TAYLOR FRY

NZ Police - Safe and ethical use of algorithms

June 2021

Document classification: Confidential

Sydney

Level 22, 45 Clarence Street
Sydney NSW 2000
P +61 2 9249 2900
P +61 2 9249 2999

Melbourne

Level 27, 459 Collins Street
Melbourne VIC 3000
P +61 3 9658 2333
P +61 3 9658 2344

Wellington

Level 3, 166 Featherston Street
Wellington 6011
P +64 4 974 5562

ACN 087 047 809
ABN 29 087 047 809
www.taylorfry.com.au

Taylor Fry Pty Ltd

Table of contents

1 Background and scope

In July 2020, the New Zealand government released the Algorithm Charter for Aotearoa New Zealand (the Charter). It positions New Zealand as a world leader in setting standards to guide the use of algorithms by public agencies. The Charter sets out several commitments for algorithm development and use in:

- Transparency
- Partnership
- People
- Data
- Privacy, ethics and human rights
- Human oversight.

The Charter also requires signatories to place their algorithms in a risk matrix. The matrix dimensions consider impact on the wellbeing of people and the likelihood of many people suffering an unintended adverse impact[1]. The Charter must be applied to high-risk algorithms and is recommended to be used for those of moderate risk.

New Zealand Police is a signatory to the Charter. You have asked us (Taylor Fry) for advice and support on matters relating to the safe and ethical use of algorithms that inform operational decision-making. The scope of this job is to:

- Perform a stocktake of algorithms that have a direct or indirect link to NZ Police operations and identify high-risk algorithms
- Discuss fairness measures that should be considered for existing moderate to high-risk algorithms
- Provide guidelines for developing or on-boarding new algorithms that conform with the Charter requirements.

In the context of the stocktake, we have:

- Worked with NZ Police representatives to identify a range of algorithms that directly and indirectly inform your operations
- Performed high-level reviews of the algorithms and how they are used to inform judgments about potential risks
- Considered what actions could be taken to mitigate risk, where risks have been identified.

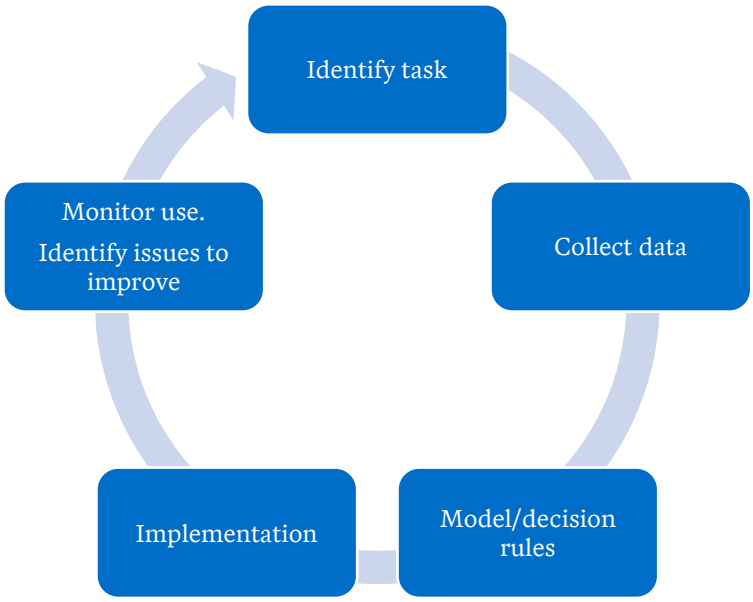
While allocating algorithms into high, moderate and low-risk categories is somewhat arbitrary and subjective, we have done so for the purposes of this exercise. However, we have rated some algorithms higher than the Charter matrix implies because of the importance of public trust in police operations. Algorithmic failings could be very damaging for NZ Police, even if the number of people adversely impacted is low.

1.1 What is an algorithm?

The concept of an algorithm means different things to different people. For the purposes of this report, we define an algorithm **as an objective system in which data is taken in, converted into a different form and returned as a set of outputs, a score or a suggested decision.** Figure 1.1 represents this diagrammatically.

[1] https://data.govt.nz/assets/data-ethics/algorithm/Algorithm-Charter-2020_Final-English-1.pdf

Figure 1.1 – An algorithmic system



For the purposes of this work, we are interested in algorithms where a subsequent operational action is informed by the information. In the case of NZ Police, the nature of the action is typically decided by a human (i.e. with human oversight), rather than being automated by the algorithm.

This is a broad definition of an algorithm in two ways:

- We view the whole system (data, design, implementation, monitoring) as being the algorithm and not just the model or decision rules part.
- While the model (or decision rules) may often be developed using machine learning (ML) or artificial intelligence (AI), they may also be developed by experts building a list of rules or risk weightings. The key feature is that the model or rules are applied in an objective manner so that the same inputs lead to the same outputs.

This broader definition is important since sources of unfairness in algorithms often arise from the data or the implementation process rather than the model itself – the benefits of developing a fair model can be lost if the overall system in which it is applied is unfair. Conversely, an unfair model may be acceptable if the overall manner in which it is applied is done fairly e.g. through some bias mitigation strategies and/or human interventions at the implementation stage. Therefore, taking a holistic view is necessary.

The converse question is – **What is not an algorithm?** Generally, we do not consider data sourcing tools as algorithms as these do not convert input data into a different form. These might include tools that scrape websites, freely available documents or legally acquired data through a warrant. That is not to say these tools can be used without due consideration – issues of privacy, trust in use and accuracy of data sourced are extremely important. However, for the purposes of this review, we do not classify them as algorithms and so they are out of scope.

1.2 General principles for algorithm design and deployment

Ethical algorithms have been discussed in many different ways in many different places, but common principles apply. Broadly speaking, these are:

- What problem are you trying to address and is this understood from a Te Ao Māori perspective? Is it appropriate to use an algorithm for this?

- When designing your algorithm, have you considered privacy, sources of bias, transparency and accountability?
- Who owns and is responsible for the algorithm? What governance is in place?
- What monitoring is in place and when will you review and revise the algorithm?
- Have you consulted with stakeholders, both in the organisation and within the population to which the algorithm will apply?

The Charter is one framework that organises many of these ideas. Another potential tool, of which you may already be aware, is the Algorithms in Policing – Take Algo-Care™ framework[2], which was developed specifically for the deployment of algorithmic assessment tools in the policing context. It covers many of the same points as the Charter but is written with policing applications in mind. A brief summary of some of the main points of this is presented below. Refer to the paper for more complete details, containing a range of questions to consider for each section and additional explanatory material.

Algo-Care™ is a mnemonic consisting of several points to consider, which are listed below. We have indicated a mapping to the Charter framework in parentheses after each point.

- Advisory – Does a human officer retain decision-making discretion? (Transparency; Human oversight)
- Lawful – Is all data acquired lawfully and is its use and benefits proportionate to its possible harms? (Data; Privacy, Ethics and Human Rights)
- Granularity – Does the algorithm make suggestions at a sufficient level of detail (granularity) for different groups? (Data)
- Ownership – Who owns the algorithm and the data it relies on? (People; Human Oversight)
- Challengeable – Are results checked for bias? Can those impacted by decisions challenge them? (Transparency; Partnership; People; Data; Privacy, Ethics and Human Rights)
- Accuracy – Does the algorithm perform sufficiently well to justify its use? (Partnership; People; Data; Privacy, Ethics and Human Rights)
- Responsible – Is the algorithm fair and used in an ethical manner for the public interest? (Partnership; People; Privacy, Ethics and Human Rights)
- Explainable – Can the developer explain why the algorithm generates certain decisions? (Transparency; Human Oversight)

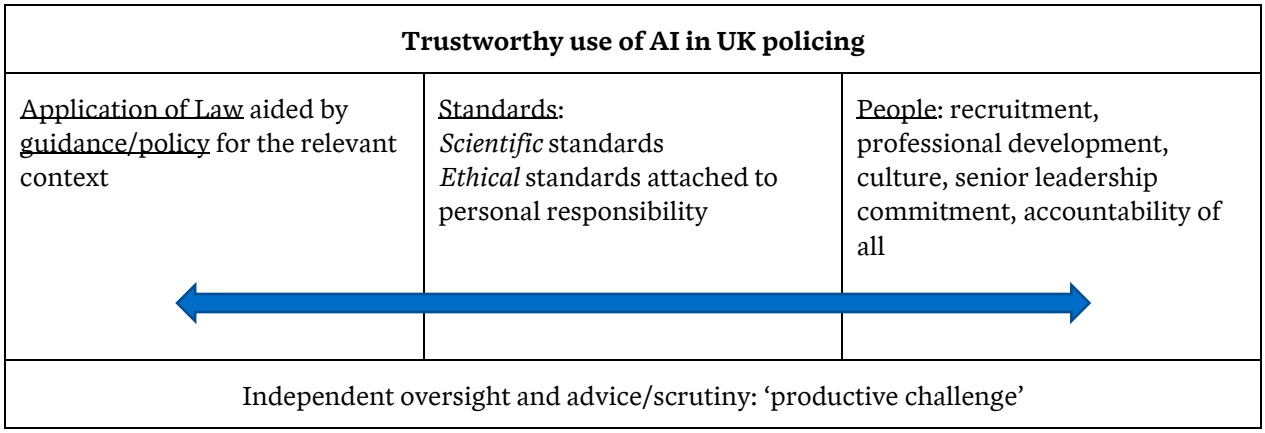
Marion Oswald, one of the designers of the Algorithms in Policing – Take Algo-Care™ framework, also proposes the use of a three-pillar approach to achieving trustworthy use of AI and emerging technology[3]

The three-pillar approach, as summarised in Figure 1.2, is designed to interlink the technical, statistical, legal, contextual, operational and ethical aspects of algorithm-informed policing. Underpinning the three pillars is independent oversight and advice. It reflects the approach taken by the data ethics committee established by the West Midlands (UK) Police and Crime Commissioner (PCC) and West Midlands Police (WMP) over the last two years.

[2] Marion Oswald, Jamie Grace, Sheena Urwin & Geoffrey C. Barnes (2018) Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality, Information & Communications Technology Law, 27:2, 223-250, DOI: 10.1080/13600834.2018.1458455 ; see Figure 1, P245 and the explanatory notes and additional considerations in Figure 2, P247

[3] Marion Oswald (2021) A three-pillar approach to achieving trustworthy use of AI and emerging technology in policing in England and Wales – Lessons from the West Midlands model

Figure 1.2 – Oswald three-pillar approach



1.3 Review process

Our starting point for this review was to arrange several interviews with representatives of police departments that use algorithms. These interviews were held between 9 November and 1 December 2020. Based on the information we gathered during these interviews, we have identified algorithms which, in our view, fall into the high or moderate-risk categories. We have also included suggestions for steps to take to mitigate some of the risks and/or better comply with the Charter. However, the review of any algorithm against the Charter commitments is necessarily a subjective exercise – in most cases, there are not definitively right or wrong processes.

We will be supplying a standalone document of guidelines for development of any new algorithms in a fair and ethical manner. We refer to this as the 'guidelines document' in this report.

2 Stocktake of existing algorithms

We discussed several algorithms with their owners and developers. It was clear from discussions that algorithms in use have been designed to:

- Limit the harm from crime – This is either because they enable efficiency gains and/or they result in more informed decision-making by humans.
- Support rather than replace decision-making – None of the algorithms we considered result in automated operational decision-making. That is not to say algorithms resulting in automated operational decision-making cannot be safe and ethical.

When considering the role an algorithm plays in supporting NZ Police's operations, it is important to think about what happens in the absence of that algorithm – usually a purely human decision-making process. Algorithms are designed to improve the decision-making process.

For example, we know purely human decision-making processes can be biased. Ideally, for an algorithm to achieve its goal of improving decision-making, the presence of bias in that algorithm will not rule its use. As a result, the decisions it supports will be less biased than those stemming from a purely human process. It follows then to take steps in minimising algorithm bias as far as reasonably possible.

We first present some general findings from the stocktake, before discussing individual algorithms and their risks in more detail.

2.1 General findings from the stocktake

In addition to the specific findings for high and moderate-risk algorithms discussed in Section 2.2, we note several general observations from the stocktake below. These include observations about existing controls employed to mitigate risk associated with algorithms e.g. application of human oversight. All algorithms we considered had some controls in place. The suggestions we make in Section 2.2 reflect opportunities to further tighten the control environment.

- Algorithms as assistive tools not decision makers – In our discussions, we noted an **emphasis on algorithms as tools to assist police officers in their day-to-day work, rather than as ultimate arbitrators of decisions**. Algorithm owners understood that 'the model says so' is not an adequate explanation for a decision. Training of users also reiterated that delegation of responsibility to the algorithm was not appropriate. In all the algorithms we discussed, final decisions on actions were made by a police officer or other responsible individual and not by the algorithm.
- Human oversight – Algorithm owners were **aware of the need to ensure all tools were subject to human oversight**. This occurred in several different ways. For instance, many tools are based on human insights and expertise (e.g. Top 5, Initial File Assessment). Another example is that most tools that apply risk scores to individuals are transparent (though not necessarily at the scene of an incident) – the weighting applied to different factors is accessible to users.
- Focus on privacy – In general, we observed a **strong awareness of privacy requirements**. Algorithm owners generally appeared to understand the need to abide by privacy standards and to discuss proposed use of data with stakeholders. With the Clearview AI example in mind, we assume appropriate processes are in place to ensure stakeholders are consulted and privacy aspects are appropriately considered when new tools are trialled or implemented.
- Proportionality – This idea is very important in policing work, since policing involves balancing individual rights against public safety and security. Consequently, any policing tools, including algorithms, need to satisfy proportionality in their use, i.e. their benefits justify any harms associated with them. **We found good awareness and acceptance of this concept in our discussions.**
- Governance – While algorithm owners generally understood sound model development principles, we did not observe a formal model governance structure. There was often a requirement to obtain

clearance from NZ Police's Chief Privacy Officer but, beyond this measure, there were no formal governance requirements.

- Monitoring – Algorithms were often subject to informal monitoring e.g. changes would be made if users raised concerns or noted issues, but there was no formal process to periodically review how the algorithm was functioning, nor to update it, nor consider if it should continue to be used.
- Loss of opportunity – The algorithms we discussed with police were generally quite simple in design. Many were based on human judgment (e.g. rules-based systems, risk weightings). More sophisticated approaches using ML techniques and AI have become popular because they are powerful at detecting patterns and generating insights not readily apparent to humans. Therefore, potential exists for much more insight to be generated by using more sophisticated tools. This point did arise in some of our discussions, so we note that algorithm owners and developers are aware of the benefits to be gained.
- Decentralised algorithms – At least one algorithm was developed and used at a district rather than national level. This introduces the potential for inconsistencies across districts and a smaller data and evidence base upon which to build the algorithms.

2.2 Specific findings for high and moderate-risk algorithms

For the purposes of this advice, we have noted in Appendix A algorithms we regard as low risk (see Table A.1). As part of the interviews, other tools were discussed but were regarded as out of scope for this exercise (usually because we considered them to be data scraping tools, as discussed in Section 1.1). These out-of-scope tools may carry risk and have ethical considerations, but they are not algorithms, so we do not cover them in this advice.

In the following sub-sections, we focus on the algorithms we regard as moderate and high risk, noting the potential risks we have identified and our thoughts on how you could mitigate them.

Note we have included algorithms that are in development and/or trial stages. Our categorisation of their risk assumes they will be rolled out nationally in due course.

2.2.1 Hot spots policing

What is it?	A police deployment tool identifying geographical areas containing public places that experience relatively high levels of harm from crime comparative to the rest of the geographic area.
Owner	Evidence Based Policing
Risk categorisation	Moderate The rating reflects that the tool may introduce bias because it is based on reported offending data. The risk is tempered by the fact that the tool appears to have a relatively small impact on allocation of police resources, and related police operations are intended to be a deterrent, not punitive. The risk rating may be higher if these factors change.
What data is it based on?	Calls for service, offending data and the Crime Harm Index.
How is it used?	The tool is currently in trial stage in the Waitemata and Bay of Plenty regions. The geographical areas identified help inform the deployment of police officers. Hot spots tend to receive more police presence e.g. visited when returning from another call out. Presence is intended to be a deterrent. One of the aims of the trial is to prevent over as well as under-policing of geographic areas, which are often targeted by officers based on their own experience and professional bias as to where and who should be targeted to prevent crime occurring.

Do current controls exist?	Yes
What are the potential key risks?	▪ Feedback loop – The tool has the potential to create a distorting feedback loop, where increased police presence in an area results in more crime being identified and recorded, which in turn makes it more likely that the area continues to be identified as a hot spot. ▪ Unintended bias – Different types of crimes may have different likelihoods of being reported in different communities. Because the tool is based on reported offending, it may be unintentionally biased towards certain communities over time. If under-reporting is significant in one community relative to another, it may ultimately not receive as much assistance through deterrence work as it should. These two risks are related in that the feedback loop may exaggerate any unintended bias.
What would help to further mitigate risk?	▪ Feedback loop – Steps currently taken to reduce the risks of a feedback loop are the omission of any police-generated crime reports (e.g. arrests for disorderly behaviour) from being included in the data. This means only reported crimes by members of the public inform the algorithm. You could also consider implementing formalised controls that impose upper limits on the amount of officers' time that can be influenced by the tool. We understand police officers' time is being tracked for the purposes of this trial using DAS (Deployment and Safety) data. There may be potential to use this data to monitor against any imposed limits. ▪ Unintended bias – Monitor the areas highlighted for hot spots policing and look for systematic differences between community types. Review if these are appropriate or not. Measures put in place to reduce feedback loops should also assist in not directing too much or too little attention to different regions.

2.2.2 Top 5 offender list

What is it?	This is a risk-prioritisation model used to rank those with a 'warrant to arrest' or a 'parole recall warrant' to enable the identification of the top 5 offenders in each district and nationally. High-risk offenders are those considered to pose the most risk to the public and/or police.
Owner	National Intelligence Centre
Risk categorisation	Moderate This risk rating reflects the public safety consideration and the potential for model inaccuracy.
What data is it based on?	The prioritisation tool uses data on all those with an outstanding warrant and who have been an offender or suspect in a violent offence in the past 10 years. The data used includes a range of information relevant to the risk associated with the person. This includes the nature of previous offence(s), firearms, gang membership, domestic violence, drug charges, age at first serious offence, and number of offences.
How is it used?	The prioritisation tool is run daily via an automatic report and identifies the top offenders in each district. Each relevant factor about the individual is assigned a risk score, which is then combined to form an

	overall risk score. From this, a top-10 list of offenders is generated and sent to the National Criminal Investigations Group (NCIG). The NCIG selects a national top 5 from this list and redistributes it to districts. Each identified district top 5 is taken from the automatic report and does not use this manual intervention like the national top 5 does. The district top-5 list allows prioritisation within a district, while the national list helps to ensure a coordinated response, particularly for those who might offend across several districts.
Do current controls exist?	Yes
What are the potential key risks?	▪ Model accuracy – As this aims to identify the most serious offenders, overall accuracy is important. The tool was initially developed by the NIC based on a small number of serious events. Risk weighting scores were set using judgment, experience and research, rather than a formal model. Subsequent updates have been based on experience and judgment. It is possible some patterns or behaviours correlated to the risk have been missed. ▪ Temporal biases – The algorithm has not been updated for some time, so it is possible that some of the risk scores may need adjusting. For example, change in profile of offenders who are responsible for most risk to the public due to an influx of deportees from Australia. We understand some testing has been performed recently.
What would help to further mitigate risk?	▪ Model accuracy – Consider whether sufficient data exists to build a more formal model to look for previously unidentified correlates of risk. Even if the available data is insufficient to build a fully tested model, an exploratory model could still be used to explore the data for patterns and insights, with the current mechanism for assigning risk scores (based on human expertise and judgment) retained but informed by the model insights. ▪ Data and temporal biases – Carry out periodic spot checks of the results of the algorithm where experts review several people identified at different levels of risk to see if there are any systematic problems with the ranking. To some extent, this occurs naturally when officers respond to daily reports if they see someone on the list they don't expect to see. It may be useful to have a retrospective look at those recently arrested following a past appearance at a high level of risk on the offenders' report. Their behaviour since being identified on the list may provide further insight into the performance of the model scores.

2.2.3 Road policing (in development)

What is it?	A potential tool in the early stages of development. One possible final result is an app used by police when dealing on the spot with speeding or other traffic infringements. The app would provide a prediction or forecast of the risk of the driver being involved in a serious road incident in the next three years. This can then be used to support officers' decision-making by coupling their response to individuals with the offenders' risk of future harm.
Owner	Evidence Based Policing

Risk categorisation	Potentially moderate to high, depending on how it is used The high-risk rating assumes ultimately widespread use. It reflects the potentially divisive concept of differentiated treatment based on an algorithm's assessment of potential future actions and the likelihood that people challenge the treatment they are receiving.
What data is it based on?	Data being considered includes the driver's licence database as well as NIA and CARD databases. Variables to be considered include infringements and offences in the past five years.
How is it used?	To be decided. An app for police is a possibility.
Do current controls exist?	Yes
What are the potential key risks (assuming widespread deployment)?	▪ Data biases, which may lead to unfairness – Prior recorded offending does not capture all offending and there may be bias in this e.g. speed cameras in fixed locations, more policing in some areas. ▪ Sensitivity in application – Potential for people to feel they are being treated unfairly. The concept of differentiated treatment based on potential future actions may be unpopular with some people unless police action can be clearly linked to objective information e.g. prior offences. There is a high risk of the tool being perceived as unfair at an individual level. ▪ Opportunities for redress are limited – People are likely to want to understand and possibly challenge the treatment they receive. ▪ Delegation of responsibility to the algorithm – While the intention is for the algorithm to support decision-making rather than replace it, it is possible users might be reluctant to override the algorithm recommendation. The risk profile of this tool should be reassessed when more is known about it and how it is intended to be used.
What would help to further mitigate risk?	▪ Careful audit of the data for potential biases as set out in our guidelines document – Depending on findings, the models may need to be adjusted or mitigation strategies put in place. ▪ Extensive stakeholder and community consultation and testing – Careful rollout and evaluation (pilots and a national scheme) to ensure it meets the principles of proportionality. ▪ Robust monitoring of outcomes to check for bias – To be done at group and individual levels, with a strong focus on errors of inclusion i.e. being identified as a risky driver, since the likely action is punitive. ▪ Transparency is likely to be very important – Even if the tool is proven to be better than human intuition, being able to understand its output for any particular person/event will help with public acceptance. ▪ Training of those using the tool – This is important to ensure users understand the tool is an aid to decision-making, not a final decision.

2.2.4 Victim history scorecard

What is it?	Scoring tool to assist police in determining how to respond to victims of crime.
Owner	Victim Services/Prevention
Risk Categorisation	Moderate The rating reflects the potential impact on victims if the output of the tool is inaccurate and police actions are misguided as a result. It also represents the risks noted below associated with potential inaccuracy.
What data is it based on?	Previous 12 months of victim history, including number of times victimised and nature of offence.
How is it used?	Model outputs a score and a colour-coded seriousness classification (green/amber/red). This then ties into the graduated response model (GRM), which specifies different actions for the different levels.
Do current controls exist?	Yes
What are the potential key risks?	▪ Classification errors – Classifying victims at a lower level of risk means they may not be given the help and protection they require. Conversely, classifying too many people at the highest levels will strain resources – some districts are unable to adequately support all those with a red score. ▪ Data biases – Considers victimisations only in the past 12 months – so could under-classify a victim of an offender who spends more than a year in prison. Also, the data on which the tool is based does not reflect victimisation not captured in police data, so may be biased if those victimisation events are skewed to certain communities.
What would help to further mitigate risk?	▪ Review the model to improve classification accuracy and assess the model against various fairness-appropriate measures. As well as provide insight about how the model handles different groups, these measures are also useful to assess how well the risk scoring is performing overall. ▪ Investigate new potential data sources (e.g. the new Crime Harm Index) and consider if it is possible to use a longer period of victimisation history. Also consider whether assessed differences between reported victimisation in police data and reported victimisation in the Crime and Victims Survey could be used to adjust the tool for bias.

2.2.5 Youth Offending Risk Screening Tool (YORST) (and mini YORST)

What is it?	YORST is a decision tree-based risk screening tool. It is based on a questionnaire about education, living situation, offending history and parental offending history. It produces a score out of 100, representing estimated risk of re-apprehension. The score determines a high, medium or low risk categorisation. Mini-YORST is a cut-down version of YORST with fewer questions.
Owner	Prevention/Youth Services

Risk categorisation	Moderate
What data is it based on?	Some of the answers to questions are automatically populated from National Intelligence Application (NIA). Others are populated by Youth Aid Officers working with the young person and their family. Mini-YORST includes only data automatically populated from NIA.
How is it used?	Youth Aid Officers use the tool to help inform the development of a plan for a youth offender. Both the risk category and individual question scores can be used.
Do current controls exist?	Yes
What are the potential key risks?	The tool is designed to inform assistive police intervention responses. In that context, it impacts the components of an intervention plan and the factors that are addressed in the plan. The primary risk then is that the tool unfairly supports distribution of resources away from people who need it. This could stem from one or a combination of the following: ▪ The risk event being estimated by the tool (re-apprehension) not being the best guide of the need for assistance ▪ The tool does not adequately estimate the risk for some or all people ▪ Errors in the data being used by the tool ▪ Inconsistent application of the tool.
What would help to further mitigate risk?	The specifics of the tool and the individual scores that aggregate to the overall score and risk grade are highly transparent. There is also a high degree of human oversight in the way the tool output is used by officers. We also note that the predictive ability of YORST was last formally evaluated in 2016. In the context of the risks highlighted, it may be of value to consider: ▪ Whether re-apprehension is the best available indicator of need for assistance. It has obvious intuition as an indicator but does not reflect severity of offending. Risk of future imprisonment might be an alternative option. ▪ Formalised and scheduled periodic reassessment of the predictive ability of YORST, with recalibration if required. The longer time between assessments the greater likelihood recalibration will be required. Ideally, frequency of reassessment reflects the extent to which the tool is used and the effect it has on operational responses. ▪ Ensuring the data input into the tool for a particular person represents that person and is accurate. While it may not be practical for this tool, one way to achieve this is for officers to confirm key elements of the data with the person.

2.2.6 Family violence risk-assessment algorithms

What is it?	Consists of two related algorithms – Static Assessment of Family Violence Recidivism (SAFVR) and Dynamic Risk Assessment (DYRA). These consider the risk of a family violence perpetrator committing a further family violence act in the next two years. SAFVR is based on NIA data and includes characteristics of the offender such as gender, past incidents of family violence and criminal history. DYRA is based on

	responses to a series of questions (generally with the main victim) at initial scene attendance. The SAFVR and DYRA measures are combined to create an overall level of concern for the safety of the people involved.
Owner	Prevention
Risk categorisation	Moderate to high The rating reflects the potential for severe consequences if the output of the algorithms are inaccurate and police actions are misguided as a result. It also represents the risks noted below associated with potential inaccuracy.
What data is it based on?	SAFVR is based on NIA data. DYRA is based on information collected at the initial scene collection.
How is it used?	SAFVR is accessed through officers' phones on scene (and at a computer for complaints made in a police station) and presented as a high, medium and low risk grading. The dynamic questions are completed (usually with the main victim) and an overall level of concern rating is presented (high/medium/low scale). A more detailed breakdown is available to the police officer in the station via the NIA system. The algorithms help the officer judge the level of risk and guide the development of a frontline response and safety plan. The output is retained in NIA data and the Family Safety system to support triaging of actions and follow-ups.
Do current controls exist?	Yes
What are the potential key risks?	The risk considerations are similar to those for YORST, except arguably the risk level is greater because the consequences of poor algorithm output and subsequent poorly targeted safety plans could be very serious. The algorithms inform assistive police responses and could be 'unfair' if they do not adequately reflect safety risk (overall and/or for specific individuals). Risk stems from the potential for one or a combination of the following: ▪ The risk event being estimated by the algorithms (recidivism) not being the best guide of safety risk ▪ The algorithms do not adequately estimate the risk for some or all people ▪ Errors in the data being used by the algorithms ▪ Inconsistent application of the algorithms.
What would help to further mitigate risk?	The algorithms clearly have good intentions in trying to prioritise assistive responses. We understand a high degree of human oversight is applied by officers in conjunction with the algorithms. We understand SAFVR has been through a validation process, while a similar process for DYRA is imminent. In the context of the risks highlighted, it may be of value to consider: ▪ Whether recidivism is the best available indicator of safety risk. Our understanding is that the algorithms do not reflect the seriousness of potential future offending. Consequently, frequent low-severity offending is prioritised by the algorithms over low-frequency, high-severity offending.

	▪ Formalised and scheduled periodic reassessment of the predictive ability of the algorithms, with recalibration if required. The longer time between assessments the greater likelihood recalibration will be required. Ideally, frequency of reassessment reflects the extent to which the algorithm is used and the impact the algorithm has on operational responses. Our understanding is that the SAFVR is intended to be assessed every two years. We recommend this be formalised and, similarly, DYRA be subject to a regular assessment. ▪ Ensuring the data input into the algorithms for a particular person represents that person and is accurate. While it may not be practical for these algorithms, one way to achieve this is for officers to confirm key elements of the data with the victim providing responses to the dynamic questions. ▪ Transparency – related to the previous point about data accuracy, we understand the officer at the scene cannot see the inputs that lead to the SAFVR high/medium/low risk categorisation. This makes it hard for the officer to judge the accuracy of the data inputs or understand why the rating is as it is.

2.2.7 AML (GoAML and development work)

What is it?	GoAML is an anti-money-laundering tool used to collect data and reports on suspicious activity from financial institutions. A new project is in development that intends to integrate multiple datasets with GoAML. This includes Companies Office data, property ownership data, address information, and vehicle registration and driver's licence data.
Owner	Financial Intelligence Unit
Risk categorisation	Moderate The rating relates to the development work, rather than GoAML as it currently operates. It reflects the scope for errors in linking datasets and likely variation in linking accuracy for different groups of people. While this is specifically a data sourcing tool (and therefore not an algorithm), the intention is to use this data in algorithms in the future, so it falls under the scope of this stocktake.
What data is it based on?	Noted above
How is it used?	Currently, the suspicious activity reports are manually checked, with activity judged as sufficiently suspicious passed to analysts for further analysis and information gathering. Note most reports are not suspicious activity reports, rather they are proscribed transaction reports.
Do current controls exist?	Yes
What are the potential key risks?	The tool informs punitive police response. This means we are primarily concerned about false positives (people incorrectly identified as participating in suspicious financial activity) and any bias in the rate of false positives. We have no particular concerns about GoAML as it currently operates.

	However, the proposed development work to integrate other datasets into GoAML is likely to increase the rate of false positives and probably introduce bias in that rate. This is because integrating the other datasets will require some form of probabilistic matching technique (using data such as name and date of birth), given there will not be a unique identifier common to all datasets. From our own experience with using probabilistic matching techniques, we know false-positive and false-negative matches tend to be biased. For example, Pacific Peoples are likely to be over-represented in false-negative matches because there is more fluidity in their name convention and ordering (and therefore less likelihood an accurate match will be identified). The linked data is likely to form the basis for new algorithms. In this case, all the risks for unfairness and bias that can arise in algorithms would apply.
What would help to further mitigate risk?	The identified false-positive bias risk can be materially mitigated by carefully considering the probabilistic matching technique and go on to have a reasonably high probability threshold for records in different datasets to be considered a match. Human oversight and verification of the appropriateness of the matching related to individual instances of suspicious activity would also help mitigate the risk. Our guidelines document for development of new algorithms sets out some measures to use to reduce bias and unfairness.

2.2.8 Offenders prioritisation tools

What is it?	Districts have tools to help prioritise investigations into different offences.
Owner	Districts
Risk categorisation	Moderate This risk rating reflects the day-to-day use of this algorithm to direct police resources and the potential for inaccuracy in the prioritisations determined by the various tools.
What data is it based on?	Uses the more serious crimes within a district. For example, the Counties Manukau District uses data on past crime data (burglaries, vehicle crimes), convictions, active charges and warrants, and recent offence history.
How is it used?	Relevant information about the offender is combined into a risk score, which in turn is used to prioritise crimes and offenders for investigation. Different models are used in different districts – some are more manual, while others are based on more objective measures.
Do current controls exist?	Yes
What are the potential key risks?	▪ Districts use different tools, meaning there is not consistency and transparency in results at a national level. ▪ The risk weightings assigned to each relevant criminal measure were developed using judgment and have not been formally validated

	against actual data. This may lead to sub-optimal use of police resources.
What would help to further mitigate risk?	▪ A national framework for offender prioritisation would help mitigate both these risks. A common approach would make districts comparable. It is likely greater resources would be available for development at a national level, meaning that a more formal development and validation process could be followed. ▪ Our guidelines document sets out some suggestions for developing algorithms in a fair and ethical matter.

2.2.9 OSINT tool

What is it?	An intelligence system tool that draws on internet-based open sources to collect and aggregate open-source information.
Owner	National Intelligence Centre
Risk categorisation	Moderate The risk rating reflects the risk of drawing attention to people who are not engaged in serious criminal activity. It also reflects public sensitivity associated with use of their social media data (even if it is publicly accessible) and the risks to community trust in policing.
What data is it based on?	Information is collected from online open sources not protected against public disclosure.
How is it used?	Deployed by the Open Source Intelligence Team (OSINT) who may decide to use it to follow up on information collected from online open sources and identify networks of people linked to the original suspicious report. Information is classed into type (national security, threat to life, child abuse, other) and passed to the investigation as additional intelligence.
Do current controls exist?	Yes
What are the potential key risks?	▪ Errors of inclusion – Many identified links may not be relevant to the investigation at hand. ▪ Unintended bias – Certain communities may be searched more often by the operator, resulting in a bias in those detected by the tool. While human oversight in the process tempers this risk (the output provides intelligence, but does not directly dictate investigation), potential exists for this to directly impact those communities. These two risks are related. Both result in potentially drawing attention to individuals who are not engaged in serious criminal activity.
What would help to further mitigate risk?	▪ For errors of inclusion – Ensure human oversight and application of privacy principles so data on uninvolved individuals is not retained ▪ For unintended bias – Use proportionality to determine when to use the tool.

2.2.10 National security portal

What is it?	Used to assist with the assessment of risk associated with tips reported to National Intelligence Centre (NIC)
Owner	NIC jointly with the National Security Group
Risk categorisation	Moderate The risk rating reflects the fact that it is concerned with high severity criminal and terrorist behaviour, but the overall risk is tempered by a high degree of human oversight.
What data is it based on?	Based on tips reported across NZ Police and subsequent research.
How is it used?	This algorithm determines the risk level of tips received. The algorithm is multifactorial and takes into account the Australia-New Zealand Counter-Terrorism Committee (ANZCTC), NZ Police learnings and international practices on assessment. Usually, a tip is supplemented by additional information (looking at capability, intent and ideology) by an analyst, prior to being run through the algorithm.
Do current controls exist?	Yes
What are the potential key risks?	Given the serious nature of the events, algorithm accuracy is a key concern. Missing a serious risk has significant consequences on society, whereas misclassifying a tip as being high risk may divert valuable police resources and impact those involved in the investigation significantly.
What would help to further mitigate risk?	Looking back retrospectively at tips classified and their outcomes where known may provide some insights into the performance of the algorithm and whether any of the decision matrices may need adjusting. It is important to retain strong human oversight to reduce the risk of missing high-risk information as well as misclassifying low-risk tips.

3 General advice for existing algorithms

Suggestions for mitigating specific risks for high and moderate-risk algorithms have been given in Section 2.2. Here, we make some general recommendations to consider for algorithms currently in use or development. We also give an overview on ways to measure fairness in an algorithm.

In the separate guidelines document, we provide advice for the fair and ethical development of any new algorithms, which may also provide useful guidance, particularly for algorithms in development, or subject to review. This expands on the concepts of measuring fairness but also touches on some other relevant points around algorithmic design, including concepts of proportionality, transparency and accountability.

3.1 Recommendations following our review

Recommendation 1 – Ensure a proper governance framework for algorithms as part of NZ Police's wider assurance work on emergent technologies

We recommend that the governance framework for algorithms that inform operational decisions (and as defined in this report) be part of NZ Police's overall governance framework. While not all algorithms relate to emergent technology (and vice versa), it may be wise to combine governance of these overlapping areas.

While ownership of the algorithms would reside in the relevant groups, a centralised governance framework would add additional checks to ensure algorithms are developed and deployed appropriately.

A robust governance function would include:

- A centralised area to keep a stocktake of all algorithms in use.
- Consolidation of existing approval processes to capture all algorithmic tools, including tools created at a district level. This involves considering justification for using an algorithm at all, as well as whether the particular algorithm developed meets required standards.
- Approval, monitoring and ongoing review processes.
- Standards around documentation and training for algorithms.
- An initial point of contact for concerns about algorithm use from the wider community as part of ensuring accountability of use.

A strong governance framework is particularly important to support future development of new and existing algorithms that use ML methods. Algorithms currently in use are generally relatively simple – for the most part they are rules-based approaches with close parallels to human intuition. ML approaches can add an additional layer of abstraction to the process, and can increase the difficulty of detecting bias and unfairness in the models, as well as increasing the risk of unintended harms (of course, they also have many potential benefits, such as greater accuracy and detecting useful insights not apparent to humans). Good governance is important in managing the risks of this type of transition.

Recommendation 2 – Set up a formal evaluation structure for algorithm developers

Just as new medicines are subject to a phased evaluation process before being approved for use, it makes sense for high and moderate-risk algorithms to be subject to a rigorous approval process. Possible steps in this process include:

1. Satisfactory performance in model building on a test data set.
2. A review by experts of outputs from the algorithm against the current system (if none, then the current decision would be human-assessed outcomes) to audit the performance. If the algorithm is intended to aid human decision-making, it would be important to also examine the final decisions reached, rather than those suggested by the algorithm alone. Based on our definition of an algorithmic system, the final human decision-maker forms part of the system.

3. Field-testing of the algorithm by running pilots in some areas and comparing outcomes with and without the new algorithm.

A successful algorithm passes all these hurdles prior to full-scale deployment. At all stages, reviewers stay alert for problems and unintended harms, and check the algorithm is returning sufficient benefit to justify its use. Comparing outcomes against the current system is very important (noting that the current system may be human-based decision-making). Consider a more complicated algorithm only if it leads to significantly better outcomes than a simpler process.

The selection of appropriate performance measures is an important part of algorithm evaluation and depends very much on the purposes of the algorithm. Some algorithms may be aimed at improving efficiencies in which similar outcomes – reached in a shorter period – would reflect good performance. In others, accuracy is the key performance indicator.

Recommendation 3 – Establish a more formal monitoring process for algorithms in use. Also consider creating warning flags that would trigger a review, particularly for high-risk algorithms

It is important algorithms are closely monitored to ensure they are operating as intended, yielding good-quality outcomes and that the harms generated by the algorithm are proportionate to the realised benefits. If an algorithm replaces an existing framework or algorithm, then, where possible, we suggest monitoring also consider the performance of the algorithm relative to its predecessor.

The level of detail in and the frequency of the monitoring process should reflect the level of risk attached to the algorithm. Furthermore, closer monitoring is warranted if the population to which the algorithm is applied is known or suspected to be changing over time.

It is also important to consider some red flags in advance that would trigger a review of the algorithm, but note that these are unlikely to be exhaustive, so monitoring should be sensitive to emerging results that indicate a review is needed, e.g. if the benefits are not proportionate to the harms. If algorithms have been through the formal evaluation process outlined above in Recommendation 2, then it is likely large adverse impacts would be identified at that stage. Therefore, adverse impacts during monitoring are likely to be smaller or more subtle.

A side-benefit of regular monitoring is that it requires data collection and validation for evaluation of outcomes, which in turn may lead to better data sets for use in future algorithmic development.

Recommendation 4 – As part of the monitoring framework, audit algorithm outcomes to ensure that they are fair across different population groups

As an institution, the police service relies on community trust and co-operation to function for the benefit of society. Consequently, it is particularly important to be sensitive to the perception of police actions in the community at large. For algorithms, this means that they should be employed in a fair and even-handed manner across different groups in society. Auditing algorithms for fairness is therefore important.

Recommendation 5 – Model approval to include a recommendation on when the algorithm is to be reviewed and revised if needed

Since people, circumstances and behaviour change over time, it is important to refit important algorithms regularly to ensure they capture how things are at the present time, rather than how they were five or 10 years ago.

Recommendation 6 – Continue to provide training to users of algorithms to ensure appropriate use

Algorithm owners have rightly placed an emphasis on training users in the correct use of the algorithm. An incorrectly used algorithm could lead to scarce police resources being misdirected, or to bias in those receiving police interventions. Furthermore, the use of algorithms could risk the delegation of decision-making to the algorithm rather than to the officer on the scene or on the investigation. It is important for appropriate discretion, control and oversight to remain in human hands. Due to the importance of this, we

have included ensuring ongoing appropriate training as advice here. This may include specific training on the use of discretion and expertise in modifying or overriding recommendations from algorithms.

For some algorithms, training was given to the relevant staff on implementation of the algorithm, and since then to new recruits at the Police College. Additionally, it is beneficial to hold refresher sessions to ensure continued correct use and understanding.

Recommendation 7 – Take advantage of the opportunities presented by ML and AI

While much of this review is focused on identifying and mitigating risk with algorithms, it would be remiss to ignore the positive consequences. Algorithms, particularly those built using ML techniques, can reveal insights not readily apparent to humans and lead to better results. They can be a tool in reducing bias – human-based decisions are usually not free from bias, rather we are more habituated to biased decisions from humans than from machines. The potential for increased accuracy from ML and an increased focus on bias may lead to algorithms that produce more accurate (at an overall level) and fairer (for different groups) outcomes.

Most of the algorithms discussed in the stocktake were relatively simple, so there is significant potential offered by more sophisticated processes.

Recommendation 8 – Develop algorithms nationally, rather than at a district level

Algorithms, particularly those based on statistical or ML techniques, benefit from large amounts of data. Therefore, developing algorithms at a national level rather than a district level may lead to higher quality models. It will also ensure consistency across districts and enable better control of any risks. It will likely be more efficient to hold a central data source and monitor algorithms against that.

3.2 General principles for algorithm design and deployment

Ethical AI has been discussed in many different ways in many different places, but common considerations apply. Broadly, these are:

- What problem are you trying to address? Is it appropriate to use an algorithm for this?
- When designing your algorithm, have you considered privacy, sources of bias, transparency and accountability?
- Who owns and is responsible for the algorithm? What governance is in place?
- What monitoring is in place and when will you review and revise the algorithm?

Most of these have been covered in the previous sections.

The Charter is one framework that organises many of these ideas. Another potential tool, of which you may already be aware, is the Algo-Care framework[4], which was developed specifically for the deployment of algorithmic assessment tools in the policing context. It covers many of the same points also covered by the Charter but is written with policing applications in mind. We present a brief summary of the main points of this below. Please refer to the paper for more complete details, containing a range of questions to consider for each section and additional explanatory material.

Algo-Care is a mnemonic consisting of several points to consider, which we list here, indicating a mapping to the Charter framework in parentheses after each point.

- Advisory – Does a human officer retain decision-making discretion? (Transparency; Human Oversight)

[4] Marion Oswald, Jamie Grace, Sheena Urwin & Geoffrey C. Barnes (2018) Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality, Information & Communications Technology Law, 27:2, 223-250, DOI: 10.1080/13600834.2018.1458455 ; see Figure 1, P245 and the explanatory notes and additional considerations in Figure 2, P247

- Lawful – Are all data acquired lawfully and is its use and benefits proportionate to its possible harms? (Data; Privacy, Ethics and Human Rights)
- Granularity – Does the algorithm make suggestions as a sufficient level of detail (granularity) for different groups? (Data)
- Ownership – Who owns the algorithm and the data it relies on? (People; Human Oversight)
- Challengeable – Are results checked for bias? Can those impacted by decisions challenge them? (Transparency; Partnership; People; Data; Privacy, Ethics and Human Rights)
- Accuracy – Does the model perform sufficiently well to justify its use? (Partnership; People; Data; Privacy, Ethics and Human Rights)
- Responsible – Is the algorithm fair and used in an ethical manner for the public interest? (Partnership; People; Privacy, Ethics and Human Rights)
- Explainable – Can the developer explain why the algorithm generates certain decisions? (Transparency; Human Oversight)

3.3 Measuring fairness

Assessing fairness of algorithms across different groups or individuals is an important (but by no means the only) component of assessing an algorithm conformance to ethical norms and the Charter. It is also an area with some technicalities and nuanced definitions, so we present an overview of it here. For more details, please consult the guidelines document.

3.3.1 Overview of fairness measures

In New Zealand, all persons are equal before the law and are entitled without any discrimination to the equal protection of the law. From the perspective of algorithm use, this means people are to be treated similarly regardless of characteristics such as gender, age or ethnicity. In practice, when measuring fairness, this means we consider a model performance measure for different population groups (e.g. gender, age) and check the measure is similar across the different groups.

There are many possible choices that consider different measures of who gets selected by the algorithm and who does not. These fall broadly into three categories. Note it is not possible to satisfy all fairness measures at the same time. Therefore, when judging fairness in an algorithm, the first step is to determine what fairness means in the context in which the algorithm will be used. Below, we give a common choice from each of the three categories to illustrate the different possibilities.

Selection parity

Selection parity means the proportion from each group selected by the algorithm for intervention is similar between the groups, regardless of whether the need for the intervention is different between the groups.

Equal opportunity

This measure calculates the proportion of all those selected for intervention from all those in each group who need the intervention. It is satisfied when this proportion is similar across groups.

Precision parity

Here, the quantity compared between groups is the proportion of those selected for intervention that do, in fact, require the intervention. As the name suggests, it considers algorithm accuracy across groups.

3.3.2 Measuring fairness in a specific algorithm

Some key points to consider when reviewing existing algorithms for fairness include the following:

- First, identify what fairness means for your algorithm – What is fair in the context of the specific algorithm and the task it is designed for? Which of the class of measures above best captures a fair and ethical application of the algorithm?
- Over which groups should you measure fairness? Fairness is usually specified in relation to one or more protected characteristics, such as gender, ethnicity and age.
- What exactly does fair treatment by group characteristic mean in the context of the algorithm?
  - Should you select formal equality (all people to be treated the same regardless of protected characteristic), or substantive equality (this recognises that there may be underlying inequalities in society, so the use of affirmative action is acceptable to reduce the underlying inequalities).
  - An example of formal equality might be that all youths are to be treated equally by the YORST algorithm regardless of ethnicity. An example of substantive equality is in the Initial File Assessment of the category 4 high-volume crimes, where older or vulnerable people may be preferentially treated.
- After selecting the type of measure to focus on (e.g. selection parity, equal opportunity, precision parity), also consider whether you want to focus on errors of exclusion or inclusion.
  - For interventions that are assistive in nature, a rule of thumb is to avoid unfairness by excluding people from the assistance in a discriminatory manner
  - Conversely, for punitive interventions the concern is including people in a discriminatory manner.
- Regardless of which fairness measure you select, also review the overall algorithm accuracy and performance. While sacrificing some accuracy may be necessary to achieve a fairer outcome, it is important not to excessively degrade the algorithm performance.

Once you have identified appropriate measures to use, the algorithm may then be reviewed for fairness.

It is also useful to supplement this with targeted testing for algorithm performance on known problematic groups and anomalous results and to do spot checks to ensure similar individuals are treated similarly.

Appendix A Low-risk algorithms

Table A.1 – Low-risk algorithms

Algorithm	Owner	Description	Comments
Initial file assessment	Service Group	Four questions to identify if case is solvable for high-volume crime types.	The four questions are filled out manually, are simple and well understood. Human judgment comes into play here as well. It was suggested by interviewees that the questions could be improved.
Online forms	Service Group	AI tool used on the Police 105 form website (for non-emergency reports) to help prioritise jobs. It scans the form for key words and assigns a priority.	All cases are still processed by a person, but the lower priority cases take longer to get attention.
Enhanced Comms Roster (ECR)	Service Group	Optimisation routine to assign shifts for staff at call centres.	90 days' notice is given to all employees of their upcoming shifts, with the ability to negotiate the system-generated shifts with their manager.
Crime Harm Index	Evidence Based Policing	Assigns a harm rating to each Australian and New Zealand Standard Offence Classification (ANZSOC) code. Although not strictly an algorithm, it does process data and is an important input to other algorithms, so it fell under our scope.	Used in other tools e.g. family violence risk-assessment tools. Use of actual sentencing data rather than guideline sentences to derive the index is different to equivalent overseas indices.
Automatic Number Plate Recognition (ANPR)	Road policing	Checks CCTV for number plates, e.g. stolen cars.	Number plate must be manually entered into the system. Pings exact matches, which are then seen by a person.
Road camera placement	Road policing	A list of potential routes on which to put fixed speed cameras.	List was provided as a one-off and is used by a person who judges this with other factors, such as type of road and visibility.
ABIS	National Criminal Investigations Group	Fingerprint matching system.	High degree of human oversight.
Image Management System	National Criminal Investigations Group	Also known as Photomanager. Image matching system. New system in development.	Assumed that a similar degree of human oversight to ABIS is applied (or higher).

Algorithm	Owner	Description	Comments
QueryMe	Service Group	Automated Identity Manager for vetting. Requires a human to complete the process unless a perfect match for the name is found and no information attaching to that name is also found.	Human oversight is used where there is any judgment required (i.e. the system has no perfect match, or there is a match and there is information attaching).
Winscribe	Service Group	Prioritises interviews for transcription based on likely need for court.	Transcribing is performed by a human.
BriefCam	National Criminal Investigations Group	Used to analyse CCTV footage. Note the broader functionality of this third-party product includes facial recognition, but this is not used by NZ Police.	Provides information to investigator who evaluates and decides future actions.
Lumi Drug Scan		Frontline drug screening tool combining near-infrared handheld device, a mobile phone app, and drug detection machine-learning models.	Currently being trialled.

TAYLOR FRY

www.taylorfry.com.au