Algorithmic Discrimination

October 20, 2024

ACLU Colorado

Artificial Intelligence

- Artificial intelligence systems are reshaping and influencing core social domains that impact our daily lives, from the criminal legal system and education, to health care and beyond.

- Artificial intelligence refers to computer models, or algorithms, that mimic the cognitive functions of the human mind, such as learning and problem-solving.

- It is widely used for automated decision making — analyzing massive amounts of data, finding correlations and then making predictions about future outcomes.

- For example, employers use AI systems to determine who to advertise job opportunities to and who to hire, and housing providers use AI to screen potential tenants.

Algorithmic Discrimination

- When AI systems are developed in ways that do not adequately take into account existing racism, sexism, and other inequities, built-in algorithmic bias can undermine predictive decisions and result in invisible but very real discrimination.
- As these systems are deployed, they exacerbate existing disparities and create new roadblocks for already marginalized groups.
- AI is built by humans, and too often bias can appear in its design, development, and implementation.
- Establishing laws and regulations that mandate robust auditing for equity, transparency, and accountability, alongside litigation to stop and remedy civil rights violations, and direct engagement with technology companies can help guarantee equity.
- The ACLU of Colorado strives to challenge AI's power to preserve and exacerbate systemic racism and other inequities. In coalition with other civil rights groups, state and local advocates, and national partners, we push for better policy and support grassroots movements to work towards building more equitable AI systems.

SB 205

- While we know that a framework exists within SB-205 to get there, we believe there is still a long way to go before Colorado statute has AI specific protections that are robust enough to offset the algorithmic discrimination that exists in many AI systems today.
  - Affirmative defense
  - Rebuttable presumption
  - Loopholes in definitions
  - Lack of access to information

Affirmative Defense

6-1-1706(3) In any action commenced by the attorney general to enforce this part 17, it is an affirmative defense that the developer, deployer, or other person:

(a) Discovers and cures a violation of this part 17 as a result of:

(i) Feedback that the developer, deployer, or other person encourages deployers or users to provide to the developer, deployer, or other person;

(ii) Adversarial testing or red teaming, as those terms are defined or used by the national institute of standards and technology; or

(iii) An internal review process; and

(b) is otherwise in compliance with:

(i) The latest version of the "artificial intelligence risk management framework" published by the national institute of standards and technology in the united states department of commerce and standard iso/iec 42001 of the international organization for standardization;

(ii) Another nationally or internationally recognized risk management framework for artificial intelligence systems, if the standards are substantially equivalent to or more stringent than the requirements of this part 17; or

(iii) Any risk management framework for artificial intelligence systems that the attorney general, in the attorney general's discretion, may designate and, if designated, shall publicly disseminate.

(4) A developer, a deployer, or other person bears the burden of demonstrating to the attorney general that the requirements established in subsection (3) of this section have been satisfied.

Rebuttable Presumption

6-1-1702. Developer duty to avoid algorithmic discrimination - required documentation.

(1) On and after February 1, 2026, a developer of a high-risk artificial intelligence system shall use reasonable care to protect consumers from any known or reasonably foreseeable risks of algorithmic discrimination arising from the intended and contracted uses of the high-risk artificial intelligence system. In any enforcement action brought on or after February 1, 2026, by the Attorney General pursuant to section 6-1-1706, there is a rebuttable presumption that a developer used reasonable care as required under this section if the developer complied with this section and any additional requirements or obligations as set forth in rules promulgated by the Attorney General pursuant to section 6-1-1707.

Loopholes in Definitions

- 6-1-1701(9) (a) “High-risk artificial intelligence system" means any artificial intelligence system that, when deployed, makes, or is a substantial factor in making, a consequential decision.

(b) "High-risk artificial intelligence system" does not include:

(i) An artificial intelligence system if the artificial intelligence system is intended to:

(a) Perform a narrow procedural task; or

(b) Detect decision-making patterns or deviations from prior decision-making patterns and is not intended to replace or influence a previously completed human assessment without sufficient human review; or

- 6-1-1701(10) (a) "Intentional and substantial modification" or "intentionally and substantially modifies" means a deliberate change made to an artificial intelligence system that results in any new reasonably foreseeable risk of algorithmic discrimination.

(b) "Intentional and substantial modification" or "intentionally and substantially modifies" does not include a change made to a high-risk artificial intelligence system, or the performance of a high-risk artificial intelligence system, if:

(i) The high-risk artificial intelligence system continues to learn after the high-risk artificial intelligence system is:

(a) Offered, sold, leased, licensed, given, or otherwise made available to a deployer; or

(b) Deployed;

(ii) The change is made to the high-risk artificial intelligence system as a result of any learning described in subsection (10)(b)(i) of this section;

(iii) The change was predetermined by the deployer, or a third party contracted by the deployer, when the deployer or third party completed an initial impact assessment of such high-risk artificial intelligence system pursuant to section 6-1-1703 (3); and

(iv) The change is included in technical documentation for the high-risk artificial intelligence system.

Lack of Access to Information

6-1-1703(4)(b) On and after February 1, 2026, a deployer that has deployed a high-risk artificial intelligence system to make, or be a substantial factor in making, a consequential decision concerning a consumer shall, if the consequential decision is adverse to the consumer, provide to the consumer:

(i) A statement disclosing the principal reason or reasons for the consequential decision, including:

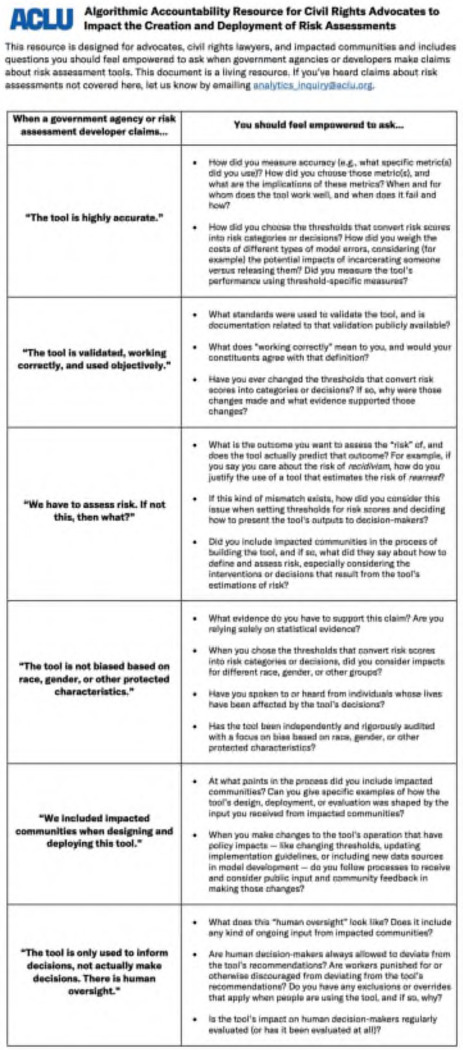
(a) The degree to which, and manner in which, the high-risk artificial intelligence system contributed to the consequential decision;

(b) The type of data that was processed by the high-risk artificial intelligence system in making the consequential decision; and

(c) The source or sources of the data described in subsection (4)(b)(i)(b) of this section;

(ii) An opportunity to correct any incorrect personal data that the high-risk artificial intelligence system processed in making, or as a substantial factor in making, the consequential decision; and

(iii) An opportunity to appeal an adverse consequential decision concerning the consumer arising from the deployment of a high-risk artificial intelligence system, which appeal must, if technically feasible, allow for human review unless providing the opportunity for appeal is not in the best interest of the consumer, including in instances in which any delay might pose a risk to the life or safety of such consumer.

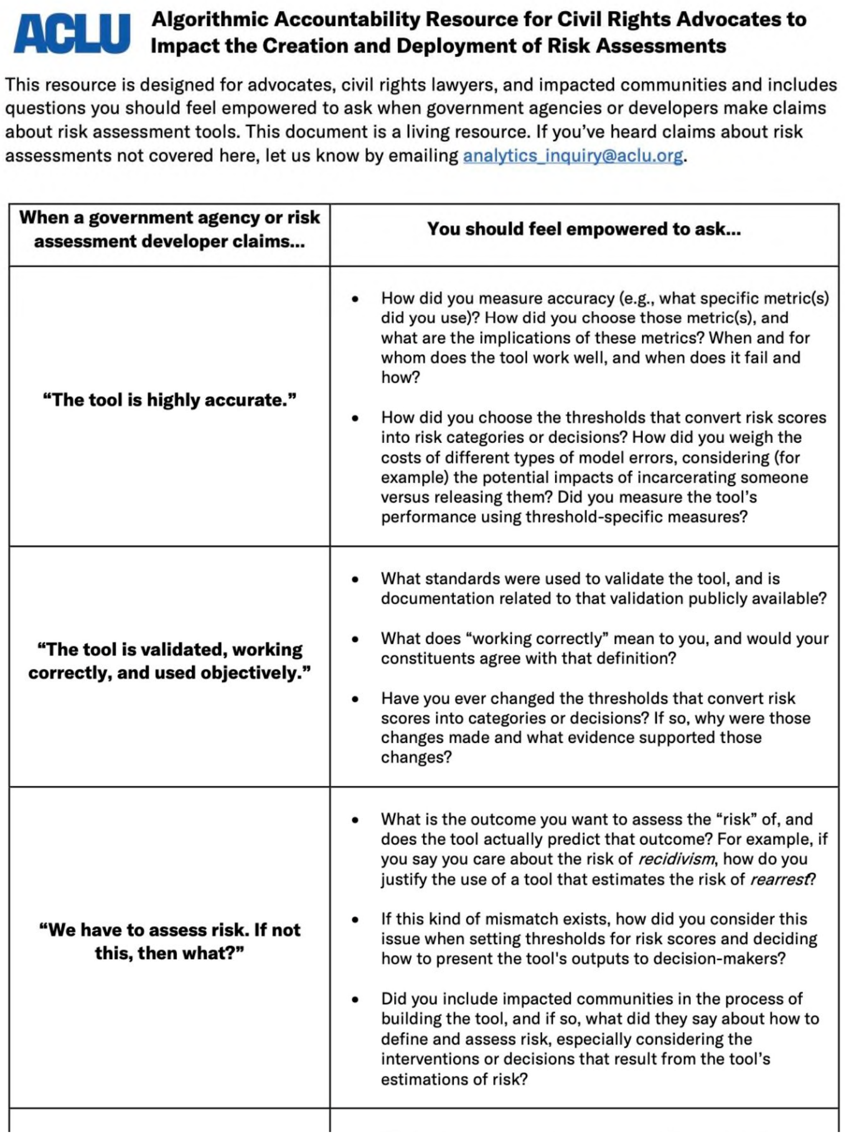


Algorithmic Accountability Resource for Civil Rights Advocates to Impact the Creation and Deployment of Risk Assessments

This resource is designed for advocates, civil rights lawyers, and impacted communities and includes questions you should feel empowered to ask when government agencies or developers make claims about risk assessment tools. This document is a living resource. If you've heard claims about risk assessments not covered here, let us know by emailing analytics_inquiry@aclu.org.

When a government agency or risk assessment developer claims…	You should feel empowered to ask…
"The tool is highly accurate."	• How did you measure accuracy (e.g., what specific metric(s) did you use)? How did you choose those metric(s), and what are the implications of these metrics? When and for whom does the tool work well, and when does it fail and how? • How did you choose the thresholds that convert risk scores into risk categories or decisions? How did you weigh the costs of different types of model errors, considering (for example) the potential impacts of incarcerating someone versus releasing them? Did you measure the tool's performance using threshold-specific measures?
"The tool is validated, working correctly, and used objectively."	• What standards were used to validate the tool, and is documentation related to that validation publicly available? • What does "working correctly" mean to you, and would your constituents agree with that definition? • Have you ever changed the thresholds that convert risk scores into categories or decisions? If so, why were those changes made and what evidence supported those changes?
"We have to assess risk. If not this, then what?"	• What is the outcome you want to assess the "risk" of, and does the tool actually predict that outcome? For example, if you say you care about the risk of *recidivism*, how do you justify the use of a tool that estimates the risk of *rearrest*? • If this kind of mismatch exists, how did you consider this issue when setting thresholds for risk scores and deciding how to present the tool's outputs to decision-makers? • Did you include impacted communities in the process of building the tool, and if so, what did they say about how to define and assess risk, especially considering the interventions or decisions that result from the tool's estimations of risk?
"The tool is not biased based on race, gender, or other protected characteristics."	• What evidence do you have to support this claim? Are you relying solely on statistical evidence? • When you chose the thresholds that convert risk scores into risk categories or decisions, did you consider impacts for different race, gender, or other groups? • Have you spoken to or heard from individuals whose lives have been affected by the tool's decisions? • Has the tool been independently and rigorously audited with a focus on bias based on race, gender, or other protected characteristics?
"We included impacted communities when designing and deploying this tool."	• At what points in the process did you include impacted communities? Can you give specific examples of how the tool's design, deployment, or evaluation was shaped by the input you received from impacted communities? • When you make changes to the tool's operation that have policy impacts — like changing thresholds, updating implementation guidelines, or including new data sources in model development — do you follow processes to receive and consider public input and community feedback in making those changes?
"The tool is only used to inform decisions, not actually make decisions. There is human oversight."	• What does this "human oversight" look like? Does it include any kind of ongoing input from impacted communities? • Are human decision-makers always allowed to deviate from the tool's recommendations? Are workers punished for or otherwise discouraged from deviating from the tool's recommendations? Do you have any exclusions or overrides that apply when people are using the tool, and if so, why? • Is the tool's impact on human decision-makers regularly evaluated (or has it been evaluated at all)?

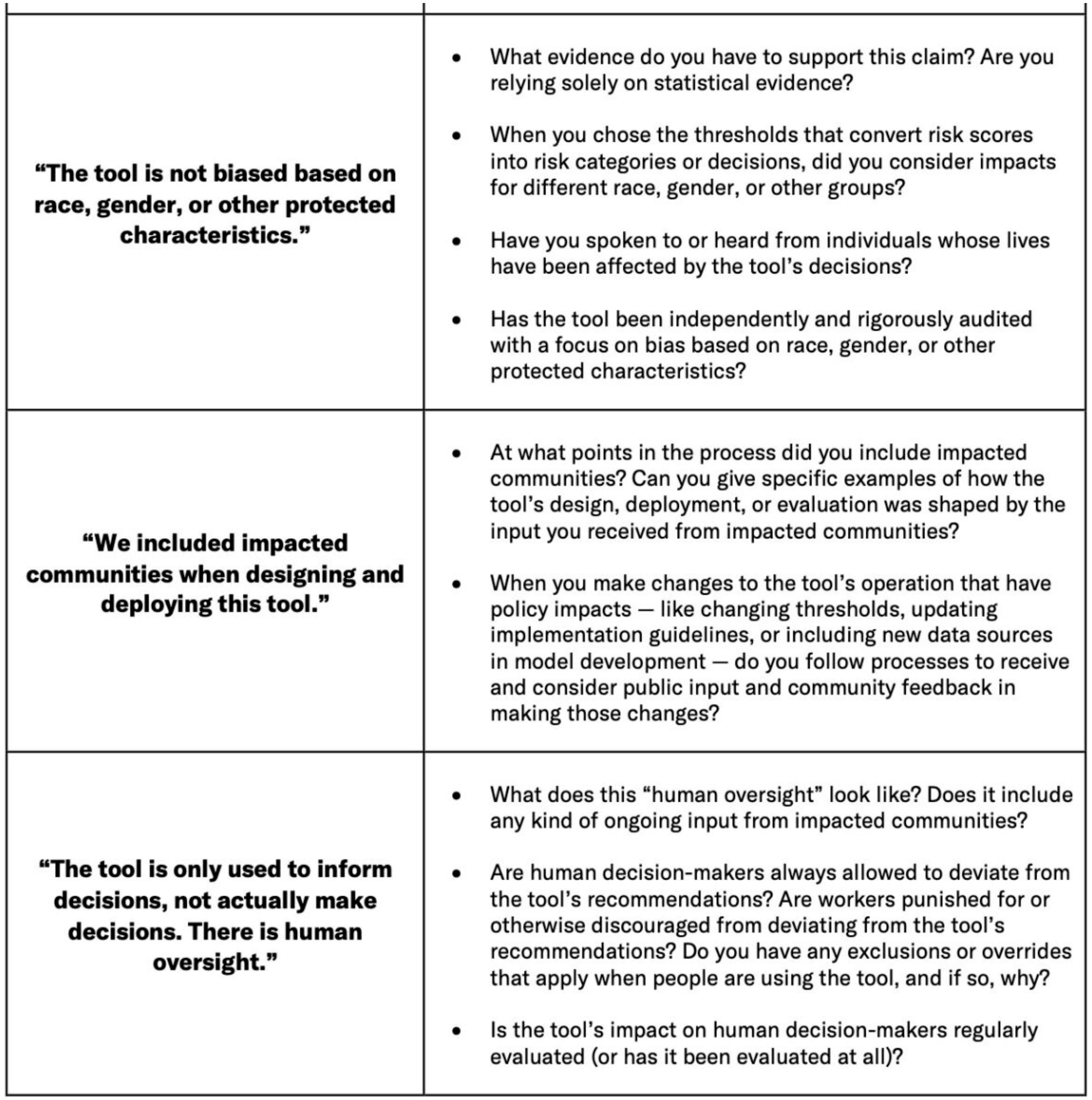
- Whether for government or private sector use, being specific about what the tool validation looks like is paramount to protecting from discrimination.



Algorithmic Accountability Resource for Civil Rights Advocates to Impact the Creation and Deployment of Risk Assessments

This resource is designed for advocates, civil rights lawyers, and impacted communities and includes questions you should feel empowered to ask when government agencies or developers make claims about risk assessment tools. This document is a living resource. If you've heard claims about risk assessments not covered here, let us know by emailing analytics_inquiry@aclu.org.

When a government agency or risk assessment developer claims...	You should feel empowered to ask...
"The tool is highly accurate."	• How did you measure accuracy (e.g., what specific metric(s) did you use)? How did you choose those metric(s), and what are the implications of these metrics? When and for whom does the tool work well, and when does it fail and how? • How did you choose the thresholds that convert risk scores into risk categories or decisions? How did you weigh the costs of different types of model errors, considering (for example) the potential impacts of incarcerating someone versus releasing them? Did you measure the tool's performance using threshold-specific measures?
"The tool is validated, working correctly, and used objectively."	• What standards were used to validate the tool, and is documentation related to that validation publicly available? • What does "working correctly" mean to you, and would your constituents agree with that definition? • Have you ever changed the thresholds that convert risk scores into categories or decisions? If so, why were those changes made and what evidence supported those changes?
"We have to assess risk. If not this, then what?"	• What is the outcome you want to assess the "risk" of, and does the tool actually predict that outcome? For example, if you say you care about the risk of *recidivism*, how do you justify the use of a tool that estimates the risk of *rearrest*? • If this kind of mismatch exists, how did you consider this issue when setting thresholds for risk scores and deciding how to present the tool's outputs to decision-makers? • Did you include impacted communities in the process of building the tool, and if so, what did they say about how to define and assess risk, especially considering the interventions or decisions that result from the tool's estimations of risk?



"The tool is not biased based on race, gender, or other protected characteristics."	- What evidence do you have to support this claim? Are you relying solely on statistical evidence? - When you chose the thresholds that convert risk scores into risk categories or decisions, did you consider impacts for different race, gender, or other groups? - Have you spoken to or heard from individuals whose lives have been affected by the tool's decisions? - Has the tool been independently and rigorously audited with a focus on bias based on race, gender, or other protected characteristics?
"We included impacted communities when designing and deploying this tool."	- At what points in the process did you include impacted communities? Can you give specific examples of how the tool's design, deployment, or evaluation was shaped by the input you received from impacted communities? - When you make changes to the tool's operation that have policy impacts — like changing thresholds, updating implementation guidelines, or including new data sources in model development — do you follow processes to receive and consider public input and community feedback in making those changes?
"The tool is only used to inform decisions, not actually make decisions. There is human oversight."	- What does this "human oversight" look like? Does it include any kind of ongoing input from impacted communities? - Are human decision-makers always allowed to deviate from the tool's recommendations? Are workers punished for or otherwise discouraged from deviating from the tool's recommendations? Do you have any exclusions or overrides that apply when people are using the tool, and if so, why? - Is the tool's impact on human decision-makers regularly evaluated (or has it been evaluated at all)?

ACLU

WE THE PEOPLE