Regulatory Sandbox Final Report: West Yorkshire Police

A summary of West Yorkshire Police's participation in the ICO's Regulatory Sandbox for their Domestic Abuse Assessment Tool

Date: October 2025

Information Commissioner's Office

Contents

1. Introduction

1.1 The Regulatory Sandbox (the Sandbox) is a service that the ICO provides to support organisations that are developing products or services which intend to use personal data in innovative and safe ways, and will deliver a potential public benefit.

1.2 The Sandbox is a free, professional service that is available to organisations of all sizes who meet our entry criteria and specified areas of focus. These criteria are assessed by the Sandbox's application processes. You can read our entry criteria in our Guide to the Sandbox and our terms and conditions.

1.3 The Sandbox specifically seeks to engage with projects operating within challenging areas of data protection. Sandbox participants have the opportunity to engage with the ICO, draw upon its expertise and receive support on mitigating risks and implementing data protection by design and default into their product or service. This helps to ensure that the participant identifies and implements appropriate protections and safeguards.

1.4 West Yorkshire Police (WYP) serve approximately 2.2 million people living in one of the five metropolitan districts of Bradford, Calderdale, Kirklees, Leeds and Wakefield. WYP have been developing a Domestic Abuse Assessment Tool (DAAT) to explore the feasibility of using emerging supervised machine learning technology to enhance the completion of risk assessments at domestic abuse incidents, helping Police Officers to make the most informed decisions about how to respond to such incidents. The intention is for Police Officers and frontline staff to use this technology to identify cases where escalation may occur so that, by collaborating with partner organisations, they can provide an early wrap-around service to support families and minimise the risk of future harm.

1.5 WYP were accepted into the Sandbox at the end of January 2024 as an exceptional innovation using AI in a policing context. WYP and the ICO developed a bespoke Sandbox plan with the following objectives:

- **Objective 1:** After a detailed product walkthrough, the Sandbox and WYP will work together to identify and address

challenges related to providing effective information about the collection and processing of personal data.

- **Objective 2:** The Sandbox and WYP will work together to consider the data protection implications of the DAAT, ie whether the project will comply with key data protection principles such as purpose limitation and data minimisation.
- **Objective 3:** WYP will explore key data protection risks (eg inaccuracy and bias) related to the DAAT and implement a 'data protection by design and default' approach to mitigate these risks. WYP will produce a Data Protection Impact Assessment (DPIA) for the Sandbox's review ahead of live processing.
- **Objective 4:** The ICO was to receive regular updates on the testing of the tool during the initial stages of WYP's shadow pilot following completion of a DPIA review (Objective 3) and any outstanding updates. However, due to changes in both the timing and nature of the pilot, transitioning from a shadow pilot to a trial deployment, this objective was revised. The ICO's role shifted to providing input during the trial's preparation phase, including feedback on WYP's approach to bias testing and the interpretation of trial results to comply with the fairness data protection principle.

1.6 The Sandbox work commenced in March 2024 and WYP's final objective was completed in July 2025. This exit report summarises the work and key learnings during WYP's time in the Sandbox. It should be noted that the report is limited to the data protection objectives described above (1.5) and does not represent all the obligations an organisation may have to comply with data protection legislation.

1.7 With consultation from participating organisations, the Sandbox publishes exit reports for all participants so that fellow innovators facing similar questions about data protection by design can benefit from the key learnings of the Sandbox. It should be noted that the views in this report are based on the ICO's specific contextual understanding of WYP's operations and, therefore, it cannot be guaranteed that other organisations will be able to apply these considerations in the same way. Furthermore, the insights provided are subject to change in response to developments in the UK's data protection landscape and any updates to WYP's organisational practices.

2. Product description

2.1 The Domestic Abuse Assessment Tool (DAAT) uses supervised machine learning technology[1] to support risk assessments in response to domestic abuse incidents and the appropriate policing response. The tool is designed to predict whether or not any serious harm offences may occur in the next 12 months following the initial incident. Currently, WYP use the Domestic Abuse, Stalking & Harassment assessment (DASH) to assess the likelihood of serious harm and have plans to roll out the College of Policing's Domestic Abuse Risk Assessment (DARA) for frontline police officers which intends to make it easier to identify coercive control. DASH is a checklist formed of questions to domestic abuse survivors that is used by police forces nationally to make referrals to the Multi-Agency Risk Assessment Conference (MARAC): this conference is used to discuss and share information about high-risk domestic abuse cases. However, multiple empirical studies made clear to WYP that DASH's predictive capability is underperforming and not always accurate at identifying the future risk of harm, especially in high risk cases.[2] To address DASH's limitations, WYP aim to develop and test whether the DAAT could be a more accurate predictive algorithmic tool. WYP will use DASH and DAAT (and later DARA) assessments together to assist officers to make effective risk assessments about the likelihood of domestic abuse occurrences. This approach is intended to support more efficient resource allocation and help reduce the incidence of domestic abuse offences.

2.2 Upon recording a domestic abuse occurrence, a member of policing staff inputs details about a survivor and suspect into an application. The application takes numerical data about those data subjects from WYP's crime data. The application then applies a pre-trained algorithm to the numerical data to forecast whether in the next year the incident is likely to be followed by one of the following predictions:

---

[1] Supervised machine learning is a machine learning task of learning a function that maps an input to an output based on examples of correctly labelled input-output pairs (AI and data protection glossary).

[2] Dashing Hopes? the Predictive Accuracy of Domestic Abuse Risk Assessment by Police; Police Attempts to Predict Domestic Murder and Serious Assaults: Is Early Warning Possible Yet?; Predicting Domestic Homicides and Serious Violence in Dorset: a Replication of Thornton's Thames Valley Analysis; Comparing Conventional and Machine-Learning Approaches to Risk Assessment in Domestic Abuse Cases

- No further domestic abuse occurrence;
- A 'less serious' domestic abuse occurrence (e.g. criminal damage under £5000 or possessing controlled drugs); or
- A 'more serious' domestic abuse occurrence (e.g. sexual assault, kidnapping, or arson endangering life).

The definitions of 'less serious' and 'more serious' are determined by the policies of WYP.

2.3 Officers dealing with the case, or safeguarding professionals, can then use the DAAT prediction score as another variable to consider in their decision making alongside other factors like their own experience and responses to the DASH checklist. Policing staff will use all the available information to determine subsequent actions such as attending the site of the incident, further risk assessments and other interventions such as protective orders, right to know disclosures and case conferencing.

2.4 WYP worked with the Cambridge Centre for Evidence Based Policing (CCEBP) to train the DAAT model on more than 140 predictor variables and a quarter of a million domestic abuse occurrences. The ability of the DAAT model to accurately predict future instances of domestic abuse was tested using another quarter of a million cases to see if the predictions were correct. CCEBP used predictor variables from the following datasets provided by WYP:

- domestic abuse occurrences in West Yorkshire – crimes and non-crime incidents tagged as domestic abuse created between 2010 and 2022. In total there were 786,632 occurrences;
- suspects linked to domestic abuse occurrences;
- victims linked to domestic abuse occurrences;
- non-domestic abuse crimes linked to domestic abuse suspects;
- non-crime incidents linked to domestic abuse suspects;
- non-domestic abuse crimes linked to domestic abuse survivors; and

- non-crime incidents linked to domestic abuse survivors.

2.5 The DAAT is an additional tool for risk-based decision making. WYP have emphasised that the algorithm will not be used for decision making and is only intended to support decision making. The implementation of the tool will be supported by training and oversight by supervising officers and a process that will require the responding officer to be able to justify their assessment and compare it with the DAAT prediction.

2.6 To test the effectiveness of the DAAT, WYP are due to start their trial of the DAAT in the Leeds district this year. The trial will work with detective officers and local district offices to monitor how the DAAT is used including increased supervision through live watching of how the DAAT runs in operations. There will be continuous monitoring during this trial applying a mixed method approach ie both qualitative and quantitative methods. The trial is intended to last for six months with the accuracy of the results measured after 12 months.

3. Key data protection considerations

3.1 WYP and the Sandbox considered a number of key data protection themes in relation to the development and conceptual implementation of the DAAT. Some of those key areas of consideration are outlined below.

Data minimisation

3.2 The third data protection principle applicable to law enforcement processing, provided for in section 37 of Part 3 of the DPA, requires that the processing of personal data be adequate, relevant and not excessive. This means the data must be limited to what is necessary for the purpose(s).

3.3 As part of the Sandbox work, the ICO and WYP discussed how WYP should apply the data minimisation principle when working out the key predictive factors for the DAAT algorithm and their relative weighting in producing an accurate DAAT assessment. The ICO advised WYP that an effective data minimisation process will need to weigh up the value of each factor

for statistical accuracy and only process the personal data that is necessary and relevant to the purpose of the DAAT algorithm. CCEBP have helped WYP identify which factors are most valuable to the accuracy of the algorithm and as a result the number of factors has been reduced in the updated algorithm.

3.4 The Sandbox work also looked at the concept of adequacy in the context of the data minimisation principle, with a focus on ensuring that the data is adequate for the purpose of predicting future risk of harm. The Sandbox confirmed the ICO's guidance on this point which states that data may be inadequate if you are making decisions about someone based on an incomplete understanding of the facts (read "When could we be processing inadequate data?" in our guidance on the data minimisation principle). To comply, WYP were advised that they may need to add new factors to the algorithm if they identify any that are likely to have a significant impact or predictive force on the score in the future. WYP were able to demonstrate the availability of a process for identifying such new factors. In the objective two workshop, WYP shared that they had already revised the categories of personal data used by the algorithm. This was achieved by applying the algorithm to updated datasets that address the historical underreporting of coercive control in crime records, alongside adjustments to the threshold for identifying high-harm cases. The workshop discussed the features of this process such as submitting new factors to an ethics and governance board which the ICO agreed was a useful method for assessing the relevance, adequacy and necessity of those new factors to deliver an accurate prediction.

3.5 WYP and the ICO also looked at how WYP could approach the balance between the data minimisation principle and their work to mitigate against bias in the algorithm in relation to protected characteristics named in the Equality Act 2010, eg. race, gender, age, etc. It should be noted that data protection law provides additional protections for sensitive processing as defined in section 35(8) Data Protection Act 2018 (DPA 2018), while UK equality law is concerned with protected characteristics and not all processing involving protected characteristics will require the additional protections for sensitive processing. This report uses the language 'protected characteristics' where it is discussing mitigating potential discriminatory effects. WYP confirmed to the ICO that information about protected characteristics was not shared with CCEBP as this was not necessary for the development of the DAAT algorithm. The ICO agreed that such a measure was likely to respect the data minimisation principle. The ICO also advised that WYP should have organisational policies to monitor the amount of variance in bias between the new DAAT tool and the existing DASH assessment, these policies should set out what the

escalation and variance investigation procedures will be. Following these steps gives WYP the means to monitor the effect on protected characteristics without including this kind of data in the algorithm testing and training process.

3.6 The ICO advised that monitoring that impact should include recognising the risks posed by "proxy variables" which are closely correlated with protected characteristics and therefore reproduce patterns of discrimination even though the explicit characteristics have not been used in the model. Therefore, removing particular attributes to mitigate the risk of discrimination will not necessarily achieve the intended outcome if proxy variables are not managed. WYP explained that they had been able to identify proxy variables such as postcodes and removed them from the data that was provided to CCEBP in order to limit bias during the development of the algorithm. WYP were advised in the Sandbox to develop this work further by ensuring they have a clear process for identifying other possible proxy variables. WYP's ethics panel should have oversight of this process when reviewing the final list of factors after the data minimisation process has been completed. This was followed by further advice in Objective 4 to use statistical correlation as a method to identify proxy variables, ie if a variable considered by the algorithm is highly correlated (for example, between 70 to 100 per cent) with another variable representing a protected characteristic. Once identified, WYP should assess if the variable explains some of the difference in outcomes amongst different characteristic groups, and whether it would have an impact on the accuracy of DAAT if it was removed. The ICO noted that any identified proxy variables should be documented in an index of data sources or features that should not be processed when making decisions about individuals to reduce the risk of direct or indirect discrimination.

Data sharing

3.7 The Sandbox workshop on data minimisation also looked at data sharing and making sure that data is only shared when it is necessary. WYP confirmed that they will be sharing the DAAT score with local authorities and statutory organisations, and may refer cases to third sector organisations and charities depending on the circumstances. WYP received feedback from the ICO that they should consider how much information, if any, about the DAAT prediction is shared and ensure a data sharing agreement is in place with each authority and organisation. This agreement should include limitations on what the third party can do with the DAAT score. In the workshop, WYP communicated that it was valuable to share information with third parties directly to reduce the risk of retraumatising survivors by having to ask them for information again. This was an

interesting example of the way in which WYP can liaise with third sector organisations to develop a shared understanding of what information would be relevant, adequate and necessary for their services to be able to help those referred. The workshop also briefly considered the fact that DAAT scores may be shared by WYP with the Crown Prosecution Service (CPS) and judges when individuals apply for domestic abuse protection orders and sexual risk orders. To respect the data minimisation principle, WYP should consider whether it is necessary to share the data and where they conclude it is, only share what is necessary for that purpose. This may include providing appropriate information about how the score is generated and the role of the DAAT so that the meaning of the result can be properly understood.

Purpose limitation

3.8 At the beginning of the Sandbox engagement, WYP were still in the process of finalising where the DAAT risk score would be recorded and in which system. The impact of such a decision was discussed in their Sandbox workshop on the application of data protection principles. In particular, the ICO highlighted that WYP should consider the purpose limitation principle when deciding where to record the DAAT score as some systems may allow the score to be used for purposes other than intended by the initial processing. Specifically, the purpose of the DAAT score is to be applied to instances of domestic abuse and WYP should consider whether how the score is recorded has the potential to act as a default flag where individuals may have a high risk of serious harm flag on their name that can be used in policing matters unrelated to domestic abuse.

3.9 When considering purpose limitation, WYP were advised that it would be important to manage access logs that can identify if the DAAT is being used in a way that does not comply with organisational policy. In addition, the DAAT score should only be generated after the responding officer has already made their own risk assessment. It was suggested in the workshop that quality assurance processes involving usage monitoring and access management be updated to cover the DAAT. These processes can be used to ensure that DAAT is limited to its purpose and is being applied after the initial risk assessment to ensure meaningful human intervention at each stage and that other factors, such as officer experience, are taken into consideration as intended.

Transparency

3.10 In contrast to the UK GDPR, Part 3 of the DPA 2018 does not include "transparency" as an explicit principle. This reflects the nature of law enforcement processing which means it would often not be possible to provide full transparency regarding processing, for example, due to the potential to prejudice an ongoing investigation in certain circumstances. The obligations to provide information to data subjects can also be restricted.

3.11 Despite these more limited transparency requirements, it is recognised that it is still important to consider what information can be shared about law enforcement processing. In this context, there has been particular public interest in police use of algorithmic processing often referred to as 'predictive policing', and it is recognised that some information relating to this processing could be subject to disclosure under the Freedom of Information Act. Therefore, there is value in police forces being proactive in considering what information can be shared with the public and data subjects about the tool.

3.12 Possible audiences for such information were identified through substantial discussion about the needs of different stakeholder groups and providing a "vision-based" explanation that outlines that DAAT is designed to be more accurate than DASH. This would provide WYP with an opportunity to explain how more accurate predictions help prevent individuals from becoming victims to future harm before it happens. WYP and the ICO agreed that there is no obligation to provide every audience with granular technical details of the DAAT and how it works. Instead, the privacy notice could contain more detail about the measures taken to improve the accuracy and fairness of the DAAT which is likely to be the general public's primary concern. The ICO highlighted the fairness, safety and performance explanations from the ICO guidance on explaining decisions made with AI as possible approaches to providing transparency information for WYP to consider. Choosing to inform the public about their approach to making the DAAT operate fairly and accurately would also be consistent with approaches taken by other police forces, such as Avon and Somerset Police, who have a separate "Ethical considerations" section in their explanation for the use of data science. The ICO was of the view that there is likely a benefit in publishing some information about the steps WYP have taken to develop the tool responsibly, including the findings from their initial and ongoing fairness testing.

3.13 As part of the work carried out under the objective on transparency, the ICO reviewed WYP's draft privacy notice for the DAAT algorithm. The review highlighted the need for a clearer description of what the DAAT is so that readers have sufficient clarity. The ICO also offered the suggestion that WYP could use the privacy notice to link to a separate and more general description of their use of machine learning, citing the Avon and Somerset Police example. The review also suggested that there was value in having a prominent "Will my personal data be used for automated decision-making, including profiling?" section in the transparency information as it is important for WYP's purposes that the public understand that the DAAT is not intended to operate as an automated decision-making tool, rather it provides officers with additional information about the potential risk of future harm following an incident.

3.14 A key conclusion from the Sandbox objective on transparency was that it would be easier for WYP to determine where the publication of information about the development, accuracy and fairness of the tool would be suitable if they completed a stakeholder mapping exercise. Such a mapping exercise can identify the needs of different groups for information that would be helpful to address their concerns, enable them to exercise their rights, and understand how the DAAT operates. WYP will want to accommodate the different needs of each group as they explain the DAAT to survivors, a police officer using the algorithm, chief officers, the general public and various interested groups. WYP worked with West Yorkshire Combined Authority to scope out who they could involve from survivor groups, charities, and third sector organisations to better understand and identify stakeholders in that mapping exercise.

3.15 WYP gave clear examples of how information would be tailored to different stakeholder groups. For example, WYP plan to produce detailed technical information about the DAAT algorithm for specialist teams to create training materials and monitor the DAAT effectively, whereas frontline officers would benefit from information about how to use the DAAT, with an emphasis that the DAAT is advisory, rather than its technical specifications.

3.16 While looking at the nature of transparency information that WYP could provide about the algorithm, the Sandbox workshop contemplated whether the provision of information could allow individuals to game or exploit the algorithm to avoid detection. From that discussion, it seemed unlikely that it would be possible to exploit the algorithm from information being made available publicly as the algorithm was developed using existing or established crime data which cannot be amended

through behaviour. However, as a result of this discussion, WYP decided they would run a focus group to confirm whether or not such exploitation was possible as they had not formally made that assessment.

Right not to be subject to solely automated decision-making (s49 DPA 2018)[3]

3.17 Section 49 DPA 2018 prohibits a controller from making a "significant decision" based solely on automated processing unless the decision is required or authorised by law. Where the controller is required or authorised by law to make a significant decision, section 50 sets out the safeguards that will apply to such a decision.

3.18 During Objective Two of WYP's Sandbox plan, the question of whether any solely automated decision-making was taking place arose. In the initial stages, WYP's intention was to monitor the accuracy and impact of the DAAT by using the DAAT algorithm on live cases alongside existing risk assessment processes (ie DASH), with the contingency that if DAAT flags a case as 'high risk' but the DASH assessment flags the case with a lower risk, there would be an additional supervisory review performed. WYP asked for feedback from the ICO as to whether the decision to carry out an additional supervisory review could amount to a "significant decision" as per s49 DPA and if there was meaningful human involvement in that decision to give an individual an extra supervisory review. In response, the ICO provided WYP with guidance on how to use their operational knowledge to make an assessment on the application of s49 DPA.

3.19 Firstly, the Sandbox prompted WYP to assess if there would be meaningful human involvement in the decision to give an additional supervisory review to a case that receives an automated score which is different to the DASH assessment score. Secondly, to determine if the decision was significant, WYP were advised to consider the impact of that extra scrutiny on the outcome for those cases; for example, whether they are obtaining a further avenue for an initial risk assessment to be overturned which otherwise would not have happened without the discrepancy with the DAAT algorithm. WYP could then use this information to weigh up whether this could have an adverse or significant effect on the data subject to bring it within

[3] Please note that Sandbox work on automated decision making was assessed under current legal framework, prior to the Data Use and Access Act 2025 provisions relating to automated decision making coming into law in 2026.

the scope of s49 DPA 2018. It may be that there is no significant effect as WYP explained that decisions about incident response are taken after multiple opportunities for human intervention and scrutiny, so that most incidents are subject to a very similar review and response process which limits the temporary effect of the extra review. As a result, it would not be significant and s49 DPA 2018 would not apply.

3.20 In making the assessment, the key question for WYP will be whether the extra review means that a case can or cannot obtain a certain outcome purely because the DAAT generated a discrepancy in the risk score. The ICO recommended that WYP use a decision tree to identify the different decisions made throughout the process of responding to an incident and if at any point the DAAT provides a difference in the review process. In the workshop, it was indicated that there are several points of intervention for all cases and therefore there might not be a significant impact as they are, in fact, receiving the same level of scrutiny regardless.

3.21 To ensure meaningful human involvement in an AI-assisted outcome, individuals (police staff) must be appropriately trained and prepared to use the model's results responsibly and fairly. Training should include conveying basic knowledge about the nature of machine learning, and about the limitations of AI and automated decision-support technologies. It should also encourage the users to view the benefits and risks of deploying these systems in terms of their role in helping humans to come to judgements, rather than replacing that judgement. In considering the above, there was discussion about whether officers should be provided with information about the weighting of each factor that goes into producing the DAAT score. The ICO and WYP worked together to weigh the pros and cons of access to this information. This included discussion about the impacts of sharing this information with some groups and not others, in particular whether it would be useful information for supervising officers rather than the responding officer. The ICO highlighted that providing information about the weighting of various factors might support an officer to consider and justify their assessment in comparison to the DAAT score. For example, understanding that a factor is a significant driver of the score may help to contextualise the difference in risk assessment where the officer knows that the factor should actually be given less weight in this particular case. The outcome of this discussion was that WYP would consider alternative solutions such as making sure this information is included in training provided to the officer so that they have a general awareness of the weightings rather than providing too much technical information at the point of use. The ICO's feedback stated that effective training should also convey basic

knowledge about the nature of machine learning, and the limitations of AI and automated decision-support, so that officers can assess the value of a DAAT score alongside other variables.

Fairness

3.22 To comply with the first data protection principle in section 35(1) DPA 2018, WYP must process personal data fairly. This means WYP must only process personal data in ways that people would reasonably expect and not use it in any way that could have adverse effects, such as unjust discrimination as a result of algorithmic processing.

3.23 WYP provided information to the ICO, to explain the nature of their bias testing which is to be carried out by CCEBP. The potential bias of the DAAT will be measured against the results of the previous approach to identifying risks in domestic abuse assessments, ie DASH assessments. Where WYP has the available demographic data, testing will measure the rate of false negatives, false positives, true positives and true negatives for different demographic groups including race, gender, sexual orientation, disability, faith etc. The ICO generally agreed with WYP's decision to give more weight to false negatives and false positives as measures of the most dangerous and/or 'wasteful' errors (noting to WYP that 'wasteful' errors can also be harmful due to implications for an individual's human rights such as freedom of movement, freedom of association and so on due to unnecessary police intervention from inaccurate labelling). The method of testing will evaluate if the error rates are more frequent than using the previous approach of DASH assessment on their own at a rate that is statistically significant. Such results would show whether the DAAT assessment is better at correctly identifying risks across most demographic groups than the use of DASH assessments alone.

3.24 In their upcoming trial, WYP intends to run the DAAT for a review period of 12 months. To limit any severe adverse effects, they will still collect month-on-month analysis data comparing the DAAT with DASH, as well as developing early warning mechanisms so that bias observed in the DAAT is not left without intervention for 12 months. WYP should also set out how their efforts to establish data quality and address any gaps in crime recording data function as mitigations to address historical biases in the crime data.

3.25 The ICO advised WYP that they should take a broad approach when checking for bias which looks at more than just the elements of algorithmic fairness that are captured through statistical approaches. A broad approach will look at the effectiveness of WYP's governance structures and compliance with anti-discriminatory legal requirements. This is particularly important as WYP have identified that they are dealing with incomplete data for some groups of protected characteristics and, as such, it may be difficult or misguided to rely on being able to mathematically measure and remove bias in the DAAT's algorithmic model using statistical measurement alone.

3.26 In addition, as the DAAT will be used to support decisions and does not make decisions itself, the ICO gave WYP the feedback that they should monitor the impact officers have on the performance of the DAAT, ie how they are using the DAAT score in their overall risk assessment. WYP would then be able to assess whether the DAAT is operating fairly or relying on the officers to mitigate the risk of bias resulting in discriminatory outcomes. Effective monitoring would include keeping a record of when officers disagree with the assessment score of the DAAT as this will enable WYP to make an informed evaluation of both their reviewers and the algorithm.

3.27 We reiterated earlier discussions on transparency and the stakeholder mapping exercise completed by WYP, including whether they would like to create a fairness explanation. If WYP chooses to do so, they could draw on the findings from their initial and ongoing fairness testing, as described above, to inform that explanation.

3.28 Once bias testing is complete, if WYP discover DAAT is resulting in discriminatory outcomes for individuals, they will have to take action to remove or minimise it. This will require discovering why those discriminatory outcomes have occurred. To investigate the cause of any discriminatory outcomes, WYP were advised that they may need to evaluate the training data that was used and assess whether it might reflect past discrimination that is impacting the DAAT. If WYP finds that the training data reflects past discrimination, the ICO advised that they could either change DAAT's learning process or modify the model after training.

3.29 Discriminatory outcomes in the model could also be driven by a relative lack of data about a statistically small minority of the population which makes them statistically 'less important' to the algorithm meaning that the DAAT is less likely to make accurate predictions for those groups. WYP have already identified these limitations in the bias testing methodology due to

certain populations being under-represented in the training and testing data. In cases where information about a specific characteristic is underrepresented, the statistical accuracy of the model can be increased by collecting more data about individuals with those characteristics. However, this is not always appropriate as taking measures to reduce bias must be balanced with the risks that collecting additional data may pose to the other rights and freedoms of those individuals. The ICO emphasised that this balancing act should consider the protection of minorities or vulnerable populations while being cautious of exacerbating pre-existing power imbalances and the obligation to comply with the data minimisation principle. The ICO's view was that, in this instance, collecting additional information about minorities in order to mitigate the impact of underrepresentation in the training data is highly unlikely to be appropriate and may risk unfair and disproportionate effects on a minority population who would be subject to greater data collection by law enforcement. WYP should focus on other interventions to mitigate discriminatory outcomes such as human intervention by officers, governance controls and greater transparency by providing an appropriate fairness explanation.

3.30 Evidence that WYP has taken steps to mitigate discriminatory outcomes will include actions such as demonstrating that the DAAT is being used by staff that are sufficiently trained to implement it fairly. This may include showing that WYP have appropriately trained the supervising officers to avoid automation bias (over-relying on the DAAT prediction scores) or automation-distrust bias (under-relying on DAAT prediction scores because of a lack of trust in them). This includes measures such as providing a clear explanation that the DAAT's role in decision making is only suggestive and is always considered alongside other information and the supervising officer's experience. A suitable explanation will cover any limitations of the DAAT such as statistical uncertainty associated with the result as well as relevant error rates and performance metrics.

3.31 In addition to data protection law, individuals are protected from direct and indirect discrimination by the UK's anti-discrimination legal framework, notably the UK Equality Act 2010, which WYP must also comply with. Although such considerations are outside of data protection law and therefore outside the scope of the Sandbox, it was noted by the ICO that these obligations are separate and additional to those relating to discrimination under data protection law. Compliance with one will not guarantee compliance with the other.

Data protection risks - accuracy

3.32 WYP set out that the purpose of the DAAT is to be a supportive tool in helping officers predict the likelihood of serious harm with more accuracy compared to existing methods. This means that WYP will need to have a clear picture of what a good level of accuracy looks like, in particular, what level of accuracy is acceptable and whether the DAAT is more accurate than the existing process. The Sandbox and WYP deliberated on the different risks associated with false negatives over false positives. WYP will use tools such as their daily meeting before the MARAC where representatives from various agencies come together to discuss and share information about high-risk domestic abuse cases to confirm gradings with domestic abuse specialists and local authorities. They will measure if the DAAT is able to provide a more accurate prediction of risk so that it addresses the limitations found in DASH.

3.33 In addition to quantitative measures of accuracy, the ICO agreed with WYP's approach to include more qualitative measures of accuracy such as direct feedback about the DAAT's performance from officers and MARAC partners in decisions to retrain DAAT. The ICO also suggested that other metrics for accuracy could include how often officers indicate that the DAAT decision should be overturned during the trial period. However, WYP should monitor this carefully as the statistical accuracy rate of the DAAT may be impacted by the fact that officers are overturning an accurate prediction.

Data Protection Impact Assessment (DPIA)

3.34 As part of Objective 4, WYP provided their data protection impact assessment (DPIA) to the ICO for review. On the whole, the ICO found that the DPIA suggested relevant mitigations to the general risk areas that were raised in workshop discussions.

3.35 However, the ICO did identify the classification of risks as an area for improvement in the DPIA itself. The review of the DPIA found that risks could more directly and specifically articulate their impact on individuals. The ICO recommended that WYP reconsider how they describe and categorise risks in the DPIA. It should be clear how a risk undermines compliance with a specific data protection principle or requirement and the impact that would have on affected individuals. WYP had

linked to a risk matrix which was not appropriate to the DPIA as it focused on budgetary, performance metrics and reputational impacts, with fairly minimal reference to impacts on data subjects. Following the review, WYP were advised to place greater emphasis on the potential impact on individuals such as any physical, emotional or material harms as part of scoring the severity of any given risk. A copy of the ICO's taxonomy of harms was provided to WYP as an example of the variety of harms that can be considered when evaluating the impact of a given risk. In the context of a DPIA, these risks should take precedence over more operational risks. In response to this feedback, WYP created a revised impact risk matrix using harms associated to individuals that was not reviewed by the ICO as separate to the initial work of the Sandbox plan.

3.36 Furthermore, the initial DPIA had no matrix for likelihood of risk, so it was unclear what considerations were being applied to inform and determine likelihood scores in the DPIA. The ICO's DPIA guidance says that to assess whether the risk is a high risk, organisations need to consider both the likelihood and severity of the possible harm. Harm does not have to be inevitable to qualify as a risk or a high risk. It must be more than remote, but any significant possibility of very serious harm may still be enough to qualify as a high risk. Equally, a high probability of widespread but more minor harm may still count as high risk. Read more in the 'How do we identify and assess risks?' section of the ICO's DPIA guidance. Following this advice, WYP revised their risk matrix for the DPIA to include likelihood scores. WYP also noted that they plan to use the outcomes from the initial DAAT trial to revise the likelihood scores in the DPIA. The trial will enable them to gather more quantitative data on potential risks by observing how the DAAT operates in practice, thereby informing more accurate assessments.

3.37 The review also noted that the scope of the DPIA should be clear and focused on the relative risks of using the DAAT in comparison to the existing processes used to assess the risk of future serious harm in domestic abuse cases. During the Objective Three workshop, WYP highlighted examples of risks that are specific to the DAAT such as the effect of significant changes to how crime is recorded. Changes to the nature of crime recording (for example, new statutory requirements, a new technical platform, introduction of health data etc) would constitute a risk that could specifically affect the DAAT as its training data would not reflect those changes, negatively impacting its accuracy. As a mitigation to this risk, WYP have documented their process for reviewing whether DAAT's accuracy is significantly impaired by these changes and, if so, committing to retrain the DAAT model.

4. Ending statement

4.1 WYP's participation in the Regulatory Sandbox has provided the ICO with a valuable opportunity to closely examine a specific example of how a police force might approach the safe implementation of machine learning for predictive policing. Both the ICO and WYP have drawn valuable insights about some of the key steps that police forces will want to consider when making sure their use of such technologies is embedded with data protection by design. In addition, this work supports the ICO's current AI strategy which aims to prevent harm and promote trust. The Sandbox project highlights the potential for police forces to innovate and build safer and more accurate prediction tools that deliver on protecting the vulnerable in society from future harm.

4.2 The Sandbox's collaborative process has provided both parties with further clarity on the challenging aspects of using AI in a policing context whilst complying with Part 3 of the DPA and mitigating any adverse effects on data subjects. WYP exits the Sandbox with a greater understanding of how to document risks associated with the DAAT, the operational considerations for compliance with data protection principles and the available opportunities for providing transparent information about the DAAT where possible. The process has informed their bias testing work which is ongoing and given WYP increased confidence as they monitor the results at an operational level during the pilot and beyond.

4.3 The Sandbox process has provided WYP with tools to engage partners, such as frontline services that provide support to survivors of domestic abuse, about the fairness and safety in the design of the tool. WYP have used the lessons from the ICO to give presentations about the DAAT to the College of Policing, International Policing Conferences and the NPCC Council so that WYP can share their knowledge to support other policing bodies that are innovating for this purpose. WYP now moves on to implement the ICO's recommendations into the trial deployment of the DAAT Tool.