthcare, and other sensitive domains, automated systems are used for many purposes, from pre-trial risk assessments and parole decisions to technologies that help doctors diagnose disease. Absent appropriate safeguards, these technologies can lead to unfair, inaccurate, or dangerous outcomes. These sensitive domains require extra protections. It is critically important that there is extensive human oversight in such settings.

What should be expected of automated systems

Implement additional human oversight and safeguards for automated systems related to sensitive domains

Automated systems used within sensitive domains, including criminal justice, employment, education, and health, should meet the expectations laid out throughout this framework, especially avoiding capricious, inappropriate, and discriminatory impacts of these technologies. Additionally, automated systems used within sensitive domains should meet these expectations:

Narrowly scoped data and inferences. Human oversight should ensure that automated systems in sensitive domains are narrowly scoped to address a defined goal, justifying each included data item or attribute as relevant to the specific use case. Data included should be carefully limited to avoid algorithmic discrimination resulting from, e.g., use of community characteristics, social network analysis, or group-based inferences.

Tailored to the situation. Human oversight should ensure that automated systems in sensitive domains are tailored to the specific use case and real-world deployment scenario, and evaluation testing should show that the system is safe and effective for that specific situation. Validation testing performed based on one location or use case should not be assumed to transfer to another.

Human consideration before any high-risk decision. Automated systems, where they are used in sensitive domains, may play a role in directly providing information or otherwise providing positive outcomes to impacted people. However, automated systems should not be allowed to directly intervene in high-risk situations, such as sentencing decisions or medical care, without human consideration.

Meaningful access to examine the system. Designers, developers, and deployers of automated systems should consider limited waivers of confidentiality (including those related to trade secrets) where necessary in order to provide meaningful oversight of systems used in sensitive domains, incorporating measures to protect intellectual property and trade secrets from unwarranted disclosure as appropriate. This includes (potentially private and protected) meaningful access to source code, documentation, and related data during any associated legal discovery, subject to effective confidentiality or court orders. Such meaningful access should include (but is not limited to) adhering to the principle on Notice and Explanation using the highest level of risk so the system is designed with built-in explanations; such systems should use fully-transparent models where the model itself can be understood by people needing to directly examine it.

Press Release

https://www.whitehouse.gov/ostp/news-updates/2022/10/04/blueprint-for-an-ai-bill-of-rightsa-vision-for-protecting-our-civil-rights-in-the-algorithmic-age/

The Blueprint for an AI Bill of Rights is designed to be used by people across American society:

…

- Policymakers can codify these measures into law or use the framework and its technical companion to help develop specific guidance on the use of automated systems within a sector.
- Parents can use the framework as a set of questions to ask school administrators about what protections exist for their children.

Fact Sheet

https://www.whitehouse.gov/ostp/news-updates/2022/10/04/fact-sheet-biden-harris-administration-announces-key-actions-to-advance-tech-accountability-and-protect-the-rights-of-the-american-public/

Today, the Biden-Harris Administration is also announcing actions across the Federal government that advance the Blueprint by protecting and supporting the American people—workers and employers, educators and students, patients and health care providers, veterans, renters and home owners, technologists, families, and communities:

…

Protecting consumers:

- To protect consumers, the Federal Trade Commission (FTC) is exploring rules to curb commercial surveillance, algorithmic discrimination, and lax data security practices that could violate section 5 of the FTC Act. [children mentioned multiple times in the FTC's ANPRM]

Protecting students and supporting educators:

- **To guide schools in the use of AI**, the Department of Education will release recommendations on the use of AI for teaching and learning by early 2023. These recommendations will: give educators, parents and caregivers, students, and communities tools to leverage AI to advance universal design for learning; define specifications for the safety, fairness, and efficacy of AI models used within education; and introduce guidelines and guardrails that build on existing education data privacy regulations as well as introduce new policies to support schools in protecting students when using AI.