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)?