the CHURCHILL fellowship

THE COMPUTER SAYS SO:

AUTOMATED RECOMMENDATION-MAKING TOOLS IN IMMIGRATION SYSTEMS

A comparative analysis between Canada, the USA and the UK

KATIE SCHWARZMANN
2023 Churchill Fellow

Field research conducted Summer 2024; written September 2024; submitted to the Churchill Fellowship October 2024.

Published December 2024.

Cover image generated by Chat GPT.

Table of Contents

Acknowledgements

Firstly, a huge thank you to the support and funding of the Churchill Fellowship for extending this opportunity to me. Thank you also to Wilson Solicitors for providing me the time and space to conduct this research alongside work.

This report would not have been possible without the generosity and insights of the people and organisations I met with in Canada and the USA who showed me such warm hospitality and were so generous with their time. I am particularly grateful to the technology companies and government departments who were willing to speak candidly to me about their work, knowing a researcher's remit is to critically assess the status quo and consider areas for improvement.

While it would be too long a list to include everyone personally who provided me with their time, insights and support during this Fellowship, I would like to extend special thanks to Lucie Audibert whose expert review made the final draft of this report far better than the first. I would also like to thank Will Tao, Mario Bellissimo and Grant Fergusson for their generosity in reviewing the Canadian and US sections respectively, and whose comments substantially improved the quality of those sections. I am grateful to Joe Tomlinson and Mia Leslie for their time and guidance when I was initially setting out on this research journey and, along with Victoria Adelmante and Niamh Leonard, for putting me in touch with so many of their invaluable contacts. I would further like to thank Meg Goulding for her encouragement and support throughout my career so far, including in applying for this Fellowship, alongside Harriet Hall, Katya Novakovic and Jen Watson for the latter.

Finally, thank you to Mum, Dad, Joel and Ed – for everything.

About the author

Katie Schwarzmann is a public law and human rights lawyer specialising in strategic litigation at the intersection of migrants' rights and emerging technologies. She primarily represents asylum seekers, victims of trafficking and other marginalised clients to assert their rights by way of judicial review and civil claims for compensation. Katie recently brought the first case in the UK to challenge the government's policy of indefinitely GPS-tagging migrants.

In 2017, Katie graduated with first-class honours in History and Philosophy from the University of Cambridge. For her examination performance, she was awarded the Rowley Mainhood Prize, Arthur Tindal Hart Prize, Owen Scholarship and the Abdul Aziz Prize. Following graduation, she trained as a lawyer with the corporate law firm Freshfields Bruckhaus Deringer, where she completed a secondment to the human rights NGO, Liberty. She then went on to work in the human rights departments of Hickman & Rose and Wilson Solicitors LLP.

In 2023, Katie was awarded a Churchill Fellowship to prepare a comparative study on the uses and regulation of automated decision-support tools in the immigration systems of the USA, Canada and the UK. Katie is also studying part-time for a master's degree in International Human Rights Law at the University of Oxford.

Foreword

The infrastructure of UK Government is increasingly digital, and the use of automation is central to this ongoing transformation. Border and immigration systems are not only not immune to this trend but have been at the forefront of it – often being a testing ground for new automated applications and systems.

As this report shows, these changes are not just 'backroom' changes to administrative practice but hold the potential to fundamentally change how people interact with, and are treated by, the immigration system. While there may be many possible advantages of automation, they come with great risks to the people who are subject to it. These are risks which we have long talked about in the abstract but, as the case studies in this report highlight, are increasingly becoming real before our eyes.

This report seeks to learn from international experiences of how to harness the advantages of automation in this context while also preventing harms and improving accountability. It makes a powerful case for more proactive regulation, updating terminology, improving monitoring and transparency, stronger independent oversight, better training, and more effective redress mechanisms. It should be widely read by everyone with an interest in this area and, in particular, those responsible for administering these systems.

Joe Tomlinson[1]
Professor of Public Law
University of York

[1] https://jptomlinson.uk/

Preface

Imagine, you've just ordered a pricey piece of furniture online – something that was supposed to be the crown jewel of your living room. When it finally arrives, it looks less 'majestic throne' and more 'wonky footstool'. It's not blatantly false advertising – it's just a generous interpretation of the truth. You want a refund.

You click on the company's website, armed with the righteous fury of a customer wronged. You click the 'contact us' button, only to discover the company has hidden its customer services' contact details like they're state secrets. No phone number. No email. Just a chatbot named something endearing like 'ChatterBot'. Fantastic.

You spend the next two hours embroiled in a digital tug of war with the bot, trying to explain the intricacies of your disappointment, which ChatterBot, of course, is ill-equipped to understand. 'Was the item damaged?' No. 'Did it arrive late?' Nope. 'Does it look like it's been assembled by a toddler?' Ah, we're getting closer.

Meanwhile, your actual job is relegated to the background as you find yourself in a one-sided battle with an algorithm, clicking 'No, that didn't help' over and over like this will help you level up in some kind of video game. Finally – FINALLY – after what feels like a lifetime of rejecting inadequate answers, ChatterBot grudgingly coughs up its most sacred possession: the phone number of a real-life human in customer services.

Victory. Two hours of your life well spent.

Now imagine, instead of trying to secure a refund for a piece of furniture, you are applying for asylum having just fled your home country for fear of your life.

Instead of trying to explain to ChatterBot why you need a refund, you are trying to convince it to consider your entire life journey that has led you to fleeing your country and requiring safety in a new country.

And ChatterBot's role is not simply to simulate human conversation, but instead to make actual recommendations or decisions which contribute to whether you are granted asylum. Unlike an outraged customer, you may never have the opportunity to speak to a human at the end to understand ChatterBot's decisions. You might not even be told that ChatterBot was involved in any stage of your application, nor might any human be able to trace or explain why ChatterBot made the recommendations it made.

Yet its recommendations may determine if your asylum application is flagged as suspicious leading to increased scrutiny or if it is deprioritised as non-urgent meaning you have to wait months or years for a decision. Instead of its recommendations affecting a few hundred pounds, it may affect your entire future or even your ability to have a future.

The disparity highlighted in this story – the difference between an inconvenience like a refund request and a life-altering decision like an asylum application – exemplifies the issue with uses of algorithmic tools in administrative decision-making processes. As Virginia Eubanks puts it, while increased use of algorithms by government bodies impacts all of us, 'they don't impact us all equally'.[2] This concern is the impetus for this research.

[2] Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor* (St Martin's Press, 2018)

Definitions

- **ADM** – Automated Decision-Making
- **AI** – Artificial Intelligence
- **AIA** – Algorithmic Impact Assessment
- **ARM** – Automated Recommendation-Making
- **ARMT** – Automated Recommendation-Making Tool
- **ATRS** – Algorithmic Transparency Recording Standard
- **Bill** – Public Authority Algorithmic and Automated Decision-Making Systems Bill
- **CIAOs** – Chief Artificial Intelligence Officers
- **DADM** – Directive on Automated Decision-Making
- **DHS** – Department of Homeland Security
- **DSIT** – Department of Science, Innovation and Technology
- **EO** – Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence
- **FOIA** – Freedom of Information Act 2000
- **GBA+** – Gender-based Analysis Plus
- **GPS** – Global Position System
- **Guide** – Guide on the Scope of the Directive on Automated Decision-Making
- **HSI** – Homeland Security Investigations
- **ICE** – Immigration and Customs Enforcement
- **ICM** – Investigative Case Management
- **ICO** – Information Commissioner's Office
- **IPL** – Immigration Policy Lab
- **IRCC** – Immigration, Refugees and Citizenship Canada
- **ITAT** – Integrity Trend Analysis Tool
- **Mandatory Policy** – Algorithmic Transparency Recording Standard Mandatory Scope and Exemptions Policy
- **Memo** – Memorandum on Advancing Governance, Innovation and Risk Management for Agency Use of Artificial Intelligence
- **OMB** – Office of Management and Budget
- **Palantir** – Palantir Technologies
- **PIA** – Privacy Impact Assessment
- **PSR** – Privately Sponsored Refugee
- **RAU** – Risk Assessment Units
- **RCA** – Risk Classification Assessment
- **SAR** – Subject Access Request
- **Senior Official** – Senior official at Immigration, Refugees and Citizenship Canada
- **TBS** – Treasury Board of Canada Secretariat

Executive summary

The impact of automated tools to assist or replace human decision-making (herein referred to as 'automated recommendation-making tools' (ARMTs)) in the public sector cannot be overstated. With their potential to improve efficiency, accuracy and reduce costs, ARMTs do seem to offer governments a solution to modern pressures. And so, the Home Office (the UK government department responsible for immigration) understandably sees ARMTs as an innovative way to deal with the immigration backlog, announcing in 2021 its plan to become 'digital by design'.[3]

However, these solutions are currently being tested on real lives and behind closed doors. The UK's regulation of ARMTs has not kept pace with its development and has been described by a UN Special Rapporteur as existing in a 'human rights-free zone'.[4] As a lawyer representing migrants, I have seen first-hand how my clients are not routinely given information about if, when or how ARMTs are involved in decisions which affect them, curtailing their ability to scrutinise these decisions adequately. This is concerning as decisions in this context can have life or death consequences for people in vulnerable situations. So it is important the UK gets the regulation of ARMTs in its immigration system right.

To assist the UK Government in regulating these tools, this report provides a comparative analysis of the uses and regulation of ARMTs in the immigration systems of the USA and Canada, concluding with policy recommendations for regulatory best practice in the UK.

The **first section** of this report begins by mapping out the potential risks and harms of using ARMTs in immigration systems identified by research participants. The risks identified fall under seven categories: i) bias, ii) transparency; iii) controversial uses; iv) terminology; v) psychological effects; vi) data; and vii) accountability. This list facilitates regulators to consider where the trade-offs must lie, including which risks present ethical red lines for technological development, and which can be mitigated with proactive regulation.

The **second section** summarises the UK policy and regulatory landscape, before highlighting areas for improvement. It explains that although the Home Office has been embracing ARMTs, the corresponding regulatory framework is fragmented and insufficient.

The **third section** considers the Canadian context including its world-first mandatory directive regulating public sector use of ARMTs which requires public bodies to publish Algorithmic Impact Assessments prior to deployment. It details other Canadian policies, including its immigration department's current bans on ARMTs that either automate refusals or use 'black box' technology (i.e. are non-explainable). It then analyses the adequacy of Canada's regulation for managing the risks of specific ARMTs in operation in its immigration system, such as risk assessment and triaging tools.

The **fourth section** maps the US regulatory context, before describing two contrasting case studies: Investigative Case Management, a controversial tool allegedly used in immigration enforcement, and GeoMatch, a predictive tool designed to optimise

[3] Home Office, 'Digital, Data and Technology Strategy' (2024): https://www.gov.uk/government/publications/home-office-digital-data-and-technology-strategy-2024/home-office-digital-data-and-technology-strategy-2024

[4] Report of the Special Rapporteur on extreme poverty and human rights (2019): https://www.ohchr.org/sites/default/files/Documents/Issues/Poverty/A_74_48037_AdvanceUneditedVersion.docx

employability in refugee resettlement. These case studies together demonstrate the need for proactive ethical regulation to ensure ARMTs do not perpetuate harm while highlighting areas where ARMTs can potentially be used to improve migrants' experiences.

The **fifth section** summarises recommendations for UK Government, lawyers and civil society arising from the lessons learned in Canada and the USA. The suggestions focus on proactive regulation, transparency, prevention of bias and accountability to help safeguard migrants' rights. They include:

- Proactive regulation: Establish binding regulations to ensure that ethical boundaries are set by public consultation rather than private interests. This includes banning 'black box' algorithms and automated refusals.

- Terminology updates: Shift language to 'automated *recommendation*-making tools' from 'automated *decision*-making tools' and replace 'human in the loop' with 'chain of decision-making' to better reflect human involvement at all stages of these decision-making processes and ensure comprehensive oversight.

- Monitoring and transparency: Mandate pre-deployment disclosure of impact assessments and transparency records, require updates at regular intervals and enforce collection of protected characteristic data to monitor for bias. Include notice and explanation requirements to inform individuals when and how ARMTs are used in their cases.

- Independent oversight: Involve human rights advisors in the development of ARMTs and set up an external, independent audit system to assess fairness, accuracy and bias, with corrective actions required before deployment.

- Training: Implement specialised training for judges and decision-makers on ARMT impacts.

- Redress mechanisms: Provide individuals with a right to appeal at every stage of ARMT involvement.

The UK is at a historical moment where the decisions made now will shape future public life. To harness the full potential of ARMTs for improving the UK's immigration system without weakening the rule of law or perpetuating human rights violations, the UK urgently needs a new regime – one grounded in research and evidence from the experiences of North America.

Introduction

UK context

In 2021, the Home Office launched its strategy to 'become digital by design'. This involved 'embracing automation'[5] and increasing its use of algorithms to replace or assist human decision-makers.[6]

It is understandable that the Home Office sees automation as an attractive way to reduce its immigration backlog, given this is a key political priority of both the Conservative and Labour parties. Automation has obvious appeal in this context: it promises increased efficiency, lower costs and the opportunity to free up civil servants' time to spend on other, more complex immigration issues. As such, the Home Office has been using algorithms in various contexts, such as to detect sham marriages for visa purposes,[7] stream visa applications[8] and process data collected by migrants' GPS tags.[9]

However, using algorithms to replace or assist human decision-making (herein referred to as 'Automated Recommendation-Making' (ARM)) has its risks including discrimination and human rights abuses. Such risks must be mitigated for the benefits of automation to be realised. Otherwise, any efficiency or cost-savings benefits will be undermined by expensive taxpayer-funded litigation, needs to recall the technology and doubling of the Home Office workload by the requirement to re-consider wrongly decided applications.

Accordingly, the UK has been gradually developing a patchwork of legal and policy initiatives to regulate public sector use of ARM tools (ARMTs). However, progress has been piecemeal and so far has failed to sufficiently mitigate the risks. As it stands, the Home Office does not routinely disclose its use of ARMTs to the public or affected individuals, nor is it legally required to assess and disclose algorithmic risks before deployment. My experience as a lawyer representing migrants aligns with this: Home Office policies increasingly reference use of ARMTs yet I have not had a client be notified of their use, leaving them without the necessary information to scrutinise decisions affecting them. As a former UN Special Rapporteur noted, the UK's use of public sector algorithms is currently operating in a 'human rights-free zone'.[10]

This is concerning because decisions in the immigration context can, at best, be the difference between someone being able to take up work, education or life opportunities, and at worst, have life or death consequences. To harness the full potential of ARMTs for improving the UK's immigration system without weakening the rule of law and human rights, the UK urgently needs a new regime – one that is grounded in evidence and research.

---

[5] Home Office, 'Digital, Data and Technology Strategy' (2024): https://www.gov.uk/government/publications/home-office-digital-data-and-technology-strategy-2024/home-office-digital-data-and-technology-strategy-2024

[6] I note these policies were introduced under the former Conservative government. At the time of writing, the new Labour government has not announced a change in policy direction.

[7] Public Law Project, 'Legal action launched over sham marriage screening algorithm' (2023): https://publiclawproject.org.uk/latest/legal-action-launched-over-sham-marriage-screening-algorithm/

[8] Foxglove, 'Home Office says it will abandon its racist visa algorithm – after we sued them' (2020): https://www.foxglove.org.uk/2020/08/04/home-office-says-it-will-abandon-its-racist-visa-algorithm-after-we-sued-them/

[9] See page 23

[10] Report of the Special Rapporteur on extreme poverty and human rights (2019): https://www.ohchr.org/sites/default/files/Documents/Issues/Poverty/A_74_48037_AdvanceUneditedVersion.docx

International context

The UK is not alone in grappling with these issues. The USA and Canada are also trialling algorithms in attempts to streamline their immigration processes: Canada to triage applications into simple cases for processing by the algorithm and complex cases for human review,[11] the USA to decide whether a migrant should be detained prior to removal,[12] and both to match migrants to areas in which they may be most employable.[13]

In comparison to the UK, consideration into how to regulate these developments has received far more attention in the USA and Canada. In Canada, since a pioneering paper in 2018 on ARMTs in its immigration system by The Citizen Lab, University of Toronto,[14] there has been an unparalleled proliferation of academic interest in the area and the introduction of the world's first directive specifically regulating public sector use of ARMTs. Similarly, litigation in the USA of these tools is more advanced with cases frequently managing to reveal information of ARMTs' role in the US immigration system and expose controversial uses, leading to discussion on where the ethical red lines in Artificial Intelligence (AI) should be drawn. Both countries therefore offer valuable insights to the UK about how ARMTs should be regulated.

The nascence of these technologies and their application to immigration processes offers the UK a historical opportunity to position itself as a leader in ethical use of ARMTs and set a positive example worldwide. To assist the UK in taking up this opportunity, this report presents comparative research on the development and regulation of ARMTs in Canada, the USA and the UK, to provide policy recommendations for best practice in the UK.

Research objectives and methodology

The following research questions guided this project:

- What risks or harms do ARMTs present in the immigration context?
- Where are ARMTs being developed and deployed in the Canadian and US immigration systems?
- How are these deployments regulated?
- How should they be regulated?

To address these questions, over the spring and summer of 2024, I met with more than 50 people working in over 25 organisations to conduct interviews about their work. In the USA, I visited New York, Washington DC and San Francisco. In Canada, I visited Toronto, Ottawa, Montreal and Vancouver. Across both countries, I met with a wide range of stakeholders including government officials, civil servants, technology developers, software engineers, academics, lawyers, researchers and migrants' organisations.

[11] See page 35
[12] Katie Evans and Robert Koulish, 'Manipulating Risk: Immigration Detention through Automation' (2020): https://scholarship.law.duke.edu/faculty_scholarship/3994/
[13] See page 42
[14] Petra Molnar and Lex Gill, 'Bots at the Gate: a human rights analysis of automated decision-making in Canada's Immigration and refugee system' (2018): https://citizenlab.ca/2018/09/bots-at-the-gate-human-rights-analysis-automated-decision-making-in-canadas-immigration-refugee-system/

A consent form was provided to each interview participant ahead of their participation in the study in which they were informed of the purpose of the research, the use and storage of their data and whether they agreed to sharing their identity and organisational affiliation. To the extent each participant consented, each meeting was recorded and transcribed.

This report takes a human rights critical perspective to the developments of new technologies, seeking to flush out their potential and actual risks or harms on migrants' rights. I am therefore grateful to all interview participants, including the candour of the technology companies and governments, since a researcher's remit is to critically assess the status quo and consider areas of improvement.

Scope and limitations

This report focuses on the following functions of ARMTs:

- **Workflow management**: Triaging or categorising applications based on perceived complexity.
- **Risk assessments**: Triaging, categorising, associating or scoring applications based on perceived risk.
- **Electronic monitoring**: Processing data from GPS tags.
- **Matching tools**: Location-based dispersal tools designed to optimise migrants' integration into new locations.

While facial recognition and biometric collection can also be classified as types of ARMTs, they are outside the scope of this report. Likewise, there are regulatory developments occurring outside of the USA, UK and Canada, such as the EU's recent AI Act,[15] which are not discussed herein.

Finally, anyone who works in the area knows how fast-paced it is with new developments being announced almost daily. The field research for this project was conducted in spring and summer 2024, and the writing process took place in mid-September 2024. The research does not account for developments after this period.

[15] European Union, 'Artificial Intelligence Act' (13 June 2024): https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689

Section 1 – Issue spotting

Prior to conducting the field research, I assumed that there was a well-rehearsed and shared set of risks that was common knowledge to everyone working in the area. From the outset it became apparent that this was not the case, and so this report begins by providing a consolidated list of the risks identified by all interviewees of ARMTs in the immigration context. Only once there is a shared understanding of the risks, can effective proactive legal policy be designed.

Before summarising these risks, it is worth briefly noting that the term 'risks' suggests that these are potential, unrealised consequences of using ARMTs. However, as Joanna Redden persuasively argues, given these technologies are already in operation, many risks should be understood as 'harms' with the damage already done. The discourse of 'risk' is enabling to the extent it implies these tools are not already creating harm.[16] Thus, the below risk (or harm) analysis should be understood as both a projection of how further harm might be caused if better regulation is not created, and a summary of harms that have already been caused by ARMTs. This analysis contains both intentional and unintentional harms, with the identified risks to be understood as overlapping rather than as existing in siloes.

Bias

It is well-documented that algorithms can exhibit multiple forms of bias. That said, some interviewees pointed out that human decision-making is no more rational, transparent or less biased than algorithms, and in fact at least with algorithms it is possible to trace decision-making more easily.[17] However, as discussed below on risks to accountability (pages 21-22), the issue is that traditional legal systems were designed to scrutinise human decision-making processes but are not yet equipped to oversee algorithmic ones, allowing such decisions to escape regulation. Additionally, bias in algorithms – even if a mirror of human bias – is amplified by the scale and speed of their application to countless cases. As such, the risk of bias in algorithms is a serious concern that must be addressed by corresponding regulation. The following forms of bias were noted by interview participants:

1. **Historical bias**: This form of bias stems from the historical patterns of bias present in the data used to train an algorithm (known as 'training data'). If past decisions were biased, those biases can be embedded into the algorithm and applied to future cases.

2. **Data bias**: Unrepresentative or incomplete data can lead to skewed outputs. For example, using data from an overwhelmingly male part of the population to draw inferences about the whole population would lead to unreliable results.

3. **Developer bias**: The design of algorithms reflects the conscious and unconscious choices, assumptions or biases of the developers who create them – choices about the datasets selected, variables chosen and outcomes optimised for. This is particularly relevant given the lack of diversity in the technology industry.

[16] Interview with Joanna Redden, Co-Directive of the Data Justice Lab in Cardiff and Co-Director of Starling Centre at Western University in Canada (11 June 2024)
[17] Interview with Sean Rehaag, Director of the Centre for Refugee Studies and the Director of the Refugee Law Laboratory at York University in Canada (13 June 2024)

4. **Proxy bias**: Algorithms may use input variables that indirectly correlate with protected characteristics under equalities legislation, such as race or religion. For example, an algorithm may start associating visa applicants from certain postcodes with higher rates of fraud which may mean, in effect, that people from areas with high proportions of certain ethnic or religious minorities may be being disproportionately and unfairly linked to fraud.

5. **Automation bias**: Human decision-makers have a proven tendency to over-rely on algorithms, assuming they are more reliable than human judgement.[18] This tendency is exacerbated when individuals lack understanding or training of how the algorithm operates, are under time or other pressure or have limited power to depart from the algorithm's recommendations.[19] This phenomenon may undermine people's rights to a fair and impartial decision-maker, since their decisions will be swayed by a prior machine recommendation. While no interviewee suggested this risk could be eradicated, it was commonly acknowledged that it can only be mitigated if the decision-maker can understand and follow why a machine has made a certain recommendation, and therefore question this process.

6. **Feedback loop bias**: An algorithm's recommendations or decisions influence the future data it receives, creating a cycle that reinforces its initial biases. For example, if an algorithm suggests that women are less employable in certain areas, this may lead to fewer women seeking and finding work in those areas, which in turn feeds back into the algorithm, leading to similar recommendations being made in the future.[20]

7. **Confirmation bias**: An algorithm may reinforce a human decision-maker's pre-existing stereotypes since it is designed to confirm trends rather than discover new insights. For instance, if a decision-maker holds a bias that Black men are more likely to commit crimes, and the algorithm predicts the same, the human reviewer may be more likely to accept the recommendation.[21] This allows algorithms to conceal and perpetuate human biases.

8. **Computational bias**: Operational errors arise from the algorithm not working properly. There are numerous systems which have been plagued with significant errors. For example, a US data-sharing initiative between the Federal Bureau of Investigations, the Department of Homeland Security and local law enforcement checked fingerprints against federal databases, but a faulty algorithm erroneously led to 5,880 US citizens being wrongfully flagged for potential detention and deportation.[22]

---

[18] Hannah Ruschemeier, 'The problems of the automation bias in the public sector – a legal perspective' (2023): https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4521474
[19] Amnesty International, 'Trapped by automation: poverty and discrimination in Serbia's welfare state': https://www.amnesty.org/en/latest/research/2023/12/trapped-by-automation-poverty-and-discrimination-in-serbias-welfare-state/
[20] Lydia T Liu, 'When bias begets bias: a source of negative feedback loops in AI systems' (2020): https://www.microsoft.com/en-us/research/blog/when-bias-begets-bias-a-source-of-negative-feedback-loops-in-ai-systems/
[21] Interview with Mario Bellissimo, founder and immigration lawyer at Bellissimo Law Group PC (11 June 2024)
[22] Data Justice Lab, 'Data Harm Record' (2020) https://datajusticelab.org/data-harm-record/

9. **Quantitative bias** – Algorithms prioritise quantitative data, which can oversimplify complex cases and miss important nuances.[23] For example, a Canadian immigration officer cited, in his view, a non-controversial ARMT: workflow management tools that triage simple cases for automatic approval and more complex ones for human review. He explained that it could simply be that an individual applying for a visa with a child would need further investigation since it is automatically a more complicated case. However, it is questionable whether this is non-controversial: it is more likely that someone with a child will be a woman, thus the tool will lead to more women's applications disproportionately receiving additional scrutiny.

Transparency

Without transparency of public sector decision-making, the democratic right of the public to hold the government accountable for how it makes decisions is eroded. Thus, transparency is the gateway to accountability.[24] The following concerns with transparency were raised:

1. **Role of technology companies**: A repeated concern was the opaque role of technology companies in the development of algorithms for public sector use. It is problematic if a request for information under the Freedom of Information Act 2000 (FOIA) or by an impacted individual is refused because it constitutes a company's trade secret, when a public body would otherwise have to disclose this information. This highlights a key issue with ARMTs: decision-making processes that would ordinarily be disclosable to the public can now be hidden behind trade secrecy law.[25]

2. **Networks of algorithms**: Algorithms do not work in siloes but are increasingly used in conjunction with each other. The output from one algorithm based on one dataset is fed into another dataset or algorithm and so on. Most transparency initiatives, such as impact assessments or algorithmic transparency recording standards, are only designed for single use algorithms. However, mechanisms promoting algorithmic transparency must account for the interaction between algorithms.[26]

3. **Collection of disaggregated data**: As far as I am aware, none of the three countries' immigration departments systematically collect data disaggregated by race and other protected characteristics in the output of algorithms. It is not possible to monitor bias without governments first collecting this information.

4. **Default transparency**: Transparency can involve disclosing various aspects of an algorithm, such as its existence, purpose, source code, training data, input variables, human involvement and impact on protected characteristics. However, when public bodies do disclose information on ARMTs, it is often piecemeal and limited. The default should be full disclosure and public bodies should be required to justify any departure from full transparency based on strict, necessary and proportionate guidelines.

[23] See footnote 21, Mario Bellissimo interview
[24] See footnote 17, Sean Rehaag interview
[25] Interview with anonymous academic at NYU (13 May 2024)
[26] Interview with Jennifer Raso, Assistant Professor at McGill University's Faculty of Law (19 June 2024)

5. **Replacement of law with institutional choices**: In traditional immigration systems, asylum decisions are based on criteria set out in the law. However, with algorithms, governments may introduce secret variables not rooted in law. This issue was at the heart of the UK case where civil society organisations Foxglove and Joint Council for the Welfare of Immigrants challenged the Home Office's use of a visa application 'streaming algorithm'. The claimants argued that the algorithm discriminatorily subjected applicants from certain nationalities to greater scrutiny, despite this not being a legal basis for distinction – and in fact breaching UK equalities law. Following the challenge, the Home Office announced it would halt the use of the algorithm and conduct a full review of its visa processing system.[27]

   Despite this controversy, authorities often argue that disclosing algorithmic variables would allow migrants to 'game the system' by reverse-engineering successful applications. However, this secrecy contradicts public law which holds that individuals are entitled to know the criteria for decisions that affect them. The UK Supreme Court case of *Lumba v Secretary of State for the Home Department*[28] ruled that secret policies which are inconsistent with public ones are unlawful. If the input variables are consistent with criteria in public policies, it is unclear why they cannot be disclosed, nor how they would lead to applicants gaining any advantage when compared to transparent manual decision systems. Further, disclosing these variables is necessary for proper public accountability to check that they are indeed consistent with public laws and policies.

Controversial uses – 'automated suspicion'

As Francesca Palmiotto argues, there is a controversial class of ARMTs that are best categorised as those that 'automate suspicion'. Specifically, these are tools that do not make final decisions but 'generate suspicion'[29] about certain applications due to their similarities or connections with other or past applications. How, and whether, these tools should be allowed in administrative decisions is rightly the subject of much debate. Types of ARMTs that automate suspicion include:

1. **Predictive algorithms**: These are algorithms designed to analyse historical data and predict future outcomes or behaviours. This is a form of automating suspicion whereby individuals or applications are flagged because they have similarities with, for example, past fraudulent applications. Being wrongfully flagged can have significant negative effects on individuals' ability to gain immigration status, employment, travel and in some cases can lead to wrongful detention and deportation. These predictive tools could therefore be contributing to the creation of a two-tier immigration system, where individuals who look and seem like people who have historically been deemed 'non-risky' experience an efficient, frictionless immigration system, while the rest of the global population increasingly finds it difficult to travel, take up opportunities or emigrate. In effect, it is a policy of automated stereotyping.

2. **Triaging, scoring or flagging risk or complexity**: This is where algorithms are used to flag, score or triage individuals based on perceived risk or complexity, often

[27] See footnote 8, Foxglove, 'Home Office says it will abandon its racist visa algorithm'
[28] Walumba Lumba (Congo) 1 and 2 v Secretary of State for the Home Department [2011] UKSC 12
[29] Francesca Palmiotto, 'When is a decision automated? A taxonomy for a fundamental rights analysis' (2024), pages 224-229: https://www.cambridge.org/core/journals/german-law-journal/article/when-is-a-decision-automated-a-taxonomy-for-a-fundamental-rights-analysis/362AF985585D28E5E762F4FEEF4719B7

leading 'higher-risk' or more 'complex' cases to face additional scrutiny, delays, and added stress or costs for applicants. While humans make the final decision, the algorithm's flagging serves as a trigger for further review. This can shift the presumption of innocence for (or in public law terminology, jeopardise decision-makers' ability to apply an open mind to) these applications, since decision-makers need to provide positive evidence to disprove that a case is not in fact risky. The path of least resistance for the human decision-maker, therefore, is to agree with the flag. This can disproportionately affect marginalised communities, due to historical, data, developer bias and so on, as discussed above.

3. **Associations**: Another way algorithms automate suspicion is by creating networks of association. Even if someone has no personal history of criminality, an algorithm may find a reason to connect individuals to other individuals with a criminal history. The London Metropolitan Police have a controversial algorithmically-based 'Gangs Matrix' which contains information about individuals who are suspected gang members in London. There is evidence that it disproportionately over-represents Black men in the matrix.[30] Such associative algorithms are also problematic when it is common in certain ethnic minorities for unrelated individuals to have similar or even the same names, leading to false connections being made.

Terminology

A significant and often overlooked battleground of this debate is over terminology. The lack of clarity over definitions risks allowing companies and government departments to describe their ARMTs in terms that avoid regulation.

1. **Automated *recommendation*-making tools**: The lack of clarity over terminology became apparent to me even when arranging meetings for my research. When I contacted software developers explaining I was researching 'automated *decision*-making' tools, I received responses asking why I had contacted them as none of their tools make *decisions* – they only make *recommendations*. This is important because how tools are defined determines whether they are within scope of regulation. Failure to define adequately these tools, therefore, risks allowing various ARMTs to not be regulated.

   However, as illustrated above in my analysis of tools that automate suspicion, algorithms often make critical interventions (such as triaging or classifying cases) in a decision-making process that can affect individuals' rights even if a human makes the final decision. Accordingly, I advocate moving away from the term 'automated *decision*-making' to 'automated *recommendation*-making' tools so that these tools are captured by regulation.

   I note that some reports seek to resolve this terminological issue by defining tools based on their technological capacities, such as whether they use artificial intelligence, machine learning and so on. However, defining tools by their technological capacities shifts the focus to the underlying technology rather than the effect these tools have on decision-making processes and individuals' rights. Further, technological definitions allow companies and governments to argue that their tools do not fall under certain regulations if they do not meet specific criteria,

[30] Amnesty International, 'Trapped in the matrix: secrecy, stigma and bias in the Met's gangs database' (2018), page 5: https://www.amnesty.org.uk/files/reports/Trapped%20in%20the%20Matrix%20Amnesty%20report.pdf

such as being 'AI-based'. 'ARMT' offers a simpler, more flexible term that avoids constant redefinition as technology advances.

2. **Bias**: A definition that is often lost in translation between software developers and civil society is 'bias'. In the field of computing, 'bias' refers to errors, so bias elimination is error elimination. Whereas to civil society and lawyers, 'bias' pertains to notions of lacking impartiality based on improper considerations such as prejudice against a person or group. It is therefore important that civil society is careful to phrase its advocacy efforts such that technologists understand bias elimination to refer to prejudice elimination as well as simply error elimination.[31]

3. ***'Human in the loop'***: As discussed below (on page 24), UK law on ARMTs states that if a human reviews the final decision (known as having a 'human in the loop'), then a tool is not considered 'solely' automated and can evade certain regulations. However, this safeguard is problematic for two reasons:

    a. Vague definition: The phrase lacks a clear definition in the regulation, meaning even minimal human involvement can be used to bypass regulation. However, it is not clear that even if a human reviews the final decision, how meaningful this is or whether the human: can understand how the machine made the decision, is sufficiently empowered to override the decisions, or has targets that agreeing with the decision will help them achieve.

    b. Misleading imagery: The imagery conjured by the phrase 'human in the loop' is one of a simplistic, unidirectional and autonomous data flow with only a single human required to review the final output. However, it is more accurate to view an algorithmic process as 'a chain of interconnected decision-making stages' stemming each from human intervention – such as in selecting inputs, evaluating variables and reviewing intermediate decisions – rather than just at the end.[32] This chain imagery reveals that an algorithmic process is not simply an objective machine process, but intrinsically a product of human decision-making at every stage. Once this is understood, further questions can then follow including: which humans make which decisions? What qualifies them to do so? How are their decisions (which then become entrenched within future automated processes) recorded and vetted? Is it possible to scrutinise and challenge each of these human interventions?

Psychological effects

Increased reliance on these technologies can have psychological effects on the impacted or potentially impacted individuals, and the institutions and decision-makers using them, as well as society at large.

1. **Psychological impact on individuals**: Migrants and their representatives noted that there is a unique sense of injustice or indignity when an administrative decision with significant effect relies, even partially, on an algorithmic output – as if it is not important enough to merit individualised human consideration.[33]

[31] Interview with Will Tao, founder and immigration lawyer at Heron Law Offices (27 May 2024)
[32] See footnote 26, Jennifer Raso interview
[33] Anonymous interview with migrants' organisation (18 June 2024)

2. **Chilling effect on people in vulnerable situations**: The knowledge that these systems exist is a substantial source of stress for people. This knowledge can affect the way they behave, including their willingness to visit certain neighbourhoods or see friends and family. This is because they fear that such associations may be misinterpreted by opaque algorithms. As a migrant seeking status in a country, they often want to ensure they do nothing that might harm their immigration application. However, the opacity of algorithms and uncertainty around what may impact their applications can have the practical effect of substantially restricting their freedom.[34]

3. **Public trust**: Without transparency, even if algorithms are being internally audited and assessed as fair, public trust is eroded. Lack of public trust leads to, perhaps ultimately unnecessary, taxpayer-funded litigation. Significant time and resources could be saved if public bodies instead proactively disclose information on ARMTs.

4. **Institutional defensiveness**: As public sector institutions increasingly adopt advanced technologies, they become more reliant on them, often leading to defensiveness and a reluctance to acknowledge flaws.[35] A prime example is the Horizon IT system, introduced by Fujitsu in 1999. Despite evidence of its faults, the Post Office defended the system for decades, dismissing complaints and wrongfully prosecuting sub-postmasters for alleged financial discrepancies.[36] This case illustrates the risks of reactive regulation – once faulty technology is embedded, it becomes harder to persuade public bodies to discontinue using it.

5. **Global and international influence**: Both immigration and technology are borderless. Potential clients for technology companies are simply different countries' governments, so it is up to governments already using this technology to lead by example and impose on itself strict ethical standards for its development and deployment. The norms they set will likely shape how societies worldwide regard the ethical boundaries of administrative uses of ARMTs.

Data

There are numerous concerns with the use, storage and collection of data which are at the heart of algorithmic systems, including:

1. **Data privacy**: There is a lack of public information about the collection, storage and use of data by algorithmic systems, especially when data is provided by a commercial entity. This leaves numerous important questions unanswered: how is data being stored? Can it be accessed by other departments for different purposes to those it was collected for? How long can data be stored for? Who else will this data be shared with? Are individuals told about each of the uses of their data?

2. **Permeability of databases**: To make algorithms more technologically powerful, they require more data. As such, there is a drive from public sector organisations and technology companies to increase the sharing of datasets across government. However, when personal data is used for purposes different to those

[34] Interview with the Electronic Privacy Information Centre (16 May 2024)
[35] Interview with Amnesty International Tech (13 May 2024)
[36] BBC news, 'Post Office Horizon scandal: why hundreds were wrongly prosecuted' (2024): https://www.bbc.com/news/business-56718036

for which it was originally collected, ethical and legal concerns arise.[37] For example, if the Home Office was granted access to asylum seekers' medical records, this could potentially influence their immigration decisions based on whether they think an individual would be costly to the NHS. Further, errors also arise when databases are merged, since data taken out of context can be outdated and/or misinterpreted, leading to dangerous and unfair outcomes.[38]

3. **Data integrity**: Relatedly, algorithmic outputs are only as good as the data that is fed into the algorithm: errors in the data leads to errors in the algorithms. There are many examples of algorithms relying on data that humans have failed to verify or keep up to date.[39]

Accountability

In a democracy there is a legitimate expectation by the public that decisions made by public bodies are made fairly and rationally. While humans may make irrational decisions, members of the public are entitled to challenge these decision-making processes and request that the decisions be remade according to the correct process. Algorithms, as they currently exist, radically alter the public's ability to hold the government to account for the following reasons:

1. **Increases the power asymmetry**: The current lack of regulation and transparency around algorithms' roles in public sector decisions means members of the public are unable to scrutinise public decision-making in the same ways they ordinarily can. They do not know what factors (or 'variables') are taken into account, their relative weightings and whether important factors are ignored. Algorithms therefore have the effect of increasing the power differential between members of the public and government, such that the normal democratic processes of holding the government to account are obfuscated and weakened.[40] This increase in the relative power of government is heightened in the immigration context, because these communities are some of the most disempowered in society.

2. **Lack of explainability**: There are forms of machine learning algorithms which are inherently inscrutable and therefore non-explainable. In simple terms, these are algorithms where the machine is fed multiple data points and asked to derive patterns to inform its output. The internal processes used and the various weighted factors remain unknown to all humans, even the designers of the algorithm. Public sector use of such algorithms at any stage of a decision-making process presents obvious problems since their reasoning can never be fully traced. In other words, people will have decisions which are wholly or partly made simply because 'the computer says so'.

3. ***Ex post facto* rationalisation**: Lawyers interviewed, especially in Canada where use of algorithms in its immigration systems is more widely known, observed that clients receiving refusals on immigration applications are given increasingly shorter and more template reasons. The lawyers suspect this is because algorithms are batch assessing applications rather than making individualised assessments. They believe

[37] See footnote 35, Amnesty International Tech interview
[38] See footnote 30, Amnesty International, 'Trapped by automation'
[39] ibid
[40] See footnote 34, Electronic Privacy Information Centre interview; Interview with the Law Commission of Ontario (14 June 2024)

this indicates a trend towards lessening the expectation that individualised decisions are required for every decision at the time the decision is made, and instead generating reasons *ex post facto* ('after the event') if challenged.[41]

4. **Lack of appeal or complaint mechanisms**: There has been a failure so far to build complementary mechanisms of accountability alongside these new processes of decision-making. It is inadequate to rely on traditional methods of judicial review or appeals, and only at the stage a final decision is made, because these do not accommodate this new mode of decision-making. There needs to be appeal or complaint rights at every stage of automated recommendation-making, including by adjudicators who are trained in the risks of these new technologies.

[41] See footnote 31, Will Tao interview

Section 2 – The UK

Case study 1 – GPS tagging

My interest in this area began when representing clients being electronically monitored by a Global Positioning System (GPS) ankle tag by the Home Office. In August 2021, the Home Office introduced a scheme to impose GPS monitoring on anyone who was subject to immigration bail and liable to being deported, unless it would be impractical or a breach of their rights under the European Convention of Human Rights.[42]

Information about how the GPS tagging scheme works is found in the Home Office's 'Immigration Bail' policy, and mentions involvement of an ARMT. It states that Home Office decision-makers have access to an ARMT 'which utilises automated business rules to provide decision recommendations' as to whether a person is suitable to be GPS tagged, or whether someone already fitted with a GPS ankle tag may be moved to a 'not-fitted device' (such as a GPS fingerprint scanner) or to no device at all.[43] It is understood the ARMT began to be involved in decision-making from the week commencing 7 November 2022.[44]

The policy provides limited information on how the ARMT works, only explaining that the data the ARMT uses includes (but is not limited to) if someone is over 18, if they have been identified as 'mentally disordered', how long they have already been tagged for, their previous compliance with immigration bail and the risk of harm posed to the public. It explains this risk of harm score is generated on the basis of the relative harm of various criminal offences. This harm score then recommends how long someone should be tagged for. It makes clear that the tool only generates a recommendation and humans make the ultimate decisions.[45] There are then several pages of redacted information for 'internal Home Office use' only.[46]

Numerous concerns arise from the limited disclosure of this ARMT for my clients. Firstly, I am not aware of the Home Office notifying any individuals if the ARMT has been involved in their case, despite its recommendations contributing to decisions to continue to restrict their liberty. Further, although some factors that the ARMT takes into account are mentioned in the policy, not all are. There is no way for individuals to assess whether any of these additional factors act as a proxy for a protected characteristic under equalities law. The lack of transparency in how the algorithm operates prevents individuals from receiving the necessary information to understand the decision-making process behind their tagging, thus limiting their ability to effectively challenge these decisions.

---

[42] Paragraph 2(5) of Schedule 10 of the Immigration Act 2016
[43] Home Office, 'Immigration Bail Policy v19.0' (2024), page 49: (https://assets.publishing.service.gov.uk/media/65f4260efa1851001a0117bf/Immigration+bail.pdf
[44] Public Law Project, Freedom of Information Act 2000 request (2023): https://gbr01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.whatdotheyknow.com%2Frequest%2F959230%2Fresponse%2F2281853%2Fattach%2F3%2F75193%2520Leslie.pdf%3Fcookie_passthrough%3D1&data=05%7C01%7CFOIRequests%40homeoffice.gov.uk%7Ce4a40acb6ecb47d9450508dbece5773c%7Cf24d93ecb2914192a08af182245945c2%7C0%7C0%7C638364243202848798%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=Kkp1yX2Ui1F8Irg2dJYY%2F8BANoDzheM2%2F6HyJuIRim4%3D&reserved=0
[45] See footnote 43, 'Immigration Bail' policy
[46] ibid, pages 52-53

As the above case study suggests, in the UK there is limited legislation governing public sector use of ARMTs, and none that is specific to the immigration context. I will briefly explain the legislation and policies that do exist, before highlighting their gaps.

DPA and UK GDPR

The key UK law addressing ARMTs is found at sections 49–50 of the Data Protection Act 2018 (DPA) and Article 22 of the UK General Data Protection Regulation (UK GDPR). These provide individuals with the right not to be subject to 'solely' automated decisions that have a 'legal or similarly significant effect' on them, except in a few circumstances. It is worth breaking this down:

- **Solely automated decisions**: This refers to fully automated processes which exclude any human involvement that is more than a token gesture such that the human has discretion to change the outcome.[47] Thus any automated process that involves a 'human in the loop' (even if only minimally) is excluded from regulation.

- **Legal effect**: A decision impacting someone's legal rights or status. For example, a decision that results in the withdrawal of a person's benefits would affect their legal rights.

- **Similarly significant effect**: A decision affecting someone's circumstances, behaviour or choices beyond simply their legal rights. For example, as well as having a legal effect, withdrawal of benefits could also affect a person's ability to rent.[48] Deciding whether a decision has such effects is left to the discretion of the public body itself.

The few decision-making processes which are captured by these regulations (i.e. only those that a public body considers are 'solely' automated AND have 'legal or similarly significant effects') are prohibited, unless they are: (i) necessary for a contract; (ii) based on explicit consent; or (iii) authorised or required by law.[49]

In cases referred to under (i) or (ii), for the public body to use an ARMT, they must implement safeguards such as ensuring individuals can obtain human intervention, express their point of view and contest the decision.[50] Specifically, section 50 DPA details that individuals must be notified in writing 'as soon as reasonably practicable' that the decision was made by an ARMT and have one month to request for the decision be remade by a human. Finally, Article 13(2)(f) UK GDPR requires the public body to provide, 'meaningful information about the logic involved, as well as the significance and envisaged consequences of such processing for the data subject'. The practical effect of these provisions, therefore, is that even the few ARTMs that do fall within the scope of the DPA and UK GDPR can still be used in many circumstances, i.e. when they are authorised by law or where certain safeguards have been complied with.

---

[47] Information Commissioner's Office, 'What does the UK GDPR say about automated decision-making and profiling?' https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/individual-rights/automated-decision-making-and-profiling/what-does-the-uk-gdpr-say-about-automated-decision-making-and-profiling/#:~:text=Article%2022(1)%20of%20the,similarly%20significant%20effect%20on%20individuals

[48] ibid

[49] Such exceptions can only be invoked if the processes do not include 'special category data' (Art 22(4) UK GDPR) i.e. sensitive personal information such as one's political opinions or racial or ethnic origin (Art 9(1) UK GPDR)

[50] Art 22(3) UK GDPR

These safeguards do little to mitigate the risks these tools pose. For example, the notification timeframe of 'as soon as reasonably practicable' is vague, meaning people are potentially left unaware until much later that significant decisions affecting their rights were made by an ARMT. Further, a one-month period to request a review may be insufficient for significant decisions or individuals with limited access to legal advice or resources. Finally, if the public body decides its use of an algorithm falls outside the scope of this legislation (i.e. is not 'solely' automated or does not have a legal or similarly significant effect) then there are no notice or disclosure requirements, leaving the public unaware of the ARMT's role in decisions affecting them.

Other relevant UK legislation

As well as specific legislation on ARMTs, there are other general public law obligations that affect a public body's ability to use these tools, including:

- **Equality Act 2010**: This prohibits indirect or direct discrimination by public bodies based on protected characteristics (e.g. age, race, gender) (ss19, 13), or the development of new policies without 'due regard' for possible discriminatory effects of their policies and practices (s149).

- **Human Rights Act 1998**: This bans public sector use of tools that breach individuals' rights such as their rights to life (Art 2) and private and family life (Art 8).

- **Public law principles**: Public law governs *how* a decision is made by a public body, and so includes obligations on public bodies to make decisions fairly, rationally and accountably. For example, a court cannot decide to imprison someone because they are wearing yellow.

However, given the limited scope of the UK GDPR and DPA 2018, it is difficult for individuals to challenge decisions made wholly or partially through automation based on public law principles, human rights or equalities law because they will mostly be unaware how, or even if, automation was used in making these decisions.

The Algorithmic Transparency Recording Standard

Although not an instrument with legally binding force, the Government has introduced a key policy on public sector use of algorithms: the Algorithmic Transparency Recording Standard (ATRS). The ATRS was introduced in November 2021 following the former Conservative government's recognition that transparency is crucial for building public trust in public sector uses of ARMTs[51] and the need to develop a cross-government AI standard.[52]

I held a meeting on 29 August 2024 with policy officers at the Responsible Technology Adoption Unit (RTA) which is co-responsible for the ATRS (along with the Central Digital Data Offices) and sits under the Department for Science, Innovation and Technology (DSIT). They explained the ATRS' purpose is to provide a template for public bodies drafting transparency records of their algorithmic tools. It has a two-tiered structure, with Tier 1 covering non-technical explanations relating to how and why an algorithmic tool is used, designed for the general public. Tier 2 covers more technical details, such as information

[51] Department for Digital, Culture, Media & Sport, 'National Data Strategy' (2020): https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy
[52] DSIT and others, 'National AI Strategy' (2021): https://www.gov.uk/government/publications/national-ai-strategy

about its technical specifications, the decision-making process and the data used. This tier is aimed at providing information to experts such as civil society, lawyers and the media.

The ATRS was initially launched as a voluntary framework. However, compliance was weak: as of September 2024, only nine records across all government departments had been published and none by the Home Office.[53] Accordingly, in a welcome development in February 2024, the Government committed to making the ATRS mandatory for central government departments throughout 2024.[54] The mandatory requirements of the ATRS are set out in an 'ATRS Mandatory Scope and Exemptions Policy' paper (Mandatory Policy). At the time of writing, the final version has not been published, but the RTA kindly gave me permission to share their latest draft as at 29 August 2024 (see Annex A).

The Mandatory Policy explains that government departments must prepare ATRS records for any 'algorithmic tool' that 'supports or solves specific problems using complex algorithms' and which either directly interacts with the public or meaningfully assists, supplements or fully automates a decision-making process with 'public effect'. It explains that to have 'public effect' a tool must: i) materially affect or have a legal, economic, or similar impact on individuals, organisations or groups; iii) affect procedural or substantive rights; or iv) impact eligibility for, receipt of, or denial of a programme.[55]

It is welcome that this policy makes clear that ATRS records should apply equally to recommendation-making tools (i.e. even where humans make the ultimate decision) as it does to 'solely' automated processes. That said, there are several areas of the Mandatory Policy that could be improved or clarified:

- **No legal enforcement**: While the ATRS is mandatory, it is a policy decision rather than a legally binding requirement, so a government department that fails to comply cannot be sued by the public.

- **Lack of independent oversight**: The ATRS relies heavily on the self-reporting and self-governance of public sector bodies, and there is no clear mandate for independent audits or oversight of ARMTs in use. This could lead to incomplete or biased information being reported.

- **Application only to 'complex' algorithms is unnecessary**: This qualifier has not been defined and its inclusion is not explained. A simple algorithm can still have significant effects on individuals.

- **Lack of update requirements**: While the policy states that an ATRS record should be created when the ARMT is in its beta, pilot or production phase,[56] it does not yet mandate that the ATRS must be completed *prior* to deployment/production or when and how often ATRS records should be updated. This means public bodies can still deploy an ARMT that affects the public before disclosing its use or use ARMTs with ATRS records that are significantly out of date.

[53] DSIT, 'Find out how algorithmic tools are used in public organisations': https://www.gov.uk/algorithmic-transparency-records
[54] DSIT, 'AI White Paper consultation response' (2024): https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response
[55] Annex A (Mandatory Policy), page 56
[56] Ibid, page 57

- **Lack of discrimination, bias or equalities requirements**: The standard does not require public bodies to address discrimination or equalities impacts in their disclosures.[57] This is significant because these concerns are at the heart of ensuring fairness of algorithms and instilling public trust. Without considering or disclosing this information, it is impossible to assess their lawfulness.

  In my meeting with the RTA, I asked why the ATRS does not include equalities or discrimination disclosure requirements. RTA policy officers responded clarifying that it is important to understand ATRS' purpose: it is a communication standard, not a technical or quality assurance standard. Its role is to ensure public bodies collect and share existing information rather than create new resources or test the quality of their algorithmic tools. Governance and quality assurance are deferred to other policies and frameworks.

  I worry this is a practically and conceptually inconsistent clarification. Firstly, the ATRS *does* include governance assurance aspects. For example, the Mandatory Policy specifically builds governance around procurement into the ATRS. It clarifies that if a government department wishes to buy an algorithm from a third-party company, the company must assure the department that it is comfortable with publishing an ATRS record.[58] This provision is placing a governance burden on public bodies to only procure ARMTs from third parties in certain circumstances, not simply ensuring transparency. Secondly, conceptually it is not possible to delimit between transparency and governance. Transparency is a form of governance. Transparency is only meaningful if certain information, such as equalities information, is collected and disclosed. Otherwise, it could be considered 'transparency-washing' while lacking in substantive content.

- **Exemptions**: Aspects of systems used in national security and some other sensitive aspects of a tool may be exempt from the Mandatory Policy, meaning decisions with serious implications could lack transparency. Additionally, there is a lack of clarity if 'dual-use' systems that are used in both national security and immigration contexts will be completely exempted.[59]

Despite these shortcomings, I welcome the fact that ATRS is an iterative process which is seeking to improve upon feedback. The purpose of the Mandatory Policy is to pilot this development, and then improve the ATRS after implementation is trialled and feedback received. This report's findings from Canada and the USA are designed to assist both the iterative development of the ATRS as well as UK lawmakers' and civil society's efforts to drive the development of ethical regulation of ARMTs.

Private Member's Bill

The final UK development to mention is the Public Authority Algorithmic and Automated Decision-Making Systems Bill (hereafter the 'Bill') which is currently passing through Parliament.[60]

[57] ATRS and template for completing a record (see under tab 2.5 'Risks'): https://www.gov.uk/government/collections/algorithmic-transparency-recording-standard-hub
[58] Annex A (Mandatory Policy), page 59
[59] ibid, page 57
[60] 'Public Authority Algorithmic and Automated Decision-Making Systems Bill' (2024): https://bills.parliament.uk/bills/3760

The Bill calls for ATRS records and impact assessments to be completed *prior* to deploying ARMTs, or even giving their development the green light.[61] It requires impact assessments to include bias assessments and to be updated when the system's functionality changes.[62] It compels authorities to provide affected individuals with meaningful and personalised explanations of how decisions are made.[63] It suggests there be processes to monitor outcomes and audit the ARMTs.[64] Additionally, it seeks to mandate training for public sector employees on the system's design, function and risks.[65] Finally, it prohibits the procurement or use of ARMTs that cannot be effectively scrutinised due to technical or contractual barriers.[66]

This is a commendable Bill. It seeks to build accountability into the development of ARMTs at the outset so that transparency is not simply an afterthought or ignored altogether. The introduction of the Bill would be a significant step towards the UK spearheading efforts to ensure the ethical use of ARMTs in immigration systems. However, since the Bill originated in the House of Lords (known as a 'Private Member's Bill'), less parliamentary time will be given to it and it is unlikely to pass into law. Nonetheless, it does indicate UK lawmakers are paying attention to public sector uses of ARMTs.

Summary

Having outlined the risks or harms arising from the use of ARMTs in immigration systems (Section 1) and the limited UK legislative and policy frameworks seeking to address them (Section 2), I will now describe my learnings from Canada (Section 3) and the USA (Section 4). I will summarise each country's legal and policy frameworks, before analysing in-depth case studies of specific ARMTs in their immigration systems. This will lead me to evaluate the utility and shortcomings of their regulation in mitigating the risks arising from these tools, before concluding with lessons for the UK's benefit (Section 5).

[61] ibid, s3(1) and 4(1)
[62] ibid, s3(6)(e) and 3(3)
[63] ibid, s5(1)(b)
[64] ibid, s5(1)(c)(1) and s5(1)(d)
[65] ibid, s6(2)
[66] ibid, s8(1)

Section 3 – Canada

It is useful to begin with Canada, who pioneered the introduction of a mandatory directive on ARMTs across the public sector, the Directive on Automated Decision-Making (DADM).[67] It was the first of its kind in the world.[68]

Directive on Automated Decision-Making

Canada's DADM first took effect in April 2019 and is currently in its third iteration. Its objective is to ensure automated systems 'are deployed in a manner that reduces risks to clients,[69] federal institutions and Canadian society, and leads to more efficient, accurate, consistent and interpretable decisions made pursuant to Canadian law'.[70]

The DADM applies to 'any system […] used to make an administrative decision or a related assessment about a client'.[71] Such a system is defined as 'any technology that either assists or replaces the judgment of human decision-makers'.[72] Thus, the umbrella term of automated system here includes ARMTs where humans still make the final decision. The TBS 'Guide on the Scope of the DADM' (Guide) also makes clear that DADM, unlike the ATRS' mandatory scope, applies to simple 'deterministic rules-based systems' not just to complex 'advanced AI systems'.[73]

Oversight of the DADM is conducted by the Treasury Board of Canada Secretariat (TBS) (the equivalent of the UK Treasury). The TBS has powers of enforcement since non-compliance can be met with withdrawal or reallocation of departmental funding.[74] That said, I was informed by TBS that it would only use these powers in cases of serious and malicious non-compliance because such enforcement actions could ultimately affect the department's ability to serve the public. To date, TBS has never considered it needed to invoke these powers in the context of the DADM.[75] Thus, while the DADM is mandatory, its enforcement is left to TBS (rather than the public) who may have extenuating reasons for being reluctant to enforce compliance.

Algorithmic Impact Assessments

Crucially, the DADM requires all public sector bodies to complete and publish an Algorithmic Impact Assessment (AIA) '*prior*' to the production of any ARMT, which must be reviewed regularly including when the ARMT's function or scope changes.[76] The AIA is a questionnaire aimed at determining the impact of a system. There are four Impact

[67] TBS, 'DADM' (2023): https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592
[68] OCED.AI, 'Canada's Directive of Automated Decision-Making' (2024): https://oecd.ai/en/dashboards/policy-initiatives/http:%2F%2Faipo.oecd.org%2F2021-data-policyInitiatives-24240
[69] 'Client' is defined as 'any person or business that receives a service from the Canadian government' as per the DADM's parent policy, 'Policy on Service and Digital', Appendix A: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32603
[70] See footnote 67, DADM, Art 4
[71] ibid, Art 5
[72] ibid, Appendix A
[73] TBS, 'Guide' (2024): https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/guide-scope-directive-automated-decision-making.html
[74] TBS, 'Framework for the Management of Compliance' (2010), Appendix C: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=17151
[75] Interview with Benoit Deshaies, Director for Responsible Data and AI at TBS who oversees the development of the DADM (6 June 2024)
[76] See footnote 67, DADM, Art 6.1.3 and 6.1.1; see footnote 73, Guide

Levels, ranging from Level I (little, reversible and brief impact) to Level IV (very high impact that is irreversible and perpetual). Impact is evaluated according to the following criteria:

> *the rights of individuals or communities; the equality, dignity, privacy and autonomy of individuals; the health or well-being of individuals or communities; the economic interests of individuals, entities, or communities; the ongoing sustainability of an ecosystem*.[77]

Depending on the Impact Level determined in the AIA, the DADM prescribes different requirements to be complied with, as set out in Appendix C of the DADM[78] and summarised here:

LEVEL	REQUIREMENTS
All	• Publish meaningful explanations for common decisions. This includes the ARMT's role in a decision-making process, its inputs and outputs, key factors behind a decision and recourse options where appropriate.
Level II and above	• Prepare and publish at least one peer review of their AIA e.g. by another government department, or a summary thereof. As of September 2024, only three have been published.[79] • Prepare a Gender-based Analysis Plus (GBA+), a tool used across Canadian Government to assess policies' impact on gender, race, etc. As of September 2024, only one has been published with an AIA.[80] • Provide notice through all service channels that the decision will be made in whole or partially by an ARMT. • Provide meaningful explanations to clients for decisions resulting in the denial of benefits or that involve regulatory actions. • Provide employee training on system design and functionality.
Level III and above	• Publish notices on relevant websites of how the ARMT works, its support in decisions, results of reviews or audits and a description or link to anonymised training data (if publicly available). • A human must make the final decision. • Use of the ARMT must be approved by the department's Deputy Head or, if Level IV, TBS itself.

[77] See footnote 67, DADM, Appendix B
[78] ibid, Art 6.1.2
[79] Open Government Portal, 'AIAs and peer reviews' (September 2024) https://search.open.canada.ca/opendata/?search_text=algorithmic+impact+assessment+peer+review&sort=metadata_modified+desc&page=1
[80] Open Government Portal, 'AIAs and GBA+' (September 2024) https://search.open.canada.ca/opendata/?search_text=algorithmic+impact+assessment+gender+based&sort=metadata_modified+desc&page=1

Immigration, Refugees and Citizenship Canada (IRCC) AIAs

As of September 2024, there were 22 AIAs published online across all Canadian government departments.[81] Promisingly, the highest number of these (10) published by one department is by Immigration, Refugees and Citizenship Canada (IRCC) (the Canadian equivalent of the UK Home Office).[82]

To understand better IRCC's use of ARMTs and the impact of the DADM on them, I met with a senior official at IRCC (Senior Official) on 6 June 2024. The Senior Official works within a team of policy advisors and data scientists focusing on responsible uses of emerging technology at IRCC, which arose in 2018. Interestingly, he explained that the vast majority of tools are developed in-house rather than contracted out to third parties, although this is not consistent across all Canadian ministries. The Senior Official further explained that IRCC currently has bans on two types of tools for being too risky at present:

1. **Automated refusals**: ARMTs that automate or recommend refusals based on predictive models. The Senior Official explained this is because any refusal must be based on an 'objective, verifiable fact'.[83] Thus, IRCC only fully automates positive decisions. This ban is set out in IRCC policy documents including its *Policy Playbook on Automated Support for Decision-Making*.[84]

2. '**Black box**': IRCC will not use any 'black box' algorithms in an administrative decision-making process. This is because IRCC wants all systems' decision-making logic to be knowable and traceable to IRCC internally, so that individuals can have a right to a meaningful explanation of a decision on their file. It is unclear if this ban applies to all stages of automated processes, including triaging decisions and other intermediary processes where humans make the final decision; clarity should be given by IRCC in this regard.

Both bans are laudable (on the assumption they apply to all stages of automated processes) especially since, as the Senior Official indicated, there are numerous applications of ARMTs that can achieve efficiencies without presenting the risks to administrative fairness that these types of tools would.

I also discussed with the Senior Official IRCC's use of the Integrity Trend Analysis Tool (ITAT); a predictive risk assessment tool based on historical data. ITAT, in my opinion, is the most controversial ARMT used by IRCC (according to publicly available information) yet it seems to have had relatively little public attention so far.

[81] Open Government Portal, 'AIAs' (September 2024) https://search.open.canada.ca/opendata/?search_text=algorithmic+impact+assessment&sort=metadata_created+desc&page=1
[82] ibid, 'AIAs and IRCC' (September 2024) https://search.open.canada.ca/opendata/?search_text=algorithmic+impact+assessment+&sort=metadata_created+desc&page=1&owner_org=cic
[83] Interview with a senior official at IRCC (6 June 2024)
[84] IRCC, 'Policy Playbook on Automated Support for Decision-Making' (2021): https://meurrensonimmigration.com/wp-content/uploads/2021/11/A202119597-AI-Playbook.pdf

Case study 2 – Integrity Trend Analysis Tool

Purpose: ITAT is a tool designed to 'streamline verification activities by simplifying the process that risk assessment units would normally require to manually review potential fraud risks in temporary resident applications.'[85]

Scope: The AIA published in December 2022 explains that ITAT will be implemented in a 'phased approach starting with temporary resident programs: students, electronic travel authorization, visitors and workers'.[86] This means anyone looking to study, work, visit family and friends, travel or temporarily live in Canada will have their applications assessed by ITAT. The wording in the AIA suggests that ITAT may in the future be (or is already) used on more permanent applications.

How it works: ITAT analyses 'large volumes of [historical] data and extracts various risk and fraud patterns within applications containing adverse characteristics, such as inadmissibility findings (e.g. criminality and misrepresentation) and other administrative or enforcement actions'.[87] As well as historical data, ITAT is also informed by officer experience. Patterns can be as specific as a phone number associated with fraud.[88] In sum, therefore, ITAT identifies risk patterns from historical data and codified officer experience, and then matches new applications that are consistent with those risk patterns.

It is unclear what constitute the 'various risk […] patterns' mentioned above, aside from fraud, that ITAT is searching for. It is possible they include risk of applying for asylum or committing crimes based on what people who look or seem like them have done previously, but not necessarily because the individual him/herself has done either.

Who uses the tool: The fraud patterns and matching applications that ITAT has identified are provided to the Risk Assessment Unit (RAU).[89] The RAU is a specialist team within IRCC whose role it is to identify fraudulent applications, separate to the frontline officers (processing officers) who ultimately decide whether to approve or reject an application.[90] I understand that while both RAU and processing officers can decide to run additional verification checks (e.g. checking proof of funds from a bank), only RAU officers benefit from the additional information ITAT produces when deciding to run these additional checks.

ITAT's AIA boasts that processing officers only see RAU's verification results (and do not know if these have resulted from ITAT flags or manual decisions)[91] as an attempt to prevent automation bias. The logic is that the ultimate decision-makers are not influenced by the algorithm's findings when making their final decision, as they are only privy to the final decision of the RAU team. However, this logic is worth scrutinising. By pushing the decision back one level to the RAU team, it does not prevent the RAU team themselves experiencing automation bias and tending to scrutinise – and in turn identify as risky more frequently – the ITAT-flagged

[85] Open Government Portal, 'AIA – ITAT' (2022): https://open.canada.ca/data/en/dataset/240f1dbc-a3b5-46b1-9b5f-d0d3cbec9378
[86] IRCC, 'AIA – ITAT' (2022), page 2 https://open.canada.ca/data/en/info/240f1dbc-a3b5-46b1-9b5f-d0d3cbec9378/resource/62473d4f-d38c-4b35-8adf-784da82bf516
[87] ibid, page 1
[88] See footnote 21, Mario Bellissimo interview
[89] See footnote 86, ITAT AIA, pages 2 and 5
[90] See footnote 83, Senior Official interview
[91] See footnote 86, ITAT AIA, page 1

applications more often than others. Indeed, after receiving an ITAT flag, if a RAU officer runs verification checks and does identify the application as risky or fraudulent, it is highly unlikely the processing officer will then approve that application. It is therefore unclear how the RAU team only being privy to ITAT results removes automation bias from the process.

Type of technology: Something not apparent from the AIA is the type of technology used by the algorithm since this is not a question included in the AIA. The Senior Official clarified that ITAT uses 'advanced analytics' since it looks for patterns in past cases for verifiable instances of fraud but does not use '*black box*' processing since, as mentioned above, IRCC has a current ban on using ARMTs that are not completely knowable and traceable to internal IRCC staff.[92]

Possible future applications: In a follow-up email received on 10 June 2024, the Senior Official explained that, 'It is possible that in the future IRCC could use ITAT as part of risk-based triaging of incoming applications, but at the outset it was felt that the tool should first be proven, and in order to proceed cautiously it was approved for use only by Risk Assessment Units. With a smaller number of officers using the tool it is easier to train them and monitor use'.[93]

Thus, processing officers may soon have access to ITAT results themselves, rather than going through the buffer of RAU offices. Instead, ITAT will triage most, if not all, visitor visa applications to Canada based on risk by matching risk patterns (arising from historical data and codified officer experience) with new applications that display characteristics in some way consistent with these patterns.

Concerns: Risk prediction tools based on historical and codified officer experience data is a controversial application of automation by IRCC for numerous reasons:

- High risk of multiple forms of bias and discrimination: As discussed on pages 14-16, historical, data, developer, proxy, feedback loop, automation, confirmation and quantitative bias plague these forms of predictive tools.

- Failure to monitor for bias according to protected characteristics: In the follow-up email received from the Senior Official, they confirmed that 'there are currently no IRCC application forms that give clients the option to identify their race or ethnicity', though they noted there are plans for pilot projects whereby 'volunteering clients' can 'self-identify their race/ethnicity'.[94] It is presumed this is the same for sexual orientation, religion, disability status and so on. It is therefore impossible for IRCC to assess, or attempt to rectify, if ITAT is displaying the forms of bias mentioned above and disproportionately flagging people with certain protected characteristics, resulting in a disproportionate number of these people having their applications ultimately refused.

- Lack of transparency: Neither the GBA+ assessment nor the peer review of ITAT have been released. The public and affected individuals have no visibility into the issues that were identified with ITAT during these processes or whether those issues have since been resolved.

[92] See footnote 83, Senior Official interview
[93] Follow-up email, Senior Official, 10 June 2024
[94] While country of citizenship is traceable, this is importantly different from race or ethnicity and does not constitute or necessarily indicate any protected characteristics.

Further, the Senior Official confirmed that neither a lawyer nor a client would be told that ITAT has flagged their specific application.[95] I understand IRCC's reasoning to be that, since ITAT itself does not 'automate [...] [or] recommend final decisions'[96] and its use is not disclosed to the processing officers who make the final decisions, IRCC is not required to disclose its existence to affected individuals either. That is, since ITAT does not *itself* result in the denial of a visa – it simply flags applications as potentially risky for further review by RAU officers – disclosure of ITAT flags in specific applications are not given to clients, even when their applications are later refused based on those flagged risks.

However, Appendix C to DADM makes clear that for Impact Level II ARMTs (such as ITAT), government departments must provide meaningful explanations for decisions that result in the denial of a benefit or service (e.g. a visa) to individuals, including the role of the ARMT in the decision-making process. Appendix C requirements cover tools that 'assist' as well as 'replace' the judgement of human decision-makers and/or which generate an 'assessment, score or classification',[97] such as ITAT. Thus, it is not apparent why IRCC does not inform individuals of ITAT's involvement in their cases under DADM.

- Presumption of innocence reversed (or in public law terminology, jeopardising decision-makers' ability to apply an open-mind): While the final decision may be made by a human, the suspicion over certain applications automated by ITAT triggers additional scrutiny over those applications. It also puts the burden on the reviewing officers to decide why those applications do *not* present a risk, which other applications do not have. This is concerning when the risk flags are generated based on historical applications that *look* and *seem* like the individual's but are not necessarily based on actions of the individual themself. Use of ARMTs like ITAT are, in effect, a way of making automated stereotyping a part of official government policy.

- AIA Impact Level: IRCC has self-assessed ITAT as only Impact Level II. IRCC answers 'no' to whether the stakes of the decision are very high and states it will have 'little to no impact' on individuals' rights and freedoms since ITAT itself 'does not make or recommend decisions on applications'. The AIA also answers 'no' to several questions including whether a Privacy Impact Assessment (PIA) has been undertaken or if IRCC has developed a process to document how data quality issues were resolved during the design process.[98]

  It is unclear how IRCC can accurately self-assess ITAT's impact if it has not conducted a PIA and does not address the several risks of the tool identified in this report. Significantly more information must be provided to justify IRCC's opinions than short 'no' or 'little to no impact' answers.[99] As Mr Bellissimo puts it, 'The rationale for finding that ITAT is not high risk is based

[95] See footnote 83, Senior Official interview
[96] See footnote 86, 'ITAT AIA', page 2
[97] See footnote 73, 'Guide'
[98] See footnote 86, 'ITAT AIA', pages 4-7 and 9
[99] Mario Bellissimo, 'Techno Centric-Decision-Making in Canadian Immigration Law and Practice: Artificial Intelligence Deployment - How Can the Existing Canadian Immigration Legal Eco-System and Immigration Advocates Respond to the Use of AI Technologies?' (2024), pages 14-15

on assumptions that are not plain to understand'. By leaving the assessment of impact to the government departments themselves, there is a lack of independence in assessing the ARMTs' impact. This means tools can be self-assessed as lower impact than they potentially are and therefore be subject to the less rigorous requirements of Appendix C to DADM. The appearance of the lack of independence also erodes public trust in the process.

ITAT is an example of an algorithm that urgently requires heavy public scrutiny given the inherent risks it presents. That said, it has only been possible to apply this level of scrutiny to ITAT because IRCC has published an AIA on the tool and been willing to be interviewed as part of this research. Further, civil society colleagues in Canada have recently been invited to review drafts of a 'Peer Review Guide' and 'AIA Guide', and expect to be consulted on the fourth review of the DADM starting late October 2024. So, while there remains significant room for improvement in the regulation of this tool (including questions over whether it should be permitted to be used at all), IRCC is setting an example in its willingness to subject its tools to scrutiny. It is hoped recommendations from this report and others on the tool will feed into future iterations of the DADM, and further, that the UK Home Office will follow Canada's example in being more transparent about the ARMTs they use and engaging with civil society on them.

Before summarising the recommendations for improvement of Canadian regulation, it is worth briefly considering the impact of another type of ARMT that the IRCC most commonly uses: triaging tools. I will examine a triaging tool that IRCC is using in the high stakes context of refugee applications.

Case study 3 – Privately sponsored refugee applications[100]

Summary: IRCC uses an ARMT to streamline the processing of privately sponsored refugee (PSR) applications. The PSR scheme allows for resettlement of refugees in Canada who have the financial and emotional support of private sponsors, rather than relying solely on government assistance. A PSR application has two parts – the sponsorship and refugee parts.

An AIA for the tool used in PSR applications was completed in December 2022. It explains that the ARMT first scans the sponsorship part to identify routine cases that can be automatically approved. All sponsorship parts that are not automatically approved by the system then go through a manual officer review process.

For the refugee part, the system triages applications to officers at migration offices overseas based on office capabilities and officer expertise. Decisions on the eligibility and admissibility of the refugee part of the application are not automated on any application and are all made by an officer.[101]

Concerns:

- Impact of triaging overlooked: The tool is used to triage assessments of the refugee part of applications based on officer expertise, but it is unclear on what basis (risk, complexity or another metric) they are being triaged. This can result in certain types of applications always being reviewed by more senior officers. Given the lack of equalities monitoring, it is possible these

[100] IRCC, 'AIA – automated tools to help process privately sponsored refugee applications' (2022) https://ouvert.canada.ca/data/info/ad4be3b8-ac97-4dc1-8dd8-231239d018f2/resource/648c460b-e69b-4cdb-8aa9-f7264bdbd4fd

[101] ibid, page 1

applications will more often involve people with certain protected characteristics. This is concerning since senior officers will likely be aware that cases have been allocated to them for a reason and so (due to automation bias) may apply to them a higher level of scrutiny and in turn reject them at a higher rate. Even if not rejected, these applications will experience delay which can cause anxiety and additional costs. Despite these concerns, IRCC has self-assessed this tool as only Impact Level II and states that the decisions will have 'little to no impact' on individuals' rights or freedoms.[102]

- Lack of transparency: IRCC has not recognised the triaging function of the tool as being subject to the AIA and, consequently, has not ensured its compliance with Appendix C requirements. For instance, when asked to describe the impacts of decisions made by the tool, the focus is solely on the positive eligibility assessment within the sponsorship process, neglecting the impacts of the triaging function.[103] As a result, it appears that the requirements of Appendix C – such as providing meaningful explanations to individuals regarding the involvement of ARMTs in decision-making that affects them – are only applied to the function that automates approvals. This interpretation overlooks the fact that, according to the DADM, automation includes both functions that replace and assist human decision-makers.

- Automation bias: The AIA explains that the processing officers will be informed of the tool's results, but not why or how the tool reached a decision.[104] A more experienced officer will know they have been allocated a more complex or riskier case but not know why. Given human tendency to over-rely on ARMT recommendations, the officers' inability to know why a decision has been reached is likely to lead them to try harder to find a reason. Instead of mitigating the risk of automation bias, IRCC's approach could be exacerbating it.

Recommendations for improvement

While it is welcomed that IRCC is spearheading the Canadian Government's transparency efforts on its use of ARMTs, it also reveals that regulation over ARMTs needs improvement – especially in the high stakes and high-risk context of immigration.

In my interview with Benoit Deshaies, the Director for Responsible Data and AI at TBS who oversees the development of the DADM, Mr Deshaies noted several areas of the DADM that his team are currently looking to improve. These include that the DADM currently does not assess the risk of discrimination based on human rights grounds, it only captures new systems[105] and it includes limited requirement for public consultation upon the release of new systems. Mr Deshaies also suggested that it could be beneficial if an external body had formal oversight functions such as TBS itself, an independent committee or a privacy commissioner, since government departments' compliance with the DADM is currently self-policed. I agree that each of these areas require further consideration. Additionally, I suggest there should be:

[102] ibid, page 4
[103] ibid, pages 4-5
[104] ibid, page 5
[105] 'Developed, procured or modified before 1 April 2020' (See footnote 67, DADM, Art 1)

1. **A ban on risk-flagging ARMTs**: Predictive risk-flagging tools like ITAT should be banned since they are problematic from a human rights' perspective due to the high likelihood of discrimination and bias on marginalised populations, as discussed above on page 17. Automation of stereotypes contravenes basic human rights, public law and equality principles.

2. **Clarity over disclosure of ARMTs**: Although DADM and its ancillary Guide defines the tools in scope as both those that replace *or* assist decision-making, the requirements of the DADM are being interpreted in practice to only apply fully to the aspects of tools that *replace* decision-making. As such, individualised notice or explanations of automated triaging, categorising and/or recommending tools are not being provided to affected individuals. Training to all departments should be given, and mandatory questions in the AIA and peer review templates should be designed to address this area of persistent non-compliance. The Guide should also be updated to make this point explicit.

3. **IRCC red lines should be mandated across government**: IRCC bans on the use of inscrutable machine learning tools and those that automate refusals should be codified in a mandatory policy or law and made applicable across government. These bans have a clear logic: the former to ensure all administrative decisions are traceable, the latter to ensure all refusals have the benefit of individualised assessments. IRCC should also clarify that these bans apply to all stages of automated processes, including triaging decisions and other intermediary processes where humans make the final decision.

4. **AIA question on type of technology**: The AIA questionnaire does not ask about the type of technology used in an ARMT, which is critical to understanding its risks and potential public impact. The AIA should add a question asking whether machine learning has been used, and in what capacity. While IRCC has an internal informal ban on inscrutable or unexplainable algorithms, this is not legislated or government-wide. It is therefore necessary to understand a tool's impact on public law.

5. **Collection of race and other data disaggregated according to protected characteristics**: It is only possible to ensure an ARMT is not perpetuating discriminatory practices if its outputs are monitored according to disaggregated data based on protected characteristics. To do so, protected characteristics data needs to be recorded in the first place. Use of technologies before such data is collected is concerning.

6. **Improvement of peer review process**: Peer reviews should be mandatory for all Impact Levels, otherwise government departments can self-assess an ARMT as the lowest Impact Level and avoid peer review of this very assessment. Further, peer reviews should only be conducted by specialist bodies or individuals external to and independent of government. Finally, there should be requirements for the government department to adapt the ARMT as a result of the peer-review findings, otherwise peer reviews can act simply as a check-box exercise.

7. **Mandate disclosure of GBA+ assessments**: Since some of civil society's chief concerns with ARMTs is their disproportionate impact on marginalised people, publication of the GBA+ assessments would improve public confidence and accountability.

Notwithstanding these recommendations for improvement, it is promising that the DADM is an iterative process and both IRCC and TBS welcome engagement with researchers on the subject. This open dialogue creates public confidence and pooling of productive ideas as governments try to grapple with this fast-moving, complex balance of reaping the benefits of new technologies without perpetuating harms.

Section 4 – The USA

Canada has been a useful example of a country seeking to regulate public sector use of ARMTs such that government departments proactively disclose information about them. In contrast, while the USA has also introduced some relevant policies, interviewees suggested they were piecemeal and have little effect on the ground. Most uses of ARMTs in the US immigration system have only been discovered through litigation and investigative journalism.

I will begin by summarising the policies interviewees suggested were most relevant, before analysing two distinct case studies of ARMTs that represent different ends of the spectrum in terms of public perception and ethical considerations.

Piecemeal policy

In October 2023, the Biden Administration released an 'Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence' [106] (EO) to guide responsible AI development and deployment by public bodies. In March 2024, to help federal departments implement the EO, the Office of Management and Budget (OMB) released a Memorandum 'Advancing Governance, Innovation and Risk Management for Agency Use of AI' (the Memo) which outlines the specific safeguards required for use of ARMTs.[107]

The Memo mandates federal agencies to cease using AI decision-making (defined as any artificial system performing a task without significant human oversight, or that can learn from experience and improve performance when exposed to datasets[108]) if the agency is unable to adequately mitigate any associated risk of unlawful discrimination.[109] It also requires agencies to maintain appropriate documentation to accompany AI decision-making and to publicly disclose it to the extent consistent with applicable law and federal policy.[110]

Such risk management practices are a positive development. However, they have significant limitations. Firstly, as in the UK and Canada, these policies are not legislative. A change in administration could dissolve them. They also do not give rise to actionable rights by the public and so rely purely on implementation efforts by the OMB and each agencies' AI oversight officers, particularly the Chief AI Officers (CIAOs). Further, an EO, unlike federal legislation, does not automatically have funding to implement it. Thus, while the OMB and CIAOs have some oversight powers, without additional specific budgets they cannot be reasonably expected to oversee implementation in each federal body.[111]

---

[106] The White House, 'EO' (2023): https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[107] OMB, 'Memo' (2023) https://emea01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.whitehouse.gov%2Fwp-content%2Fuploads%2F2024%2F03%2FM-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf&data=05%7C02%7C%7C3e43454947304dbc5a0b08dc729ff758%7C84df9e7fe9f640afb435aaaaaaaaaaaa%7C1%7C0%7C638511278581566520%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C0%7C%7C%7C&sdata=5CBLW01BoDyxGNsSOcb%2FJniGv0cN2jZWCIHIx16Ib7E%3D&reserved=0

[108] ibid, para 6

[109] ibid, para 5(c)(v)(A)(4)

[110] ibid

[111] Anonymous civil society interview (13 May 2024)

The Memo also has significant carve-outs in two respects. Firstly, it does not cover AI when used as a component in a national security system.[112] Interviewees believe that the Government will interpret this exemption as covering the entirety of 'dual-use' systems including those used by the immigration department simply because they also have a national security aspect.[113]

Secondly, and more sweepingly, the Memo contains a vaguely worded wavier whereby an agency may waive the requirements to manage risks of AI if it determines that fulfilling the requirement(s) would increase risks to safety or rights overall or would create an unacceptable impediment to critical agency operations.[114] This is a broad discretionary power permitting all agencies to waive their obligations under the Memo. For this reason, interviewees told me the Memo and EO will likely make minimal practical difference on the ground to public sector use of ARMTs. That said, agencies have until 1 December 2024 to begin implementation of the EO and Memo, so we will not have a clear picture as to whether these documents affect agencies' AI use until spring or summer 2025.

Despite the lack of regulation over public sector use of ARMTs, the USA has been an early adopter of them in its immigration system. I will focus on two distinct uses: the Investigative Case Management system (ICM) and GeoMatch. ICM, developed by Palantir Technologies (Palantir), is a tool employed by Homeland Security Investigations (HSI) to gather and analyse data from multiple sources for law enforcement purposes, and has sparked controversy about its alleged use in immigration enforcement. On the other hand, GeoMatch, developed by the Immigration Policy Lab (IPL) at Stanford University seeks to optimise refugee resettlement by using predictive algorithms to match refugees to locations where they are most likely to find employment. This tool demonstrates the potential of ARMTs to improve migrants' lives but re-emphasises the need to proactively regulate ARMTs to safeguard against possible harms.

Case study 4 – Investigative Case Management

A famous ARMT used in the US immigration system is ICM developed by Palantir. I met with the Global Director of Privacy and Civil Liberties Engineering at Palantir, Courtney Bowman, on 13 May 2024. I am grateful to Palantir for engaging on the record with civil society on its work and its surrounding media attention.

ICM is a case management software used primarily by HSI agents, a division of Immigration and Customs Enforcement (ICE).[115] ICM has various functions which include an ability to pool data from an array of federal and private law enforcement entities. This function reportedly provides HSI agents access to a wide range of personal information including schooling, personal relationships, biometric

[112] See footnote 107, Memo, para 2(c)
[113] See footnote 34, Electronic Privacy Information Centre interview
[114] See footnote 107, Memo, para 5(c)(iii)
[115] Homeland Security, 'Privacy Impact Assessment for the ICM', (2021): https://emea01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.dhs.gov%2Fsites%2Fdefault%2Ffiles%2Fpublications%2Fprivacy-pia-ice045a-icm-august2021.pdf&data=05%7C02%7C%7Cabfb1e34fca34f88cddb08dc73a03216%7C84df9e7fe9f640afb435aaaaaaaaaaaa%7C1%7C0%7C638512379080132538%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C0%7C%7C%7C&sdata=biW1AyYYmCczlv%2Fh1uRuWHWFnalO3lxSyZd1glmx820%3D&reserved=0; Palantir, 'HSI renews partnership with Palantir for case management software' (2022): https://www.palantir.com/newsroom/press-releases/homeland-security-investigations-renews-partnership-with-palantir/

traits, employment and phone records.[116] Another 'of the tool's capabilities', as explained by Mr Bowman, is 'to do additional analytics'. He clarified this means 'analytics in the broadest sense', i.e. 'finding associated pieces of information and putting them together into a case profile'.[117]

ICM's use has sparked significant controversy. The Intercept reported that ICM helped immigration officials in identifying targets for investigation and then administering cases against them.[118] Specifically, ICM was allegedly used in an operation involving the deportation of relatives of migrant children apprehended at the border.[119] When sufficient information was gathered during an investigation on the family of a child, ICM was allegedly used to send a 'collateral case' to the relevant team for action.[120] In some cases, this resulted in charges being brought against the child's family members. ICE data provided to The Intercept suggested this operation resulted in 443 arrests, 35 of which were criminal arrests.[121]

I asked Mr Bowman about ICM's alleged involvement in this operation. Mr Bowman explained that a confusion often arises because HSI is part of ICE, but that HSI primarily focuses on multinational criminal investigations rather than immigration processing. That said, Mr Bowman accepted that there was 'a kernel of truth in some of the reporting' since 'ICM was probably part of the system that was used to conduct the criminal investigative part of that operation'. Mr Bowman suggested that, as in this case, 'where things get messy and problematic' is when the US Government is politically motivated to use ICM for different purposes than expected, such as 'interior enforcement' or administrative immigration arrests.

Given this risk, Mr Bowman noted that Palantir was 'constantly evaluating' the scope of its work to 'minimise the risk' of ICM being used by ICE for these purposes. This includes reviewing requests to integrate new data sources and assessing whether they present 'outsized risk' that the platform could be used in an 'untoward way if the agency veered in a specific direction'.[122] In essence, my understanding is that Palantir continually assesses the risk that ICM could be used for risky or unethical purposes, such as politically driven administrative arrests by ICE.

Concerns:

1. Lack of regulatory safeguards: This conversation indicates that it is possible that ICM *can* be used in immigration enforcement contexts if the US Government decides to use it for that function, and there are no technical or legal safeguards preventing such operational choices. This

[116] The Intercept, 'Palantir Provides the Engine for Donald Trump's Deportation Machine' (2017): https://theintercept.com/2017/03/02/palantir-provides-the-engine-for-donald-trumps-deportation-machine/; MIT Technology Review, 'Amazon is the invisible backbone of ICE's immigration crackdown' (2018): https://www.technologyreview.com/2018/10/22/139639/amazon-is-the-invisible-backbone-behind-ices-immigration-crackdown/

[117] Interview with Courtney Bowman, Global Director of Privacy and Civil Liberties Engineering at Palantir (13 May 2024)

[118] See footnote 116, The Intercept (2017)

[119] The Intercept, 'Peter Thiel's Palantir Was Used to Bust Relatives of Migrant Children, New Documents Show' (2019): https://theintercept.com/2019/05/02/peter-thiels-palantir-was-used-to-bust-hundreds-of-relatives-of-migrant-children-new-documents-show/

[120] ICE, 'Unaccompanied alien children human smuggling disruptive initiative: concept of operations' (2017), p4: https://www.documentcloud.org/documents/5980596-Smuggling-Initiative-ConOP.html

[121] See footnote 119, The Intercept (2019)

[122] See footnote 117, Interview with Courtney Bowman

highlights the problem with the absence of regulation of ARMTs: the ethical boundaries for using new technologies are determined internally by the Government or technology companies, with limited legal guidelines in place.

2. Terminology: While Mr Bowman explained that, to his knowledge, he does not think that ICM has 'any automated or AI-orientated capabilities', he did note that it uses some kind of 'analytics' to find associations between pieces of information and puts them together in a case profile. It became apparent that the lack of clarity over definitions was an issue. While I understand ICM to be an ARMT, Palantir does not. And definitions have consequences for whether technologies are defined as falling within the scope of regulation, or not.

3. Associative algorithms: As mentioned on page 18 above, associative algorithms raise various ethical and legal concerns for their potential role in 'automating suspicion'.

4. Transparency: As far as I am aware, there is minimal public information about how, or on what basis, ICM determines how different pieces of information are associated. The lack of transparency around this means the public and affected individuals are unable to scrutinise whether these processes are fair.

During my trip to the USA, I found that these concerns arising from the absence of regulated ethical boundaries of ARMTs used in controversial fields, such as immigration enforcement, persist even when ARMTs are designed to improve refugee outcomes. For example, there is a suite of ARMTs designed to recommend optimal resettlement locations for refugees based on metrics like employability and proximity to support services. These tools follow research showing that where a refugee is first resettled plays a critical role in their success integrating.[123] They represent the positive potential of ARMTs in immigration and, as such, have received comparatively less scrutiny. However, as with any technology affecting human lives, I learnt that realising their benefits still requires clearer ethical guidelines. I will illustrate the insights I gleaned through the case study of one such tool: GeoMatch.

Case study 5 – GeoMatch

What is GeoMatch? GeoMatch is an AI-driven algorithmic tool designed to help governments and non-profits match refugees to locations in which they are most likely to thrive.[124] The algorithm can be adapted to promote any integration priority of the government's or non-profit's choosing, for example, employment or language acquisition.

Where is GeoMatch in operation? Versions of GeoMatch are being trialled in the USA and Switzerland, with applications in Canada and the Netherlands being developed. I am focusing on the US application where GeoMatch is partnering with the resettlement agency, Global Refuge, to make recommendations about which locations might improve a refugee's chances of finding employment within 90 days.

[123] Kirk Bansak et al, 'Improving refugee integration through data-drive algorithmic assignment' (2018) https://www.science.org/doi/10.1126/science.aao4408
[124] Immigration Policy Lab, 'GeoMatch': https://immigrationlab.org/geomatch/

Who designs GeoMatch? GeoMatch is a tool designed by the IPL, a research group at Stanford University, involving experts in academia, data science, and social policy. During a visit to Stanford in May 2024, I met with some of the team to discuss the algorithm's applications. I am grateful to them for giving me so much of their time to discuss how their tool works, including providing me with a live demonstration. It is encouraging that software companies are keen to seek insights and collaborate with researchers.

The problem that GeoMatch is designed to address: When an individual has been approved for refugee resettlement by the US Government, they are assigned to one of 10 national resettlement agencies which then selects an affiliate location for resettlement. If the refugee has no ties to a particular area, the agency decides where they should be placed considering factors like family size, language ability and the location's capacity. While refugee preferences are not formally collected, there is nothing to prevent refugees expressing their location preferences to resettlement officers (regardless of whether the resettlement officer chooses to take these into account). Refugees can move shortly after being relocated but they risk losing the government-backed support they receive from the resettlement agency. Traditionally, these initial placement decisions were made by resettlement agency officers using fragmented data across various spreadsheets and anecdotal evidence, which was time-consuming and inconsistent. GeoMatch seeks to streamline this process by making data-driven recommendations about where a refugee is likely to find employment, while considering different locations' capacity.

How it works: there are four stages to GeoMatch:

1. Modelling stage: GeoMatch is fed historical data about the success of past refugees in finding employment within 90 days in different areas. Machine learning models are used to identify patterns between those refugees' personal characteristics (including gender, age, ethnicity, education, work history, religion etc) and their employability in each location.

2. Prediction stage: The tool then uses these models to make predictions for newly arriving refugees. In other words, it makes predictions about where new refugees have the best chance of finding employment within 90 days, based on their similarities with the profiles of past refugees. For families who need relocating together, employment predictions are made for each adult household member.

3. Mapping stage: GeoMatch then turns individual level predictions into family-level predictions. In families, the algorithm considers the employment predictions made for each adult and optimises for (i.e. makes a decision for the whole family based on) which adult has the highest prospect of employability within the family.

4. Matching stage: Finally, the algorithm identifies the best locations for the refugee family, optimising for employability while also considering each location's capacity. With regards to capacity, the algorithm considers both the numbers of current cases and projected future arrivals, ensuring recommendations are distributed evenly across locations. By balancing these factors, the algorithm helps the resettlement agencies avoid

overwhelming any single location, distributing cases more evenly, optimising resources, and accommodating future needs.

GeoMatch then produces a ranked list of locations according to the likelihood of one of the family's adult members finding employment within 90 days. In the list, some locations appear in red due to being incompatible with specific other constraints such as a family's size, the availability of interpreters or of certain medical facilities. The decision-maker at the resettlement agency can then compare the different locations' scores and any other factors they wish to consider before making a final decision.

Explainability of results: GeoMatch uses machine learning to analyse patterns between a refugee's personal characteristics ('covariables') and their likely employability across different locations based on historical data. Although humans provide the input data, the algorithm processes it holistically, comparing current cases to past ones to identify patterns and use those to make case-specific location recommendations. As a result, the specific reasoning behind why a particular location is suggested for a particular individual is unknowable, making it impossible to explain the recommendations. Any attempt to do so risks oversimplification and misinterpretation.

Concerns:

- Refugees as economic objects: By optimising for employability, there is a risk that refugees are being treated primarily as economic objects rather than as subjects of humanitarian protection. Refugee status is granted for having a well-founded fear of persecution from one's home country, not for being a source of wealth generation for a new country. While finding employment quickly may well be a refugee's priority, there are numerous other short- and long-term goals that may be more important to them, such as availability of therapy services or location in relation to diaspora communities.

- Refugee consent or preferences: Relatedly, refugees have no choice or ability to opt out of their involvement in this tool. Nor are their preferences considered by the tool. The counter-argument is that refugees did not have this ability when human decision-makers were making the decisions alone, and now the decision-maker just makes more data-informed decisions. However, previously the decision-maker may be more likely to make decisions in the round, considering various competing factors for each location rather than primarily considering employability.

  An example of a matching tool that does seek to prioritise refugees' agency and preferences has been developed by a Canadian and European organisation, Pairity.[125] In its recent '*Re:Match*'[126] pilot project, which relocated displaced Ukrainians from Poland to different municipalities in Germany, refugees could determine which factors – such

[125] Ahmed Ezzeldin Mohamed and Craig Damian Smith, 'Ethically informed algorithmic matching and refugee resettlement' (2024) https://www.fmreview.org/digital-disruption/ezzeldinmohamed-smith/
[126] Berlin Governance Platform and Pairity, 'Re:Match – Relocation via Matching: An algorithm-based & equitable solution for refugees and welcoming municipalities – Pilot Project Interim Evaluation' (2024): https://pairity.ca/wp-content/uploads/2024/01/ReMatch_Interim-Evaluation-Report_2023_english_web.pdf

as proximity to diaspora communities, education opportunities, or employability – should be prioritised in their resettlement recommendation. The reason for this is because, as Pairity writes:

> *Introducing preferences-as-data can build algorithms that limit bias and minimise reliance on unverified assumptions and stereotypes. Similar to labour-market assumptions, the common and seemingly innocuous assumption that refugees prefer relocation near co-nationals or co-religionists could have ethical repercussions, especially for those fleeing discrimination due to their identity factors like ethnicity, religion, or sexual orientation and gender expression. Including refugees' preferences in algorithms minimises these potential pitfalls.*[127]

- Quality of employment: The tool does not necessarily consider the quality of the employment opportunities relative to refugee skill level. For instance, finding employment quickly in a new country where one does not speak the language is likely to be difficult unless it involves low-skilled roles where local language skills are unimportant.

- Optimisation for the most employable member of a household: There are risks in deciding location of a family based predominantly on employability of the most employable family member. In a heteronormative couple, it is not unreasonable to suspect that recommendations will be more frequently made based on the man's employability. This is because historical data will likely suggest men find employment more quickly, and so such patterns will be replicated with GeoMatch more frequently optimising for the man's employability. I asked about this risk during the interview and was informed that the team have been monitoring for this and have not observed this pattern yet or any other unintended biases or harms. Such findings would need to be published and peer reviewed to be proven.

  The team also raised the philosophical point that even if it was observed that some subgroups (e.g. men) make proportionately higher employment gains than others (e.g. women), provided both are making gains compared to not using the tool, then it is still helping both subgroups. However, this could inadvertently disadvantage women in the long term, as their employability may not receive the same level of optimisation, potentially contributing to widening of the relative inequality between genders of social and economic outcomes. The tool's success should be evaluated not just by whether both subgroups benefit, but by how equitably those benefits are distributed.

- Explainability of results: As mentioned above at page 21, I have concerns about the suitability of 'black box' machine learning tools in administrative decision-making. While I appreciate these tools yield more accurate results (assuming they are trained on appropriate and representative data), this comes at the cost of individuals being able to understand them. This alters the power relationship between the public and the state since the public cannot effectively hold government decision-making to account. While the stakes are lower in the field of refugees who have already been granted legal status, the state is still making significant decisions about where

[127] See footnote 125, Mohamed and Smith

someone should live. A trade-off of a slightly less accurate tool but one that is explainable is a worthwhile one in the interests of upholding democracy.

- Automation bias: Relatedly, since no human can understand GeoMatch's recommendation-making process, it raises questions about the possibility of mitigating decision-makers' propensity to 'automation bias' (see page 15). This is because it is impossible for the decision-maker to scrutinise why the recommendation has been made or if it has been made accurately.

- Amplification of historical biases: One possible consequence of automating predictions about employability based on historical data is that it will magnify historical trends. If certain areas have always been hostile to employing women or certain ethnicities, GeoMatch may magnify those trends by not recommending locating them in those areas. This in turn would then potentially create further data confirming this bias that would then be input back into the algorithm, creating a 'feedback loop bias' and reinforcing such enclaves over time.

  I raised this concern with the GeoMatch team. They had several counter-arguments. Firstly, it is outside GeoMatch's remit to change the status quo. Instead, GeoMatch is simply seeking to help refugees find employment, taking the world as it is with its existing levels of discrimination, and trying to optimise within it. Secondly, that indeed it might be ethically questionable to prioritise the potential improvement of long-term trends in the future (which GeoMatch might not even have a large enough impact to effect) over actual gains to individuals and families in the present.

  While GeoMatch suggests its role is not to change the status quo, it is important to recognise that by merely reflecting historical biases, the system is not neutral – it becomes a participant in perpetuating those biases. Public sector entities should ensure that tools like GeoMatch actively address and correct for historical discrimination, rather than passively optimise within a flawed system. The goal should not only be to improve individual outcomes but also to avoid reinforcing patterns of exclusion or inequality.

- Use of data: GeoMatch currently uses refugees' demographic data including age, gender and ethnicity to inform its resettlement recommendations. The use of demographic data raises concerns similar to those raised in a 2019 case in which civil rights groups including the American Civil Liberties Union argued Facebook's ad-targeting practices were discriminatory. Facebook's tools allegedly allowed advertisers, including in housing and employment, to exclude users based on factors like age and gender, preventing older workers and women from seeing certain job ads. It was argued that this practice restricted access to key opportunities for marginalised groups and violated civil rights law. Facebook ultimately settled the case and implemented changes to its advertising platform.[128]

  This case underscores the risks to human rights posed by algorithms that use sensitive demographic data to make predictions in crucial areas like housing and employment. In GeoMatch's context, there is a concern that

[128] ACLU, 'Facebook Settles Civil Rights Cases by Making Sweeping Changes to Its Online Ad Platform' (2019): https://www.aclu.org/news/womens-rights/facebook-settles-civil-rights-cases-making-sweeping

the algorithm could unintentionally exclude women or other marginalised individuals from certain opportunities, perpetuating inequalities rather than mitigating them.

In contrast to GeoMatch's application in the USA where issues of consent, bias and the use of demographic data arise, media reports indicate that IPL is developing another version of GeoMatch with Canada's IRCC which offers a different approach. There, GeoMatch is being developed as an opt-in tool for skilled economic migrants. This version aims to predict how migrants' qualifications and skillsets relate to their potential salaries in various locations.[129] GeoMatch's recommendations will then be offered as a free optional product to the migrants themselves so that some may be encouraged to move to locations where they will earn more, not just to those areas with which they are familiar.

This Canadian application addresses many of the concerns seen in the USA as participation will be voluntary.[130] Additionally, migrants, as the end user (rather than a resettlement agency officer) retain full autonomy in deciding whether to follow GeoMatch's recommendations, allowing them to prioritise their own preferences and needs. That said, to harness the potential of this tool to truly empower migrants, informed consent must be taken from them at the start such that they fully understand the applicable risks outlined above.

In sum, this exposition of data driven location-matching tools, like GeoMatch, demonstrate that even with good intentions, ARMTs in administrative decision-making processes require tight regulation to ensure they do not harm migrants' rights. It is hoped that the benefits of these tools, for example to distribute migrants according to areas' capacity as well as migrants' preferences and needs, will be realised when such regulation is in place.

Recommendations for improvement

1. **Proactive regulation**: It is an obvious point, but currently the drawing of ethical red lines is left to profit-driven software companies or politically-motivated government departments. In a democratic society, regulation that has undergone public consultation is required urgently for these red lines to be drawn more fairly before further real-life harms are caused.

2. **Explainability**: ARMTs' processes should be fully explainable to humans to protect fairness in administrative decision-making. Otherwise, such tools should only be used when the affected individuals are the end users, have given informed consent and have an ability to opt out of both their involvement in the tool and of compliance with its recommendations.

3. **Informed consent**: ARMTs should only be used where the affected individual has given genuinely informed consent about the use of their data and the tool's limitations and risks, and in circumstances where they are able to opt out of being a subject of the tool. For individuals who opt out, there should be an alternative

[129] Toronto Star, 'New tool could point immigrants to spot in Canada where they're most likely to succeed' (2021) https://www.thestar.com/news/canada/new-tool-could-point-immigrants-to-spot-in-canada-where-they-re-most-likely-to/article_d79e1ff5-a2e0-5a67-96c8-baa71bfed205.html; Immigration.ca 'Canada working on use of AI to help immigrants prosper' (2021), https://immigration.ca/canada-working-on-use-of-artificial-intelligence-to-help-immigrants-prosper/

[130] Although scrutiny will need to be given to the Canadian use of this tool in light of IRCC's current ban on using 'black box' algorithms

human decision-making process available so that those who opt out are not at a disadvantage or, in effect, discriminated against.

4. **Refugees' preferences and agency**: As Pairity suggests, ARMTs should consider refugees' preferences and be a tool to platform their agency, rather than compound their lack of it. This will mean that refugees cannot be located in areas primarily on the basis of their short-term employability without their consent but according to the factor(s) which they think they require to begin their new lives in safety.

5. **Human rights advisors**: Conversations with technology companies revealed that data often serves as the primary driver, while human rights and ethical considerations are less familiar territory. This is not to imply that technologists lack ethics, but rather that ethical theory is not their area of expertise. To address this gap, it is crucial to involve independent and empowered human rights advisors throughout the development process – from conceptualisation and design to monitoring and coding. These advisors should have direct experience working with affected communities or come from those communities themselves.

6. **Independent auditors**: The accuracy, efficacy and fairness of these tools should be audited by external, technical specialists according to strict criteria. Only after tools have met the minimum thresholds of these criteria, as determined by the auditor, should it be permitted for them to be deployed.

Section 5 – Recommendations for the UK

Based on the above analysis of the developments and regulation of ARMTs in the Canadian, the US and the UK immigration systems, I set out the below summary of my recommendations for best practice in the UK.

Recommendations for government

Proactive regulation: Crucially, the Government urgently needs to proactively establish clear, legally binding regulation of public sector use of ARMTs to ensure ethical boundaries are defined by regulations derived from public consultations, rather than by profit-driven companies or politically motivated government departments. Such regulation should provide for the following:

1. **Ethical boundaries**

    a. Ban 'black box' algorithms: Following Canada's example, there should be a ban on Home Office use of 'black box' machine learning models at any stage of a decision-making process. This is to ensure that ARMTs used in immigration processes are fully explainable to human decision-makers. The only exception to this ban should be if the affected individual is the end user of the tool, has given informed consent and can opt out of both their involvement in the tool and compliance with its recommendations. For individuals who opt out, there should be an alternative human decision-making process available so that those who opt out are not at a disadvantage or, in effect, discriminated against.

    b. Ban automated refusals: Also following Canada's example, there should be a ban on Home Office use of ARMTs that automate refusals, ensuring all refusals are subject to individualised human review.

    c. Refugees' preferences and agency: As Pairity suggests, ARMTs should consider refugees' preferences and be a tool to platform their agency, rather than compound their lack of it. This means tools should: i) directly consider refugees' preferences; ii) make the migrants the end users; iii) obtain migrants' informed consent; and/or iv) give migrants the ability to opt out. Empowering migrants in these ways would help balance the power dynamic between individuals and the state, ensuring that technology serves to improve their experiences rather than undermine their rights.

2. **Terminology**

    a. 'Automated recommendation-making tools': Update language in policies and communications on the topic to refer to *recommendation*-making tools instead of *decision*-making ones. This will ensure all tools influencing decision-making are subject to oversight, including tools that simply make recommendations, categorise, associate and/or triage cases, even if a human makes the final decision. It will also ensure definitions are not weighed down by changing and granular technological definitions.

    b. Make explicit that triaging, associating and categorising tools are included in regulatory oversight: Relatedly, while Canada's DADM applies to decision *recommendation*-making tools, in practice government departments only

interpret it to cover fully the decision-*making* aspects of automated tools. Canada's DADM and the UK's Mandatory Policy (and any related regulation) should make explicit that triaging, categorising and other ARMTs are within the scope of algorithmic regulation to clear up confusion and ensure full compliance by public bodies.

c. 'Chain of decision-making': Replace language in regulation and communications regarding a 'human in the loop' to a 'chain of decision-making'. This will help ensure that all human interactions with automated tools from the data input, optimisation choices and decision-making are subject to scrutiny. Regulatory oversight can then be applied throughout the decision-making chain, not just at the final stage.

d. Expand the scope of the ATRS beyond 'complex' algorithms: As Canada's Guide makes clear, the UK's Mandatory Policy, and any regulation in this context, should apply to all algorithms regardless of their complexity, as even simple algorithms can have profound impacts on individuals.

3. **Monitoring and transparency**

a. Mandate pre-deployment disclosure of AIAs and ATRS records, and regularly thereafter: Following the DADM's example, mandate the disclosure of impact assessments and ATRS records *before* deploying ARMTs. It should be required that these are updated at specified (e.g. six-monthly) intervals, every time new datasets are incorporated and if the function is adapted.

b. Default full transparency: Transparency records should by default include disclosure of technical details of algorithms including source code, training data and type of technology used and how it works. It should also include the algorithm's use of data, role in the decision-making process, associated training materials for decision-makers, audits and reviews. The burden should be on the public body to justify any departure from full transparency, with specific and clear reasons provided.

c. AIAs: Like Canada's DADM, ATRS records should include impact assessments which evaluate 'the rights of individuals or communities, the equality, dignity, privacy and autonomy of individuals, the health or well-being of individuals or communities, the economic interests of individuals, entities, or communities, the ongoing sustainability of an ecosystem.' They should also evaluate the potential risks to human rights. Finally, they should be designed in a way that accounts for the interaction between different ARMTs, for example, which (and how) databases or algorithmic processes are merged.

d. Equalities impact assessment: Similar to Canada's GBA+ impact assessments, specific equalities impact assessments should be mandated that assess tools' risks on certain protected characteristics. This is necessary under s149 Equality Act 2010 which requires public bodies to ensure, in the exercise of its functions, that it has taken '*due regard*' to the need to eliminate discrimination and advance equality.

e. Collect protected characteristic data: Immediately begin collecting and disclosing race, ethnicity, gender and other disaggregated protected characteristic data to monitor potential discrimination in ARMTs' outcomes.

f. Procurement guidelines: Involvement of third-party companies should not be a valid reason for departing from full transparency. Instead, companies wanting to develop ARMTs for public sector use should be required to contractually agree to transparency obligations from the outset. This goes beyond merely requiring government departments to indicate an expectation of transparency in invitations to tender, as suggested in the Mandatory Policy.[131]

g. Transparency for 'dual-use' and national security systems: Ensure that dual-use ARMTs (e.g. systems that serve both national security and immigration purposes) remain as transparent as possible. A clear distinction should be made between national security and immigration-related functions of such ARMTs, with transparency requirements for the immigration aspects.

h. Notice and explanation requirements: Implement rules mandating proactive notice requirements to affected individuals about the fact that an ARMT has been used in their case. This notice should include providing detailed information about how the tool influenced the decision, the factors considered and individuals' rights to challenge the decision. Such notice should be given either along with the notification of the substantive decision, or in cases where notifications are not usually given (e.g. when a triaging recommendation is made) within 48 hours of the recommendation.

4. **Oversight**

a. Human rights advisors: Mandate the involvement of independent human rights advisors in the development of ARMTs, ensuring ethical considerations are prioritised at every stage of the development and deployment of ARMTs.

b. Independent auditing: Establish an external audit system where ARMTs are assessed for fairness, accuracy and bias *before* they are implemented and regularly thereafter, including when new datasets are added or their function is adapted. Mandate that audit recommendations be incorporated before implementation. Require transparency of these audits.

5. **Training**

a. Mandate training for decision-makers: Mandate training for all public sector decision-makers who use ARMTs. This training must ensure that decision-makers understand the limits of ARMTs, how to identify and mitigate biases and the importance of transparency in their use. Additionally, decision-makers should be educated on how to properly exercise their discretion when reviewing ARMT recommendations and feel empowered to override these recommendations when necessary.

b. Mandate training for ARMT developers: Developers of ARMTs should be trained on diversity, inclusion and equality in ARMT development, as well algorithms' limitations and risks to peoples' human and non-discrimination rights.

c. Train judges and adjudicators: Implement specialised training for judges and adjudicators on the impact of ARMTs on the decision-making process and public rights.

[131] Annex A (Mandatory Policy), page 59

6. **Redress**

    a. Right of appeal: Affected individuals should have a right of appeal at every stage of automated intervention in the decision-making process, including triaging decisions. While this may create additional litigation in the beginning, it will significantly reduce costs and court time in the long run by ironing out issues with ARMTs.

    b. Actionable rights: Actionable rights should be created so that members of the public can sue public bodies for failure to comply with regulation on ARMTs, rather than leaving this to self-regulation by the public bodies themselves.

Recommendations for lawyers

1. **Request disclosure**: Since many Home Office decisions will likely have been influenced by ARMTs, most immigration and public law challenges should be requesting disclosure of any ARMTs involved in decision-making processes. For example, at the pre-action stage, in reliance on the duty of candour during public law proceedings or through subject access requests (SARs).

2. **Challenge lack of notice**: Challenge the Home Office's failure to notify individuals about the use of ARMTs in decision-making processes for potential breach of public law principles, human rights, data protection or equalities legislation.

3. **Training for legal professionals**: Provide training on ARMTs and their role in immigration decisions to ensure lawyers are equipped to identify and challenge algorithmic issues.

Recommendations for civil society

1. **Public awareness campaigns**: Continue to gather information about Home Office use of ARMTs to raise awareness about their risks. Public Law Project's '*Tracking Automated Government*' register[132] is an excellent example of the power of civil society to push for greater transparency.

2. **Gather affected individuals' experiences**: Conduct qualitative research gathering affected or potentially affected individuals' experiences of the role of ARMTs in their cases. Record these experiences, including whether they suggest any additional risks or benefits of these tools to inform further advocacy.

3. **Engagement with the Information Commissioner's Office (ICO)**: Bring complaints or challenges to the ICO (the UK's independent regulatory offices dealing with data and privacy).[133] The ICO has a strong general mandate to investigate complaints from members of the public who believe a public body has failed to respond correctly to a request for information under FOIA or a SAR, including in relation to ARMTs.[134]

[132] Public Law Project, 'TAG register' https://publiclawproject.org.uk/resources/the-tracking-automated-government-register/
[133] With thanks to Lucie Audibert for this suggestion
[134] The ICO, 'FOI complaints and ICO enforcement powers', https://ico.org.uk/for-organisations/foi/foi-complaints-and-ico-enforcement-powers/

Final remarks

The experiences of the USA and Canada deliver a clear mandate: the UK Government, civil society, technology companies and the public must work urgently and collaboratively to ensure ARMTs are deployed in ways that realise their potential without sacrificing migrant rights. By raising public awareness and fostering informed debate, this report aims to support the UK in navigating this new frontier in immigration decision-making, setting standards for fairness, transparency and innovation on the global stage.

Annex A – 'ATRS Mandatory Scope and Exemptions Policy – FINAL'[135]

Summary
This is the mandatory scope and exemptions policy for the Algorithmic Transparency Recording Standard (ATRS). It sets out which organisations and algorithmic tools are in scope of the mandatory requirement to publish ATRS records, as announced in government's response to the consultation on the AI White Paper "A pro-innovation response to AI regulation" in February 2024. It also sets out the required steps to ensure that sensitive information is handled appropriately.

1. Introduction

Since the beginning of 2022, the Algorithmic Transparency Recording Standard (ATRS) has been an established mechanism for the proactive publication of information on the use of algorithmic tools in the public sector. The ATRS has been piloted with various organisations, enhanced and iterated multiple times, and on 6 Feb 2024 it was mandated across all government departments, with a stated intent to extend the requirement to the broader public sector over time. To implement this mandatory rollout, we need to establish clear lines about which organisations and tools are in scope, and the type of information that, for various reasons, may be too sensitive for publication on Gov.uk.

Transparency around how the public sector is using algorithmic tools is useful and appropriate in most circumstances and should be our default position. However, there is some need for caution to make sure that information that is sensitive or confidential is handled properly.

This document sets out the scope for mandatory publication of ATRS records, implementing the cross-government policy set out in the AI White Paper consultation response on 6 Feb 2024.

It sets out:

- The organisations in scope of this policy, which currently comprise central government (with an intent to extend this more broadly across the public sector in future)
- The algorithmic tools that are in scope
- The steps to be taken to make sure that sensitive information is handled appropriately

2. Organisations in scope

2.1 Overall scope

The mandatory requirement to complete ATRS records currently applies to central government. For the purposes of this policy, this consists of the following:

- Ministerial departments, and
- Non-ministerial departments, and
- Arm's-length-bodies (ALBs), meaning executive agencies and non-departmental public bodies, **which provide public or frontline services, or routinely interact with the general public**.

[135] Reproduced with the kind permission of the RTA

This scope is intended to capture the majority of central government uses of algorithmic tools, without placing a disproportionate burden on large numbers of very small or non-frontline arm's-length bodies that are unlikely to be responsible for any in-scope algorithmic tools. We will work with departments to finalise which arm's-length bodies fall in or out of scope, but as examples we would anticipate the following:

Examples of organisations in mandatory scope:

- All ministerial departments e.g. MoJ, DfE, DSIT
- All non-ministerial departments e.g. HMRC, the National Archives, Competitions and Markets Authority
- All other arm's-length bodies (ALBs) that provide public or frontline services, or routinely interact with the general public e.g. HM Land Registry, HM Prisons and Probation Service

Examples of organisations out of mandatory scope:

- Arm's-length bodies that do not provide public or frontline services, or routinely interact with the general public. Likely examples may include:
    - Shared Business Services Limited
    - National Infrastructure Commission
    - Biometric & Forensics Ethics Group
- Any organisations that are not in scope of the Freedom of Information Act.

2.2 Initial rollout

The initial rollout of the mandatory policy is proceeding in two phases:

- Phase 1: Most ministerial departments and HMRC, specifically:
    - Cabinet Office
    - Department for Business and Trade
    - Department for Culture Media & Sport
    - Department for Education
    - Department for Energy Security & Net Zero
    - Department for Environment Food and Rural Affairs
    - Department for Levelling Up, Housing and Communities
    - Department for Science, Innovation & Technology
    - Department for Transport
    - Department for Work & Pensions
    - Department for Health and Social Care
    - Foreign, Commonwealth & Development Office
    - HM Revenue & Customs
    - HM Treasury
    - Home Office
    - Ministry of Defence
    - Ministry of Justice
- Phase 2: The remaining ministerial and non-ministerial departments, and arm's-length bodies that fall in the scope listed above.

As a DSA-endorsed Standard, the ATRS remains recommended across the broader public sector. Hence, the sections below will also be useful to other organisations in determining for which tools it would be good practice to publish ATRS records.

3. Algorithmic tools in scope

Organisations determined to be in mandatory scope above are required to publish ATRS records for algorithmic tools they are currently using *in relevant use cases*.

3.1 What is an 'algorithmic tool'?

An algorithmic tool is a product, application, or device that supports or solves a specific problem using complex algorithms.

We use 'algorithmic tool' as an intentionally broad term that covers different applications of artificial intelligence (AI), statistical modelling and complex algorithms. An algorithmic tool might often incorporate a number of different component models integrated as part of a broader digital tool.

3.2 For which tools must I complete an algorithmic transparency record?

The mandatory requirement to publish an ATRS record applies to algorithmic tools that either:

1. have a significant influence on a decision-making process with public effect, or
2. directly interact with the general public.

'Significant influence' may mean that an algorithmic tool meaningfully assists, supplements, or fully automates a decision-making process.

By 'public effect' we mean a decision-making process having an impact on members of the public, where the latter are understood as any individuals or groups of individuals, irrespective of their nationality or geographical location.

To decide whether a decision-making process has a public effect, you might want to consider whether usage of the tool could:

- materially affect individuals, organisations or groups
- have a legal, economic, or similar impact on individuals, organisations or groups
- affect procedural or substantive rights
- impact eligibility for, receipt of, or denial of a programme

Note that this is intended to apply to situations where an algorithmic tool is influencing specific operational decisions about individuals, organisations or groups, not where a tool is an analytical model supporting broad government policy-making. Analytical models in scope of the guidance in the Aqua Book will typically be outside of ATRS scope (though it is possible to envisage some specific circumstances where both would be applicable, see examples below).

Examples of tools that could fall within the scope of these criteria are:

- a machine learning algorithm providing members of the public with a score to help a government department determine their eligibility for benefits (impact on decision-making with public effect)
- a chatbot on a government website interacting directly with the public which responds to individual queries and directs members of the public to appropriate content on the website (direct interaction with the public)

Examples of tools that would likely not fall within the scope of the criteria include:

- A tool being used by a government department to transform image to text (e.g. used in digitisation of handwritten documents) as part of an archiving process (no significant decision or direct public interaction)
- An automated scheduling tool which sends out internal diary invites from a mailbox (doesn't have public effect)

Further examples are listed below in Annex 1.

To emphasise, **the context of use of the algorithmic tool matters here**. The same image to text algorithm above might be relevant if being used instead to digitise paper application forms for a government service (e.g. poor performance of the algorithm on some handwriting styles might well have an influence on success rates for individual applicants).

Note that the algorithmic tool scope listed above is that of the policy for **mandatory ATRS adoption in central government**. If you are using an algorithmic tool that does not strictly meet these criteria but you would like to provide the general public with information about it, you can still fill out and publish an algorithmic transparency record.

3.3 At which stage in a tool's development lifecycle should an ATRS record be created?

The mandatory requirement to publish an ATRS record applies to tools that are in Beta/Pilot or Production phase.

Teams are welcome to submit records for tools in earlier stages of the lifecycle, but it is not mandatory to do so.

For tools that have previously been in use and had a record created for them and which are later being retired, the responsible team should submit an updated record changing the information in the phase field to 'Retired'. This update will be reflected on the published record on Gov.uk.

4. Exemptions

4.1 What information should organisations not publish?

The ATRS has been designed to minimise likely risks that could arise from publication of records (e.g. to security, privacy or intellectual property).

Situations where no information can be safely published are expected to be unusual (e.g. in cases where even the existence of a tool cannot be made public for security reasons).

More commonly, for some tools, there may be particular information requested in the ATRS that you may be concerned about releasing into the public domain, even if the majority of the information about the tool is publishable. This may relate to a risk of gaming the tool, risks to national security, infringing intellectual property or releasing commercially sensitive information.

In most such instances, the appropriate response is to reduce the level of information supplied for relevant fields, for example giving a broad description of the type of data used by a tool instead of specific details of individual data sources, or a broad summary of how a tool works instead of precise information about the system architecture.

In developing the rationale behind the exemptions for this policy, we align with those set out in the Freedom of Information Act 2000 ("FOIA"). Although FOIA is a reactive means of providing information to the public while the publication of ATRS records is proactive, we settled on using the FOIA exemptions as a basis for our exemptions policy since the

logic around which types of information are too sensitive to publish openly remains the same. Moreover, the FOIA policy is firmly established in the public sector as business as usual, thus this will reduce the administrative burden for organisations to comply with the ATRS policy mandate.

As a general rule, this ATRS Scope and Exemptions Policy **does not require the publication of information that would be subject to an exemption under access to information legislation**, i.e. the FOIA, Environmental Information Regulations and data protection legislation.

To understand how this applies in practice, imagine that you created a full internal version of an ATRS record and (hypothetically) received an FOI request to publish that record. How would you respond to such a request?

- If you would release the record in full, then the same applies to proactive publication of the record.
- If you would release some of the information in the record, but would need to redact some of it under FOI exemptions, then you should remove or de-sensitise the exempt content from the ATRS record prior to publication.
- In (rare) circumstances where you were not able to confirm the existence of the tool, e.g. you would issue a neither-confirm-or-deny response to the hypothetical FOI request, then it would be inappropriate to publish the ATRS record at all.

Not all FOIA exemptions are relevant here. Specifically:

- The ATRS is designed to capture *tool-level information* rather than personal information. **Concerns about publishing personal data should therefore not apply** when considering whether to remove or de-sensitise partial or whole ATRS records (FOIA section 40).
- There are limitations on cost of responses, vexatious queries and information already in the public domain, that are necessary for a reactive duty such as the FOIA to avoid disproportionate effort in responding to an unbounded number of incoming requests (FOIA sections 12, 14, 21, 22). They are not relevant to the publication of ATRS records which is inherently limited to one record per tool.
- Exemptions for reasons of commercial sensitivity (FOIA section 23) need to be applied with care (see Section 4.3 below)

4.2. How to exempt types of information within a transparency record

As mentioned above, in most cases it will be sufficient to give higher-level information within particular fields. However, where entire fields are fully exempt from publication, or you wish to indicate explicitly why the information in a record is limited, we recommend recording this within a record in the following format, with the third column giving a general description of the reason. For example:

2.4.2.9	**Dataset purposes**	Indicate how each dataset was used during the model development process (e.g. training, validation, testing).	EXEMPT: National security

2.4.2.9	**Dataset purposes**	Indicate how each dataset was used during the model development process (e.g. training, validation, testing).	EXEMPT: After considering factors for and against disclosure of this information, we have concluded that it is not in the public interest to disclose information on the specific datasets. This is based on the specific information contained within them having the potential to give malicious actors insights on how to evade/game the tool and therefore endanger the UK's defence capabilities (section 26 of the FOIA 2000).

If you have any concerns about publishing information that are not covered above, we would ask you to get in touch with the ATRS team to discuss (algorithmic-transparency@dsit.gov.uk).

4.3 Dealing with commercially sensitive information

Many algorithmic tools used in the public sector will involve an external supplier in soe form, and hence publication of an ATRS record will require some consideration of commercially sensitive information.

There is a need for care in applying commercial exemptions, i.e. in applying Section 43 of the Freedom of Information Act.

If this exemption is applied too broadly, it would undermine the overall intent of this policy to increase transparency on the government's use of algorithmic tools and limit its benefits. Commercial suppliers that wish to sell algorithmic solutions to public bodies that are then used in processes that impact members of the public should be comfortable with this level of transparency that is expected of the public sector. Public bodies that are procuring solutions from vendors should make this expectation clear in their invitation to tender or other route to market.

The primary focus of the ATRS is the use case that a tool is deployed into, and the steps taken to ensure **that a tool is appropriate in that context. Though there is some information required about the** technical aspects of the tool, there is flexibility on how much information is provided here, and the Standard is designed to be consistent with emerging industry good practice on model cards (and indeed information that might often be published by vendors in a white paper).

As such, public authorities are encouraged to work with their supply chains at the start of the process to minimise the amount of information being withheld for commercial reasons.

Annex 1: additional examples of in and out of scope tools

Type of tool	Example use case(s) in mandatory ATRS scope *(rationale)*	Example use case(s) out of mandatory ATRS scope *(rationale)*
Large language model	An LLM developed/used as a digital assistant to suggest services or resources someone may be eligible for or benefit from after they explain their circumstances to it *(direct interaction with the public)*	Ad-hoc usage of large language models, such as Microsoft Copilot being used by individual civil servants internally within an organisation/department to transcribe and summarise meetings *(no significant decision or direct public interaction)*
Chatbots	A chatbot on a website interacting directly with the public which responds to individual queries and directs members of the public to appropriate content on the website *(direct interaction with the public)*	A similar chatbot for purely internal purposes e.g. as part of an internal IT support channel *(no significant decision or direct public interaction)*
Text recognition	A tool being used by a government department to transform image to text to digitise paper application forms for a government service *(potential influence on decisions, e.g. poor performance of the algorithm on some handwriting styles might well have an influence on success rates for individual applicants)*	A tool being used by a government department to transform image to text (e.g. used in digitisation of handwritten documents) as part of an archiving process *(no significant decision or direct public interaction)*
Productivity	An AI-based automated scheduling tool applying prioritisation criteria to schedule medical appointments *(direct interaction with public)*	An automated scheduling tool which sends out internal diary invites from a mailbox *(no significant decision or direct public interaction)* An algorithmic tool that uses written input to generate financial reports, emails and charts, or autofill internal admin documents *(no significant decision or direct public interaction)*
Scoring and risk assessment	A machine learning algorithm providing members of the public with a score to help a government department determine their relative priority for accessing a public service *(impact on decision making)*	A statistical model estimating overall demand for a public service to inform policy-making and overall capacity planning *(not decisions about individuals)* A predictive analysis tool that uses previous data, predictive

	A tool used to score the risk of harm associated with a detainee and recommends the level of precaution required. *(impact on decision making)* A tool used to calculate the complexity of a case and assign a different caseworker depending on the score assigned. *(If assessment determines that how this is done could create an impact on decision-making)* A real time analysis tool that uses readily available data with newly obtained on the spot data to score or risk profile a person, object, location or event, e.g. customs scan *(potential to be used in decisions such as who or what to search or check, impacting individuals)*	data to inform requirements and aid in risk assessment e.g. future energy grid, road and rail usage, flood risk, staffing requirements *(not decisions about individuals)*
Biometrics	A facial recognition model used to verify the identity of an applicant for a product (e.g. a passport) against a face held on file *(potential to influence decisions, e.g. Poor performance on matching particular faces could have an influence on success rates for individual applicants)*	Use of facial or fingerprint recognition by public sector employee to unlock a corporate mobile device *(no significant decision or direct public interaction)*
Image recognition	An image recognition algorithm reviewing images and labelling them when spotting a pre-trained identifier in the image e.g. driving while on the phone *(potential to influence decisions, e.g. individual drivers receiving fines/penalties)* A machine learning algorithm that reviews x-ray images using pre-trained images to aid with diagnosis *(aids in decision making in dictating a healthcare response, plus direct public interaction with patients)*	A matching algorithm that takes multiple data sources of varying quality to decide if two plus records of varying data types are the same individual, e.g. to clean up a database *(no significant decision or direct public interaction)*