Policing Project
NYU School of Law

NYU School of Law
40 Washington Square South
New York, NY 10012

info@policingproject.org
@policingproject
212.992.6950

March 10, 2022

Dear Acting Attorney General Platkin,

I submit this letter on behalf of the Policing Project at New York University School of Law, in response to your request for public input on regulating law enforcement's use of facial recognition technology (FRT).[1] The mission of the Policing Project is to partner with communities, policymakers, police, and technology companies across the country to bring democratic accountability to policing. By democratic accountability we mean that the public has a voice in setting transparent, ethical, and effective policing policies **before** policing agencies act. Sometimes we refer to this as "front-end accountability," as distinguished from seeking a remedy after harm has been done. This opportunity for public comment is an admirable example of front-end accountability in action, and the Policing Project is grateful to participate.

From the thorough prompts provided, it is clear that your Office already has thought deeply about the thorny issues raised by this technology. We have as well. Over the past year we convened a diverse group of stakeholders (including law enforcement, civil rights and racial justice advocates, and technology vendors) for hours of discussion on law enforcement use of FRT. The second half of our letter responds to your specific prompts, but we begin with one overarching recommendation that we believe should guide your development of a statewide FRT policy: *law enforcement use of FRT only should be authorized by your Office for a limited pilot phase, with restrictive safeguards in place, and the means and intention of evaluating whether FRT makes the public safer – and at what cost.*

I. Limit FRT to a Pilot Phase with Use Centralized in a Single State Agency

At the Policing Project, our evaluation of any policing technology starts with a basic question: will the public benefit from the use of this tool? We preach—and it is a bedrock principle of cost-benefit analysis—that before one even considers the costs of government action, the burden is on government to show there is some identifiable, concrete benefit that will be obtained. If there is no benefit, it is unnecessary to consider costs. And if a benefit is established, then the goal is to mitigate or eliminate any accompanying costs.

Current law enforcement use of FRT has inverted this analytical process – applying instead a deploy first, assess benefit later (if ever) approach. We believe this is a huge mistake. We understand that sometimes trial is needed to learn about benefits, so that is what we recommend. That if you move forward you do it with the intention of addressing unanswered questions through a limited pilot phase. The sections that follow elaborate on the type of information that should be collected and the restrictions that should be in place during this pilot phase.

We also ***strongly recommend*** you consider centralizing FRT review in a single state agency, rather than permitting individual agencies to conduct searches. Centralizing FRT review would facilitate consistent

[1] *Facial Recognition Technology*, Office of the Attorney General, State of New Jersey, https://www.njoag.gov/facialrecognition.

standards around training and use, ensure separation between the human reviewers and investigating officers, and consolidate data collection and reporting in a single, well-resourced agency.

1. Comprehensive Data Collection Is Needed to Assess Impact

The potential benefits from police use of FRT are not difficult to conceptualize—as you note, FRT offers the promise of identifying those who otherwise might remain unidentified, or of doing so more quickly. But we as a society have little idea of the scope of these benefits. We lack answers to even the most basic of questions, such as how often do (or will) agencies use FRT? For which types of crimes? How many suspects are identified that could not have been otherwise, or at what time expenditure, and for what sorts of offenses?

The speculative nature of FRT's benefits is especially concerning given the severity of the harms posed, particularly to those communities already most impacted by the criminal system, especially Black communities. The publicly known false arrests stemming from face recognition identifications—including one in New Jersey—were all of Black men. And law enforcement has used FRT to target individuals exercising their First Amendment rights at Black Lives Matter protests and even during Juneteenth celebrations.[2]

Your Office has "unique authority to issue statewide policy directives" for New Jersey's 38,0000 law enforcement officials.[3] This authority means you are uniquely positioned to facilitate meaningful assessment of FRT's public safety impact. You can do this by taking two key steps:

> ***First***, require (and help facilitate) agency collection of comprehensive data on their use of FRT, including how often searches are run, for which types of crimes, and to what result.[4] As noted above, centralizing FRT use in a single state agency would make this data collection easier.
>
> ***Second***, use the data collected to conduct a transparent assessment of public safety impact. This assessment will allow you to see which aspects of the policy framework are working, which require modification, whether the program should be scrapped because benefits do not outweigh the costs, or whether the most serious costs – such as those to racial justice interests – can be mitigated with careful safeguards.

New Jersey is hardly unique in not knowing the answers to these questions. But with your guidance, New Jersey can be a testing ground for the use of this technology in a responsible way.

2. Operational Testing Is Required to Meaningfully Assess Accuracy

FRT is a powerful and expensive tool that raises serious risks for civil rights, civil liberties, and racial justice concerns. The public deserves to know whether it actually works.

---

[2] Joanne Cavanaugh Simpson & Marc Freeman, South Florida police quietly ran facial recognition scans to identify peaceful protestors. Is that legal?, South Florida Sun-Sentinel (June 26, 2021), https://www.sun-sentinel.com/local/broward/fl-ne-facial-recognition-protests-20210626-7sll5uuaqfbeba32rndlv3xwxi-htmlstory.html.

[3] Policing Policy, State of New Jersey, https://www.njoag.gov/programs/policing-policy.

[4] We've included a more comprehensive list of information agencies should track in the attached appendix.

And to know whether FRT works in practice requires testing it for accuracy and bias in actual uses contexts – i.e., assessing FRT as actually deployed with a human-in-the-loop, on the quality of images actually searched, the size of database searched and so on. This type of assessment is called "operational testing."

Your office recognizes the important role played by the National Institute of Standards and Technology's (NIST) technical benchmark tests. To be sure, NIST testing is the gold standard for assessing facial recognition *algorithms*; as such, its evaluations can serve an important gatekeeping function when vetting vendors.[5]

But NIST doesn't conduct operational testing. For example, NIST does not evaluate the humans-in-the-loop part of FRT, so its tests don't tell us about actual system performance. In addition, most of the images NIST tests (86%) are "excellent" portrait quality photos and not the low-quality surveillance camera images that law enforcement typically uses for FRT searches.[6] This discrepancy matters because image quality has a huge impact on accuracy. As NIST itself has explained, "While publicly available test data from NIST and elsewhere can inform owners, it will usually be informative to *specifically measure accuracy of the operational algorithm on the operational image data*."[7]

Truly meaningful evaluation requires operational testing. Luckily, your office is uniquely well-positioned to implement a robust evaluative process that includes operational testing:

> ***First,*** you should spearhead the development of an operational testing protocol. There are some resources to work from, such as the NIST-backed Facial Identification Scientific Working Group's operational testing protocol.[8] Better yet, you could commission a group of diverse subject matter experts—including computer science academics, technologists, and public defenders—to develop an operational testing protocol.
>
> ***Second***, you should institute a vendor vetting process that incorporates both NIST evaluation and operational testing. Specifically, vendors should first be required to demonstrate high accuracy across demographic groups on NIST testing before they are eligible to submit to operational testing.
>
> ***Third***, you should develop a comprehensive and mandatory training program for all face examiners.

Together, these steps will help ensure an accurate process.

---

[5] *See, e.g.*, Kate Kaye, This little-known facial-recognition accuracy tests has big influence, iapp (Jan. 7, 2019), https://iapp.org/news/a/this-little-known-facial-recognition-accuracy-test-has-big-influence.

[6] Patrick Grother et al., Face Recognition Vendor Test (FRVT) Part 2: Identification, NIST (Feb. 23, 2022), https://pages.nist.gov/frvt/reports/1N/frvt_1N_report.pdf at 7, 19; *see e.g.,* IJIS Institute & IACP, Law Enforcement Facial Recognition Use Case Catalog (Mar. 2019), https://www.theiacp.org/sites/default/files/2019-10/IJIS_IACP%20WP_LEITTF_Facial%20Recognition%20UseCasesRpt_20190322.pdf.

[7] Patrick Grother et al., Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects, NIST (Dec. 2019), https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf at 3 (emphasis added).

[8] Understanding and Testing for Face Recognition Systems Operational Assurance, FISWG, https://fiswg.org/fiswg_understanding_&_testing_for_frs_operatnl_assur_v1.0_2020.12.11.pdf.

3. Categorical and Procedural Limits Should Serve as Additional Safeguards

At least during the initial pilot phase, we recommend limiting the scope of FRT's reach to mitigate potential harm. In particular, we urge limits designed to protect vulnerable populations and communities who already have borne the brunt of many unfortunate policing practices.

For example, you should limit FRT searches to serious felony crimes – like murder, rape, and child kidnapping. This would help ensure that FRT's use does not further exacerbate overcriminalization, an epidemic that disproportionately affects people of color, particularly Black and brown communities.

In addition, a court order should be required to initiate an FRT search. This will ensure that FRT is only being used as intended and will create an audit trail to monitor use. Applications for a court order should identify the officer making the application; include a description of the person the officer seeks to identify; the probe images and database to be searched; and the predicate to believe that the person in the probe photo has committed a qualifying crime.

There are other safeguards that your Office clearly is considering, and that we address below. But these strike us as crucial.

II. Response to Prompts

We now turn to the specific prompts presented on your website and answer in order. For questions that included scaled responses or dropdown selections, we have translated our answers into short sentence responses. We also have copied our responses to short answer questions to the website.

A. Prohibitions

We agree with the spirit of the three general principles you have identified, but would extend the reach of the prohibitions on surveillance and arrests a bit further. When it comes to surveillance—i.e., using real-time or stored video footage to track people, with or without individualized suspicion, allowing their whereabouts to be traced—we favor a ban without exception, at least until we learn through a pilot phase more about what is needed to regulate FRT effectively. Regarding the prohibition on using FRT results as the sole basis to justify arrests, we would extend this to other enforcement actions such as stops or stop-and-frisks.

B. Procedures Under Consideration for Human Analysis

- *Do you agree with the measures presented? Any measures you would add or take away?*

We agree with the measures presented and especially are heartened that you include a requirement for search results to be produced to the accused. We would urge you to consult with local public defender offices to shape the scope of this disclosure requirement. In the attached appendix, we include disclosure recommendations suggested to us by a leading expert in public defense and digital forensics as well as some additional specific training suggestions, including that possible matches be subject to two levels of **blinded** review (i.e., two face examiners independently reviewing possible matches).

- *Please share your opinion on law enforcement using facial recognition software to generate leads only.*

Limiting law enforcement use of facial recognition to lead generation is a sensible restriction, but only is meaningful so long as additional independent evidence is required to corroborate a lead. Statewide policy must provide specific guidance on what constitutes sufficient, independent corroboration and should require that the steps taken to corroborate a lead are recorded in the case file. Absent these safeguards, there is a real risk that natural biases—like the human tendency to over-trust machine output—or sloppy investigative tactics may taint the process.

- *If facial recognition software generates a lead regarding an unknown suspect, what standard should be applied requiring additional independent evidence that corroborates the lead before an arrest warrant may be sought?*

The decision to seek an arrest warrant should be made just as it would in the absence of a lead generated by a facial recognition search. Any other approach allows the algorithm to have a say in who is arrested.

C. Additional Questions

- *Rating the importance of no false positives and no false negatives*

You present two prompts that, taken together, ask whether it is more important for a face recognition system to minimize false positives or false negatives. It is difficult to answer this question in the abstract because the impact of each type of error depends on the use case. If an agency is using face recognition to try to identify a missing person, we might want them to be able to cast a wide net, which would mean that minimizing false negatives is important. By contrast, in criminal investigations where the goal is to identify a suspect, we might be more concerned about the risk of ensnaring an innocent person in the criminal legal system, in which case minimizing false positives may be more important. For a general rule of thumb, minimizing false positive is more important when there are criminal-legal consequences.

- *In your opinion, is a false positive rate of 3 in 1,000 an appropriate limit? If no, what would an acceptable false positive rate look like?*

We agree with the idea animating this question – that maintaining low algorithmic error rates is a necessary part of an accurate FRT process that mitigates harm. Still, absent the sort of operational testing we call for at the start of our letter, it is difficult to know whether a false positive rate of 3 in 1,000 is the right target.

- *Results from facial recognition technology are dependent on the gallery of images searched by the software. In particular, larger galleries and those with lower-quality images will increase false positive rates. Would it improve your confidence in the facial recognition software if the vendor was required to certify that it calibrated the software to a target false positive rate, such as 3 in 1,000, when used on the exact same gallery the law enforcement agency will be using?*

We agree with the basic premise that vendors should need to demonstrate low error rates on operationally representative data. This requirement would be a step in the right direction, but it is not enough. Additional steps, such as certifying low error rates on the same types of probe images actually searched, also would be necessary. This is why we call for operational testing.

- *What additional measures would improve your confidence that demographic discrepancies will not result from investigative lead generation using facial recognition software?*

The following four measures would help mitigate the risks of demographic disparities:

(1) **Operational testing** that assesses and reports error rates by demographic subgroups.

(2) **Standards for training requirements and investigative corroboration**: Comprehensive training and a standardized, documented process for investigative follow up will reduce the opportunity for discretionary decisionmaking that may be impacted by bias.

(3) **Crime limits**: Categorically limiting FRT use to serious, violent crimes will help ensure racial disparities in the criminal legal system are not exacerbated.

(4) **Data collection and transparent reporting on use**: Accurate reporting on FRT use and impact that includes demographic data can facilitate an accounting of whether some communities are being targeted or disproportionately burdened by the deployment of FRT.

- *How could results from the NIST Face Recognition Vendor Test be used when evaluating systems for law enforcement agencies to procure?*

As discussed at the start, NIST's FRVT tests only should serve a gatekeeping function when evaluating systems for law enforcement agencies to procure. Only vendors that demonstrate low error rates across demographic groups on NIST testing should be eligible for operational testing. But operational testing nonetheless is needed.

- *Please share anything else you think we should know on this topic.*

Law enforcement use of FRT has leapfrogged both independent and rigorous scientific research and data on actual benefits. This is exactly wrong. Use of this technology should be permitted only on a conditional basis, with robust safeguards in place, while you collect the data sufficient to conduct a clear-eyed and honest analysis of its utility and costs.

Thank you for this opportunity to provide public comment.

Best regards,

Katie Kinsey

Katie Kinsey

Staff Attorney, Policing Project

APPENDIX

I. Sample FRT Record-Keeping List

Agencies should track information about FRT use in a particular case and include this information in the case file. At minimum, agencies should track:

1. Name of officer requesting search and name of examiner(s) reviewing search results;
2. Case number;
3. Date/time of incident and when search was requested and run;
4. Type of incident and person searched for (suspect, victim or deceased identification) including demographic information;
5. Original probe image(s) used; any modified versions of probe image(s) searched; description and reasons for any modifications that were made;
6. Threshold used (if any);
7. Examiner's report that includes: the results of the FRT search; time spent on the comparison; notes describing how the examiner conducted the face comparison and reached the decision to return or not return candidates to the investigator;
8. Candidate list information: if candidates are returned, the report should contain copies of the images contained in the candidate list and the percentage of the match, rank number, similarity score and/or confidence score assigned to each image in the list; and, if applicable, which probe photo produced which candidate list;
9. Investigation outcome: whether the FRT search produced an investigative lead; and if so, the outcome of that investigation.

II. Sample Discovery Disclosure List

1. Result reports and paperwork (possible match, no match, image rejection, etc.)
2. The original submitted probe photo(s)
3. Possible match photo(s)
4. Sources of the possible match photo(s) and the probe photo(s)
5. Any edited or altered version of the probe photo that was used to try to make a match
6. Detailed notes and logs of edits or alteration that were made to the probe photo(s) (e.g. the software used, lighting changes, mirroring, etc.)
7. Full candidate lists for each photo run through the system for recognition, including the ranking and confidence scores, noting which photos produced which candidate lists
8. Names, roles, credentials, and qualifications of any person involved in the facial recognition process
9. Name, manufacturer, and version number(s) of the facial recognition system(s) utilized
10. Name, manufacturer, and version(s) of the facial recognition algorithm(s) used

11. Results of any internal and third-party testing of the facial recognition system(s) and algorithm(s)
12. Image quality information and demographic breakdowns of training data
13. Error rates and documentation of how the error rates were calculated
14. Measurements, nodal points, or other unique identifying marks used by the system in creating facial feature vectors. If weighted differently, the scores given to each respective mark.

III. Sample Additional Training Requirements

Currently, the Facial Identification Scientific Working Group has developed minimum training and competency requirements for face recognition examiners that can serve as a model for New Jersey's training program.[9]

In addition to the topics presented in your prompts, some specific topics that officers should receive training and instruction on include:

- Technical capabilities and limitations of the particular face recognition software itself, including the system's logging and reporting specifications, how error rates may be affected by image quality, demographic factors and other environmental factors;
- Biases that affect face comparison, such as "other race effect";
- Best practices in human-machine interaction including biases that affect human-machine interaction (such as "automation bias," when humans overtrust machine output).

[9] Facial Identification Scientific Working Group, FISWG Documents, https://fiswg.org/documents.html.