FINDINGS AND RECOMMENDATION: Field Testing of Law Enforcement AI Tools

The National Artificial Intelligence Advisory Committee (NAIAC)
Law Enforcement Subcommittee (NAIAC-LE Subcommittee)

September 2024

NAIAC MEMBERS

Miriam Vogel, *Chair*
President and CEO of EqualAI

James Manyika, *Vice Chair*
Senior Vice President, Google, President for Research, Technology & Society

Amanda Ballantyne
Director of the AFL-CIO Technology Institute

Aneesh Chopra
Chief Strategy Officer of Arcadia

Jack Clark
Co-founder of Anthropic

Beth Cobert
Board Member, CBRE Group, Inc., Former Acting Director for the U.S. Office of Personnel Management

Angie Cooper
Executive Vice President of Heartland Forward

David Danks
Professor of Data Science and Philosophy at the University of California, San Diego

Victoria A. Espinel
President and CEO of BSA | The Software Alliance

Paula Goldman
Chief Ethical and Humane Use Officer at Salesforce

Susan Gonzales
Founder and CEO of AIandYou

Janet Haven
Executive Director of Data & Society

Daniel E. Ho
William Benjamin Scott and Luna M. Scott Professor of Law, Professor of Political Science, and Professor of Computer Science (by courtesy), Senior Fellow, Stanford Institute for Human-Centered AI, Stanford University

Ayanna Howard
Dean of Engineering at The Ohio State University

Christopher Howard
EVP, COO of ASU Enterprise

Jon Kleinberg
Professor in the Departments of Computer Science and Information Science at Cornell University

Ramayya Krishnan
W. W. Cooper and Ruth F. Cooper Professor of Management Science and Information Systems, Carnegie Mellon University

Ashley Llorens
Vice President, Distinguished Scientist, and Managing Director at Microsoft Research

Haniyeh Mahmoudian
Global AI Ethicist at DataRobot, Inc.

Christina Montgomery
Chief Privacy & Trust Officer and Vice President at IBM

Liz O'Sullivan
CEO of Vera

Fred Oswald
Professor and Herbert S. Autrey Chair in Social Sciences, Rice University

Trooper Sanders
Founder, Wise Whisper

Navrina Singh
Founder and CEO of Credo AI

Swami Sivasubramanian
Vice President for Data and Machine Learning Services at Amazon Web Services

Keith Strier
Senior Vice President of Global AI Markets, AMD

Reggie Townsend
Vice President of Data Ethics at SAS Institute

ACKNOWLEDGEMENTS

The NAIAC-LE Subcommittee members participated in the preparation of this document. The full membership of NAIAC reviewed and approved this Findings and Recommendation.

Armando Aguilar
Assistant Chief of Police, Miami Police Department

Anthony Bak
Head of AI, Palantir

Amanda Ballantyne
Director of the AFL-CIO Technology Institute

Jane Bambauer
Director - Marion B. Brechner First Amendment Project, Brechner Eminent Scholar at the College of Journalism and Communications and at Levin College of Law, University of Florida

Esha Bhandari
Deputy Director of the American Civil Liberties Union's Speech, Privacy, and Technology Project

Jennifer Eberhardt
Professor of Organizational Behavior and Psychology, Stanford University

Farhang Heydari
Assistant Professor of Law, Vanderbilt Law School

Benji Hutchinson
Chief Revenue Officer of Rank One Computing

Rashawn Ray
Vice-President and Executive Director of the AIR Equity Initiative

Cynthia Rudin
Professor of Computer Science, Electrical and Computer Engineering, Statistical Science, Mathematics, Biostatistics & Bioinformatics at Duke University

INTRODUCTION

The responsible use of AI in law enforcement requires AI developers to train, test, and audit their AI tools to ensure that the results of a predictive tool are sufficiently accurate, non-discriminatory, rights-respecting, and cost-effective. But the true value and risks of an AI tool will depend on how it operates in the real world. The White House now requires all federal agencies to test an AI tool for performance in real-world settings (OMB Memo M-24-10 §5 (c)(iv)(B)).) Very few resources are available to help guide the AI industry, law enforcement departments, and independent researchers through the process of testing AI tools when they are provisionally used in the field. This report and set of recommendations provide the infrastructure for AI field testing in the context of policing.

RECOMMENDATIONS

Support the use of field testing through recommendations, funding, and required disclosures.

A. Background and Motivation:

When law enforcement agencies adopt a new technology, they often have to rely on testing performed under relatively sterile conditions. Law enforcement may be justifiably concerned that their particular use of the tool in its operational context will lead to different performance characteristics than either published tests or as reported by other agencies. Also, the testing performed by producers of an AI tool sometimes have not been independently verified, and this simultaneously can create too much optimism for a poor-performing tool or too much skepticism of a useful tool. As a result, law enforcement (as well as the public) often don't have good information about whether the tool is as accurate, fair, high-performing, and cost-saving as expected.

This memorandum provides field testing guidance for law enforcement agencies to test the performance of an AI tool before it is fully adopted and integrated into normal use. We have synthesized a range of empirical testing methods and adapted them to the context of policing using the NIST AI Risk Management Framework (NIST RMF). Specifically, the guidance below will take field test designers through best practices for the "MAP" and "MEASURE" stages of AI risk management. The "MANAGEMENT" phase of trustworthy implementation of AI is not addressed in this project, but the evidence derived from field testing will allow decision-makers to make informed decisions as they manage and tradeoff multiple risks and objectives. The results of the real-world testing should be made public so that they may contribute to an informed conversation and debate about the responsible use of AI.

In addition to the example checklist below, we make three recommendations to create the support, incentives, and access to field testing.

B. Specific recommendations:

Recommendation 1:

Promote a standardized approach to field testing.

Consistent with OMB Guidance 5(c)(iv)(B)-(C), which require the testing of AI for performance in a real-world context and independently evaluating the AI, OMB should recommend that federal law enforcement agencies undergo a form of field testing that:

- provides transparency of process and results;
- establishes a baseline for operations without the tool;
- identifies a testing method; and
- establishes meaningful metrics that are of greatest relevance to the community.

The field testing requirement may be waived if the agency's use policy restricts the tool's use to the same use policy, and substantially similar conditions, under which it has been previously field tested by another agency.

Recommendation 2:

Require that the plans and results of real-world testing be made public.

OMB should revise OMB Guidance to clarify that field testing plans and results must be published in the relevant AI inventory or on another public government website. This should occur even if the AI application is not adopted following the field test.

Recommendation 3:

Provide funding and research support for field testing in state and local law enforcement agencies.

The Office of the President should charge NIST and the Department of Justice to create incentives and infrastructure for coordinated field studies of law enforcement AI tools. The program should allow AI companies to propose, and law enforcement agencies to opt into, multi-site field tests consistent with standardized field testing.

Selected proposals should be supported through equipment purchases, law enforcement grants, IT supports, and research team funding. The Office of the White House should administer special-purpose grants through the Department of Justice.

FINDINGS: THE FIELD TEST CHECKLIST

Field testing is essential to the government's and the public's understanding of AI applications in law enforcement. However, a good field test will need to be designed carefully to fit the context, needs, and practical limitations of a particular AI application. Researchers, police departments, and technology vendors will have to work together to create the conditions for high-quality field testing. This example checklist can be used and made public to craft a field testing plan. What follows is an annotated version of a field test checklist. Explanatory language is marked in blue. A (non-annotated) version of the checklist appears at the end of the document in Appendix A.

Description of the AI Tool: What is the AI tool, and how does it work?

Intended Use: Check all that apply.

Use Category		Description of Use(s)
Event Detection		
Person Identification		
AI-Assisted Surveillance		
Investigation of an Identified Subject		
Risk Assessment / Scoring as a Basis for Adverse Action		
Dot Connecting Methods Not Involving Personal Information		
Resource Allocation Decisions		
Accountability Technology		
Robotics		
Other (Please Describe)		

Explanatory Note: For further information on each of the use categories, and to see how a single AI tool may be used across multiple use categories, *see* NAIAC-LE Findings: Year 1 Roadmap.

Use Limitation Plan:

Provide here a link to existing use limitation plans.

Criminal investigations for which the tool may not be used (e.g., misdemeanors, non-violent crimes, traffic crimes):

Restrictions on staff who may not access or use the tool:

List all training or other prerequisites for users of the tool:

Will the output of the AI tool be used as evidence or justification for a search, seizure, or warrant application? __ Yes __ No

List all restrictions on the evidentiary use of the tool:

List all other constraints on the authorized use of this technology:

AI Impact Assessment: Place a link here to the current version of the department's AI Impact Assessment for this technology.

Identifying the Baseline(s):

How will the police department conduct change as a result of the introduction of the AI tool? What will they do, not do, or do differently when the tool is available?

____

____

____

Explanatory Notes: Consistent with the NIST AI RMF, the research team must identify a baseline (or "control condition") against which the performance, risks, and benefits of an AI tool will be measured. The testing methods described next will help the team collect metrics on the baseline/control condition in the process of studying AI in the field. Identifying the control condition up front will help the research team better understand the nature and limitations of the study that they will perform.

Testing Method: Mark the method you plan to use.

Explanatory Notes: The following testing methods are listed in the order that is typically associated with validity, from most rigorous (blind randomized controlled trials) to least (matched case studies). All of these tests, when designed properly, can produce useful information that improves the available evidence base. But the methods listed higher in the hierarchy are more likely to suggest causal relationships by removing the influence of external factors ("confounders").

From the menu below, what is the highest ranking methodology that your department, research team, and testing context can support? Refer to Appendix B for an explainer on threats to validity.

		Requirements	**Threats to validity**
Matched Case Study		Identifying one or more cases/incidents from the past or presently under investigation, possibly from another jurisdiction, that is factually similar to the case/incident treated with the AI tool	Very low power/inadequate sample size; External confounders
Pre/Post Testing		The ability to access or collect data on the chosen metrics from a sufficient period before the introduction of the AI tool	Low power/inadequate sample size; External confounders
Difference-in-Difference Testing ("Diff-in-Diff")		All requirements above plus access to the same type of data from another jurisdiction that is not adopting the tool	Spillover effects; Dissimilar comparison jurisdictions; Low power/inadequate sample size; External confounders

		Requirements	**Threads to Validity**
Staggered Rollout Testing		All requirements for "Pre/Post Testing" plus the introduction of the technology to different precincts, jurisdictions, or departments at different times (whether planned or unplanned)	Spillover effects; Dissimilar comparison jurisdictions; Low power/inadequate sample size; External confounders
Randomized Controlled Trials (also known as A/B Testing) ("RCTs")		All requirements for "Pre/Post Testing" plus an ability to randomly assign cases or officers to treatment and control conditions	Inadequate randomization; Spillover effects; Low power/ inadequate sample size; Ethical restrictions on random trials
Blind Randomized Controlled Trials ("Blind RCTs")		All requirements for RCTs plus an ability to prevent the law enforcement officers and staff from knowing whether the recommendation received is from the AI tool or from the control source	Inadequate blinding; Inadequate randomization; Spillover effects; Low power/ inadequate sample size
Other Please describe your methods:			

Describe here any plans to address and mitigate the threats to validity and to collect additional data on potential confounding factors:

____

____

____

Metrics:

Explanatory Notes: Any time a field test is designed in advance, it creates an opportunity to discover information about a wide range of effects. Each output metric typically adds only a minimal amount of extra cost or effort. For this reason, we recommend considering and collecting data on the widest range of outcomes that could plausibly be useful.

We have designed this questionnaire to help you brainstorm and identify metrics of two different sort: what might be called the "micro" metrics related to how a new tool performs on a per-use or per-case basis, and the "macro" metrics that attempt to measure the impact of the tool on the law enforcement system as a whole. Use the table below to identify as many metrics as possible that are either already routinely collected or that could, with reasonable effort, be collected in the future. Designers should keep in mind that in most cases, they will want to consider metrics that can be measured not only when the AI tool is in use, but also under similar situations when the tool is not used, and when other tools or techniques are used instead. To illustrate the process, we use examples based on existing studies of recidivism risk scoring systems (1, 2), of Miami's Real Time Crime Center (1), and of body-worn cameras (about which there are conflicting results — e.g., 1, 2, 3).

Note: A combined list of potential metrics discussed in this section is available in the unannotated Field Test Checklist in Appendix A

Keep in mind: all metrics must be observable and measurable for both the AI treatment and the control conditions.

Accuracy/Performance Metrics

When the tool is used, how will you know whether it has worked? Accuracy is the "closeness of results of observations, computations, or estimates to the true values or the values accepted as being true." (ISO/IEC TS 5723:2022. See also the discussion of "Valid and Reliable" characteristics in the NIST AI RMF). Field researchers must select an outcome metric that is going to be a stand-in for truth– something that can be accepted as representing the ground truth that is independent from inputs or results of the AI tool.

Micro Metrics
AI is trained for a specific quantified objective. This allows the AI to improve with more and more test cases tied to "true" answers. In the field, this outcome data isn't always available. For example, if a tool is going to be used to detect whether a bag is

concealing weapons, it can be trained using a series of bags that either are or are not pre-loaded with weapons. In the field, the accuracy will have to be assessed based on the outcome of subsequent searches if a search is permissible under the law and if the physical search is sensitive enough to find weapons when they exist. A tool used to identify an individual can be assessed based on later confirmation (or disconfirmation) of the identity.

To prepare to collect micro metrics related to accuracy, the research team will need to identify the **unit of analysis**, select the **population** under study, and select the **measures of performance** that can be assessed.

Selecting a **unit of analysis** is not always straightforward. The unit of analysis may be individuals when a tool is used to identify a suspect, or could be individual objects if the tool is used at a screening checkpoint for vehicles or luggage. The unit of analysis for an AI tool that generates reports based on body cam footage might be man-hours of service. An AI tool that attempts to find new leads for cold cases could be analyzed by the unit of case or victim. Other AI tools, such as those meant to prioritize tips and information, may require some creativity for setting the unit of analysis.

Depending on the AI application, it may also be necessary to select the **population** for field study in advance. This will often be a straightforward application of the use limitation policy established above.

Once a unit has been selected and the population identified, accuracy can be assessed using standard measures of performance and error. These include:

Binary measures: false positives, false negatives, true positives, true negatives
Continuous measures: sign and scale of calibration error, area under the curve
Non-response rates

Macro Metrics

Ultimately, the goal of an accurate tool is to achieve success solving or at least progressing a case. Thus, the system basic performance metrics attempt to observe the effect of the tool on these ultimate or intermediate goals. Macro measures of performance man include the following:

Clearance rate(s)
False search/arrest rate(s)
Secondary outcomes (e.g. finding witnesses)

For example, the study of Miami Police Department's Real Time Crime Center compared cases investigated using the center to similar cases investigated without the center, and found the following:

The use of MRTCC technologies has significantly improved the ability to clear violent crime cases. In the quasi-experiment that compared MRTCC-assisted case clearances with those of a stratified randomly drawn control sample, it was found that MRTCC-assisted cases had significantly greater odds of being cleared compared to similar cases without MRTCC support. After controlling for the neighborhood, crime type, and case-level characteristics, the MRTCC-assisted cases had 66 percent better odds of being cleared compared to those cases not receiving MRTCC support.

Bias/Disparity Metrics

Each of the accuracy metrics selected above should be used to detect and measure unintended disparities. Law enforcement and the public will want to be aware of any risk that the various forms of performance error identified in the last step are disproportionately common for one or more demographic groups. Disparate rates of accuracy or of error are not the only measures of AI bias (*see* Mayson (2019)), so researchers should consider using metrics that can also detect differences in discretionary decisions related to geographic location, types of crimes investigated, or other factors that may create disparities.

Demographics of Interest

Researchers will begin by identifying the demographic groups that need to be studied. The list of legally protected categories (race, gender, sexual orientation, national origin, religious affiliation) provide a good starting point, but not every legally protected class needs to be studied depending on the context and frequency of use of the AI tool. It will also not always be possible to collect accurate information about, e.g., religion or sexual orientation. Conversely, there may be demographic variables that are not among the subgroups recognized in Equal Protection law and other nondiscrimination laws that may nevertheless warrant careful study. Thus, a non-exhaustive list of demographic categories that researchers could study include:

- Race and Ethnicity (See U.S. Census Bureau and OMB race/ethnicity categories here)
- Sex or Gender
- National Origin
- Religion
- Sexual Orientation
- Age
- Income / Socioeconomic Status
- Zip code / Neighborhood Attributes

Micro Metrics

Each of the accuracy metrics selected above should be analyzed for disparities across the selected demographic groups. This may require some work mapping the demographic categories onto the selected unit of analysis. (Cars, for example, do not have a race or gender. So if an AI is used to select vehicles for inspection at a checkpoint, researchers will need to select one or more ways to code the demographic status, such as by including the race of the driver, of the owner, or of all passengers. We will use the term "unit of analysis status" for research plans that use something other than an individual as a unit of analysis.

We also recommend considering developing alternative units of analysis that will allow the research team to determine whether the adoption or use of AI differs based on the geolocation or the demographics of the victim. A non-exhaustive list of micro metrics includes the following:

Differential error rates (using errors selected for accuracy metrics)
Differential non-response rate
Differential AI use rates by crime victim status
Differential AI use rates by suspect/unit of analysis status

Macro Metrics

As with performance, the ultimate goal of guarding against AI bias is to ensure that the community as a whole can have confidence that new policing tools improve equity and fairness rather than exacerbating existing disparities. Researchers should consider some of the following macro measures to detect disparities at the community or population level:

Differential clearance rates by victim status
Differential false search or false arrest rates by suspect status
Differential investigation rates by victim status
Differential crime rates by victim status
Differential complaints of abuse rates by complainant status
Differential privacy costs by status

Note: The measures of disparities described here do not necessarily and automatically indicate a discrimination or inequitable outcomes. Differences in error rates, clearance rates, and other measures that appear across race, gender, and demographic lines may be explained by confounding factors such as age or gang presence. Research teams should collect data on potential confounding factors as frequently as possible. More generally, there should be care when interpreting the results that measures of bias are not necessarily measures of injustice.

Civil Rights, Efficiency, and Community Impact Metrics

Research teams should also decide in advance how they can measure additional risks and benefits related to civil rights (lost privacy, lost autonomy, and lost trust), police department efficiency (duration, officer hours, other costs), and community impact (crime rates, trust measures, perceptions of safety, and the subjective experiences of officers, suspects, witnesses, and community members). Possible micro and macro measures include but are not restricted to:

	Micro Metrics	**Macro Metrics (Key System Performance Indicators)**
Civilian Costs	Privacy costs (access or use of information by police or by others) Please describe:	Privacy costs (access or use of information by police or by others) Please describe:
	Time and autonomy costs (time spent for questioning, queuing in lines, witnesses/interviews)	Use of force rates
	Emotional costs (fear/intimidation)	Complaints of abuse rates
	Financial costs (e.g., fines and fees)	
	Collateral consequences (e.g., suspended driver's license)	Impact on First Amendment activities (e.g., chilling effects)
Efficiency Metrics	Time to solve, arrest, etc (duration)	Costs (price, compute costs, man hours)
	Officer hours to completion	Officer activity time distributions (how officers spend their time across different tasks)
Community Impact	Experience of officer	Crime rate(s)
	Experience of witnesses and suspects	Trust measures (surveys, focus groups, other)

Test Duration and Retest Plan

Explanatory Notes: Finally, the research team must determine how long the test will run (measured either in time or cases/units) and whether/when a field test will be conducted again. The duration of the test is likely to be determined based on the research needs (to ensure that there is enough information related to both the AI use and the control) and based on practical necessities (the needs of the public and the department).

The cadence of re-testing may depend on: (a) the initial field test results (a high performing tool may not need to be retested as soon as a moderately performing tool); (b) the likely rate of performance degradation; (c) the likely rate of performance improvements and upgrades; (d) the likelihood that the tool will be tested in the field elsewhere, by other departments; and (e) the costs and hassle of conducting the field test.

Planned Test Duration:

Expected Re-Test Plan if AI Tool Is Adopted (may be revised after initial results have been analyzed):

Appendix A: Complete Checklist

<u>Description of the AI Tool:</u> What is the AI tool, and how does it work?

<u>Intended Use:</u> Check all that apply.

Use Category		**Description of Use(s)**
Event Detection		
Person Identification		
AI-Assisted Surveillance		
Investigation of an Identified Subject		
Risk Assessment / Scoring as a Basis for Adverse Action		
Dot Connecting Methods Not Involving Personal Information		
Resource Allocation Decisions		
Accountability Technology		
Robotics		
Other (Please Describe)		

<u>Use Limitation Plan:</u>

Provide a link here to existing use limitation plans.

1. Criminal investigations for which the tool **<u>may not</u>** be used (e.g. misdemeanors, non-violent crimes, traffic crimes, etc.):

2. Restrictions on staff who may not access or use the tool:

List all training or other prerequisites for users of the tool:

3. Will the output of the AI tool be used as evidence or justification for a search, seizure, or warrant application? __ Yes __ No

List all restrictions on the evidentiary use of the tool:

4. List all other constraints on the authorized use of this technology:

AI Impact Assessment: Place a link here to the current version of the department's AI Impact Assessment for this technology.

Identifying the Baseline(s):

How will the police department conduct change as a result of the introduction of the AI tool? What will they do, not do, or do differently when the tool is available?

____

____

____

Testing Method: Mark the method you plan to use.

		Requirements	**Threats to validity**
Matched Case Study		Identifying one or more cases/incidents from the past or presently under investigation, possibly from another jurisdiction, that is factually similar to the case/incident treated with the AI tool	Very low power/inadequate sample size; External confounders

		Requirements	**Threats to validity**
Pre/Post Testing		The ability to access or collect data on the chosen metrics from a sufficient period before the introduction of the AI tool	Low power/inadequate sample size; External confounders
Difference-in-Difference Testing (“Diff-in-Diff”)		All requirements above plus access to the same type of data from another jurisdiction that is not adopting the tool	Spillover effects; Dissimilar comparison jurisdictions; Low power/inadequate sample size; External confounders
Staggered Rollout Testing		All requirements for “Pre/Post Testing” plus the introduction of the technology to different precincts, jurisdictions, or departments at different times (whether planned or unplanned)	Spillover effects; Dissimilar comparison jurisdictions; Low power/inadequate sample size; External confounders
Randomized Controlled Trials (also known as A/B Testing) (“RCTs”)		All requirements for “Pre/Post Testing” plus an ability to randomly assign cases or officers to treatment and control conditions	Inadequate randomization; Spillover effects; Low power/ inadequate sample size; Ethical restrictions on random trials
Blind Randomized Controlled Trials (“Blind RCTs”)		All requirements for RCTs plus an ability to prevent the law enforcement officers and staff from knowing whether the recommendation received is from the AI tool or from the control source	Inadequate blinding; Inadequate randomization; Spillover effects; Low power/inadequate sample size;
Other Please describe your methods:			

Describe here any plans to address and mitigate the threats to validity and to collect additional data on potential confounding factors:

Combined List of Potential Metrics

	Micro Metrics	**Macro Metrics (Key System Performance Indicators)**
Accuracy Metrics	**Unit of analysis: __________** **Study population: __________**	
	Binary measures (false positives, false negatives, true positives, true negatives)	Clearance rate(s)
	Continuous measures (sign and scale of calibration error, area under the curve)	False search/arrest rate(s)
	Non-response rate	Secondary outcomes (e.g. finding witnesses)
	Other (please describe)	Other (please describe)
Bias Metrics	**Demographic categories of concern: _________________**	
	Differential accuracy and error rates (using the accuracy metrics established above)	Differential clearance rates by victim status
	Differential non-response rate	Differential false search or false arrest rates by suspect status
	Differential AI use rates by crime victim status	Differential investigation rates by victim status
	Other (please describe)	Differential crime rates by victim status
		Differential complaints of abuse rates by complainant status

	Micro Metrics	**Macro Metrics (Key System Performance Indicators)**
		Differential privacy costs by status
		Other (please describe)
Civilian Costs	Privacy costs (access to private information)	Privacy (access to information)
	Time and autonomy costs (time spent for questioning, queuing in lines, witnesses/interviews)	Use of force rates
	Emotional costs (fear/intimidation)	Complaints of abuse rates
	Financial costs (e.g., fines and fees)	
	Collateral consequences (e.g., suspended drivers license)	Impact on First Amendment activities (e.g., chilling effects)
Efficiency Metrics	Time to solve, arrest, etc (duration)	Costs (price, compute costs, man hours)
	Officer hours to completion	Officer activity time distribution
Community Impact	Experience of officer	Crime rate(s)
	Experience of witnesses and suspects	Trust measures (surveys, focus groups, other)
	Experience of other community members	

Planned Test Duration:

____

____

Expected Re-Test Plan if AI Tool Is Adopted (may be revised after initial results have been analyzed):

____

____

Appendix B: Threats to Validity

Inadequate randomization (RCTs)
If assignment to the treatment or control groups are presumed to be randomized but are actually *not* random, there may be selection bias that researchers do not attempt to control against. This can occur, for example, if the researchers effectively allow police officers to decide whether they will or will not be part of the experimental group since those who are eager to use the new tool may be different in a range of ways from those who are not. *See* this explainer for failures of randomization.

Spillover effects (RCTs)
Sometimes, it is impossible to keep an experimental treatment from affecting the control group. For example, if use of an AI tool leads to an insight about an area of town or a time of day when crime is more likely to occur, it is plausible that ordinary conversation between police officers will allow that insight to spill over into the control group, potentially affecting the control group indirectly. *See* the Wikipedia summary here.

Ethical limitations (RCTs)
If a law enforcement department has an investigation tool that may provide a valuable lead, it may be unethical to refrain from using the tool for a case that has been assigned to the control group. See this summary from bioethics or this skeptical take on the topic.)

External confounders (difference-in-difference and pre/post studies)
In a pre/post study, the period during which an AI is used instead of the control method may be very different for reasons that have nothing to do with the tool. Imagine, for example, that a department introduced an AI tool in December of 2019, immediately before the world-wide impact of the COVID-19 pandemic. The data from the "pre" period may be very different from the "post" period due to the wide range of social and economic changes, and as a result the AI tool may receive unfair credit or lack of credit. While major pandemics are obvious confounders for empirical validity, other factors tend to affect crime and investigation rates as well. For example, election years cause known changes in crime reporting and investigation, and changes in economic trends (e.g. recessions) and changes in crime trends (e.g. a sudden increase in gang violence) can also affect test outcomes. Difference-in-difference models can reduce the problems of confounders to some extent, but not entirely if the trend affects the comparison jurisdiction differently. *See* this explainer and this article on correction methods for more detail.

Small sample size / low power (all)
If researchers have only a small number of cases to assess, they will not have confidence that an AI tool has or has not made a difference unless the AI tool happens to be wildly effective as compared to the baseline/control method. Differences between test and control cases might be a matter of random chance. *See* this explainer for more detail.

Appendix C: Illustrative Studies and Field Tests

Matched Case Study	Carr, Jillian and Jennifer L. Doleac. "The Geography, Incidence, and Underreporting of Gun Violence: New Evidence Using Shotspotter Data." SSRN, April 2016. https://ssrn.com/abstract=2770506. For general descriptions, see: Loftin, Colin and David McDowall. "The analysis of case-control studies in criminology." *J Quant Criminol* 4 (1988): 85–98. https://link.springer.com/article/10.1007/BF01066886. Rose, Sheri and Mark J van der Laan. "Why match? Investigating matched case-control study designs with causal effect estimation." *Int J Biostat* 5 (1) (2009). https://pubmed.ncbi.nlm.nih.gov/20231866/. Dehejia, Rajeev H. and Sadek Wahba. "Propensity Score Matching Methods for Nonexperimental Causal Studies." *The Review of Economics and Statistics* 84 (2002): 151–61. https://direct.mit.edu/rest/article-abstract/84/1/151/57311/Propensity-Score-Matching-Methods-for.
Pre/Post Testing	Guerette, Rob and Kimberly Przeszlowski "Does the Rapid Deployment of Information to Police Improve Crime Solvability? A Quasi-Experimental Impact Evaluation of Real-Time Crime Center (RTCC) Technologies on Violent Crime Incident Outcomes." *Justice Quarterly* 40 (7) (2023): 950-974. https://www.tandfonline.com/doi/full/10.1080/07418825.2023.2264362#:~:text=087.
Difference-in-Difference Testing ("Diff-in-Diff")	Weisburd, David et al. "Does Crime Just Move Around the Corner? A Controlled Study of Spatial Displacement and Diffusion of Crime Control Benefits." *Criminology* 44 (2006): 549–92. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-9125.2006.00057.x. Doleac, Jennifer. "The Effects of DNA Databases on Crime." *American Economic Journal: Applied Economics* 9 (1) (2017):165-201. https://www.aeaweb.org/articles?id=10.1257/app.20150043. Stevenson, Megan and Jennifer Doleac. "Algorithmic Risk Assessment in the Hands of Humans." SSRN, December 2019. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3489440.

Staggered Rollout Testing	Braga, Anthony et al. “An Ex Post Facto Evaluation Framework for Place-Based Police Interventions.” *Evaluation Review*, 35 (6) (2011): 592-626. https://pubmed.ncbi.nlm.nih.gov/22238369/. Anker, Anne Sofie Tegner. “The effects of DNA databases on the deterrence and detection of offenders.” *American Economic Journal: Applied Economics* 13 (4) (2021): 194-225. https://www.aeaweb.org/articles?id=10.1257/app.20190207.
Randomized Controlled Trials (also known as A/B Testing) (“RCTs”)	Braga, Anthony et al. “Do body-worn cameras improve community perceptions of the police? Results from a controlled experimental evaluation.” *J Exp Criminol* 19 (2023): 279–310. https://link.springer.com/article/10.1007/s11292-021-09476-9. Ratcliffe, J. et al. 2011. “The Philadelphia Foot Patrol Experiment: A Randomized Controlled Trial of Police Patrol Effectiveness in Violent Crime Hot Spots.” *Criminology* 49 (2011): 795–831. Yokum, David et al. “A Randomized Control Trial Evaluating the Effects of Police Body-Worn Cameras.” *PNAS* 116 (21) (2019): 10329-10332. https://pubmed.ncbi.nlm.nih.gov/31064877/.
Blind Randomized Controlled Trials (“Blind RCTs”)	Ariel, Barak et al. “Can the police cool down quality-of-life hotspots? A double-blind national randomized control trial of policing low-harm hotspots.” *Policing: A Journal of Policy and Practice* 17 (2023). https://www.researchgate.net/publication/376585255_Can_the_police_cool_down_quality-of-life_hotspots_A_double-blind_national_randomized_control_trial_of_policing_low-harm_hotspots.

ABOUT NAIAC-LE SUBCOMMITTEE

The Law Enforcement Subcommittee of the National Artificial Intelligence Advisory Committee (NAIAC) has the responsibility to make recommendations and provide advice on matters relating to the development, adoption, or use of AI in the context of law enforcement.

The Subcommittee was established in Section 5104 (e) of the National Artificial Intelligence Initiative Act of 2020. It is charged with providing advice to the President, through recommendations that will be considered by the full NAIAC, on a range of legal and ethical issues that will arise as law enforcement increases its use of AI tools. These issues include AI bias, data security, adoption protocols, and legal standards. (Section 5104 (e)(2).)

The Law Enforcement Subcommittee was established in the summer of 2023 and began its work in August 2023.

ABOUT NAIAC

The National Artificial Intelligence Advisory Committee (NAIAC) advises the President and the White House National AI Initiative Office (NAIIO) on the intersection of AI and innovation, competition, societal issues, the economy, law, international relations, and other areas that can and will be impacted by AI in the near and long term. Their work guides the U.S. government in leveraging AI in a uniquely American way — one that prioritizes democratic values and civil liberties, while also increasing opportunity.

NAIAC was established in April 2022 by the William M. (Mac) Thornberry National Defense Authorization Act. It first convened in May 2022. It consists of leading experts in AI across a wide range of domains, from industry to academia to civil society. https://www.ai.gov/naiac/

###