AiLECS lab
Ai for law enforcement and community safety

Briefing Note BN24/01. Date of publication: 17 May 2024.

Australia New Zealand Police Artificial Intelligence Principles - a Primer

Leading Senior Constable Janis Dalins, PhD and Associate Professor Campbell Wilson

AiLECS Lab, Melbourne, Australia
Email: janis.dalins@afp.gov.au; campbell.wilson@monash.edu

1. Background

Australia New Zealand Policing Advisory Agency (ANZPAA) released its Policing Artificial Intelligence Principles in July 2023. Originally informed through background research by ANZPAA officers, the design was finalised over a two day focus group consisting of representatives from each Australian and New Zealand policing organisation. Given its academic and policing membership, AiLECS was in the unique position to contribute to their design throughout the entire process.

In addition to this primer, further dicussion of the principles is included within an episode of ANZPAA's *Police Horizons* podcast[1]

The principles themselves are intended to be as simple as possible, with the decision actively made to make them fit to a single page, in readily accessible, non-technical language. Critically, the method of *implementation* is not dictated, reflecting each jurisdiction's autonomy.

2. The Principles Themselves

The Principles[2] consist of nine individual concepts:

- Transparency
- Human Oversight
- Proportionality and Justifiabiilty
- Explainability
- Fairness
- Reliability
- Accountability
- Skills and Knowledge
- Privacy and Security

The full text for each item is not included for brevity. The reader is advised to read this document in parallel with the principles themselves.

2.1 What is Artificial Intelligence? Should I care?

The principles do not attempt to define AI, instead quoting a previous definition from the (Australian) Department of Industry, Science, Energy and Resources.

In fact, these principles readily apply to *any* law enforcement task undergoing automation.

Takeout: Any data driven project should consider these principles. Don't get distracted by people calling it "AI" or otherwise.

3. An explanatory scenario

The principles document is intentionally abstract, primarily for reasons of brevity and clarity. For the purposes of this document, we will use a scenario of where 'AI' has 'gone wrong' in a law enforcement context.

[1] https://open.spotify.com/episode/50gqOdXKYcoEcsyldahg8v?si=UXrgkEdeRYaLxMdijFXrJA
[2] https://www.anzpaa.org.au/resources/publications/australia-new-zealand-police-artificial-intelligence-principles

Note: Our summary of this case is based upon media reportage of a lawsuit related to this incident. We do not claim first hand knowledge, nor claim that all elements are fact - only that they are alleged.

In October 2023 Harvey Eugene Murphy Jr (MURPHY) was arrested for an alleged armed robbery of a Sunglass Hut retail outlet in the Houston, Texas area in January 2002. He was incarcerated for several weeks before his alibi, being that he lived in California at the time of the alleged incident, was confirmed and charges dropped. In fact, his arrest occurred when he identified himself to the Texas Department of Motor Vehicles in order to renew his Texan driver licence upon his return to the State[3]. MURPHY alleges that during his incarceration, he was bashed and raped by other prisoners.

According to the lawsuit, the footage used as the basis for MURPHY's misidentification was of low quality, and may have been based upon a 1980s era mugshot[4]. It occurred after a Sunglass Hut employee shared the CCTV footage with Macy's (a separate retail store chain), who in turn jointly notified police of the identification. Given the circumstances of the case, MURPHY alleges that facial recognition software is the only reasonable explanation as to how he came to be involved in the matter. As at January 2024, no confirmation or otherwise of this aspect has been observed.

4. Transparency

Transparency is a simple concept - be as open as possible without undermining your core mission, including victims' rights.

However, this is not just being open about using AI - go more broad:

- What are you using it for?
- How did you procure it?
- How did you test it?
- What data did you use to train/test it? Whose was it, and how did you procure it?
- How is all data involved in this system stored?
- Who has access to the data and the system itself?
- How are you monitoring performance?
- What will you do when things go wrong?
- Is there a right of appeal for affected parties?

There is a definitive lack of transparency alleged across our scenario (Section 3). Whereas a company/product name is *mentioned* in some reports, the basis for MURPHY's identification itself is not confirmed. We may not expect a specific product or tool to be publicly named, but it does posit the question as to whether the investigating police, judicial officer issuing the arrest warrant, or the arresting officer were aware as to how MURPHY was identified. Would all of these parties have undertaken their duties in exactly the same way if they'd known?

The CCTV footage was shared between non-law enforcement parties, and the data used to train this unidentified system is completely unknown - though we note MURPHY's inference that a potentially 40 year old mugshot may have been accessible to the system. How much further did this information sharing go? Were customers aware?

Takeout: Be as open as possible about the entire project - not just the use of AI (or otherwise). Often it's the process itself as a whole that needs to be understood.

5. Human Oversight

Again, go further with this principle - don't only think about oversight of the algorithm. Think about oversight *throughout* the entire system. For our scenario:

- Was the training/test data gathered automatically? Is it accurate?
- Is the output accurate?
- What are the consequences for outputs? Are your users blindly following results, or are they applying suitable levels of scrutiny?

Our scenario, as reported, includes human oversight in the form of judicial review - presumably at the point of warrant issue, but definitely after arrest. However, this does not cover the vast bulk of the overall process. As mentioned previously, there is no indication as to whether any of the people in the loop understood the potential nature of the hypothetical facial recognition system, including the possibility for false positives. Given the completely opaque

[3] https://www.theguardian.com/technology/2024/jan/22/sunglass-hut-facial-recognition-wrongful-arrest-lawsuit
[4] https://www.vice.com/en/article/3akekk/man-jailed-raped-and-beaten-after-false-facial-recognition-match-dollar10m-lawsuit-alleges

nature of the system, it is impossible for us to provide *any* indication of the oversight present during development and operation.

Takeout: The human needs to be in the loop throughout the entire project, not just outputs.

6. Skills and Knowledge

Having a human in the loop is completely meaningless if they don't have the appropriate skills and knowledge to perform their role. Not everyone needs to be a data scientist, in fact, that'll possibly be your project's smaller demographic. Developers, maintainers, management and users (both immediate and downstream) need to understand the aspects relevant to their roles.

For example, in our aforementioned scenario, this requires:

- Understanding of the facial recognition system's limitations, particularly as they apply to low quality imagery;
- Assuming the human in the loop validated their identification in some way (presumably by checking against a photo held by the system), how strong was their view of a match? Was it sufficient for its purpose in this case?
- Familiarity of the issuing judicial officer and the arresting police officer with automated facial recognition and its limitations? Did they even know it was used in this case?

You will note that many of the required aspects *aren't* technical, and they're not issues of mathematics. If anything, they're business, legal and process concerns.

Whilst not seeking to criticise the parties involved in the incident, we do ask - would the entire process, including the arrest and subsequent incarceration, have occurred differently if the entire chain of authorities involved had known of the use of automated facial recognition, its limitations, and the imagery's alleged low quality? If the answer is 'yes', was there really a human in the loop?

Takeout: Your human is not in the loop if they don't understand their position. Ensure skills commensurate to the role, plus knowledge of where the role fits in the wider system.

7. Proportionality and Justifiability

Now, this is where things become less clear-cut, and experience and judgement take the lead.

A 2020 report by Monash University [5] gives us some hints regarding community expectations - when presented with the question "What are the main reasons for your distrust in the development and use of facial recognition technology in the federal government?", the strongest responses related to invasions of privacy and being watched (25% and 28% respectively), yet issues around technical accuracy and reliability were regarded by a larger proportion as 'not very important' or 'unimportant'.

Switching over to application, the same study sees an interesting result - when ranked by strength of support for social use cases, *every* scenario with greater than 50% 'support' relates to policing. In fact, the only law enforcement scenario with *less* than 50% is "To identify people for minor offences", with 47%.

Justifications for the collection of data come down to the mission, **not** an issue with AI - rather, it comes down to justifications for the bulk collection, storage and access to sensitive data such as people's faces, the privacy implications, and the risks of misuse by both internal (i.e. trusted) users and external parties.

Our scenario raises several issues of proportionality - armed robbery is a serious offence, so if aligning with the aforementioned survey, the use of automated facial recognition in this context would most likely be seen as reasonable by most members of the general public. The capture and sharing of CCTV by private entities perhaps less so, particularly if used for offences such as shoplifting. What the survey does not consider, though, is the justifiability of the outcome based on the AI system, rather than use case. In our scenario, it is alleged the identification was used as the main (if not only) grounds for arrest and several weeks' incarceration, rather than merely a starting point for further police enquiries.

Takeout: Your choice (or otherwise) to use AI may not be the most relevant factor. Measure the risk/rewards of your mission itself, and remember to consider outputs *beyond* the first step, and across all possible paths

8. Explainability

As with any explanation of technology, it is important to consider the relevant audience. Describing how an AI system works from a technical perspective is only one component of explaining its use. Many AI algorithms are resistant to simple explanations along the lines of "this is exactly how the output was generated" because of their inherent

[5] "Australian Attitudes to Facial Recognition: A National Survey https://www.monash.edu/__data/assets/pdf_file/0011/2211599/Facial-Recognition-Whitepaper-Monash,-ASWG.pdf

complexity. However, in many cases this is probably not the most important aspect of interest. For instance, it may be as crucial to explain how the data that trained the AI was collected, how results from the AI were interpreted, how the AI integrates with other aspects (both human and automated) of an investigation and/or what safeguards are in place to govern its use.

If our scenario's nature is as reported, it is reasonable to say that beyond the process of incarceration and subsequent release, the overall explainability of how MURPHY came to be arrested in the first place is rather poor - in fact, the lack of explainability appears very much a key point of contention. One can only imagine that if his alibi was unable to be established, a subsequent trial would have focused almost exclusively on this aspect - not just how the hypothetical algorithm came to its conclusion, but what the overall system and process actually were.

Takeout: Consider for whom explanations are being made and avoid unnecessary complexity. Understand and explain how the overall system works, *including* the safeguards you installed to keep things reliable

9. Fairness

If a hypothetical facial recognition system used to identify and arrest wanted criminals was 99% accurate, is it fair? To a reasonable person, sure, but it comes down to why that 1% of mistakes is occurring. If it's due to some inherent variability in the system and seems effectively random, well then, the 1 person from every 100 being stopped for formal identification encounters a near one-off inconvenience. But what if that error is due to the system having a bias against something inherent to you? Then *you* become that 1%, and if your life involves regularly passing by a checkpoint related to our theoretical facial recognition system, you're going to get stopped each and every time until someone works out a guardrail to prevent such harassment.

In our scenario, do you regard it as 'fair' if even one person ends up being incarcerated for several weeks on the basis of an incorrect identification? Do you think the storing of involuntary data such as mugshots in private repositories in perpetuity is fair? Your answer may well be 'yes', and in your circumstances, most if not all may well agree, but you need to understand your system's further impacts in order to understand the question.

Takeout: Think beyond performance in percentage accuracy terms. Will your system somehow impact upon or harm people, even if working exactly as designed?

10. Accountability

Accountability is incorporated into all legislation related to policing, either explicitly or via case law. To our knowledge no legislation specifically placing burdens of responsibility onto technology (and away from humans) currently exists, nor is planned in the foreseeable future.

Persons affected by any policing activities have a right of appeal. Automation by any means does not remove this right, necessitating clear roles for any such appeal or internal review. This should not be limited to prosecution actions or other outputs. Ownership needs to be established throughout the project lifecycle, hence the need to ensure responsibility for aspects such as data collection, storage and deletion.

We would suggest most (if not all) of these roles already exist in most policing organisations, just not in this context.

Our scenario shows that some rights of appeal exist, both in criminal court (the alibi resulting in MURPHY's release) and civil, through the lawsuit's existence. In this instance, the lawsuit is aimed against Macy's and Sunglass Hut rather than law enforcement, though one could imagine a similar suit being launched against police (and potentially the judiciary) if evidence of negligence or malpractice was established in their acceptance and use of the identification provided to them. The reputation costs of such a case establishing a lack of accountability and ownership over a process including incarceration could ultimately make any financial settlement pale into insignificance.

Takeout: Establish ownership over every project element, not just the technology and data, and do it early.

11. Privacy and Security

Data privacy and security are key tenets of policing, though have evolved of late. Issues around AI confidentiality, integrity and accessibility remain unchanged. What *has* changed, however, is data is now used to train models (aka algorithms) to make the inferences and decisions we want to automate.

What does this mean? Algorithms adapt to data they're trained on. Being mathematics, every change can theoretically be reversed or at least analysed, without necessarily requiring full access to the model itself. As an example - private phone numbers from ChatGPT[6]. So two things - does the data still exist if it can theoretically be reconstructed, and could your model become a back door to your sensitive information?

Beyond our technical type attacks, our scenario raises multiple questions, such as:

[6]Scalable Extraction of Training Data from (Production) Language Models, https://arxiv.org/abs/2311.17035

- Under what authority were staff able to pass CCTV footage between their organisations?
- Where was the data used in the hypothetical facial recognition system sourced and stored? Is it accurate, i.e. was the photo used to identify MURPHY actually of him?
- Is there a right to be forgotten? If, as alleged, the system held a nearly 40 year old mug shot photo of MURPHY, *should* it, particularly if it's a privately owned initiative?

If our system learned from matches (presumably as approved/rejected by users), what happens when there's a mistake? Is the model actually learning incorrect data, effectively making your quality assurance process an actual threat?

Takeouts:

- **Your project will involve data throughout its lifecyle. Make sure you know where it's coming from, how you're treating it, who has access to it, how you've secured it, and how it's destroyed.**
- **Does your system adapt as it operates? If so, is your data actually deleted?**