ResearchCore: Core AI Series

Likely Risks of Machine Learning and AI tools in the Public Sector

Tori P. A. Olphin, MBE

To Reference:

Olphin, T.P.A. (2025) *Core AI Series: Likely Risks of Machine Learning and AI tools in the Public Sector*, Manchester, UK: ResearchCore

Core AI Series: Likely Risks of Machine Learning and AI tools in the Public Sector © 2025 by Tori Olphin is licensed under Creative Commons Attribution 4.0 International. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

Address for Correspondence

Tori Olphin, MBE

ResearchCore, 4 Olive St, Greater Manchester, OL11 2TX, UK

Tel: +447921 808988

Email: tori@researchcore.co.uk

What is the ResearchCore: Core AI Series

Welcome to the ResearchCore: Core AI Series; a collection of discussion pieces based on our learning from machine learning and AI projects in the public sector, aimed at helping policy makers to construct more effective policies for delivery of AI and Machine Learning tools in the public sector.

A key role of ResearchCore is translation of evidence and experience from delivery of AI and machine learning tools, data projects and research into usable information that can inform policy decisions, and help public sector organisations to avoid implementation pitfalls. Our team have built and developed numerous AI and ML tools, and draw upon experience from delivery of real-world solutions to ensure that public sector organisations can succeed in implementation of transparent and effective tools.

Through the ResearchCore: Core Research, Core Data and Core AI series, we provide open source publication of information and findings from research and implementation in live public sector environments. Our team have worked on projects across a range of public sector agencies, from police and criminal justice, to social care, education and health.

ResearchCore

Why Look at Risks of AI and ML Implementation?

Implementation of AI and machine learning tools is becoming increasingly common across the public sector, yet many tools do not provide transparent information about the risks they pose, their accuracy, bias and outcome measures, nor do they publish information about how risks are mitigated during implementation.

Our team have developed AI and machine learning tools in the UK Public Sector for resource allocation in crime investigation, and for risk assessment in domestic abuse, and have been involved in national work around algorithmic transparency. Through this work it became apparent that it would be much easier and cheaper for companies to create tools that did not mitigate risks that they pose, and that there was no imperative to account for the risks that are posed, and little to no comeback if tools fail. This poses a significant technological and reputational risk that public sector organisations may not know they are taking.

Therefore we set about examining a range of tools and areas where tools were likely to be developed and implemented. This, along with learning gained from development and implementation of tools in the UK Public Sector, allowed us to identify a list of risks that most, if not all, tools will take or create, and which should be taken into account in delivery of any tool.

The aim of this is to provide a usable list of likely risks, so that public sector organisations can be more informed in purchasing and procurement of these tools, and should prevent more work in the long term, by taking a pre-mortem approach to delivery of AI tools. This list can also be used by providers of AI tools, to deliver better service to the public sector. It is our recommendation that providers are asked to account for all of these risks, as well as any additional risks that are identified, to allow for better delivery of reliable and transparent AI tools.

ResearchCore

Likely Risks of Machine Learning and AI tools

This section identifies risks, and descriptions of those risks, that are likely to be relevant in most, if not all, implementations of AI and machine learning tools in the public sector.

It would be beneficial if all tools that are developed for public sector organisations provided details of how they were mitigating these risks, as well as how serious the outcome of the risk is likely to be for their tool if it does go wrong, and how likely it is that risk would occur.

This will hopefully lead to better implementation of AI and machine learning tools, and therefore fewer risks being taken in a manner where the risks are unknown to the organisations building, and using the tools. This list has been published as an open source resource, and can be used for commercial purposes, to allow for organisations that are developing tools for the public sector to provide appropriate information about the tools they are developing.

RISK	DESCRIPTION OF RISK	EXAMPLE OF RISK
Tool used in a manner it is not meant to be	It is possible for police to take actions that the data ethics committee and public would not deem appropriate for an algorithm to lead to	Facial recognition tool approved for serious crime, instead used for more minor offences
Model Bias	There is bias in data held by public sector organisations, and this will create bias in any model that is produced from these data. These biases can lead to differential treatment and provision of services, or to differential enforcement	Policing happens more in poorer neighbourhoods, so more crime is found there. This can be hard-wired into data systems meaning that where you live can be seen by models as a risk factor
Model Unfairness	Unfairness can occur through bias of data, or through inappropriate use of features	Tools may make more mistakes in one ethnic group than another

RISK	DESCRIPTION OF RISK	EXAMPLE OF RISK
Model Oversensitivity	If the model is too sensitive to any individual piece of information, it may be majorly affected if data for that part is missing, or erroneous	Some predictions are more sensitive to individual factors, and data for these factors becoming erroneous would dramatically change model outcome
Fairness Gerrymandering	It can be possible to increase fairness in wider groups whilst reducing fairness in combined subgroups	Tuning a model to be more fair overall in terms of ethnic background may make it less fair in some groups than an untuned model
Failure of the Tech Stack	If parts of the technical solution fail, it would cause the model not to run correctly	Database containing the data fails to update
New Crime (or other information) Categories	The list of crimes that can be added into the system is not retained in a consistent manner, and so is not currently in a position to be used indefinitely	Non-fatal strangulation offence was created, an offence that did not previously exist
Missing Data	Data can be missing for various reasons, and this can affect the validity of models if not dealt with appropriately	Crime data does not exist for unreported offences, even if they occurred
COVID-19 impact on data and outcomes	COVID-19 and lockdowns have changed the way that crimes have occurred during 2020 and 2021, and the mechanism by which this has occurred is not entirely known. Therefore it is imperative that the model is tracked continuously once implemented in order to ensure that the accuracy and bias are not negatively impacted by a return to non-lockdown conditions	Some outcomes became incredibly rare during COVID-19, due to lockdowns and closure of venues

RISK	DESCRIPTION OF RISK	EXAMPLE OF RISK
Unseen Telephone Game between Tools	Where multiple tools or models are chained together, errors or hallucinations in one may be passed through other tools and errors may be built upon	A chat bot is used to gain information from a victim, and is then used in a risk assessment, but the chat bot hallucinated some information
Data Input Inaccuracy	There has been a lot of care taken to clean up data that informs the building of the model. It is therefore also necessary that data that is used to obtain risk decisions from the model be as clean as possible	Any human-entered information can be mistyped, or incorrectly entered
Person Linkage errors	There are issues with persons having multiple PERSON_ID numbers (it is not a golden nominal system), and this means that there is a possibility for occurrences to be missed for people both when building the model, and when searching based on a new person.	Where one person has multiple reports of them being missing, but under different unique IDs, meaning their records are not matched up when identifying risk
Delays in Data Import Process	Any delay in getting the information to the decision maker increases the likelihood of the model either being ignored, or of the model losing legitimacy in the eyes of the police as delays would lead to additional requirement for risk assessment which not only increases resource cost, but also reduces motivation of officers who made assessments earlier in the process	Delay in update of a data system causing part of the data used by the tool to be unavailable at the time the tool runs
Imprisonment or death prevents offending, causing downgrade in outcome variable	If a person who would have committed a high harm offence was imprisoned, and therefore unable to commit an offence, this would be recorded as a standard risk erroneously in training data	Model was trained on data without imprisonment removed, meaning that some records could not have led to harm being caused, where they might have if not in prison

RISK	DESCRIPTION OF RISK	EXAMPLE OF RISK
Model deployed at the wrong time in the process	Where human officials have to act on information that is created, it is possible for confusion to be caused if the model is deployed either before the information is available, or too late to change behaviour	A risk assessment tool that is deployed either before the visit to the victim has been completed, or after the supervisor would have to ratify the risk grading
Unknown whether previous action changed outcomes	Where previous cases were recorded as being high risk, it is possible that treatment by police and partners had an effect on the outcome	Where a regular missing person was given an intervention to reduce risk, but was part of the training group
Model performance changes the data that may later be used to retrain it or future models	Models go stale over time, and it is necessary to retrain them. However, any cases that have gone through this model may have been changed in terms of outcome, as there will be more information relating to what works gained through use of a model. This change in outcome would affect the new model that was trained on these data	Not having a hold out set means that all of the data has been potentially treated differently due to running through the model already
Lack of trust in the model	Some professionals may choose to override the model and go with professional judgement regardless of the evidence. This may result in less accurate predictions	A professional who does not believe that the model can predict outcomes for domestic abuse, so they deliberately choose to go with older less accurate methods
Model changes actions of professionals in cases where they should have used discretion	It is also possible that professionals turn to just relying on the model without making their own decisions to override it when they should do so. This would potentially also lead to less accurate predictions	A professional knows some important information the model doesn't, but they blindly trust the model even though the model did not have access to that information

RISK	DESCRIPTION OF RISK	EXAMPLE OF RISK
Deliberate manipulation of the model	If you know how a model works, it is possible to manipulate the output through provision of erroneous data. This could be used to manipulate police actions if done effectively	Knowing what is used to allocate resources to a job could allow a person to change how they act so their crime does not get investigated or safeguarded as much
Requirement to remove an individual from the model	If an individual's data is required to be removed from the data retained by the organisation for any reason, it may be necessary to retrain the model without that individual's data to ensure that there are no residual traces of that data remaining in the trained model	GDPR requirement to have the right to be forgotten
Model becomes stale	Over time, models may become stale, slowly becoming less accurate due to slow drift in all of the environment that predictions are made in. This could be seen as a generalised chronic data drift occurring slowly over time	This will happen in almost all examples of a model being used for any length of time
Technical Debt Build Up	Technical debt is built up in many ways during the machine learning development process. Choices made during model design can be hard coded into the machine learning pipeline, and if other parts of the process are built on top of these, it can lead to slowing in model performance, reduction in decision making quality, or increase in compute costs over time. There are many other impacts of technical debt build up that are compounded as more tools are built	Anywhere in the system where poor design leads to something that would delay a tool being fixed if something that looks simple breaks in future.

RISK	DESCRIPTION OF RISK	EXAMPLE OF RISK
Model concept drift	If the outcome concept changes, this would likely render the algorithm unable to function in the manner it was designed to. Any acute change in the outcome variable would likely lead to this issue in some way	Definition of spam changes for a spam detector, and it therefore becomes immediately less accurate
Previous performance bias can be hard-coded	If there is biased provision of services, or biased recording of variables, this might lead to a bias that is picked up by the model, which would then be hard coded into bias in future decisions	Less time being taken with some people than others, would mean the effects of time could be hard coded into the system
Outliers may influence policy	It is possible for algorithms to pick up on outliers and hard code these into decision making, in a manner that may unknowingly affect policy.	If a crime solvability and resourcing algorithm was built on data that showed one criminal offence as always being unsolved, it is possible that the algorithm could code that crime type as unsolvable, and therefore lead to accidental decriminalisation of offences
Obfuscation of data for future projects	Manipulation of data for the purpose of an algorithmic tool can change the recording of data, or can add new data or cause other data to be removed or obfuscated. This has the potential to limit future projects that might have found the obfuscated data useful	A model decision could be used instead of data that used to be collected by a human. Any other benefits of recording the old information could be lost

RISK	DESCRIPTION OF RISK	EXAMPLE OF RISK
Lack of understanding of, or attention to, training data	If the model is being designed with insufficient understanding of, or attention to, the features that are going into the training data, it may lead to features being created inappropriately, or bias being introduced unknowingly through inclusion of features that would not be desirable	Public sector agencies often use jargon, and this could create mis-labelling in the data, for example when identifying the difference between a suspension and an exclusion
Lack of understanding of, or attention to, desired outcome	If the model is being designed with insufficient understanding of, or attention to, the outcome variable that is chosen, it may lead to predictions being made that are not aligned with human values or requirements of the organisation	Risk outcomes are often not labelled neatly, so require professional judgement to know what classifications would be different levels of risk
Data drift – Acute change in feature variables	Acute changes in data received as inputs by the model could dramatically impact the accuracy of the model and could cause dramatic variance in decisions	If text analysis were used to form a feature, and then a copy-paste script containing previously impactful words were implemented, this would cause all cases to answer yes to this feature which would dramatically change the outcome
Data drift – due to model use	It is possible for features that make up a model to be altered by the use of the model; either by differential treatment of a previous incident which then alters the path that incident would have taken, or through inclusion of a feature that is directly affected through an unwanted loop	Using previous outcomes of risk assessment as part of the assessment of risk this time

RISK	DESCRIPTION OF RISK	EXAMPLE OF RISK
Data drift – change in input accuracy	If there is a change in the level of accuracy of recording of features, this might affect the accuracy of the predictions of the model	Where an improvement in recording occurs due to a change in system input method
Data drift – Rare event changes data	As with Covid-19 above, rare large scale events that alter the environment in which the model is performing can lead to the model being inaccurate in the new environment, or at least mistuned	War, changes in legislation, natural disasters or economic depressions
Data drift – New categories, definitions or classifications	Introduction of new entries or categories into existing data structures can lead either to model drift, or to the model ceasing to function due to a break in the pipeline logic	Out of court disposals being introduced, changed the outcomes and inputs for future offences
Data drift – change in measurement resolution	Any change in the resolution of data that is going into the model would likely lead to an alteration in how the model performs	Increased sensitivity of drug detection may lead to false positives
Data drift – Tool built upon other tools	In cases where multiple models exist, and outputs from one model make up part of the input to another, this can lead to a massive compounding of technical debt, and can entangle predictions and recommendations, making them almost impossible to disentangle. In addition, changing anything changes everything, meaning that there is an increased risk of changes to one tool causing drift in another	Where an outcome from a social services model is then used as part of a model to assess risk by police

RISK	DESCRIPTION OF RISK	EXAMPLE OF RISK
Misalignment with human values	It is possible for a model to very accurately predict something that is not aligned well with human values, thus leading to decision makers being misled, or making decisions based on logic that they might not have agreed with	A solvability algorithm could be trained to optimise resources and clearance rate, or could be trained to minimise caseload of certain crimes. These would have vastly different outcomes for policing, which could also have knock on effects in relation to differential levels of public confidence, perceptions of legitimacy, or even levels of deterrence which could actually lead to more crime
Technical issues – Delays in data import process	Delays in the data reaching the model could lead to the model output not being available in a timely manner, and not being available at a time that would be useful to prevent harm	Bandwidth issues, or equipment failure
Technical issues – Timeliness in delivery of output	Due to the fact that person matching has to be conducted each time data is run through the model, as well as other modelling steps that will be pre-coded, there will be an amount of time that is taken to execute the code. This is a delay in getting the information back to officers	Inefficient code, or code that needs to be run at a particular time that does not fit with when information is needed
Technical issues – Lack of testing provision	Untested code and data can introduce problems that are unseen, and if built upon, can result in issues throughout the modelling process, inconsistent application of models, and unexplained errors	Any untested code

RISK	DESCRIPTION OF RISK	EXAMPLE OF RISK
Differential levels of information available for different occurrences	Where some cases have more detail than others, it can cause errors in the measurement of outcome. This risk can also apply if some features would be differentially affected for different persons	In this dataset, there are some persons who reside outside the area, and therefore crime information relating to these persons are not available for follow-up crimes if they were in the area only once or sporadically
Model Building Decisions – Missing data treatment or unintended hidden feedback loop creation	If any features of the model link directly to data created by the model this would create unintended and unwanted feedback loops in the dataset. These feedback loops can cause significant issues for model performance and reliability	A link is formed in the database, causing the result of the model to be fed to the model as an input, creating significant errors
Model Building Decisions – Inappropriate feature creation	It is important that all features are appropriate for use in the model in question, as it would be possible to create features that may indirectly increase the level of bias or unfairness in a dataset.	The inclusion of postcodes could actually lead to the model discriminating against certain populations that are geographically identifiable
Unintended consumers can use model scoring without training, or can create unintended hidden feedback loops	It is possible that unintended and untrained consumers of the model score could lead to unwanted feedback loops if they then record information from the model decision in a way that can then be used by the model in future. These feedback loops can cause significant issues for model performance and reliability	A model for prediction of abuse in police that provides outputs to social services that then gets used in social services without knowledge or oversight could lead to significant problems. In addition, just having a model isn't enough, training is needed to know how to use it, when to override it, and what the outputs mean

RISK	DESCRIPTION OF RISK	EXAMPLE OF RISK
Model used by bad actor to gain insight into data the model was trained on	Given enough access to the model, it might be possible to gain insight into the dataset that was used to train the model. This could potentially be used to predict people's personal data if they were known to be part of the build set	An organised crime group could use the model to gain insight into data about their organisation or their competitors
Training data manipulation by bad actor	It is possible to inject erroneous data into a training set, either through bad actors, or through mistakes in the data acquisition stage. Either of these occurring could lead to the model being trained to do something differently from the original intent	Knowing a model will be built that would impact on criminal capability, it would be possible to input data that leads a model to be trained in a direction of a criminal group's choosing
Breaches in the data pipeline	Increasing the complexity of data pathways to incorporate usage of an algorithmic tool could expose the data pipeline to additional risks of breach. In addition, retention of additional datasets for rebuilding of algorithms or maintenance also carries this same risk	Hackers obtain personal data relating to people vulnerable to financial crimes
Loss of public trust	If the tool is not presented to the public in a manner that shows that it is fair and legitimate, it would be possible for this to lead to loss of public trust	Police using models which only come to public attention when a news article reports on them
Tool output causes offence, loss of confidence or fear/anger	If the outputs of the tool are not managed appropriately, it may be possible for the tool to make recommendations or provide answers that would decrease public confidence, or that could lead to people making poor decisions	A chat bot that makes recommendations that are inappropriate, incorrect, or harmful

RISK	DESCRIPTION OF RISK	EXAMPLE OF RISK
Tool input becomes manipulated	In areas where sensors, whether video or other sensor types, are used to produce data that is used as part of the decision making process, manipulation of the signal received by these sensors or provided by them can change the output of the system	ANPR cameras can be manipulated through use of false license plates
Human in the loop does not have sufficient training or understanding to use the tool appropriately	Where a human has to make the final decision, it is necessary for the human to understand where they know information that the model does not. If they do not have sufficient training or understanding, any decision may not truly be a human one	A human is asked to make a risk assessment decision and the model says medium, but the professional doesn't know what went into that decision or how to override it
Lack of explainability	Some models produce outputs that are difficult or impossible to interpret, especially in deep learning systems	Where a human is expected to be the final decision maker, use of a complicated model may lead to an automated decision
Emergent Behaviour	Where models are more complex, they can exhibit unexpected behaviours that did not occur during training, and may be entirely unwanted	Chat bots reacting with strange answers, or advice that is not aligned with the organisation's views
Model may be overfitted to noise	Models may learn patterns that are irrelevant or spurious, especially in data that has a lot of features without domain knowledge being used to identify spurious features	Some areas, like festival venues, have occasional significant levels of crime, a model could identify those areas to receive more policing all year round
Membership inference attacks	It can be possible for bad actors with access to a model to ascertain whether specific data points were part of the training set, either giving away private information, or checking whether their data is known about	Organised crime groups could use the model outputs to identify whether information about their organisation is known

RISK	DESCRIPTION OF RISK	EXAMPLE OF RISK
Shadow model creation	Bad actors can create replication models to probe vulnerabilities	Using a model to identify ways around being identified as an offender
Model can't differentiate between real information and AI generated information	Generative models are becoming capable of generating huge amounts of content, and could overwhelm systems. In addition, inability to differentiate artificial content may lead to inappropriate use of public resources	Chat bot used to create crime reports could be overrun due to a targeted attack by creating millions of reports that are AI generated, preventing police from responding to crimes that are happening
Model is less accurate or performs less well than other options	Some solutions are better than others, and this also includes the human decision maker. It is necessary to check whether models do outperform the status quo, and whether this model is as good as others	Agencies continue to use DASH when it is clear that algorithmic solutions can outperform it for domestic abuse risk assessment
Environmental cost of training can be high	Use of, and training of large models can consume significant energy, leading to environmental damage	Training a new model could be expensive in terms of energy usage and environmental impact
Job displacement	Tools can lead to structural unemployment, which in the public sector can create additional problems, where staff can not be retrained rapidly to be redeployed	Police implementation of AI tools could lead to dramatic changes in workload in different areas
Amplification of erroneous information	Where AI tools produce information that sounds feasible, it can be trusted when it should not be	AI transcription tools produce a convincing transcription of a criminal interview which then causes a case to be thrown out at court

RISK	DESCRIPTION OF RISK	EXAMPLE OF RISK
Violation of intellectual property	Models may have been build using information that had intellectual property violated, which may cause legal issues later	Implementation of a model into a core function with no backup, which then gets removed from service for IP breaches
Transfer learning may not be effective	Models that are pre-trained on data that is not specific to the field in question may carry over biases or irrelevant features, and may lead to incorrect information being provided	A chat bot trained for financial services may not do a good job at recording reported crimes and gathering information
Change of deployment context can fail	Deploying a model that works well in one context into another may well not work to the same degree, and may fail	Speech recognition tools trained on American English may not work well in areas with regional accents
A problem that looks the same may not be the same	Two areas may have a problem that looks similar, but the underlying data may be significantly different and a separate model, trained on data from that context, may be needed to perform well	Factors that relate to a crime being solved may be different in rural and urban areas
Lack of accountability	It is not clear who is responsible when AI systems cause harm	A risk assessment decision goes wrong and it is not known who is responsible for the decision
Procurement without knowledge	AI tools may appear to be of major benefit to the public sector, but without sufficient transparency or scrutiny, the tools may not do what they were sold to do	A strong sales team convinces an agency to buy a system which does not have any tracking of outcome accuracy and does not actually do what it is supposed to
Ethics washing	Light touch ethical oversight may not be sufficient to determine whether the use of the tool in question is right for the agency or community	A public sector agency publishes a superficial ethics statement, but without having engaged with communities its use would affect most

ResearchCore

Summary

This provides a usable list of risks that we believe are present in most implementations of AI and machine learning tools that make decisions or recommendations about interactions with members of the public. While there are some types of tool that will be affected by some risks and not others, it is likely that most tools would at least need to consider the majority of these factors, and it is likely that as the public become more highly educated about the AI tools that exist around us, more is going to be expected of public sector agencies in relation to how they handle risks such as these.

When they go right and are implemented well, these tools can improve diagnosis of risk, and allow resources to go where they are most needed. However, when shortcuts are taken, they could undermine public confidence and even create crippling technical debt which becomes unmanageable for organisations.

It is therefore imperative that organisations manage and understand the risks that are being taken, and put measures in place to mitigate for these risks. It is also a major part of transparent AI delivery to show that risks are being considered and mitigated for. This document provides a basis to consider the tools that are being implemented.

Authors and Referencing

Lead Researcher: Tori Olphin, MBE

Director of Research and Data Science

ResearchCore

To reference, please use:

Olphin, T.P.A. (2025) *Core AI Series: Likely Risks of Machine Learning and AI tools in the Public Sector*, Manchester, UK: ResearchCore

Core AI Series: Likely Risks of Machine Learning and AI tools in the Public Sector © 2025 by Tori Olphin is licensed under Creative Commons Attribution 4.0 International. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

Contact Us

If you have any questions, please contact the author via tori@researchcore.co.uk

Our website has more publications that our team have produced, please visit us at https://www.researchcore.co.uk

For any enquiries, or to ask us to conduct work for your organisation, please contact us at info@researchcore.co.uk