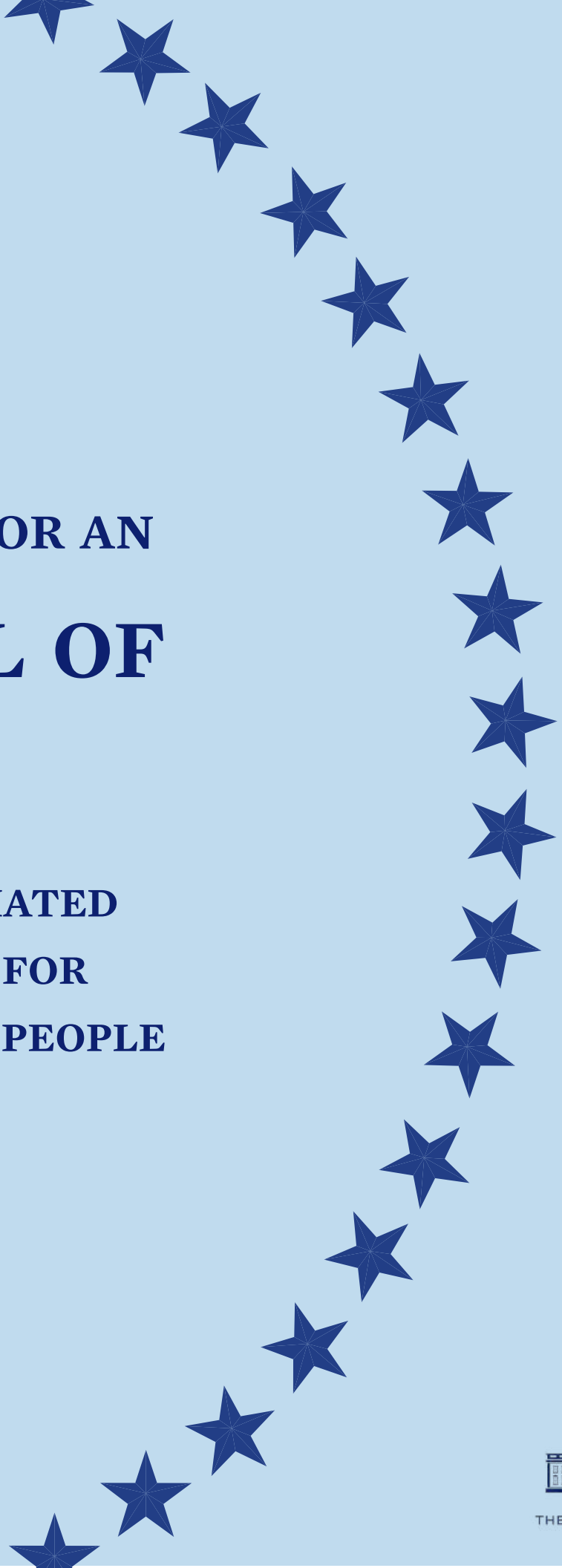
BLUEPRINT FOR AN

AI BILL OF RIGHTS

MAKING AUTOMATED SYSTEMS WORK FOR THE AMERICAN PEOPLE

OCTOBER 2022



THE WHITE HOUSE
WASHINGTON

About this Document

The Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People was published by the White House Office of Science and Technology Policy in October 2022. This framework was released one year after OSTP announced the launch of a process to develop "a bill of rights for an AI-powered world." Its release follows a year of public engagement to inform this initiative. The framework is available online at: https://www.whitehouse.gov/ostp/ai-bill-of-rights

About the Office of Science and Technology Policy

The Office of Science and Technology Policy (OSTP) was established by the National Science and Technology Policy, Organization, and Priorities Act of 1976 to provide the President and others within the Executive Office of the President with advice on the scientific, engineering, and technological aspects of the economy, national security, health, foreign relations, the environment, and the technological recovery and use of resources, among other topics. OSTP leads interagency science and technology policy coordination efforts, assists the Office of Management and Budget (OMB) with an annual review and analysis of Federal research and development in budgets, and serves as a source of scientific and technological analysis and judgment for the President with respect to major policies, plans, and programs of the Federal Government.

Legal Disclaimer

The Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People is a white paper published by the White House Office of Science and Technology Policy. It is intended to support the development of policies and practices that protect civil rights and promote democratic values in the building, deployment, and governance of automated systems.

The *Blueprint for an AI Bill of Rights* is non-binding and does not constitute U.S. government policy. It does not supersede, modify, or direct an interpretation of any existing statute, regulation, policy, or international instrument. It does not constitute binding guidance for the public or Federal agencies and therefore does not require compliance with the principles described herein. It also is not determinative of what the U.S. government's position will be in any international negotiation. Adoption of these principles may not meet the requirements of existing statutes, regulations, policies, or international instruments, or the requirements of the Federal agencies that enforce them. These principles are not intended to, and do not, prohibit or limit any lawful activity of a government agency, including law enforcement, national security, or intelligence activities.

The appropriate application of the principles set forth in this white paper depends significantly on the context in which automated systems are being utilized. In some circumstances, application of these principles in whole or in part may not be appropriate given the intended use of automated systems to achieve government agency missions. Future sector-specific guidance will likely be necessary and important for guiding the use of automated systems in certain settings such as AI systems used as part of school building security or automated health diagnostic systems.

The *Blueprint for an AI Bill of Rights* recognizes that law enforcement activities require a balancing of equities, for example, between the protection of sensitive law enforcement information and the principle of notice; as such, notice may not be appropriate, or may need to be adjusted to protect sources, methods, and other law enforcement equities. Even in contexts where these principles may not apply in whole or in part, federal departments and agencies remain subject to judicial, privacy, and civil liberties oversight as well as existing policies and safeguards that govern automated systems, including, for example, Executive Order 13960, Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government (December 2020).

This white paper recognizes that national security (which includes certain law enforcement and homeland security activities) and defense activities are of increased sensitivity and interest to our nation's adversaries and are often subject to special requirements, such as those governing classified information and other protected data. Such activities require alternative, compatible safeguards through existing policies that govern automated systems and AI, such as the Department of Defense (DOD) AI Ethical Principles and Responsible AI Implementation Pathway and the Intelligence Community (IC) AI Ethics Principles and Framework. The implementation of these policies to national security and defense activities can be informed by the *Blueprint for an AI Bill of Rights* where feasible.

The *Blueprint for an AI Bill of Rights* is not intended to, and does not, create any legal right, benefit, or defense, substantive or procedural, enforceable at law or in equity by any party against the United States, its departments, agencies, or entities, its officers, employees, or agents, or any other person, nor does it constitute a waiver of sovereign immunity.

Copyright Information

Foreword

Among the great challenges posed to democracy today is the use of technology, data, and automated systems in ways that threaten the rights of the American public. Too often, these tools are used to limit our opportunities and prevent our access to critical resources or services. These problems are well documented. In America and around the world, systems supposed to help with patient care have proven unsafe, ineffective, or biased. Algorithms used in hiring and credit decisions have been found to reflect and reproduce existing unwanted inequities or embed new harmful bias and discrimination. Unchecked social media data collection has been used to threaten people's opportunities, undermine their privacy, or pervasively track their activity—often without their knowledge or consent.

These outcomes are deeply harmful—but they are not inevitable. Automated systems have brought about extraordinary benefits, from technology that helps farmers grow food more efficiently and computers that predict storm paths, to algorithms that can identify diseases in patients. These tools now drive important decisions across sectors, while data is helping to revolutionize global industries. Fueled by the power of American innovation, these tools hold the potential to redefine every part of our society and make life better for everyone.

This important progress must not come at the price of civil rights or democratic values, foundational American principles that President Biden has affirmed as a cornerstone of his Administration. On his first day in office, the President ordered the full Federal government to work to root out inequity, embed fairness in decision-making processes, and affirmatively advance civil rights, equal opportunity, and racial justice in America.[1] The President has spoken forcefully about the urgent challenges posed to democracy today and has regularly called on people of conscience to act to preserve civil rights—including the right to privacy, which he has called "the basis for so many more rights that we have come to take for granted that are ingrained in the fabric of this country."[2]

To advance President Biden's vision, the White House Office of Science and Technology Policy has identified five principles that should guide the design, use, and deployment of automated systems to protect the American public in the age of artificial intelligence. The Blueprint for an AI Bill of Rights is a guide for a society that protects all people from these threats—and uses technologies in ways that reinforce our highest values. Responding to the experiences of the American public, and informed by insights from researchers, technologists, advocates, journalists, and policymakers, this framework is accompanied by a technical companion—a handbook for anyone seeking to incorporate these protections into policy and practice, including detailed steps toward actualizing these principles in the technological design process. These principles help provide guidance whenever automated systems can meaningfully impact the public's rights, opportunities, or access to critical needs.

About this Framework

The Blueprint for an AI Bill of Rights is a set of five principles and associated practices to help guide the design, use, and deployment of automated systems to protect the rights of the American public in the age of artificial intel-ligence. Developed through extensive consultation with the American public, these principles are a blueprint for building and deploying automated systems that are aligned with democratic values and protect civil rights, civil liberties, and privacy. The Blueprint for an AI Bill of Rights includes this Foreword, the five principles, notes on Applying the The Blueprint for an AI Bill of Rights, and a Technical Companion that gives concrete steps that can be taken by many kinds of organizations—from governments at all levels to companies of all sizes—to uphold these values. Experts from across the private sector, governments, and international consortia have published principles and frameworks to guide the responsible use of automated systems; this framework provides a national values statement and toolkit that is sector-agnostic to inform building these protections into policy, practice, or the technological design process. Where existing law or policy—such as sector-specific privacy laws and oversight requirements—do not already provide guidance, the Blueprint for an AI Bill of Rights should be used to inform policy decisions.

Listening to the American Public

The White House Office of Science and Technology Policy has led a year-long process to seek and distill input from people across the country—from impacted communities and industry stakeholders to technology developers and other experts across fields and sectors, as well as policymakers throughout the Federal government—on the issue of algorithmic and data-driven harms and potential remedies. Through panel discussions, public listening sessions, meetings, a formal request for information, and input to a publicly accessible and widely-publicized email address, people throughout the United States, public servants across Federal agencies, and members of the international community spoke up about both the promises and potential harms of these technologies, and played a central role in shaping the Blueprint for an AI Bill of Rights. The core messages gleaned from these discussions include that AI has transformative potential to improve Americans' lives, and that preventing the harms of these technologies is both necessary and achievable. The Appendix includes a full list of public engagements.

SAFE AND EFFECTIVE SYSTEMS

YOU SHOULD BE PROTECTED FROM UNSAFE OR INEFFECTIVE SYSTEMS. Automated systems should be developed with consultation from diverse communities, stakeholders, and domain experts to identify concerns, risks, and potential impacts of the system. Systems should undergo pre-deployment testing, risk identification and mitigation, and ongoing monitoring that demonstrate they are safe and effective based on their intended use, mitigation of unsafe outcomes including those beyond the intended use, and adherence to domain-specific standards. Outcomes of these protective measures should include the possibility of not deploying the system or removing a system from use. Automated systems should not be designed with an intent or reasonably foreseeable possibility of endangering your safety or the safety of your community. They should be designed to proactively protect you from harms stemming from unintended, yet foreseeable, uses or impacts of automated systems. You should be protected from inappropriate or irrelevant data use in the design, development, and deployment of automated systems, and from the compounded harm of its reuse. Independent evaluation and reporting that confirms that the system is safe and effective, including reporting of steps taken to mitigate potential harms, should be performed and the results made public whenever possible.

ALGORITHMIC DISCRIMINATION PROTECTIONS

YOU SHOULD NOT FACE DISCRIMINATION BY ALGORITHMS AND SYSTEMS SHOULD BE USED AND DESIGNED IN AN EQUITABLE WAY. Algorithmic discrimination occurs when automated systems contribute to unjustified different treatment or impacts disfavoring people based on their race, color, ethnicity, sex (including pregnancy, childbirth, and related medical conditions, gender identity, intersex status, and sexual orientation), religion, age, national origin, disability, veteran status, genetic information, or any other classification protected by law. Depending on the specific circumstances, such algorithmic discrimination may violate legal protections. Designers, developers, and deployers of automated systems should take proactive and continuous measures to protect individuals and communities from algorithmic discrimination and to use and design systems in an equitable way. This protection should include proactive equity assessments as part of the system design, use of representative data and protection against proxies for demographic features, ensuring accessibility for people with disabilities in design and development, pre-deployment and ongoing disparity testing and mitigation, and clear organizational oversight. Independent evaluation and plain language reporting in the form of an algorithmic impact assessment, including disparity testing results and mitigation information, should be performed and made public whenever possible to confirm these protections.

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. You should be protected from violations of privacy through design choices that ensure such protections are included by default, including ensuring that data collection conforms to reasonable expectations and that only data strictly necessary for the specific context is collected. Designers, developers, and deployers of automated systems should seek your permission and respect your decisions regarding collection, use, access, transfer, and deletion of your data in appropriate ways and to the greatest extent possible; where not possible, alternative privacy by design safeguards should be used. Systems should not employ user experience and design decisions that obfuscate user choice or burden users with defaults that are privacy invasive. Consent should only be used to justify collection of data in cases where it can be appropriately and meaningfully given. Any consent requests should be brief, be understandable in plain language, and give you agency over data collection and the specific context of use; current hard-to-understand notice-and-choice practices for broad uses of data should be changed. 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. In sensitive domains, your data and related inferences should only be used for necessary functions, and you should be protected by ethical review and use prohibitions. You and your communities should be free from unchecked surveillance; surveillance technologies should be subject to heightened oversight that includes at least pre-deployment assessment of their potential harms and scope limits to protect privacy and civil liberties. 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. Whenever possible, you should have access to reporting that confirms your data decisions have been respected and provides an assessment of the potential impact of surveillance technologies on your rights, opportunities, or access.

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. Designers, developers, and deployers of automated systems should provide generally accessible plain language documentation including clear descriptions of the overall system functioning and the role automation plays, notice that such systems are in use, the individual or organization responsible for the system, and explanations of outcomes that are clear, timely, and accessible. Such notice should be kept up-to-date and people impacted by the system should be notified of significant use case or key functionality changes. You should know how and why an outcome impacting you was determined by an automated system, including when the automated system is not the sole input determining the outcome. Automated systems should provide explanations that are technically valid, meaningful and useful to you and to any operators or others who need to understand the system, and calibrated to the level of risk based on the context. Reporting that includes summary information about these automated systems in plain language and assessments of the clarity and quality of the notice and explanations should be made public whenever possible.

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. You should be able to opt out from automated systems in favor of a human alternative, where appropriate. Appropriateness should be determined based on reasonable expectations in a given context and with a focus on ensuring broad accessibility and protecting the public from especially harmful impacts. In some cases, a human or other alternative may be required by law. You should have access to timely human consideration and remedy by a fallback and escalation process if an automated system fails, it produces an error, or you would like to appeal or contest its impacts on you. Human consideration and fallback should be accessible, equitable, effective, maintained, accompanied by appropriate operator training, and should not impose an unreasonable burden on the public. 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. Reporting that includes a description of these human governance processes and assessment of their timeliness, accessibility, outcomes, and effectiveness should be made public whenever possible.

Definitions for key terms in The Blueprint for an AI Bill of Rights *can be found in Applying the Blueprint for an AI Bill of Rights. Accompanying analysis and tools for actualizing each principle can be found in the* Technical Companion.

Applying The Blueprint for an AI Bill of Rights

While many of the concerns addressed in this framework derive from the use of AI, the technical capabilities and specific definitions of such systems change with the speed of innovation, and the potential harms of their use occur even with less technologically sophisticated tools. Thus, this framework uses a two-part test to determine what systems are in scope. **This framework applies to (1) automated systems that (2) have the potential to meaningfully impact the American public's rights, opportunities, or access to critical resources or services.** These rights, opportunities, and access to critical resources of services should be enjoyed equally and be fully protected, regardless of the changing role that automated systems may play in our lives.

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.

A list of examples of automated systems for which these principles should be considered is provided in the Appendix. The Technical Companion, which follows, offers supportive guidance for any person or entity that creates, deploys, or oversees automated systems.

Considered together, the five principles and associated practices of the Blueprint for an AI Bill of Rights form an overlapping set of backstops against potential harms. This purposefully overlapping framework, when taken as a whole, forms a blueprint to help protect the public from harm. The measures taken to realize the vision set forward in this framework should be proportionate with the extent and nature of the harm, or risk of harm, to people's rights, opportunities, and access.

Relationship to Existing Law and Policy

The Blueprint for an AI Bill of Rights is an exercise in envisioning a future where the American public is protected from the potential harms, and can fully enjoy the benefits, of automated systems. It describes principles that can help ensure these protections. Some of these protections are already required by the U.S. Constitution or implemented under existing U.S. laws. For example, government surveillance, and data search and seizure are subject to legal requirements and judicial oversight. There are Constitutional requirements for human review of criminal investigative matters and statutory requirements for judicial review. Civil rights laws protect the American people against discrimination.

Applying The Blueprint for an AI Bill of Rights

Relationship to Existing Law and Policy

There are regulatory safety requirements for medical devices, as well as sector-, population-, or technology-specific privacy and security protections. Ensuring some of the additional protections proposed in this framework would require new laws to be enacted or new policies and practices to be adopted. In some cases, exceptions to the principles described in the Blueprint for an AI Bill of Rights may be necessary to comply with existing law, conform to the practicalities of a specific use case, or balance competing public interests. In particular, law enforcement, and other regulatory contexts may require government actors to protect civil rights, civil liberties, and privacy in a manner consistent with, but using alternate mechanisms to, the specific principles discussed in this framework. The Blueprint for an AI Bill of Rights is meant to assist governments and the private sector in moving principles into practice.

The expectations given in the Technical Companion are meant to serve as a blueprint for the development of additional technical standards and practices that should be tailored for particular sectors and contexts. While existing laws informed the development of the Blueprint for an AI Bill of Rights, this framework does not detail those laws beyond providing them as examples, where appropriate, of existing protective measures. This framework instead shares a broad, forward-leaning vision of recommended principles for automated system development and use to inform private and public involvement with these systems where they have the potential to meaningfully impact rights, opportunities, or access. Additionally, this framework does not analyze or take a position on legislative and regulatory proposals in municipal, state, and federal government, or those in other countries.

We have seen modest progress in recent years, with some state and local governments responding to these problems with legislation, and some courts extending longstanding statutory protections to new and emerging technologies. There are companies working to incorporate additional protections in their design and use of automated systems, and researchers developing innovative guardrails. Advocates, researchers, and government organizations have proposed principles for the ethical use of AI and other automated systems. These include the Organization for Economic Co-operation and Development's (OECD's) 2019 Recommendation on Artificial Intelligence, which includes principles for responsible stewardship of trustworthy AI and which the United States adopted, and Executive Order 13960 on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government, which sets out principles that govern the federal government's use of AI. The Blueprint for an AI Bill of Rights is fully consistent with these principles and with the direction in Executive Order 13985 on Advancing Racial Equity and Support for Underserved Communities Through the Federal Government. These principles find kinship in the Fair Information Practice Principles (FIPPs), derived from the 1973 report of an advisory committee to the U.S. Department of Health, Education, and Welfare, Records, Computers, and the Rights of Citizens.[4] While there is no single, universal articulation of the FIPPs, these core principles for managing information about individuals have been incorporated into data privacy laws and policies across the globe.[5] The Blueprint for an AI Bill of Rights embraces elements of the FIPPs that are particularly relevant to automated systems, without articulating a specific set of FIPPs or scoping applicability or the interests served to a single particular domain, like privacy, civil rights and civil liberties, ethics, or risk management. The Technical Companion builds on this prior work to provide practical next steps to move these principles into practice and promote common approaches that allow technological innovation to flourish while protecting people from harm.

Applying The Blueprint for an AI Bill of Rights

Definitions

ALGORITHMIC DISCRIMINATION: "Algorithmic discrimination" occurs when automated systems contribute to unjustified different treatment or impacts disfavoring people based on their race, color, ethnicity, sex (including pregnancy, childbirth, and related medical conditions, gender identity, intersex status, and sexual orientation), religion, age, national origin, disability, veteran status, genetic information, or any other classification protected by law. Depending on the specific circumstances, such algorithmic discrimination may violate legal protections. Throughout this framework the term "algorithmic discrimination" takes this meaning (and not a technical understanding of discrimination as distinguishing between items).

AUTOMATED SYSTEM: An "automated system" is any system, software, or process that uses computation as whole or part of a system to determine outcomes, make or aid decisions, inform policy implementation, collect data or observations, or otherwise interact with individuals and/or communities. Automated systems include, but are not limited to, systems derived from machine learning, statistics, or other data processing or artificial intelligence techniques, and exclude passive computing infrastructure. "Passive computing infrastructure" is any intermediary technology that does not influence or determine the outcome of decision, make or aid in decisions, inform policy implementation, or collect data or observations, including web hosting, domain registration, networking, caching, data storage, or cybersecurity. Throughout this framework, automated systems that are considered in scope are only those that have the potential to meaningfully impact individuals' or communi-ties' rights, opportunities, or access.

COMMUNITIES: "Communities" include: neighborhoods; social network connections (both online and offline); families (construed broadly); people connected by affinity, identity, or shared traits; and formal organizational ties. This includes Tribes, Clans, Bands, Rancherias, Villages, and other Indigenous communities. AI and other data-driven automated systems most directly collect data on, make inferences about, and may cause harm to individuals. But the overall magnitude of their impacts may be most readily visible at the level of communities. Accordingly, the concept of community is integral to the scope of the Blueprint for an AI Bill of Rights. United States law and policy have long employed approaches for protecting the rights of individuals, but existing frameworks have sometimes struggled to provide protections when effects manifest most clearly at a community level. For these reasons, the Blueprint for an AI Bill of Rights asserts that the harms of automated systems should be evaluated, protected against, and redressed at both the individual and community levels.

EQUITY: "Equity" means the consistent and systematic fair, just, and impartial treatment of all individuals. Systemic, fair, and just treatment must take into account the status of individuals who belong to underserved communities that have been denied such treatment, such as Black, Latino, and Indigenous and Native American persons, Asian Americans and Pacific Islanders and other persons of color; members of religious minorities; women, girls, and non-binary people; lesbian, gay, bisexual, transgender, queer, and intersex (LGBTQI+) persons; older adults; persons with disabilities; persons who live in rural areas; and persons otherwise adversely affected by persistent poverty or inequality.

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.

SURVEILLANCE TECHNOLOGY: "Surveillance technology" refers to products or services marketed for or that can be lawfully used to detect, monitor, intercept, collect, exploit, preserve, protect, transmit, and/or retain data, identifying information, or communications concerning individuals or groups. This framework limits its focus to both government and commercial use of surveillance technologies when juxtaposed with real-time or subsequent automated analysis and when such systems have a potential for meaningful impact on individuals' or communities' rights, opportunities, or access.

UNDERSERVED COMMUNITIES: The term "underserved communities" refers to communities that have been systematically denied a full opportunity to participate in aspects of economic, social, and civic life, as exemplified by the list in the preceding definition of "equity."

FROM PRINCIPLES TO PRACTICE

A Techincal Companion to the Blueprint for an AI Bill of Rights

Table of Contents

Using this Technical Companion

The Blueprint for an AI Bill of Rights is a set of five principles and associated practices to help guide the design, use, and deployment of automated systems to protect the rights of the American public in the age of artificial intelligence. This technical companion considers each principle in the Blueprint for an AI Bill of Rights and provides examples and concrete steps for communities, industry, governments, and others to take in order to build these protections into policy, practice, or the technological design process.

Taken together, the technical protections and practices laid out in the Blueprint for an AI Bill of Rights can help guard the American public against many of the potential and actual harms identified by researchers, technologists, advocates, journalists, policymakers, and communities in the United States and around the world. This technical companion is intended to be used as a reference by people across many circumstances – anyone impacted by automated systems, and anyone developing, designing, deploying, evaluating, or making policy to govern the use of an automated system.

Each principle is accompanied by three supplemental sections:

1 **Why this principle is important:**

This section provides a brief summary of the problems that the principle seeks to address and protect against, including illustrative examples.

2 **What should be expected of automated systems:**

- **The expectations for automated systems are meant to serve as a blueprint for the development of additional technical standards and practices that should be tailored for particular sectors and contexts.**
- This section outlines practical steps that can be implemented to realize the vision of the Blueprint for an AI Bill of Rights. The expectations laid out often mirror existing practices for technology development, including pre-deployment testing, ongoing monitoring, and governance structures for automated systems, but also go further to address unmet needs for change and offer concrete directions for how those changes can be made.
- Expectations about reporting are intended for the entity developing or using the automated system. The resulting reports can be provided to the public, regulators, auditors, industry standards groups, or others engaged in independent review, and should be made public as much as possible consistent with law, regulation, and policy, and noting that intellectual property, law enforcement, or national security considerations may prevent public release. Where public reports are not possible, the information should be provided to oversight bodies and privacy, civil liberties, or other ethics officers charged with safeguard ing individuals' rights. These reporting expectations are important for transparency, so the American people can have confidence that their rights, opportunities, and access as well as their expectations about technologies are respected.

3 **How these principles can move into practice:**

This section provides real-life examples of how these guiding principles can become reality, through laws, policies, and practices. It describes practical technical and sociotechnical approaches to protecting rights, opportunities, and access.

The examples provided are not critiques or endorsements, but rather are offered as illustrative cases to help provide a concrete vision for actualizing the Blueprint for an AI Bill of Rights. Effectively implementing these processes require the cooperation of and collaboration among industry, civil society, researchers, policymakers, technologists, and the public.

SAFE AND EFFECTIVE SYSTEMS

You should be protected from unsafe or ineffective systems. Automated systems should be developed with consultation from diverse communities, stakeholders, and domain experts to identify concerns, risks, and potential impacts of the system. Systems should undergo pre-deployment testing, risk identification and mitigation, and ongoing monitoring that demonstrate they are safe and effective based on their intended use, mitigation of unsafe outcomes including those beyond the intended use, and adherence to domain-specific standards. Outcomes of these protective measures should include the possibility of not deploying the system or removing a system from use. Automated systems should not be designed with an intent or reasonably foreseeable possibility of endangering your safety or the safety of your community. They should be designed to proactively protect you from harms stemming from unintended, yet foreseeable, uses or impacts of automated systems. You should be protected from inappropriate or irrelevant data use in the design, development, and deployment of automated systems, and from the compounded harm of its reuse. Independent evaluation and reporting that confirms that the system is safe and effective, including reporting of steps taken to mitigate potential harms, should be performed and the results made public whenever possible.

Why this principle is important

This section provides a brief summary of the problems which the principle seeks to address and protect against, including illustrative examples.

While technologies are being deployed to solve problems across a wide array of issues, our reliance on technology can also lead to its use in situations where it has not yet been proven to work—either at all or within an acceptable range of error. In other cases, technologies do not work as intended or as promised, causing substantial and unjustified harm. Automated systems sometimes rely on data from other systems, including historical data, allowing irrelevant information from past decisions to infect decision-making in unrelated situations. In some cases, technologies are purposefully designed to violate the safety of others, such as technologies designed to facilitate stalking; in other cases, intended or unintended uses lead to unintended harms.

Many of the harms resulting from these technologies are preventable, and actions are already being taken to protect the public. Some companies have put in place safeguards that have prevented harm from occurring by ensuring that key development decisions are vetted by an ethics review; others have identified and mitigated harms found through pre-deployment testing and ongoing monitoring processes. Governments at all levels have existing public consultation processes that may be applied when considering the use of new automated systems, and existing product development and testing practices already protect the American public from many potential harms.

Still, these kinds of practices are deployed too rarely and unevenly. Expanded, proactive protections could build on these existing practices, increase confidence in the use of automated systems, and protect the American public. Innovators deserve clear rules of the road that allow new ideas to flourish, and the American public deserves protections from unsafe outcomes. All can benefit from assurances that automated systems will be designed, tested, and consistently confirmed to work as intended, and that they will be proactively protected from foreseeable unintended harmful outcomes.

- A proprietary model was developed to predict the likelihood of sepsis in hospitalized patients and was implemented at hundreds of hospitals around the country. An independent study showed that the model predictions underperformed relative to the designer's claims while also causing 'alert fatigue' by falsely alerting likelihood of sepsis.[6]

- On social media, Black people who quote and criticize racist messages have had their own speech silenced when a platform's automated moderation system failed to distinguish this "counter speech" (or other critique and journalism) from the original hateful messages to which such speech responded.[7]

- A device originally developed to help people track and find lost items has been used as a tool by stalkers to track victims' locations in violation of their privacy and safety. The device manufacturer took steps after release to protect people from unwanted tracking by alerting people on their phones when a device is found to be moving with them over time and also by having the device make an occasional noise, but not all phones are able to receive the notification and the devices remain a safety concern due to their misuse.[8]

- An algorithm used to deploy police was found to repeatedly send police to neighborhoods they regularly visit, even if those neighborhoods were not the ones with the highest crime rates. These incorrect crime predictions were the result of a feedback loop generated from the reuse of data from previous arrests and algorithm predictions.[9]

Why this principle is important

This section provides a brief summary of the problems which the principle seeks to address and protect against, including illustrative examples.

- AI-enabled "nudification" technology that creates images where people appear to be nude—including apps that enable non-technical users to create or alter images of individuals without their consent—has proliferated at an alarming rate. Such technology is becoming a common form of image-based abuse that disproportionately impacts women. As these tools become more sophisticated, they are producing altered images that are increasingly realistic and are difficult for both humans and AI to detect as inauthentic. Regardless of authenticity, the experience of harm to victims of non-consensual intimate images can be devastatingly real—affecting their personal and professional lives, and impacting their mental and physical health.[10]

- A company installed AI-powered cameras in its delivery vans in order to evaluate the road safety habits of its drivers, but the system incorrectly penalized drivers when other cars cut them off or when other events beyond their control took place on the road. As a result, drivers were incorrectly ineligible to receive a bonus.[11]

What should be expected of automated systems

The expectations for automated systems are meant to serve as a blueprint for the development of additional technical standards and practices that are tailored for particular sectors and contexts.

In order to ensure that an automated system is safe and effective, it should include safeguards to protect the public from harm in a proactive and ongoing manner; avoid use of data inappropriate for or irrelevant to the task at hand, including reuse that could cause compounded harm; and demonstrate the safety and effectiveness of the system. These expectations are explained below.

Protect the public from harm in a proactive and ongoing manner

Consultation. The public should be consulted in the design, implementation, deployment, acquisition, and maintenance phases of automated system development, with emphasis on early-stage consultation before a system is introduced or a large change implemented. This consultation should directly engage diverse impacted communities to consider concerns and risks that may be unique to those communities, or disproportionately prevalent or severe for them. The extent of this engagement and the form of outreach to relevant stakeholders may differ depending on the specific automated system and development phase, but should include subject matter, sector-specific, and context-specific experts as well as experts on potential impacts such as civil rights, civil liberties, and privacy experts. For private sector applications, consultations before product launch may need to be confidential. Government applications, particularly law enforcement applications or applications that raise national security considerations, may require confidential or limited engagement based on system sensitivities and preexisting oversight laws and structures. Concerns raised in this consultation should be documented, and the automated system developers were proposing to create, use, or deploy should be reconsidered based on this feedback.

Testing. Systems should undergo extensive testing before deployment. This testing should follow domain-specific best practices, when available, for ensuring the technology will work in its real-world context. Such testing should take into account both the specific technology used and the roles of any human operators or reviewers who impact system outcomes or effectiveness; testing should include both automated systems testing and human-led (manual) testing. Testing conditions should mirror as closely as possible the conditions in which the system will be deployed, and new testing may be required for each deployment to account for material differences in conditions from one deployment to another. Following testing, system performance should be compared with the in-place, potentially human-driven, status quo procedures, with existing human performance considered as a performance baseline for the algorithm to meet pre-deployment, and as a lifecycle minimum performance standard. Decision possibilities resulting from performance testing should include the possibility of not deploying the system.

Risk identification and mitigation. Before deployment, and in a proactive and ongoing manner, potential risks of the automated system should be identified and mitigated. Identified risks should focus on the potential for meaningful impact on people's rights, opportunities, or access and include those to impacted communities that may not be direct users of the automated system, risks resulting from purposeful misuse of the system, and other concerns identified via the consultation process. Assessment and, where possible, measurement of the impact of risks should be included and balanced such that high impact risks receive attention and mitigation proportionate with those impacts. Automated systems with the intended purpose of violating the safety of others should not be developed or used; systems with such safety violations as identified unintended consequences should not be used until the risk can be mitigated. Ongoing risk mitigation may necessitate rollback or significant modification to a launched automated system.

What should be expected of automated systems

The expectations for automated systems are meant to serve as a blueprint for the development of additional technical standards and practices that are tailored for particular sectors and contexts.

Ongoing monitoring. Automated systems should have ongoing monitoring procedures, including recalibration procedures, in place to ensure that their performance does not fall below an acceptable level over time, based on changing real-world conditions or deployment contexts, post-deployment modification, or unexpected conditions. This ongoing monitoring should include continuous evaluation of performance metrics and harm assessments, updates of any systems, and retraining of any machine learning models as necessary, as well as ensuring that fallback mechanisms are in place to allow reversion to a previously working system. Monitoring should take into account the performance of both technical system components (the algorithm as well as any hardware components, data inputs, etc.) and human operators. It should include mechanisms for testing the actual accuracy of any predictions or recommendations generated by a system, not just a human operator's determination of their accuracy. Ongoing monitoring procedures should include manual, human-led monitoring as a check in the event there are shortcomings in automated monitoring systems. These monitoring procedures should be in place for the lifespan of the deployed automated system.

Clear organizational oversight. Entities responsible for the development or use of automated systems should lay out clear governance structures and procedures. This includes clearly-stated governance procedures before deploying the system, as well as responsibility of specific individuals or entities to oversee ongoing assessment and mitigation. Organizational stakeholders including those with oversight of the business process or operation being automated, as well as other organizational divisions that may be affected due to the use of the system, should be involved in establishing governance procedures. Responsibility should rest high enough in the organization that decisions about resources, mitigation, incident response, and potential rollback can be made promptly, with sufficient weight given to risk mitigation objectives against competing concerns. Those holding this responsibility should be made aware of any use cases with the potential for meaningful impact on people's rights, opportunities, or access as determined based on risk identification procedures. In some cases, it may be appropriate for an independent ethics review to be conducted before deployment.

Avoid inappropriate, low-quality, or irrelevant data use and the compounded harm of its reuse

Relevant and high-quality data. Data used as part of any automated system's creation, evaluation, or deployment should be relevant, of high quality, and tailored to the task at hand. Relevancy should be established based on research-backed demonstration of the causal influence of the data to the specific use case or justified more generally based on a reasonable expectation of usefulness in the domain and/or for the system design or ongoing development. Relevance of data should not be established solely by appealing to its historical connection to the outcome. High quality and tailored data should be representative of the task at hand and errors from data entry or other sources should be measured and limited. Any data used as the target of a prediction process should receive particular attention to the quality and validity of the predicted outcome or label to ensure the goal of the automated system is appropriately identified and measured. Additionally, justification should be documented for each data attribute and source to explain why it is appropriate to use that data to inform the results of the automated system and why such use will not violate any applicable laws. In cases of high-dimensional and/or derived attributes, such justifications can be provided as overall descriptions of the attribute generation process and appropriateness.

What should be expected of automated systems

The expectations for automated systems are meant to serve as a blueprint for the development of additional technical standards and practices that are tailored for particular sectors and contexts.

Derived data sources tracked and reviewed carefully. Data that is derived from other data through the use of algorithms, such as data derived or inferred from prior model outputs, should be identified and tracked, e.g., via a specialized type in a data schema. Derived data should be viewed as potentially high-risk inputs that may lead to feedback loops, compounded harm, or inaccurate results. Such sources should be carefully validated against the risk of collateral consequences.

Data reuse limits in sensitive domains. Data reuse, and especially data reuse in a new context, can result in the spreading and scaling of harms. Data from some domains, including criminal justice data and data indicating adverse outcomes in domains such as finance, employment, and housing, is especially sensitive, and in some cases its reuse is limited by law. Accordingly, such data should be subject to extra oversight to ensure safety and efficacy. Data reuse of sensitive domain data in other contexts (e.g., criminal data reuse for civil legal matters or private sector use) should only occur where use of such data is legally authorized and, after examination, has benefits for those impacted by the system that outweigh identified risks and, as appropriate, reasonable measures have been implemented to mitigate the identified risks. Such data should be clearly labeled to identify contexts for limited reuse based on sensitivity. Where possible, aggregated datasets may be useful for replacing individual-level sensitive data.

Demonstrate the safety and effectiveness of the system

Independent evaluation. Automated systems should be designed to allow for independent evaluation (e.g., via application programming interfaces). Independent evaluators, such as researchers, journalists, ethics review boards, inspectors general, and third-party auditors, should be given access to the system and samples of associated data, in a manner consistent with privacy, security, law, or regulation (including, e.g., intellectual property law), in order to perform such evaluations. Mechanisms should be included to ensure that system access for evaluation is: provided in a timely manner to the deployment-ready version of the system; trusted to provide genuine, unfiltered access to the full system; and truly independent such that evaluator access cannot be revoked without reasonable and verified justification.

Reporting.[12] Entities responsible for the development or use of automated systems should provide regularly-updated reports that include: an overview of the system, including how it is embedded in the organization's business processes or other activities, system goals, any human-run procedures that form a part of the system, and specific performance expectations; a description of any data used to train machine learning models or for other purposes, including how data sources were processed and interpreted, a summary of what data might be missing, incomplete, or erroneous, and data relevancy justifications; the results of public consultation such as concerns raised and any decisions made due to these concerns; risk identification and management assessments and any steps taken to mitigate potential harms; the results of performance testing including, but not limited to, accuracy, differential demographic impact, resulting error rates (overall and per demographic group), and comparisons to previously deployed systems; ongoing monitoring procedures and regular performance testing reports, including monitoring frequency, results, and actions taken; and the procedures for and results from independent evaluations. Reporting should be provided in a plain language and machine-readable manner.

How these principles can move into practice

Real-life examples of how these principles can become reality, through laws, policies, and practical technical and sociotechnical approaches to protecting rights, opportunities, and access.

Executive Order 13960 on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government requires that certain federal agencies adhere to nine principles when designing, developing, acquiring, or using AI for purposes other than national security or defense. These principles—while taking into account the sensitive law enforcement and other contexts in which the federal government may use AI, as opposed to private sector use of AI—require that AI is: (a) lawful and respectful of our Nation's values; (b) purposeful and performance-driven; (c) accurate, reliable, and effective; (d) safe, secure, and resilient; (e) understandable; (f) responsible and traceable; (g) regularly monitored; (h) transparent; and, (i) accountable. The Blueprint for an AI Bill of Rights is consistent with the Executive Order. Affected agencies across the federal government have released AI use case inventories[13] and are implementing plans to bring those AI systems into compliance with the Executive Order or retire them.

The law and policy landscape for motor vehicles shows that strong safety regulations—and measures to address harms when they occur—can enhance innovation in the context of complex technologies. Cars, like automated digital systems, comprise a complex collection of components. The National Highway Traffic Safety Administration,[14] through its rigorous standards and independent evaluation, helps make sure vehicles on our roads are safe without limiting manufacturers' ability to innovate.[15] At the same time, rules of the road are implemented locally to impose contextually appropriate requirements on drivers, such as slowing down near schools or playgrounds.[16]

From large companies to start-ups, industry is providing innovative solutions that allow organizations to mitigate risks to the safety and efficacy of AI systems, both before deployment and through monitoring over time.[17] These innovative solutions include risk assessments, auditing mechanisms, assessment of organizational procedures, dashboards to allow for ongoing monitoring, documentation procedures specific to model assessments, and many other strategies that aim to mitigate risks posed by the use of AI to companies' reputation, legal responsibilities, and other product safety and effectiveness concerns.

The Office of Management and Budget (OMB) has called for an expansion of opportunities for meaningful stakeholder engagement in the design of programs and services. OMB also points to numerous examples of effective and proactive stakeholder engagement, including the Community-Based Participatory Research Program developed by the National Institutes of Health and the participatory technology assessments developed by the National Oceanic and Atmospheric Administration.[18]

The National Institute of Standards and Technology (NIST) is developing a risk management framework to better manage risks posed to individuals, organizations, and society by AI.[19] The NIST AI Risk Management Framework, as mandated by Congress, is intended for voluntary use to help incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems. The NIST framework is being developed through a consensus-driven, open, transparent, and collaborative process that includes workshops and other opportunities to provide input. The NIST framework aims to foster the development of innovative approaches to address characteristics of trustworthiness including accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security (resilience), and mitigation of unintended and/or harmful bias, as well as of harmful uses. The NIST framework will consider and encompass principles such as transparency, accountability, and fairness during pre-design, design and development, deployment, use, and testing and evaluation of AI technologies and systems. It is expected to be released in the winter of 2022-23.

How these principles can move into practice

Real-life examples of how these principles can become reality, through laws, policies, and practical technical and sociotechnical approaches to protecting rights, opportunities, and access.

Some U.S government agencies have developed specific frameworks for ethical use of AI systems. The Department of Energy (DOE) has activated the AI Advancement Council that oversees coordination and advises on implementation of the DOE AI Strategy and addresses issues and/or escalations on the ethical use and development of AI systems.[20] The Department of Defense has adopted Artificial Intelligence Ethical Principles, and tenets for Responsible Artificial Intelligence specifically tailored to its national security and defense activities.[21] Similarly, the U.S. Intelligence Community (IC) has developed the Principles of Artificial Intelligence Ethics for the Intelligence Community to guide personnel on whether and how to develop and use AI in furtherance of the IC's mission, as well as an AI Ethics Framework to help implement these principles.[22]

The National Science Foundation (NSF) funds extensive research to help foster the development of automated systems that adhere to and advance their safety, security and effectiveness. Multiple NSF programs support research that directly addresses many of these principles: the National AI Research Institutes[23] support research on all aspects of safe, trustworthy, fair, and explainable AI algorithms and systems; the Cyber Physical Systems[24] program supports research on developing safe autonomous and cyber physical systems with AI components; the Secure and Trustworthy Cyberspace[25] program supports research on cybersecurity and privacy enhancing technologies in automated systems; the Formal Methods in the Field[26] program supports research on rigorous formal verification and analysis of automated systems and machine learning, and the Designing Accountable Software Systems[27] program supports research on rigorous and reproducible methodologies for developing software systems with legal and regulatory compliance in mind.

Some state legislatures have placed strong transparency and validity requirements on the use of pretrial risk assessments. The use of algorithmic pretrial risk assessments has been a cause of concern for civil rights groups.[28] Idaho Code Section 19-1910, enacted in 2019,[29] requires that any pretrial risk assessment, before use in the state, first be "shown to be free of bias against any class of individuals protected from discrimination by state or federal law", that any locality using a pretrial risk assessment must first formally validate the claim of its being free of bias, that "all documents, records, and information used to build or validate the risk assessment shall be open to public inspection," and that assertions of trade secrets cannot be used "to quash discovery in a criminal matter by a party to a criminal case."

Algorithmic Discrimination Protections

You should not face discrimination by algorithms and systems should be used and designed in an equitable way. Algorithmic discrimination occurs when automated systems contribute to unjustified different treatment or impacts disfavoring people based on their race, color, ethnicity, sex (including pregnancy, childbirth, and related medical conditions, gender identity, intersex status, and sexual orientation), religion, age, national origin, disability, veteran status, genetic infor-mation, or any other classification protected by law. Depending on the specific circumstances, such algorithmic discrimination may violate legal protections. Designers, developers, and deployers of automated systems should take proactive and continuous measures to protect individuals and communities from algorithmic discrimination and to use and design systems in an equitable way. This protection should include proactive equity assessments as part of the system design, use of representative data and protection against proxies for demographic features, ensuring accessibility for people with disabilities in design and development, pre-deployment and ongoing disparity testing and mitigation, and clear organizational oversight. Independent evaluation and plain language reporting in the form of an algorithmic impact assessment, including disparity testing results and mitigation information, should be performed and made public whenever possible to confirm these protections.

Why this principle is important

This section provides a brief summary of the problems which the principle seeks to address and protect against, including illustrative examples.

There is extensive evidence showing that automated systems can produce inequitable outcomes and amplify existing inequity.[30] Data that fails to account for existing systemic biases in American society can result in a range of consequences. For example, facial recognition technology that can contribute to wrongful and discriminatory arrests,[31] hiring algorithms that inform discriminatory decisions, and healthcare algorithms that discount the severity of certain diseases in Black Americans. Instances of discriminatory practices built into and resulting from AI and other automated systems exist across many industries, areas, and contexts. While automated systems have the capacity to drive extraordinary advances and innovations, algorithmic discrimination protections should be built into their design, deployment, and ongoing use.

Many companies, non-profits, and federal government agencies are already taking steps to ensure the public is protected from algorithmic discrimination. Some companies have instituted bias testing as part of their product quality assessment and launch procedures, and in some cases this testing has led products to be changed or not launched, preventing harm to the public. Federal government agencies have been developing standards and guidance for the use of automated systems in order to help prevent bias. Non-profits and companies have developed best practices for audits and impact assessments to help identify potential algorithmic discrimination and provide transparency to the public in the mitigation of such biases.

But there is much more work to do to protect the public from algorithmic discrimination to use and design automated systems in an equitable way. The guardrails protecting the public from discrimination in their daily lives should include their digital lives and impacts—basic safeguards against abuse, bias, and discrimination to ensure that all people are treated fairly when automated systems are used. This includes all dimensions of their lives, from hiring to loan approvals, from medical treatment and payment to encounters with the criminal justice system. Ensuring equity should also go beyond existing guardrails to consider the holistic impact that automated systems make on underserved communities and to institute proactive protections that support these communities.

- 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.[32]

- A hiring tool that learned the features of a company's employees (predominantly men) rejected women applicants for spurious and discriminatory reasons; resumes with the word "women's," such as "women's chess club captain," were penalized in the candidate ranking.[33]

- 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.[34]

- A risk assessment tool designed to predict the risk of recidivism for individuals in federal custody showed evidence of disparity in prediction. The tool overpredicts the risk of recidivism for some groups of color on the general recidivism tools, and underpredicts the risk of recidivism for some groups of color on some of the violent recidivism tools. The Department of Justice is working to reduce these disparities and has publicly released a report detailing its review of the tool.[35]

Why this principle is important

This section provides a brief summary of the problems which the principle seeks to address and protect against, including illustrative examples.

- An automated sentiment analyzer, a tool often used by technology platforms to determine whether a statement posted online expresses a positive or negative sentiment, was found to be biased against Jews and gay people. For example, the analyzer marked the statement "I'm a Jew" as representing a negative sentiment, while "I'm a Christian" was identified as expressing a positive sentiment.[36] This could lead to the preemptive blocking of social media comments such as: "I'm gay." A related company with this bias concern has made their data public to encourage researchers to help address the issue[37] and has released reports identifying and measuring this problem as well as detailing attempts to address it.[38]

- Searches for "Black girls," "Asian girls," or "Latina girls" return predominantly[39] sexualized content, rather than role models, toys, or activities.[40] Some search engines have been working to reduce the prevalence of these results, but the problem remains.[41]

- Advertisement delivery systems that predict who is most likely to click on a job advertisement end up delivering ads in ways that reinforce racial and gender stereotypes, such as overwhelmingly directing supermarket cashier ads to women and jobs with taxi companies to primarily Black people.[42]

- Body scanners, used by TSA at airport checkpoints, require the operator to select a "male" or "female" scanning setting based on the passenger's sex, but the setting is chosen based on the operator's perception of the passenger's gender identity. These scanners are more likely to flag transgender travelers as requiring extra screening done by a person. Transgender travelers have described degrading experiences associated with these extra screenings.[43] TSA has recently announced plans to implement a gender-neutral algorithm[44] while simultaneously enhancing the security effectiveness capabilities of the existing technology.

- 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.[45]

- An algorithm designed to identify patients with high needs for healthcare systematically assigned lower scores (indicating that they were not as high need) to Black patients than to those of white patients, even when those patients had similar numbers of chronic conditions and other markers of health.[46] In addition, healthcare clinical algorithms that are used by physicians to guide clinical decisions may include sociodemographic variables that adjust or "correct" the algorithm's output on the basis of a patient's race or ethnicity, which can lead to race-based health inequities.[47]

What should be expected of automated systems

The expectations for automated systems are meant to serve as a blueprint for the development of additional technical standards and practices that are tailored for particular sectors and contexts.

Any automated system should be tested to help ensure it is free from algorithmic discrimination before it can be sold or used. Protection against algorithmic discrimination should include designing to ensure equity, broadly construed. Some algorithmic discrimination is already prohibited under existing anti-discrimination law. The expectations set out below describe proactive technical and policy steps that can be taken to not only reinforce those legal protections but extend beyond them to ensure equity for underserved communities[48] even in circumstances where a specific legal protection may not be clearly established. These protections should be instituted throughout the design, development, and deployment process and are described below roughly in the order in which they would be instituted.

Protect the public from algorithmic discrimination in a proactive and ongoing manner

Proactive assessment of equity in design. Those responsible for the development, use, or oversight of automated systems should conduct proactive equity assessments in the design phase of the technology research and development or during its acquisition to review potential input data, associated historical context, accessibility for people with disabilities, and societal goals to identify potential discrimination and effects on equity resulting from the introduction of the technology. The assessed groups should be as inclusive as possible of the underserved communities mentioned in the equity definition: Black, Latino, and Indigenous and Native American persons, Asian Americans and Pacific Islanders and other persons of color; members of religious minorities; women, girls, and non-binary people; lesbian, gay, bisexual, transgender, queer, and intersex (LGBTQI+) persons; older adults; persons with disabilities; persons who live in rural areas; and persons otherwise adversely affected by persistent poverty or inequality. Assessment could include both qualitative and quantitative evaluations of the system. This equity assessment should also be considered a core part of the goals of the consultation conducted as part of the safety and efficacy review.

Representative and robust data. Any data used as part of system development or assessment should be representative of local communities based on the planned deployment setting and should be reviewed for bias based on the historical and societal context of the data. Such data should be sufficiently robust to identify and help to mitigate biases and potential harms.

Guarding against proxies. Directly using demographic information in the design, development, or deployment of an automated system (for purposes other than evaluating a system for discrimination or using a system to counter discrimination) runs a high risk of leading to algorithmic discrimination and should be avoided. In many cases, attributes that are highly correlated with demographic features, known as proxies, can contribute to algorithmic discrimination. In cases where use of the demographic features themselves would lead to illegal algorithmic discrimination, reliance on such proxies in decision-making (such as that facilitated by an algorithm) may also be prohibited by law. Proactive testing should be performed to identify proxies by testing for correlation between demographic information and attributes in any data used as part of system design, development, or use. If a proxy is identified, designers, developers, and deployers should remove the proxy; if needed, it may be possible to identify alternative attributes that can be used instead. At a minimum, organizations should ensure a proxy feature is not given undue weight and should monitor the system closely for any resulting algorithmic discrimination.

What should be expected of automated systems

The expectations for automated systems are meant to serve as a blueprint for the development of additional technical standards and practices that are tailored for particular sectors and contexts.

Ensuring accessibility during design, development, and deployment. Systems should be designed, developed, and deployed by organizations in ways that ensure accessibility to people with disabilities. This should include consideration of a wide variety of disabilities, adherence to relevant accessibility standards, and user experience research both before and after deployment to identify and address any accessibility barriers to the use or effectiveness of the automated system.

Disparity assessment. Automated systems should be tested using a broad set of measures to assess whether the system components, both in pre-deployment testing and in-context deployment, produce disparities. The demographics of the assessed groups should be as inclusive as possible of race, color, ethnicity, sex (including pregnancy, childbirth, and related medical conditions, gender identity, intersex status, and sexual orientation), religion, age, national origin, disability, veteran status, genetic information, or any other classification protected by law. The broad set of measures assessed should include demographic performance measures, overall and subgroup parity assessment, and calibration. Demographic data collected for disparity assessment should be separated from data used for the automated system and privacy protections should be instituted; in some cases it may make sense to perform such assessment using a data sample. For every instance where the deployed automated system leads to different treatment or impacts disfavoring the identified groups, the entity governing, implementing, or using the system should document the disparity and a justification for any continued use of the system.

Disparity mitigation. When a disparity assessment identifies a disparity against an assessed group, it may be appropriate to take steps to mitigate or eliminate the disparity. In some cases, mitigation or elimination of the disparity may be required by law. Disparities that have the potential to lead to algorithmic discrimination, cause meaningful harm, or violate equity[49] goals should be mitigated. When designing and evaluating an automated system, steps should be taken to evaluate multiple models and select the one that has the least adverse impact, modify data input choices, or otherwise identify a system with fewer disparities. If adequate mitigation of the disparity is not possible, then the use of the automated system should be reconsidered. One of the considerations in whether to use the system should be the validity of any target measure; unobservable targets may result in the inappropriate use of proxies. Meeting these standards may require instituting mitigation procedures and other protective measures to address algorithmic discrimination, avoid meaningful harm, and achieve equity goals.

Ongoing monitoring and mitigation. Automated systems should be regularly monitored to assess algorithmic discrimination that might arise from unforeseen interactions of the system with inequities not accounted for during the pre-deployment testing, changes to the system after deployment, or changes to the context of use or associated data. Monitoring and disparity assessment should be performed by the entity deploying or using the automated system to examine whether the system has led to algorithmic discrimination when deployed. This assessment should be performed regularly and whenever a pattern of unusual results is occurring. It can be performed using a variety of approaches, taking into account whether and how demographic information of impacted people is available, for example via testing with a sample of users or via qualitative user experience research. Riskier and higher-impact systems should be monitored and assessed more frequently. Outcomes of this assessment should include additional disparity mitigation, if needed, or fallback to earlier procedures in the case that equity standards are no longer met and can't be mitigated, and prior mechanisms provide better adherence to equity standards.

What should be expected of automated systems

The expectations for automated systems are meant to serve as a blueprint for the development of additional technical standards and practices that are tailored for particular sectors and contexts.

Demonstrate that the system protects against algorithmic discrimination

Independent evaluation. As described in the section on Safe and Effective Systems, entities should allow independent evaluation of potential algorithmic discrimination caused by automated systems they use or oversee. In the case of public sector uses, these independent evaluations should be made public unless law enforcement or national security restrictions prevent doing so. Care should be taken to balance individual privacy with evaluation data access needs; in many cases, policy-based and/or technological innovations and controls allow access to such data without compromising privacy.

Reporting. Entities responsible for the development or use of automated systems should provide reporting of an appropriately designed algorithmic impact assessment,[50] with clear specification of who performs the assessment, who evaluates the system, and how corrective actions are taken (if necessary) in response to the assessment. This algorithmic impact assessment should include at least: the results of any consultation, design stage equity assessments (potentially including qualitative analysis), accessibility designs and testing, disparity testing, document any remaining disparities, and detail any mitigation implementation and assessments. This algorithmic impact assessment should be made public whenever possible. Reporting should be provided in a clear and machine-readable manner using plain language to allow for more straightforward public accountability.

How these principles can move into practice

Real-life examples of how these principles can become reality, through laws, policies, and practical technical and sociotechnical approaches to protecting rights, opportunities, and access.

The federal government is working to combat discrimination in mortgage lending. The Department of Justice has launched a nationwide initiative to combat redlining, which includes reviewing how lenders who may be avoiding serving communities of color are conducting targeted marketing and advertising.[51] This initiative will draw upon strong partnerships across federal agencies, including the Consumer Financial Protection Bureau and prudential regulators. The Action Plan to Advance Property Appraisal and Valuation Equity includes a commitment from the agencies that oversee mortgage lending to include a nondiscrimination standard in the proposed rules for Automated Valuation Models.[52]

The Equal Employment Opportunity Commission and the Department of Justice have clearly laid out how employers' use of AI and other automated systems can result in discrimination against job applicants and employees with disabilities.[53] The documents explain how employers' use of software that relies on algorithmic decision-making may violate existing requirements under Title I of the Americans with Disabilities Act ("ADA"). This technical assistance also provides practical tips to employers on how to comply with the ADA, and to job applicants and employees who think that their rights may have been violated.

Disparity assessments identified harms to Black patients' healthcare access. A widely used healthcare algorithm relied on the cost of each patient's past medical care to predict future medical needs, recommending early interventions for the patients deemed most at risk. This process discriminated against Black patients, who generally have less access to medical care and therefore have generated less cost than white patients with similar illness and need. A landmark study documented this pattern and proposed practical ways that were shown to reduce this bias, such as focusing specifically on active chronic health conditions or avoidable future costs related to emergency visits and hospitalization.[54]

Large employers have developed best practices to scrutinize the data and models used for hiring. An industry initiative has developed Algorithmic Bias Safeguards for the Workforce, a structured questionnaire that businesses can use proactively when procuring software to evaluate workers. It covers specific technical questions such as the training data used, model training process, biases identified, and mitigation steps employed.[55]

Standards organizations have developed guidelines to incorporate accessibility criteria into technology design processes. The most prevalent in the United States is the Access Board's Section 508 regulations,[56] which are the technical standards for federal information communication technology (software, hardware, and web). Other standards include those issued by the International Organization for Standardization,[57] and the World Wide Web Consortium Web Content Accessibility Guidelines,[58] a globally recognized voluntary consensus standard for web content and other information and communications technology.

NIST has released Special Publication 1270, *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*.[59] The special publication: describes the stakes and challenges of bias in artificial intelligence and provides examples of how and why it can chip away at public trust; identifies three categories of bias in AI – systemic, statistical, and human – and describes how and where they contribute to harms; and describes three broad challenges for mitigating bias – datasets, testing and evaluation, and human factors – and introduces preliminary guidance for addressing them. Throughout, the special publication takes a sociotechnical perspective to identifying and managing AI bias.

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. You should be protected from violations of privacy through design choices that ensure such protections are included by default, including ensuring that data collection conforms to reasonable expectations and that only data strictly necessary for the specific context is collected. Designers, developers, and deployers of automated systems should seek your permission and respect your decisions regarding collection, use, access, transfer, and deletion of your data in appropriate ways and to the greatest extent possible; where not possible, alternative privacy by design safeguards should be used. Systems should not employ user experience and design decisions that obfuscate user choice or burden users with defaults that are privacy invasive. Consent should only be used to justify collection of data in cases where it can be appropriately and meaningfully given. Any consent requests should be brief, be understandable in plain language, and give you agency over data collection and the specific context of use; current hard-to-understand notice-and-choice practices for broad uses of data should be changed. 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. In sensitive domains, your data and related inferences should only be used for necessary functions, and you should be protected by ethical review and use prohibitions. You and your communities should be free from unchecked surveillance; surveillance technologies should be subject to heightened oversight that includes at least pre-deployment assessment of their potential harms and scope limits to protect privacy and civil liberties. 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. Whenever possible, you should have access to reporting that confirms your data decisions have been respected and provides an assessment of the potential impact of surveillance technologies on your rights, opportunities, or access.

Why this principle is important

This section provides a brief summary of the problems which the principle seeks to address and protect against, including illustrative examples.

Data privacy is a foundational and cross-cutting principle required for achieving all others in this framework. Surveillance and data collection, sharing, use, and reuse now sit at the foundation of business models across many industries, with more and more companies tracking the behavior of the American public, building individual profiles based on this data, and using this granular-level information as input into automated systems that further track, profile, and impact the American public. Government agencies, particularly law enforcement agencies, also use and help develop a variety of technologies that enhance and expand surveillance capabilities, which similarly collect data used as input into other automated systems that directly impact people's lives. Federal law has not grown to address the expanding scale of private data collection, or of the ability of governments at all levels to access that data and leverage the means of private collection.

Meanwhile, members of the American public are often unable to access their personal data or make critical decisions about its collection and use. Data brokers frequently collect consumer data from numerous sources without consumers' permission or knowledge.[60] Moreover, there is a risk that inaccurate and faulty data can be used to make decisions about their lives, such as whether they will qualify for a loan or get a job. Use of surveillance technologies has increased in schools and workplaces, and, when coupled with consequential management and evaluation decisions, it is leading to mental health harms such as lowered self-confidence, anxiety, depression, and a reduced ability to use analytical reasoning.[61] Documented patterns show that personal data is being aggregated by data brokers to profile communities in harmful ways.[62] The impact of all this data harvesting is corrosive, breeding distrust, anxiety, and other mental health problems; chilling speech, protest, and worker organizing; and threatening our democratic process.[63] The American public should be protected from these growing risks.

Increasingly, some companies are taking these concerns seriously and integrating mechanisms to protect consumer privacy into their products by design and by default, including by minimizing the data they collect, communicating collection and use clearly, and improving security practices. Federal government surveillance and other collection and use of data is governed by legal protections that help to protect civil liberties and provide for limits on data retention in some cases. Many states have also enacted consumer data privacy protection regimes to address some of these harms.

However, these are not yet standard practices, and the United States lacks a comprehensive statutory or regulatory framework governing the rights of the public when it comes to personal data. While a patchwork of laws exists to guide the collection and use of personal data in specific contexts, including health, employment, education, and credit, it can be unclear how these laws apply in other contexts and in an increasingly automated society. Additional protections would assure the American public that the automated systems they use are not monitoring their activities, collecting information on their lives, or otherwise surveilling them without context-specific consent or legal authority.

Why this principle is important

This section provides a brief summary of the problems which the principle seeks to address and protect against, including illustrative examples.

- An insurer might collect data from a person's social media presence as part of deciding what life insurance rates they should be offered.[64]
- A data broker harvested large amounts of personal data and then suffered a breach, exposing hundreds of thousands of people to potential identity theft. [65]
- A local public housing authority installed a facial recognition system at the entrance to housing complexes to assist law enforcement with identifying individuals viewed via camera when police reports are filed, leading the community, both those living in the housing complex and not, to have videos of them sent to the local police department and made available for scanning by its facial recognition software.[66]
- Companies use surveillance software to track employee discussions about union activity and use the resulting data to surveil individual employees and surreptitiously intervene in discussions.[67]

What should be expected of automated systems

The expectations for automated systems are meant to serve as a blueprint for the development of additional technical standards and practices that are tailored for particular sectors and contexts.

Traditional terms of service—the block of text that the public is accustomed to clicking through when using a website or digital app—are not an adequate mechanism for protecting privacy. The American public should be protected via built-in privacy protections, data minimization, use and collection limitations, and transparency, in addition to being entitled to clear mechanisms to control access to and use of their data—including their metadata—in a proactive, informed, and ongoing way. Any automated system collecting, using, sharing, or storing personal data should meet these expectations.

Protect privacy by design and by default

Privacy by design and by default. Automated systems should be designed and built with privacy protected by default. Privacy risks should be assessed throughout the development life cycle, including privacy risks from reidentification, and appropriate technical and policy mitigation measures should be implemented. This includes potential harms to those who are not users of the automated system, but who may be harmed by inferred data, purposeful privacy violations, or community surveillance or other community harms. Data collection should be minimized and clearly communicated to the people whose data is collected. Data should only be collected or used for the purposes of training or testing machine learning models if such collection and use is legal and consistent with the expectations of the people whose data is collected. User experience research should be conducted to confirm that people understand what data is being collected about them and how it will be used, and that this collection matches their expectations and desires.

Data collection and use-case scope limits. Data collection should be limited in scope, with specific, narrow identified goals, to avoid "mission creep." Anticipated data collection should be determined to be strictly necessary to the identified goals and should be minimized as much as possible. Data collected based on these identified goals and for a specific context should not be used in a different context without assessing for new privacy risks and implementing appropriate mitigation measures, which may include express consent. Clear timelines for data retention should be established, with data deleted as soon as possible in accordance with legal or policy-based limitations. Determined data retention timelines should be documented and justified.

Risk identification and mitigation. Entities that collect, use, share, or store sensitive data should attempt to proactively identify harms and seek to manage them so as to avoid, mitigate, and respond appropriately to identified risks. Appropriate responses include determining not to process data when the privacy risks outweigh the benefits or implementing measures to mitigate acceptable risks. Appropriate responses do not include sharing or transferring the privacy risks to users via notice or consent requests where users could not reasonably be expected to understand the risks without further support.

Privacy-preserving security. Entities creating, using, or governing automated systems should follow privacy and security best practices designed to ensure data and metadata do not leak beyond the specific consented use case. Best practices could include using privacy-enhancing cryptography or other types of privacy-enhancing technologies or fine-grained permissions and access control mechanisms, along with conventional system security protocols.

What should be expected of automated systems

The expectations for automated systems are meant to serve as a blueprint for the development of additional technical standards and practices that are tailored for particular sectors and contexts.

Protect the public from unchecked surveillance

Heightened oversight of surveillance. Surveillance or monitoring systems should be subject to heightened oversight that includes at a minimum assessment of potential harms during design (before deployment) and in an ongoing manner, to ensure that the American public's rights, opportunities, and access are protected. This assessment should be done before deployment and should give special attention to ensure there is not algorithmic discrimination, especially based on community membership, when deployed in a specific real-world context. Such assessment should then be reaffirmed in an ongoing manner as long as the system is in use.

Limited and proportionate surveillance. Surveillance should be avoided unless it is strictly necessary to achieve a legitimate purpose and it is proportionate to the need. Designers, developers, and deployers of surveillance systems should use the least invasive means of monitoring available and restrict monitoring to the minimum number of subjects possible. To the greatest extent possible consistent with law enforcement and national security needs, individuals subject to monitoring should be provided with clear and specific notice before it occurs and be informed about how the data gathered through surveillance will be used.

Scope limits on surveillance to protect rights and democratic values. Civil liberties and civil rights must not be limited by the threat of surveillance or harassment facilitated or aided by an automated system. Surveillance systems should not be used to monitor the exercise of democratic rights, such as voting, privacy, peaceful assembly, speech, or association, in a way that limits the exercise of civil rights or civil liberties. Information about or algorithmically-determined assumptions related to identity should be carefully limited if used to target or guide surveillance systems in order to avoid algorithmic discrimination; such identity-related information includes group characteristics or affiliations, geographic designations, location-based and association-based inferences, social networks, and biometrics. Continuous surveillance and monitoring systems should not be used in physical or digital workplaces (regardless of employment status), public educational institutions, and public accommodations. Continuous surveillance and monitoring systems should not be used in a way that has the effect of limiting access to critical resources or services or suppressing the exercise of rights, even where the organization is not under a particular duty to protect those rights.

Provide the public with mechanisms for appropriate and meaningful consent, access, and control over their data

Use-specific consent. Consent practices should not allow for abusive surveillance practices. Where data collectors or automated systems seek consent, they should seek it for specific, narrow use contexts, for specific time durations, and for use by specific entities. Consent should not extend if any of these conditions change; consent should be re-acquired before using data if the use case changes, a time limit elapses, or data is transferred to another entity (including being shared or sold). Consent requested should be limited in scope and should not request consent beyond what is required. Refusal to provide consent should be allowed, without adverse effects, to the greatest extent possible based on the needs of the use case.

Brief and direct consent requests. When seeking consent from users short, plain language consent requests should be used so that users understand for what use contexts, time span, and entities they are providing data and metadata consent. User experience research should be performed to ensure these consent requests meet performance standards for readability and comprehension. This includes ensuring that consent requests are accessible to users with disabilities and are available in the language(s) and reading level appropriate for the audience. User experience design choices that intentionally obfuscate or manipulate user choice (i.e., "dark patterns") should be not be used.

What should be expected of automated systems

The expectations for automated systems are meant to serve as a blueprint for the development of additional technical standards and practices that are tailored for particular sectors and contexts.

Data access and correction. People whose data is collected, used, shared, or stored by automated systems should be able to access data and metadata about themselves, know who has access to this data, and be able to correct it if necessary. Entities should receive consent before sharing data with other entities and should keep records of what data is shared and with whom.

Consent withdrawal and data deletion. Entities should allow (to the extent legally permissible) withdrawal of data access consent, resulting in the deletion of user data, metadata, and the timely removal of their data from any systems (e.g., machine learning models) derived from that data.[68]

Automated system support. Entities designing, developing, and deploying automated systems should establish and maintain the capabilities that will allow individuals to use their own automated systems to help them make consent, access, and control decisions in a complex data ecosystem. Capabilities include machine readable data, standardized data formats, metadata or tags for expressing data processing permissions and preferences and data provenance and lineage, context of use and access-specific tags, and training models for assessing privacy risk.

Demonstrate that data privacy and user control are protected

Independent evaluation. As described in the section on Safe and Effective Systems, entities should allow independent evaluation of the claims made regarding data policies. These independent evaluations should be made public whenever possible. Care will need to be taken to balance individual privacy with evaluation data access needs.

Reporting. When members of the public wish to know what data about them is being used in a system, the entity responsible for the development of the system should respond quickly with a report on the data it has collected or stored about them. Such a report should be machine-readable, understandable by most users, and include, to the greatest extent allowable under law, any data and metadata about them or collected from them, when and how their data and metadata were collected, the specific ways that data or metadata are being used, who has access to their data and metadata, and what time limitations apply to these data. In cases where a user login is not available, identity verification may need to be performed before providing such a report to ensure user privacy. Additionally, summary reporting should be proactively made public with general information about how peoples' data and metadata is used, accessed, and stored. Summary reporting should include the results of any surveillance pre-deployment assessment, including disparity assessment in the real-world deployment context, the specific identified goals of any data collection, and the assessment done to ensure only the minimum required data is collected. It should also include documentation about the scope limit assessments, including data retention timelines and associated justification, and an assessment of the impact of surveillance or data collection on rights, opportunities, and access. Where possible, this assessment of the impact of surveillance should be done by an independent party. Reporting should be provided in a clear and machine-readable manner.

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.[69] 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.

Extra Protections for Data Related to Sensitive Domains

- Continuous positive airway pressure machines gather data for medical purposes, such as diagnosing sleep apnea, and send usage data to a patient's insurance company, which may subsequently deny coverage for the device based on usage data. Patients were not aware that the data would be used in this way or monitored by anyone other than their doctor.[70]

- A department store company used predictive analytics applied to collected consumer data to determine that a teenage girl was pregnant, and sent maternity clothing ads and other baby-related advertisements to her house, revealing to her father that she was pregnant.[71]

- School audio surveillance systems monitor student conversations to detect potential "stress indicators" as a warning of potential violence.[72] Online proctoring systems claim to detect if a student is cheating on an exam using biometric markers.[73] 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.[74]

- Location data, acquired from a data broker, can be used to identify people who visit abortion clinics.[75]

- 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.[76] 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.

- Many employers transfer employee data to third party job verification services. This information is then used by potential future employers, banks, or landlords. In one case, a former employee alleged that a company supplied false data about her job title which resulted in a job offer being revoked.[77]

What should be expected of automated systems

The expectations for automated systems are meant to serve as a blueprint for the development of additional technical standards and practices that are tailored for particular sectors and contexts.

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

Real-life examples of how these principles can become reality, through laws, policies, and practical technical and sociotechnical approaches to protecting rights, opportunities, and access.

The Privacy Act of 1974 requires privacy protections for personal information in federal records systems, including limits on data retention, and also provides individuals a general right to access and correct their data. Among other things, the Privacy Act limits the storage of individual information in federal systems of records, illustrating the principle of limiting the scope of data retention. Under the Privacy Act, federal agencies may only retain data about an individual that is "relevant and necessary" to accomplish an agency's statutory purpose or to comply with an Executive Order of the President. The law allows for individuals to be able to access any of their individual information stored in a federal system of records, if not included under one of the systems of records exempted pursuant to the Privacy Act. In these cases, federal agencies must provide a method for an individual to determine if their personal information is stored in a particular system of records, and must provide procedures for an individual to contest the contents of a record about them. Further, the Privacy Act allows for a cause of action for an individual to seek legal relief if a federal agency does not comply with the Privacy Act's requirements. Among other things, a court may order a federal agency to amend or correct an individual's information in its records or award monetary damages if an inaccurate, irrelevant, untimely, or incomplete record results in an adverse determination about an individual's "qualifications, character, rights, ... opportunities..., or benefits."

NIST's Privacy Framework provides a comprehensive, detailed and actionable approach for organizations to manage privacy risks. The NIST Framework gives organizations ways to identify and communicate their privacy risks and goals to support ethical decision-making in system, product, and service design or deployment, as well as the measures they are taking to demonstrate compliance with applicable laws or regulations. It has been voluntarily adopted by organizations across many different sectors around the world.[78]

A school board's attempt to surveil public school students—undertaken without adequate community input—sparked a state-wide biometrics moratorium.[79] 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.[80] 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.

Federal law requires employers, and any consultants they may retain, to report the costs of surveilling employees in the context of a labor dispute, providing a transparency mechanism to help protect worker organizing. Employers engaging in workplace surveillance "where an object there-of, directly or indirectly, is [...] to obtain information concerning the activities of employees or a labor organization in connection with a labor dispute" must report expenditures relating to this surveillance to the Department of Labor Office of Labor-Management Standards, and consultants who employers retain for these purposes must also file reports regarding their activities.[81]

Privacy choices on smartphones show that when technologies are well designed, privacy and data agency can be meaningful and not overwhelming. These choices—such as contextual, timely alerts about location tracking—are brief, direct, and use-specific. Many of the expectations listed here for privacy by design and use-specific consent mirror those distributed to developers as best practices when developing for smart phone devices,[82] such as being transparent about how user data will be used, asking for app permissions during their use so that the use-context will be clear to users, and ensuring that the app will still work if users deny (or later revoke) some permissions.

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. Designers, developers, and deployers of automated systems should provide generally accessible plain language documentation including clear descriptions of the overall system functioning and the role automation plays, notice that such systems are in use, the individual or organization responsible for the system, and explanations of outcomes that are clear, timely, and accessible. Such notice should be kept up-to-date and people impacted by the system should be notified of significant use case or key functionality changes. You should know how and why an outcome impacting you was determined by an automated system, including when the automated system is not the sole input determining the outcome. Automated systems should provide explanations that are technically valid, meaningful and useful to you and to any operators or others who need to understand the system, and calibrated to the level of risk based on the context. Reporting that includes summary information about these automated systems in plain language and assessments of the clarity and quality of the notice and explanations should be made public whenever possible.

Why this principle is important

This section provides a brief summary of the problems which the principle seeks to address and protect against, including illustrative examples.

Automated systems now determine opportunities, from employment to credit, and directly shape the American public's experiences, from the courtroom to online classrooms, in ways that profoundly impact people's lives. But this expansive impact is not always visible. An applicant might not know whether a person rejected their resume or a hiring algorithm moved them to the bottom of the list. A defendant in the courtroom might not know if a judge denying their bail is informed by an automated system that labeled them "high risk." From correcting errors to contesting decisions, people are often denied the knowledge they need to address the impact of automated systems on their lives. Notice and explanations also serve an important safety and efficacy purpose, allowing experts to verify the reasonableness of a recommendation before enacting it.

In order to guard against potential harms, the American public needs to know if an automated system is being used. Clear, brief, and understandable notice is a prerequisite for achieving the other protections in this framework. Likewise, the public is often unable to ascertain how or why an automated system has made a decision or contributed to a particular outcome. The decision-making processes of automated systems tend to be opaque, complex, and, therefore, unaccountable, whether by design or by omission. These factors can make explanations both more challenging and more important, and should not be used as a pretext to avoid explaining important decisions to the people impacted by those choices. In the context of automated systems, clear and valid explanations should be recognized as a baseline requirement.

Providing notice has long been a standard practice, and in many cases is a legal requirement, when, for example, making a video recording of someone (outside of a law enforcement or national security context). In some cases, such as credit, lenders are required to provide notice and explanation to consumers. Techniques used to automate the process of explaining such systems are under active research and improvement and such explanations can take many forms. Innovative companies and researchers are rising to the challenge and creating and deploying explanatory systems that can help the public better understand decisions that impact them.

While notice and explanation requirements are already in place in some sectors or situations, the American public deserve to know consistently and across sectors if an automated system is being used in a way that impacts their rights, opportunities, or access. This knowledge should provide confidence in how the public is being treated, and trust in the validity and reasonable use of automated systems.

- A lawyer representing an older client with disabilities who had been cut off from Medicaid-funded home health-care assistance couldn't determine why, especially since the decision went against historical access practices. In a court hearing, the lawyer learned from a witness that the state in which the older client lived had recently adopted a new algorithm to determine eligibility.[83] The lack of a timely explanation made it harder to understand and contest the decision.

- 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.[84] 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.

Why this principle is important

This section provides a brief summary of the problems which the principle seeks to address and protect against, including illustrative examples.

- A predictive policing system claimed to identify individuals at greatest risk to commit or become the victim of gun violence (based on automated analysis of social ties to gang members, criminal histories, previous experiences of gun violence, and other factors) and led to individuals being placed on a watch list with no explanation or public transparency regarding how the system came to its conclusions.[85] Both police and the public deserve to understand why and how such a system is making these determinations.

- A system awarding benefits changed its criteria invisibly. Individuals were denied benefits due to data entry errors and other system flaws. These flaws were only revealed when an explanation of the system was demanded and produced.[86] The lack of an explanation made it harder for errors to be corrected in a timely manner.

What should be expected of automated systems

The expectations for automated systems are meant to serve as a blueprint for the development of additional technical standards and practices that are tailored for particular sectors and contexts.

An automated system should provide demonstrably clear, timely, understandable, and accessible notice of use, and explanations as to how and why a decision was made or an action was taken by the system. These expectations are explained below.

Provide clear, timely, understandable, and accessible notice of use and explanations

Generally accessible plain language documentation. The entity responsible for using the automated system should ensure that documentation describing the overall system (including any human components) is public and easy to find. The documentation should describe, in plain language, how the system works and how any automated component is used to determine an action or decision. It should also include expectations about reporting described throughout this framework, such as the algorithmic impact assessments described as part of Algorithmic Discrimination Protections.

Accountable. Notices should clearly identify the entity responsible for designing each component of the system and the entity using it.

Timely and up-to-date. Users should receive notice of the use of automated systems in advance of using or while being impacted by the technology. An explanation should be available with the decision itself, or soon thereafter. Notice should be kept up-to-date and people impacted by the system should be notified of use case or key functionality changes.

Brief and clear. Notices and explanations should be assessed, such as by research on users' experiences, including user testing, to ensure that the people using or impacted by the automated system are able to easily find notices and explanations, read them quickly, and understand and act on them. This includes ensuring that notices and explanations are accessible to users with disabilities and are available in the language(s) and reading level appropriate for the audience. Notices and explanations may need to be available in multiple forms, (e.g., on paper, on a physical sign, or online), in order to meet these expectations and to be accessible to the American public.

Provide explanations as to how and why a decision was made or an action was taken by an automated system

Tailored to the purpose. Explanations should be tailored to the specific purpose for which the user is expected to use the explanation, and should clearly state that purpose. An informational explanation might differ from an explanation provided to allow for the possibility of recourse, an appeal, or one provided in the context of a dispute or contestation process. For the purposes of this framework, 'explanation' should be construed broadly. An explanation need not be a plain-language statement about causality but could consist of any mechanism that allows the recipient to build the necessary understanding and intuitions to achieve the stated purpose. Tailoring should be assessed (e.g., via user experience research).

Tailored to the target of the explanation. Explanations should be targeted to specific audiences and clearly state that audience. An explanation provided to the subject of a decision might differ from one provided to an advocate, or to a domain expert or decision maker. Tailoring should be assessed (e.g., via user experience research).

What should be expected of automated systems

The expectations for automated systems are meant to serve as a blueprint for the development of additional technical standards and practices that are tailored for particular sectors and contexts.

Tailored to the level of risk. An assessment should be done to determine the level of risk of the automated system. In settings where the consequences are high as determined by a risk assessment, or extensive oversight is expected (e.g., in criminal justice or some public sector settings), explanatory mechanisms should be built into the system design so that the system's full behavior can be explained in advance (i.e., only fully transparent models should be used), rather than as an after-the-decision interpretation. In other settings, the extent of explanation provided should be tailored to the risk level.

Valid. The explanation provided by a system should accurately reflect the factors and the influences that led to a particular decision, and should be meaningful for the particular customization based on purpose, target, and level of risk. While approximation and simplification may be necessary for the system to succeed based on the explanatory purpose and target of the explanation, or to account for the risk of fraud or other concerns related to revealing decision-making information, such simplifications should be done in a scientifically supportable way. Where appropriate based on the explanatory system, error ranges for the explanation should be calculated and included in the explanation, with the choice of presentation of such information balanced with usability and overall interface complexity concerns.

Demonstrate protections for notice and explanation

Reporting. Summary reporting should document the determinations made based on the above considerations, including: the responsible entities for accountability purposes; the goal and use cases for the system, identified users, and impacted populations; the assessment of notice clarity and timeliness; the assessment of the explanation's validity and accessibility; the assessment of the level of risk; and the account and assessment of how explanations are tailored, including to the purpose, the recipient of the explanation, and the level of risk. Individualized profile information should be made readily available to the greatest extent possible that includes explanations for any system impacts or inferences. Reporting should be provided in a clear plain language and machine-readable manner.

How these principles can move into practice

Real-life examples of how these principles can become reality, through laws, policies, and practical technical and sociotechnical approaches to protecting rights, opportunities, and access.

People in Illinois are given written notice by the private sector if their biometric information is used. The Biometric Information Privacy Act enacted by the state contains a number of provisions concerning the use of individual biometric data and identifiers. Included among them is a provision that no private entity may "collect, capture, purchase, receive through trade, or otherwise obtain" such information about an individual, unless written notice is provided to that individual or their legally appointed representative. [87]

Major technology companies are piloting new ways to communicate with the public about their automated technologies. For example, a collection of non-profit organizations and companies have worked together to develop a framework that defines operational approaches to transparency for machine learning systems.[88] This framework, and others like it[89] inform the public about the use of these tools, going beyond simple notice to include reporting elements such as safety evaluations, disparity assessments, and explanations of how the systems work.

Lenders are required by federal law to notify consumers about certain decisions made about them. Both the Fair Credit Reporting Act and the Equal Credit Opportunity Act require in certain circumstances that consumers who are denied credit receive "adverse action" notices. Anyone who relies on the information in a credit report to deny a consumer credit must, under the Fair Credit Reporting Act, provide an "adverse action" notice to the consumer, which includes "notice of the reasons a creditor took adverse action on the application or on an existing credit account."[90] In addition, under the risk-based pricing rule,[91] lenders must either inform borrowers of their credit score, or else tell consumers when "they are getting worse terms because of information in their credit report." The CFPB has also asserted that "[t]he law gives every applicant the right to a specific explanation if their application for credit was denied, and that right is not diminished simply because a company uses a complex algorithm that it doesn't understand."[92] Such explanations illustrate a shared value that certain decisions need to be explained.

A California law requires that warehouse employees are provided with notice and explanation about quotas, potentially facilitated by automated systems, that apply to them. Warehousing employers in California that use quota systems (often facilitated by algorithmic monitoring systems) are required to provide employees with a written description of each quota that applies to the employee, including "quantified number of tasks to be performed or materials to be produced or handled, within the defined time period, and any potential adverse employment action that could result from failure to meet the quota."[93]

Across the federal government, agencies are conducting and supporting research on explainable AI systems. The NIST is conducting fundamental research on the explainability of AI systems. A multidisciplinary team of researchers aims to develop measurement methods and best practices to support the implementation of core tenets of explainable AI.[94] The Defense Advanced Research Projects Agency has a program on Explainable Artificial Intelligence that aims to create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance (prediction accuracy), and enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.[95] The National Science Foundation's program on Fairness in Artificial Intelligence also includes a specific interest in research foundations for explainable AI.[96]

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. You should be able to opt out from automated systems in favor of a human alternative, where appropriate. Appropriateness should be determined based on reasonable expectations in a given context and with a focus on ensuring broad accessibility and protecting the public from especially harmful impacts. In some cases, a human or other alternative may be required by law. You should have access to timely human consideration and remedy by a fallback and escalation process if an automated system fails, it produces an error, or you would like to appeal or contest its impacts on you. Human consideration and fallback should be accessible, equitable, effective, maintained, accompanied by appropriate operator training, and should not impose an unreasonable burden on the public. 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. Reporting that includes a description of these human governance processes and assessment of their timeliness, accessibility, outcomes, and effectiveness should be made public whenever possible.

Why this principle is important

This section provides a brief summary of the problems which the principle seeks to address and protect against, including illustrative examples.

There are many reasons people may prefer not to use an automated system: the system can be flawed and can lead to unintended outcomes; it may reinforce bias or be inaccessible; it may simply be inconvenient or unavailable; or it may replace a paper or manual process to which people had grown accustomed. Yet members of the public are often presented with no alternative, or are forced to endure a cumbersome process to reach a human decision-maker once they decide they no longer want to deal exclusively with the automated system or be impacted by its results. As a result of this lack of human reconsideration, many receive delayed access, or lose access, to rights, opportunities, benefits, and critical services. The American public deserves the assurance that, when rights, opportunities, or access are meaningfully at stake and there is a reasonable expectation of an alternative to an automated system, they can conveniently opt out of an automated system and will not be disadvantaged for that choice. In some cases, such a human or other alternative may be required by law, for example it could be required as "reasonable accommodations" for people with disabilities.

In addition to being able to opt out and use a human alternative, the American public deserves a human fallback system in the event that an automated system fails or causes harm. No matter how rigorously an automated system is tested, there will always be situations for which the system fails. The American public deserves protection via human review against these outlying or unexpected scenarios. In the case of time-critical systems, the public should not have to wait—immediate human consideration and fallback should be available. In many time-critical systems, such a remedy is already immediately available, such as a building manager who can open a door in the case an automated card access system fails.

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.

These critical protections have been adopted in some scenarios. Where automated systems have been introduced to provide the public access to government benefits, existing human paper and phone-based processes are generally still in place, providing an important alternative to ensure access. Companies that have introduced automated call centers often retain the option of dialing zero to reach an operator. When automated identity controls are in place to board an airplane or enter the country, there is a person supervising the systems who can be turned to for help or to appeal a misidentification.

The American people deserve the reassurance that such procedures are in place to protect their rights, opportunities, and access. People make mistakes, and a human alternative or fallback mechanism will not always have the right answer, but they serve as an important check on the power and validity of automated systems.

- An automated signature matching system is used as part of the voting process in many parts of the country to determine whether the signature on a mail-in ballot matches the signature on file. These signature matching systems are less likely to work correctly for some voters, including voters with mental or physical disabilities, voters with shorter or hyphenated names, and voters who have changed their name.[97] A human curing process,[98] which helps voters to confirm their signatures and correct other voting mistakes, is important to ensure all votes are counted,[99] and it is already standard practice in much of the country for both an election official and the voter to have the opportunity to review and correct any such issues.[100]

Why this principle is important

This section provides a brief summary of the problems which the principle seeks to address and protect against, including illustrative examples.

- An unemployment benefits system in Colorado required, as a condition of accessing benefits, that applicants have a smartphone in order to verify their identity. No alternative human option was readily available, which denied many people access to benefits.[101]

- A fraud detection system for unemployment insurance distribution incorrectly flagged entries as fraudulent, leading to people with slight discrepancies or complexities in their files having their wages withheld and tax returns seized without any chance to explain themselves or receive a review by a person.[102]

- A patient was wrongly denied access to pain medication when the hospital's software confused her medication history with that of her dog's. Even after she tracked down an explanation for the problem, doctors were afraid to override the system, and she was forced to go without pain relief due to the system's error.[103]

- A large corporation automated performance evaluation and other HR functions, leading to workers being fired by an automated system without the possibility of human review, appeal or other form of recourse.[104]

What should be expected of automated systems

The expectations for automated systems are meant to serve as a blueprint for the development of additional technical standards and practices that are tailored for particular sectors and contexts.

An automated system should provide demonstrably effective mechanisms to opt out in favor of a human alternative, where appropriate, as well as timely human consideration and remedy by a fallback system, with additional human oversight and safeguards for systems used in sensitive domains, and with training and assessment for any human-based portions of the system to ensure effectiveness.

Provide a mechanism to conveniently opt out from automated systems in favor of a human alternative, where appropriate

Brief, clear, accessible notice and instructions. Those impacted by an automated system should be given a brief, clear notice that they are entitled to opt-out, along with clear instructions for how to opt-out. Instructions should be provided in an accessible form and should be easily findable by those impacted by the automated system. The brevity, clarity, and accessibility of the notice and instructions should be assessed (e.g., via user experience research).

Human alternatives provided when appropriate. In many scenarios, there is a reasonable expectation of human involvement in attaining rights, opportunities, or access. When automated systems make up part of the attainment process, alternative timely human-driven processes should be provided. The use of a human alternative should be triggered by an opt-out process.

Timely and not burdensome human alternative. Opting out should be timely and not unreasonably burdensome in both the process of requesting to opt-out and the human-driven alternative provided.

Provide timely human consideration and remedy by a fallback and escalation system in the event that an automated system fails, produces error, or you would like to appeal or contest its impacts on you

Proportionate. The availability of human consideration and fallback, along with associated training and safeguards against human bias, should be proportionate to the potential of the automated system to meaningfully impact rights, opportunities, or access. Automated systems that have greater control over outcomes, provide input to high-stakes decisions, relate to sensitive domains, or otherwise have a greater potential to meaningfully impact rights, opportunities, or access should have greater availability (e.g., staffing) and oversight of human consideration and fallback mechanisms.

Accessible. Mechanisms for human consideration and fallback, whether in-person, on paper, by phone, or otherwise provided, should be easy to find and use. These mechanisms should be tested to ensure that users who have trouble with the automated system are able to use human consideration and fallback, with the understanding that it may be these users who are most likely to need the human assistance. Similarly, it should be tested to ensure that users with disabilities are able to find and use human consideration and fallback and also request reasonable accommodations or modifications.

Convenient. Mechanisms for human consideration and fallback should not be unreasonably burdensome as compared to the automated system's equivalent.

What should be expected of automated systems

The expectations for automated systems are meant to serve as a blueprint for the development of additional technical standards and practices that are tailored for particular sectors and contexts.

Equitable. Consideration should be given to ensuring outcomes of the fallback and escalation system are equitable when compared to those of the automated system and such that the fallback and escalation system provides equitable access to underserved communities.[105]

Timely. Human consideration and fallback are only useful if they are conducted and concluded in a timely manner. The determination of what is timely should be made relative to the specific automated system, and the review system should be staffed and regularly assessed to ensure it is providing timely consideration and fallback. In time-critical systems, this mechanism should be immediately available or, where possible, available before the harm occurs. Time-critical systems include, but are not limited to, voting-related systems, automated building access and other access systems, systems that form a critical component of healthcare, and systems that have the ability to withhold wages or otherwise cause immediate financial penalties.

Effective. The organizational structure surrounding processes for consideration and fallback should be designed so that if the human decision-maker charged with reassessing a decision determines that it should be overruled, the new decision will be effectively enacted. This includes ensuring that the new decision is entered into the automated system throughout its components, any previous repercussions from the old decision are also overturned, and safeguards are put in place to help ensure that future decisions do not result in the same errors.

Maintained. The human consideration and fallback process and any associated automated processes should be maintained and supported as long as the relevant automated system continues to be in use.

Institute training, assessment, and oversight to combat automation bias and ensure any human-based components of a system are effective.

Training and assessment. Anyone administering, interacting with, or interpreting the outputs of an automated system should receive training in that system, including how to properly interpret outputs of a system in light of its intended purpose and in how to mitigate the effects of automation bias. The training should reoccur regularly to ensure it is up to date with the system and to ensure the system is used appropriately. Assessment should be ongoing to ensure that the use of the system with human involvement provides for appropriate results, i.e., that the involvement of people does not invalidate the system's assessment as safe and effective or lead to algorithmic discrimination.

Oversight. Human-based systems have the potential for bias, including automation bias, as well as other concerns that may limit their effectiveness. The results of assessments of the efficacy and potential bias of such human-based systems should be overseen by governance structures that have the potential to update the operation of the human-based system in order to mitigate these effects.

What should be expected of automated systems

The expectations for automated systems are meant to serve as a blueprint for the development of additional technical standards and practices that are tailored for particular sectors and contexts.

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.

Demonstrate access to human alternatives, consideration, and fallback

Reporting. Reporting should include an assessment of timeliness and the extent of additional burden for human alternatives, aggregate statistics about who chooses the human alternative, along with the results of the assessment about brevity, clarity, and accessibility of notice and opt-out instructions. Reporting on the accessibility, timeliness, and effectiveness of human consideration and fallback should be made public at regular intervals for as long as the system is in use. This should include aggregated information about the number and type of requests for consideration, fallback employed, and any repeated requests; the timeliness of the handling of these requests, including mean wait times for different types of requests as well as maximum wait times; and information about the procedures used to address requests for consideration along with the results of the evaluation of their accessibility. For systems used in sensitive domains, reporting should include information about training and governance procedures for these technologies. Reporting should also include documentation of goals and assessment of meeting those goals, consideration of data included, and documentation of the governance of reasonable access to the technology. Reporting should be provided in a clear and machine-readable manner.

How these principles can move into practice

Real-life examples of how these principles can become reality, through laws, policies, and practical technical and sociotechnical approaches to protecting rights, opportunities, and access.

Healthcare "navigators" help people find their way through online signup forms to choose and obtain healthcare. A Navigator is "an individual or organization that's trained and able to help consumers, small businesses, and their employees as they look for health coverage options through the Marketplace (a government web site), including completing eligibility and enrollment forms."[106] For the 2022 plan year, the Biden-Harris Administration increased funding so that grantee organizations could "train and certify more than 1,500 Navigators to help uninsured consumers find affordable and comprehensive health coverage."[107]

The customer service industry has successfully integrated automated services such as chat-bots and AI-driven call response systems with escalation to a human support team.[108] Many businesses now use partially automated customer service platforms that help answer customer questions and compile common problems for human agents to review. These integrated human-AI systems allow companies to provide faster customer care while maintaining human agents to answer calls or otherwise respond to complicated requests. Using both AI and human agents is viewed as key to successful customer service.[109]

Ballot curing laws in at least 24 states require a fallback system that allows voters to correct their ballot and have it counted in the case that a voter signature matching algorithm incorrectly flags their ballot as invalid or there is another issue with their ballot, and review by an election official does not rectify the problem. Some federal courts have found that such cure procedures are constitutionally required.[110] Ballot curing processes vary among states, and include direct phone calls, emails, or mail contact by election officials.[111] Voters are asked to provide alternative information or a new signature to verify the validity of their ballot.