**Evaluating the use of facial recognition in UK policing**

**Malak Sadek, Sam Stockwell and Marion Oswald**

A report of a workshop jointly hosted by the Alan Turing Institute and the Metropolitan Police on 25 October 2023 exploring the aims, methods and challenges of evaluating the use of facial recognition (‘FR’) systems by UK police forces. It was noted that there were many different purposes for facial recognition and thus a variety of potential benefits, risks and issues. The workshop presented different perspectives on the evaluation of facial recognition in the policing context and aimed to raise awareness of, and provoke engagement with, the ongoing debates in this area. 

**Executive Summary**

* A number of potential benefits to the effectiveness of police deployments and activities from the use of facial recognition were identified.
* The NPL study commissioned by the Metropolitan Police to evaluate their FR system is a positive step towards a better understanding of FR and ensuring that it is equitable and effective from a technical perspective.
* It also allowed for a deeper understanding of how different factors (environmental, historical, geographical, etc.) can affect the effectiveness of FR, and especially LFR.
* However, additional considerations are needed regarding the wider ecosystem surrounding FR, including any potential biases that may be introduced by humans-in-the-loop, the accessibility of the topic to the wider public and the indirect impacts of FR on society regarding surveillance, trust in policing, etc.
* The assessment of proportionality cannot be limited either to a technical evaluation nor to a purely legal one. More work is needed to bring these aspects together within policies, procedures – including those relating to police decision-making - and evaluation mechanisms.
* To identify these risks and limitations, other entities must support the police – including civil society organisations, legal experts and academics – while the police must also actively seek the input and involvement of these groups.
* The [Structured Framework for Assessing Proportionality of Privacy Intrusion of Automated Analytics](https://cetas.turing.ac.uk/publications/privacy-intrusion-and-national-security-age-ai) is one example of how relevant factors can be identified and assessed in a systematic way, encouraging the considerations of aspects that are not only technical (such as data used and algorithms designed), but also socio-technical (such as human inspection and data management practices).

**Presentation Summaries**

The workshop was split into four sessions. Presentations in sessions one and two were followed by an expert commentary and a discussion while sessions three and four was followed by an interactive group exercise.

**Session One: Introducing facial recognition in a policing context + Evaluating demographic differences: Lessons from the National Physical Laboratory (‘NPL’) ‘Facial recognition technology in law enforcement equitability study’**

*Introduction to Police Use of FR*

FR is currently used in two ways by the Metropolitan police force:

* *Retrospective* FR (RFR) hosts a gallery of custody images which can be used to compare ‘probe images’ (eg. facial images collected from CCTV footage, images from the scene of an investigation or images from doorbell cameras) to match against a known identity. A potential match will be flagged if the result exceeds a set threshold.
* *Live* FR (LFR) is an overt operational tactic which helps locate wanted individuals. Images are captured via cameras on a static van which will isolate what is known as a ‘zone of recognition’. Those images are searched in real-time against a watchlist of custody images. If someone walks past the camera and does not meet the threshold for a potential match, their data is automatically deleted after a few seconds. If they are meeting the face-match threshold, the image will be flagged to an officer on a handheld device/ so they can assess whether further action should be taken. In order to operate in near real-time (and to ensure that police officers are able to follow up alerts immediately), the algorithm may limit the number of faces/video frames processed. The operational environment and threshold settings can also affect the accuracy of who is flagged.

*The NPL Study - summary of key points presented:*

* The NPL was commissioned by the Metropolitan Police to evaluate their facial recognition systems and test the equitability (demographic bias) and performance of the algorithm used by the system. Three use cases were tested: Live FR (LFR), Retrospective FR (RFR) and operator-initiated FR (OIFR). OIFR is a process of FR performed when the data subject is present, with the officer taking a photograph of the subject using a mobile device, with the image then searched against police databases.
* The study used an empirical (black box) testing methodology, with a focus on International Standard (ISO) performance metrics for true positive identification rates (e.g., someone who is on the database that has their image scanned *should* be recognised by the system) and false positive identification rates (e.g., someone is *incorrectly* flagged by the system as being a match to the database).
* Demographics in the NPL study were classified based on genders (e.g. male and female, but not LGBTQ+ categories), ethnicities (e.g. Asian/Asian British, Black/African/Caribbean/Black British and White), and ages (12-20, 21-30, 31-42, 43-75). The database included a cohort of 400+ subjects and a filler dataset of 180,000 face images. The testing watchlist was balanced between demographics unlike an operational watchlist which may have more limited demographic spread.
* The study showed that Asian Females had the highest true positive identification rates (91%) compared to the lowest which were Black Females (83%). On age, the oldest category had the highest true positive identification rate (93%) compared to the lowest category, which was the youngest age group (under 21s) at 84%.
* In relation to environmental factors, shorter subjects were less likely to be recognised by FR – particularly when locations are extremely busy such as Oxford Street on 16 July, thus reducing the overall true positive rate. There may also be a correlation with age i.e. the smallest are likely to be the youngest (who are also one of the demographic categories with poorer detection rates).
* The system’s face-match threshold baseline is typically set at 0.6. When lowered, however (e.g. to 0.56), the number of false positives increased with variations across demographics. For instance, at 0.6 setting, 4 Black individuals were incorrectly flagged, but 22 Black individuals were incorrectly flagged at a setting of 0.56.
* The study demonstrated that the MPS FR algorithm can be operated equitably at a threshold of 0.6 and above. The watchlist size and composition could have a small impact, but not to the same degree as the threshold. A number of study limitations were acknowledged – the evaluation addressed algorithmic accuracy and bias only, with results specific to the operational use of the system. Further assurance will require assessment of other aspects, such as performance at night and in crowded conditions.

*Discussion*

It was noted that scientific evaluation, operational use and the concept of proportionality are closely connected. As the *Bridges* case emphasised, the use of LFR should be assessed through the lens of proportionality and the public sector equality duty, requiring focused due diligence and the development of specific policies. This assessment should not only relate to potential bias within training datasets, but also to the assessment of protected characteristics within watchlists and in the use of OIFR, which was described as being akin to stop and search. FR systems were intrinsically different from the identification of a subject by a police officer from their memory, due to the use of sensitive biometric information and (some workshop participants felt) the pervasive surveillance of large groups.

Questions were raised regarding the standard deviation of the study’s results, and their statistical significance. Concern was noted regarding the face-match threshold of 0.6. If the threshold was set too low then everyone will be flagged regardless, while if it is set too high, then no one – even criminals – will be flagged. It was also noted that a match will always be assessed by a human officer. It was suggested that the threshold scores were on a spectrum where deviations will influence the positive or false recognition rates. It was noted that more accessible examples were needed to contextualise this for the public.

It was noted that human operators can introduce their own biases when responding to flagged images. The extent to which a human operator feels comfortable or confident to push back against a decision made by the FR system if they have doubts will be important to understand.

It was suggested that such evaluations should be continued and repeated in future deployments. There have been some discussions about how to measure the combined effects of bias. It was pointed out that all police officers who are operating FR receive EDI training to try and further reduce these issues arising.

Furthermore, it was noted that if an officer engages with a person flagged by an FR system, they would use their own policing powers to engage, and these may all vary depending on experience.

Regardless of the findings, there can be a mismatch between how results are actually interpreted – between technical statistics and qualitative questions on proportionality, ethics, etc. There needs to be a way of bringing these together to ensure a holistic perspective is taken.

**Session Two: Evaluating the wider impact of facial recognition for policing purposes**

*How is FR used by the Metropolitan Police and what are the policing benefits?*

The presentation highlighted the following points regarding the use of FR by the force:

* The MPS does not retain data if a match is not made by the FR system.
* Watchlists are bespoke to each operational case and deleted within 24 hours after use.
* The ML system is closed, so the underlying algorithm does not continuously learn based on new data processed.
* The system does not make any authoritative decisions.
* The algorithms are accurate – it was stated that 13 deployments in the last year resulted in no false alerts.
* The intelligence case must be necessary and proportionate for FR to be deployed, and when no other alternative technology can be used that may be less intrusive to the privacy of individuals.
* Deterrence is not sufficient on its own as a justification to warrant the use of these systems.
* The MPS has never deployed LFR at protests – it would require a solid intelligence case and be considered necessary and proportionate.
* For every operation where the force deploys LFR, the MPS makes a judgement on how much detail is made available to the public depending on the intelligence and the use case. Signage for informing the public potentially making suspects aware of the technology being used and environmental factors affecting accuracy currently limit the police’s capability. However, signage aims to allow civilians to go on their way undeterred. Arguably, the intrusion is fleeting since the data is immediately destroyed if a match is not made and the initial facial scan is pixelated to add another level of privacy.
* If there are occasions when the MPS could deploy the covert use of LFR (e.g. a small group of extremely dangerous suspects operating in a vulnerable area), this would fall under different regulatory frameworks (e.g. the IPA).
* While police leadership changes all the time, there are common law considerations, College of Policing guidance and other regulatory frameworks – senior officers or police chiefs cannot decide on a whim to deploy FR, so promises made against certain worrying uses of LFR (e.g. during protests) are not in danger of being reneged over time.

*What are the risks of FR systems to human rights?*

* FR engages the very same human rights (e.g. Article 8, 10 and 11) that both fingerprint and DNA biometric data processing involves. However, all these rights are qualified rights, which can be infringed upon if there is deemed a necessary and proportionate justification to do so. FR is often used on grounds of national security, which can be a justifiable reason to infringe on these rights. But is it both necessary and proportionate?
* How can technology be designed for the public good? Looking at these issues, accountability is important. Any system used needs to be legal when deployed, there is due process, it is proportionate and competent and there is transparency around how it is used.
* Consent is another important dimension – is there legitimated opt-out mechanisms in place for citizens when they are subjected to these technologies?
* When it comes to LFR, have deployments increased trust in the police? It is too early to say based on previous deployments, but attitudinal surveys seem to show relatively high levels of support.
* There is a struggle with the law being static in relation to the technical trends. There is a tendency now to substitute legal rules for ethical frameworks (e.g., best practices) – while ethical frameworks have benefits, there needs to be accountability here as well. Are the values embedded in these guidance documents the ones shared by citizens?

*What are the risks of FR watchlists to human rights?*

* The FR dataset and watchlist composition should be considered in accordance with demographics to determine that it is equitable and not disproportionate in relation to minority groups. At the same time, the processes leading to a person being arrested by a facial image initially flagged by the FR must be thought out.
* Retrospective FR may not be as benign and non-intrusive as some may think, considering the recent suggestions that the police may have access to passport data for comparison to facial watchlists. Human rights proportionality needs to be considered in all these debates.
* How are watchlists constructed in relation to what types of criminal offences enable inclusion in the database?
* How much do the public know on the LFR watchlists? ANPR was used originally for countering the terrorist activities of the IRA and is now being applied to everyone. It is not just about the size of the watchlists, but also regulatory oversight of these databases. There is always concern of mission creep involved with these types of processes. What would be reassuring to hear from the police is not an explanation of the benefits from FR, but to see a list of all the potential misuse cases and understand the severity of each and corresponding mitigation strategies.

*Discussion*

From a police perspective, in comparison to a human officer seeking to identify numerous suspects at a location, LFR enables both: 1) an accurate database to be stored of multiple suspects or wanted criminals that may be in that physical location; 2) the accurate detection of these potential suspects if they are within a busy environment. Additionally, LFR can capture individuals who are connected to a crime but still at large in the public and may therefore pose a danger to civilians if not detained. Such detection efforts are much more challenging when done by a human officer alone.

From a legal perspective, the wider impact of FR on society and in the future must be considered. There are discussions about banning technology like FR, but these can be unhelpful. How is FR contributing to the normalisation of surveillance in society (e.g. the chilling effect)? With the normalising of surveillance, police forces are also subject to immense scrutiny when using these technologies. What authority do the humans in the loop have to make decisions if these contradict those of the FR system? The force would welcome the help of others in identifying these risks.

**Session Three: Probabilities in the context of live facial recognition**

*What are the factors affecting the amount and success of matches flagged by an FR system?*

This session revolved around factors affecting the effectiveness of LFR. This included strategies for increasing the number of true alerts, factors that can affect the probability of getting true alerts, data needed, and places to deploy LFR. The impact of these factors reflect that FR systems do not work in a consistent manner across different environments and such considerations must be factored into police deployment of the technology to ensure that the risks from environmental changes are not impacting human rights.

* Increasing the *number* of true alerts between faces held on a watchlist and faces fed through a video feed by LFR could be achieved by adding more faces in the video feed, adding a longer time interval or choosing inspection areas with a higher likelihood of a criminal passing through.
* The *probability* of true alerts increasing or decreasing will vary on the time of year, time of the day, weather, transport nearby, shops and entertainment, and the historical attributes of pedestrians.
* Data needed in order to assess likelihood of effective matches could include mobile phone data, transport data, etc. But if data was available, how would it be used? What types of models, and what questions for the models?
* Once the best places to deploy LFR have been identified, suspects may go elsewhere to evade detection. However, it was suggested that criminals tend to not take these considerations into account based on hotspot policing research. Research on DNA databases in the US found that criminality increased in bordering states when a suspect was searched for in the adjacent one using these biometric data. A lot of literature on this topic is from the US, but they have different issues and contexts to the UK.

**Session Four: Assessing proportionality in the context of the use of facial recognition: a structured factors-led approach**

The Structured Framework for Assessing Proportionality of Privacy Intrusion of Automated Analytics presents factors to consider regarding the intrusiveness of automated analytic methods. The framework was developed as part of a CETaS research project ([Janjeva, Calder and Oswald 2023](https://cetas.turing.ac.uk/publications/privacy-intrusion-and-national-security-age-ai)).

The legal framework provides us with a structured proportionality test (***Bank Mellat v HM Treasury* [2013] UKSC 39):**

* + Is the objective *sufficiently important* to justify limiting a fundamental right, such as a pressing societal need?
  + Are the measures which have been designed to meet it *rationally connected* to it?
  + Are the means used to impair the right or freedom *no more than necessary* to accomplish the objective (e.g. is there a less intrusive measure available)?
  + Does the measure strike a *fair balance* between the rights of individuals and the interests of the community?

Relevant factors to the assessment of proportionality revolve around the datasets used for analysis, training, and testing; the probabilistic nature of the results produced; human inspection of those results; algorithm design and inscrutability (e.g. black box issue); data management; and resources available. The framework provides a systematic way of assessing a wide range of risks to human rights that FR systems could infringe upon and steps that can be taken by police and national security bodies to reduce them. Additional information can be found in the [full report](https://cetas.turing.ac.uk/publications/privacy-intrusion-and-national-security-age-ai).

**Group exercise relating to evaluation**

The following questions were discussed in small groups:

* It has been suggested that police forces could be given access to non-policing datasets in order to improve FR. Would this be effective or proportionate? How should this be evaluated?
* What additional police policies, procedures, processes, and rules might be needed to support the evaluation of the effectiveness and proportionality of FR?

**Summary**

This workshop consisted of four sessions with the aim of understanding: 1) how FR is used by the MET for policing purposes; 2) technical evaluations of the MET’s FR systems; 3) factors affecting the outcomes of FR systems; and 4) how a structured approach to proportionality can help improve the use of FR systems from a human rights perspective.

Overall, the use of FR in policing could have several benefits, including the accurate detection of potential suspects or wanted persons if they are within a busy environment, where such capabilities would be extremely difficult and potentially impossible for human officers to achieve through their own means. The Metropolitan Police has also sought to better understand the technical merits and limitations of their FR systems through scientific evaluations, reducing the opacity of the algorithms that underpin them.

Nevertheless, a number of themes identified during the workshop provides insight into ways that police forces could improve their approach to FR systems. First, increasing the accessibility of technical evaluations and what these mean from a citizen perspective, since existing examples rooted in statistics do not provide sufficient contextual detail.

Secondly, the police should continue to consider the broader implications of FR beyond algorithmic performance and factor in social concerns, including indirect consequences on surveillance and ‘chilling’ effects in public spaces (especially from LFR), the representativeness and equitability of watchlists underpinning the FR systems and how to mitigate against environmental factors that could skew FR match rates and pose risks to human rights.

Thirdly, many of these aforementioned considerations can be achieved through dialogue with other stakeholders, and the value of continued dialogue with stakeholders, including legal experts, civil society organisations and academics, was recognised, as was engagement of others in this debate.

Finally, the group sessions highlighted that above all else, clear legal boundaries on when and how FR systems can be used in a proportionate manner, in respect of which categories of crime or public issue, as well as clarity over the role of consent in being subject to FR systems, will play a fundamental role in the acceptance of citizens towards future deployment.

*18/12/23*