Human Rights Impact Assessment Tool: AI-informed Decision-making Systems in Banking

SEPTEMBER 2023

nab

Australian Human Rights Commission

Human Rights Impact Assessment Tool: AI-informed Decision-making Systems in Banking

ISBN 978-1-925917-85-7

Acknowledgments

The Australian Human Rights Commission thanks the National Australia Bank (NAB) for their collaboration with the development of this resource, and to all NAB staff members who contributed their knowledge, expertise and experience towards this project. In particular Jade Haar, Kobi Leins, George Drymonis, Alysia Abeyratne, Phillip Ward, Daniel Loden and Kathy Cena are thanked for their contributions.

The Human Rights Commissioner thanks President Rosalind Croucher and the following staff of the Australian Human Rights Commission for their contributions: Bruce Alston, Darren Dick, Patrick Hooton, and Llewellyn Spink.

This publication can be found in electronic format on the Australian Human Rights Commission's website at **https://humanrights.gov.au/our-work/technology-and-human-rights/publications/hria-tool-ai-informed-decision-making-systems**.

For further information about the Australian Human Rights Commission or copyright in this publication, please contact:

Australian Human Rights Commission
GPO Box 5218
SYDNEY NSW 2001
Telephone: (02) 9284 9600
Email: **communications@humanrights.gov.au**

Design and layout Dancingirl Designs

Cover image and internal photography Adobe Stock

Human Rights Impact Assessment Tool: AI-informed Decision-making Systems in Banking

September 2023

Australian Human Rights Commission 2023

Australian Human Rights Commission

nab

Contents

Lorraine Finlay

Human Rights Commissioner
Australian Human Rights Commission

FOREWORDS

Australian Human Rights Commission foreword

The Australian Human Rights Commission (Commission) welcomes the opportunity to have partnered with NAB to develop and produce this human rights impact assessment tool for artificial intelligence-informed decision-making systems in banking (HRIA Tool).

Artificial intelligence (AI) has the capabilities to improve efficiency and increase customer satisfaction when engaging with banking services. However, given the important role banks play in storing the wealth of Australians, it is important that when integrating AI into decision-making it is done ethically and with human rights at the forefront.

The aim of the HRIA Tool is to assist banks consider and measure the risk to human rights posed by AI systems, implement strategies to address those risks, and support the availability of remedies for any human rights violations. If AI is not integrated responsibly into decision-making systems, there may be serious and adverse consequences for everyone – including both customers and the banking industry itself.

The Commission developed this HRIA Tool (in collaboration with NAB) following on from the Commission's recommendations in its Final Report: Human Rights and Technology that private sector bodies be encouraged to undertake HRIAs before using AI systems, and that tools should be developed to assist them in doing so.

This HRIA Tool also builds upon the Commission's work to develop practical guidance for the ethical use of AI systems for various sectors and businesses, such as the Guidance Resource on Artificial Intelligence (AI) and Discrimination in Insurance Pricing and Underwriting (AI in Insurance Guidance).

On behalf of the Commission, I thank NAB for their valued contributions to this work. I also thank all the AI, technology and banking industry experts who provided advice and feedback throughout its development. I look forward to banks using the HRIA Tool to assist them in developing and deploying ethical AI-informed decision-making systems.

Lorraine Finlay

Human Rights Commissioner

NAB foreword

NAB has partnered with the Commission to develop a HRIA Tool. The purpose of this HRIA Tool is to specifically help banks consider and measure the risk to human rights posed by AI systems. The development of the HRIA Tool follows on from the Commission's recommendations in its Human Rights and Technology Final Report – which stated that *"private sector bodies be encouraged to undertake HRIAs before using AI systems and that tools should be developed to assist them in doing so"*.

The purpose of the alliance between NAB and the Commission, formed in 2021, was to facilitate the development of a HRIA Tool to help banks consider and measure the risk to human rights posed by AI systems. NAB recognises it is in both banks' and their customers' interests to ensure that we measure the risk to human rights posed by AI activities and implement strategies to address those risks. The concept is for an open-source tool to be made available as a central piece of guidance for banks to conduct their own HRIAs.

Traditionally, banks have a number of risk assessment frameworks, and the intention is not to add another assessment, but design a specific tool which could be tailored to ensure that the backbone of the guidance worked optimally for each bank that incorporated it. To date, NAB has included various human rights specific questions into its own data ethics assessment process when reviewing more general data use cases.

AI informed decision making that has a significant effect on individuals is the target use case to apply the HRIA Tool. NAB's approach to AI is to ensure that it is used for the betterment of its customers. NAB already has a set of Data Ethics Principles which has been in place since 2019 via its Data Ethics Framework. NAB sees this tool as another aspect to those foundations of responsible data use.

NAB piloted facial recognition technology (FRT) to assist customers to digitally verify their identification during COVID-19, using the Federal Government's AI Ethics Framework. NAB has embedded a process for reviewing all Data Analytics and AI projects from an ethical viewpoint before implementation. These reviews ensure that the technology is responsible, sustainable and justified. The HRIA Tool will add to that review process and supplement our development of strategies and policies.

NAB sees the tool as a relevant and timely questionnaire to assist the conversation rather than be viewed as another compliance check box. The format includes a number of questions with commentary for consideration, with many of the questions about algorithmic bias taken from the 2020 Commission Technical Paper and refined with feedback from our Data Science team.

1 Background, development and use

1.1 Background

The Commission has partnered with NAB to develop this HRIA Tool to help banks consider and measure the risk to human rights posed by AI systems, implement strategies to address those risks, and support the availability of remedies for any human rights violations.

The Commission is Australia's national human rights institution. The Commission is independent and impartial. It aims to promote and protect human rights.

The development of the HRIA Tool follows from the Commission's recommendations in its Final Report that private sector bodies be encouraged to undertake HRIAs before using AI systems, and that tools should be developed to assist them in doing so.

It also builds upon the Commission's work to develop practical guidance for the ethical use of AI systems for business such as the Guidance Resource on Artificial Intelligence (AI) and Discrimination in Insurance Pricing and Underwriting (AI in Insurance Guidance).

The content of the HRIA Tool is informed by the Commission's expertise, previous work on human rights and technology, research, and consultation with NAB. The Commission's previous work in this area includes the AI in Insurance Guidance, Final Report and Technical Paper on addressing algorithmic bias. The latter two documents are especially influential in the language and approach of the HRIA Tool.

The Technical Paper had a particular focus on algorithmic bias. Algorithmic bias can result in an AI system producing outputs that result in unfairness and can sometimes have the effect of obscuring and entrenching unfairness or even unlawful discrimination in decision making. Some of the questions and terms used in the HRIA Tool are derived from the approach to addressing the problem of algorithmic bias suggested by the Technical Paper.

In particular, the HRIA Tool adopts the following definitions from the Final Report:

- Artificial Intelligence or AI: As AI is not a universally accepted definition, the term is broadly used here to refer to a cluster of technologies and techniques, which include some forms of automation, machine learning, algorithmic decision making or neural network processing.
- AI-informed decision making is where AI is a material factor in the decision, and where the decision has a legal or similarly significant effect for an individual.

1.2 Human rights standards

Human rights standards constitute a benchmark for the HRIA Tool and should guide the impact assessment process.

In practice, the human rights most likely to be affected by AI-informed decision making in banking are those concerning privacy, non-discrimination and equality of treatment.

All human rights should be enjoyed by everyone regardless of factors such as race, sex, or disability.

Human rights standards include all those contained in international human rights treaties to which Australia is a party including the:

- International Covenant on Civil and Political Rights;
- International Covenant on Economic, Social and Cultural Rights;
- International Convention on the Elimination of All Forms of Racial Discrimination;
- Convention on the Elimination of All Forms of Discrimination against Women;
- Convention on the Rights of Persons with Disabilities; and
- United Nations' Guiding Principles on Business and Human Rights.

While human rights treaties create direct legal obligations for Australia, they do not impose obligations directly on businesses, such as banks. However, Australia has created protections for human rights through domestic law, such as federal anti-discrimination legislation and the *Privacy Act 1988* (Cth).

Federal anti-discrimination laws prohibit discrimination on the basis of protected attributes, including age, disability, race, including colour, national, or ethnic origin or immigrant status, sex pregnancy, martial or relationship status, family responsibilities or breastfeeding, or sexual orientation, gender identity of intersex status. State and territory laws generally also offer anti-discrimination protection.

Even where human rights have not been directly incorporated into domestic law, they are still relevant for businesses. The actions of businesses can still affect people's human rights. The United Nations Guiding Principles on Business and Human Rights makes clear that businesses have a responsibility to respect human rights.

1.3 Conducting a HRIA

Successful implementation of AI systems will involve a diverse set of stakeholders. Within banks, a multi-disciplinary team approach is recommended, including data scientists, legal and specialist teams covering areas like privacy, diversity and inclusion or vulnerability.

The HRIA Tool is not intended to be prescriptive about exactly how an HRIA should be conducted, provided that the bank takes full responsibility for the analysis and outcome.

The conduct of the HRIA may be an iterative, rather than linear, process. That is, for example, the assessment of human rights impacts may need to be continually reviewed as data is piloted and machine learning developed. More generally, the conduct of different aspects of the HRIA may need to be staged as an AI system is developed and involve different participants at different stages.

The following comments are intended to provide some guidance as to best practice.

1.4 The HRIA team

The team using the HRIA Tool should have the relevant interdisciplinary skills and expertise, including in anti-discrimination and human rights law and policy, and technical expertise in AI systems.

It is not necessary that a new team be created within the bank to conduct the HRIA. The responsibility for conducting the HRIA may be allocated in accordance with existing organisational structures.

Relevant senior management should be meaningfully and appropriately engaged in the HRIA process. Banks should also consider including independent experts and stakeholders in the team using the HRIA Tool.

It is possible that many aspects of the HRIA are already covered by existing assessment teams, such as privacy impact assessment teams, data ethics assessment teams, data risk teams and change risk teams. In such cases, the HRIA Tool may be used to augment existing assessments and identify gaps between these specialised assessments and a comprehensive HRIA. If the sum of these specialised assessments is equivalent to the HRIA Tool, then it is left to identify whether all these assessments are triggered by internal processes for each applicable activity.

1.5 External stakeholder engagement

Where practicable, integrating the participation of affected or potentially affected external stakeholders (including in impact assessment and mitigation) will enhance the HRIA process. However, the technical nature of AI systems and commercial confidentiality may limit the scope of consultation.

Depending on the AI system being developed or implemented, banks should consider what engagement may be appropriate by identifying the main stakeholder groups potentially affected by the AI system. For example, do vulnerable or marginalised individuals and groups need special consideration?

Engagement with customers, customer advocates, employees, contractors, investors, analysts, industry bodies, regulators and government, suppliers, and the broader community may need to be considered, as well as aligning this review with Impact Assessments required by ISO Standards, Privacy Impact Assessments, Security Assessments and others.

Competition law and other commercial interests may influence the degree of engagement. In particular, the extent to which information can be shared with stakeholders may depend on a range of factors, including limitations on the disclosure of proprietary and commercial in confidence information.

2 Human Rights Impact Assessment Tool

2.1 Purpose

The HRIA Tool is intended to help banks assess the human rights impact of the use of AI-informed decision-making systems (AI systems) in banking.

The HRIA Tool should be used to assess whether an AI system is lawful, transparent, explainable, used responsibly, and subject to appropriate human oversight, review, and intervention.

Banks have a responsibility to respect human rights, in accordance with the United Nations Guiding Principles on Business and Human Rights (UNGPs). The UNGPs establish a global standard for preventing and addressing the risk of adverse human impacts linked to business activity.

The human rights most likely to be affected by AI-informed decision making in banking are privacy, non-discrimination, and equality of treatment.

The HRIA Tool is intended to:

- strengthen knowledge and understanding of human rights impacts;
- provide practical guidance on specific human rights impacts, particularly in relation to non-discrimination and equality of treatment; and
- identify practical mitigation strategies and remedies to address any adverse human rights impacts from AI systems in banking.

The team using the HRIA Tool should be supported by adequately resourced human rights expertise with clear roles and responsibilities. Broad internal engagement with business owners, project managers, data scientists and other experts will be necessary to effectively review any tools for UNGPs considerations.

The HRIA Tool should assist banks when considering the impact of AI systems on human rights in banking. A bank's use of, or compliance with, the HRIA Tool is not mandatory and is at the bank's discretion.

The HRIA Tool does not constitute legal advice and does not provide a definitive legal answer regarding any adverse human rights impacts, including breaches of federal anti-discrimination or other relevant legislation. Organisations and individuals should seek independent legal advice if they have concerns regarding their compliance with applicable legislation and their legal obligations.

An organisation or individual will not be protected from liability for adverse human rights impacts, including unlawful discrimination, if they claim they complied with or relied on the HRIA Tool. However, the use of this HRIA Tool may help banks identify, address and remedy potential adverse human rights impacts.

The HRIA Tool applies to AI-informed decision-making, but only where AI is a material factor in a decision or decision-making process, and where the decision has a legal or similarly significant effect for an individual.

The HRIA Tool takes the form of questions designed to draw out relevant information for consideration by banks when assessing the human rights impact of their use of AI systems. The questions are accompanied by commentary to explain and contextualise the questions.

The decision to use an AI system must comply with the specific bank's risk appetite and governance frameworks, including all other reviews required by the bank's governance and internal policies.

The content of this HRIA Tool focuses on human rights concerning non-discrimination and equality of treatment, and associated rights to an effective remedy for those affected. Other forms of impact assessment may also be needed.

The development and implementation of AI systems may also need to be assessed against banks' corporate social responsibility standards more generally.

It is important to note that privacy rights in Australia relate to information rights (rather than broader coverage like a tort of privacy), and while an element of human rights, privacy is best dealt with through specific privacy impact assessment methodologies. The HRIA Tool should still assist you in focussing on group privacy protections.

3 Pre-screening

The questions in this section concern the scope of the HRIA Tool and whether an AI system needs to be subject to a human rights impact assessment in the first place.

If AI is not a material factor in decisions, or the use of AI does not have a legal or similarly significant effect on individuals, a bank may not need to consider the full questionnaire. Records should be kept documenting the reasons behind a pre-screening determination.

Not all uses of AI systems meaningfully engage with people's human rights. The HRIA Tool has been designed to be applied to AI-informed decision-making. This is where AI is a material factor in a decision or decision-making process, and where the decision has a legal or similarly significant effect for an individual.

The phrase 'legal or similarly significant effect' means the decision affects an individual's legal status or legal rights or has an equivalent level of impact on an individual's circumstances, behaviour, or choices, such as the automatic refusal of an online credit application.

Q1.1 Does the AI system involve the use of AI as a factor in a decision or decision-making process that is material?

This first criterion goes to whether AI has more than a minor, or assistive, role in decision-making.

Sometimes it is clear that the use of AI is material in a decision because all key elements of the decision-making process are automated. This is likely to be the case, for example, where AI is used in credit decisions, biometric fraud detection or cybersecurity. In other cases, the materiality of AI in the decision-making process can be more difficult to assess.

AI can be used to generate a data point that a human decision maker then relies on to make the ultimate decision. Here, the specific context is important. For example, a human decision maker may record their decision using a sophisticated word processing application that was developed using AI. The application simply records the decision, so this use of AI would not be material.

Regardless of materiality, a bank may wish to consider if any of the HRIA questions assist in the risk assessment process.

Q1.2 Does the AI system involve making decisions about:

- **the creditworthiness of individuals**
- **the eligibility of individuals for banking products e.g., credit cards, loans, insurance**
- **the pricing of banking products**
- **the identification of customers experiencing vulnerability**
- **automated customer advice**
- **collections**
- **any other matter with a legal or similarly significant effect for an individual.**

This second criterion concerns the kind of decisions that are being made and the effects for an individual. That is, whether the decision has a significant effect for an individual or individuals.

Without a detailed understanding of a particular context, it is not possible to be determinative about what uses of AI systems in banking should be subject to the HRIA Tool. However, some uses of AI systems are clearly more likely than others to have significant effects for individuals.

For example, an AI system that decides whether a bank customer should be sent direct marketing communications would generally not be subject to the HRIA. A simple chat bot that uses AI but only to provide basic assistance, like links to help and guidance, rather than to give financial advice, may not need assessment. When making your assessment of applicability, you may still need to consider where/how the data provided is collected and how appropriate responses are (e.g. acknowledging the risk that automated systems may encourage individuals to provide intimacies that otherwise would not be shared).

On the other hand, an AI system that decides on eligibility for a banking product or on product pricing should be subject to assessment. Similarly, where marketing is based on sending pricing offers only to a specific cohort, it may be desirable to subject this to a HRIA.

4 Identifying impacts

The questions in this section are designed to help identify the actual and potential impacts on human rights (non-discrimination and equality of treatment) of an AI system.

The scope of relevant impacts should be broadly interpreted. Businesses with strong ethical principles and a concern for their reputation seek to act fairly. They will consider it important, for example, to assess possible algorithmic bias regardless of whether this amounts to unlawful discrimination under anti-discrimination law.

4.1 Characteristics of the AI system

These questions are designed to help describe and analyse the type of product or service concerned, and related data flows and data processing purposes.

Q2.1 What is the purpose of the AI system? What problem is it solving? What are the main components of the AI system, including the data sets upon which it was trained, and the related product or service?

Q2.2 Where, and to whom, will the product or service be offered?

Q2.3 Who are the operators or users of the AI system?

Q2.4 What types of data are used (personal, non-personal, sensitive information)? How has it been categorised? Has it been collected with consent?

Q2.5 What are the main purposes of data processing?

Q2.6 Who is responsible for data management and processing? Where is the data from? With what other data will it be connected? How will it be stored?

4.2 Analysis of impacts

These questions are designed to ensure that all impacts are considered, and assessment is informed by an understanding of contextual issues (political, economic, regulatory, and social).

In analysing the impact of an AI system, the views of stakeholders should be considered, where appropriate and practicable. For example, could the system impact individuals, or a group, who find it difficult to manage finances and credit or do their everyday banking?

Information from parallel assessment processes may be taken into account in analysing impact.

Q2.7 Has compliance with all applicable legislation and regulations been considered?

Q2.8 Has the impact of the AI system on anti-discrimination rights been considered specifically?

Q2.9 Have potential cumulative impacts affecting the same individuals or groups been considered?

Q2.10 Can potential legacy impacts associated with the AI system be identified (e.g. lack of provision of banking services to disadvantaged communities)?

Q2.11 Does the assessment consider the scope, scale and irremediability of impacts, including for the individuals affected?

Q2.12 What policies and procedures are already in place to assess the potential impact of the AI system? For example, has a previous impact assessment been conducted in relation to specific issues or some features of the AI system (privacy, use of biometrics, data ethics)?

Q2.13 Is there a whistleblowing provision for affected communities or colleagues to raise concerns?

4.3 Acquiring and processing data

AI systems may give rise to algorithmic bias, where one group is treated less favourably than others without justification. AI systems producing outputs that result in unfairness, can sometimes have the effect of obscuring and entrenching unfairness or even unlawful discrimination in decision making.

For example, if AI is used to make home loan decisions and is trained on previous human decisions that were prejudiced against female loan applicants, the outputs of the AI system may replicate or reinforce this discrimination. If AI is trained using the addresses or postcodes of applicants, this data may act as a proxy for ethnicity and constitute discrimination on that protected attribute.

Algorithmic bias may be caused by problems attributed to the data set, the use of AI itself, societal inequality, or a combination of these sources.

These questions are technical in nature and need to be addressed by experts in data analysis in banking uses.

Q2.14 How was training data for the AI system or the third party pre-trained model acquired? Is there potential for biases or patterns in the data collection that are specific to protected groups?

Q2.15 Is there any identifiable risk of label bias in the training data?

'Label' means the value of the target variable in the training data set for a particular person.

Q2.16 What are the protected attributes that are contained in a data set, such as race or ethnic background, gender, age or disability? What protected attributes may be inferred through proxy variables?

'Protected attribute' means an attribute of a person (including age, disability, race, sex), the basis of which is unlawful to discriminate in certain areas of public life, protected under federal, state and territory anti-discrimination legislation.

'Proxy' for a variable (proxy variable) means a feature that is distinct from the variable in question but potentially contains some information about it. For instance, postcode may be a proxy for socio-economic status because certain neighbourhoods are wealthier than others.

Q2.17 Does the data include data points that may act as a proxy for other attributes? Are there any indicators that may identify an individual or group of individuals in a way that might be harmful, or used in a way that might be harmful?

Q2.18 Is the number of data points sufficient to achieve an acceptable level of confidence in the accuracy and stability? Are any groups of people under-represented or under sampled?

Q2.19 Is pre-processing of the training data required to address potential bias? Is there any reason to 'mask' protected attributes in pre-processing?

Q2.20 Are there groups within the data set that may be underrepresented in the data and require special considerations?

Q2.21 Does the data reflect, or is likely to reflect, historical social inequalities?

4.4 Designing the AI system

In many areas of science, the complexity of the solution must be justified against the problem. This ensures a model can be justified and understood, and helps avoid unintended consequences and unexpected behaviour.

Conversely, a simple model may be unable to accurately represent all the heterogeneity in the population and lead to reduced accuracy for under-represented groups.

In AI it is necessary to make a trade-off between these two, and justify the level of complexity.

Q2.22 Is there a simpler way of achieving the same goal? Will increasing the complexity of the model assist in achieving greater equality/fairness/ accuracy?

Q2.23 What is the AI system designed to predict? Given the target is generally a non-numerical value (such as 'profitability') what proxy variables are being used to determine the target?

Q2.24 What type of AI system is appropriate for the intended outcome? Are the impacts of inaccurate decisions appropriately coded into the error function and is it suitable for achieving the intended outcome?

Q2.25 How will the prediction from this model be used to inform the decision-making process? Will there be human oversight? If so, what explanations will be required and by whom (level of seniority and responsibility)?

4.5 Testing and monitoring

Q2.26 What ongoing testing and monitoring is planned to ensure that fairness continues to be assessed?

Q2.27 What fairness measures are used in testing and monitoring? Is fairness assessed though various measures? Have a diverse set of perspectives been considered in deciding on the fairness measures to be used?

Fairness measures such as selection parity, equal opportunity and precision parity may be used as potential indicators of algorithmic bias or discrimination. However, these can rarely be achieved simultaneously, and do not directly measure the fairness of treatment, impact, or opportunity. It is important to go beyond these mathematical constructs and consider the human subject when examining the fairness of a model.

Q2.28 What potential unfairness is revealed by the various fairness measures? Are these results expected or do they demonstrate an unreasonable disparity?

4.6 Algorithmic bias - risk of unlawful discrimination

The use of AI can make it more difficult to determine whether banks have complied with anti-discrimination legislation. Further systems rarely stand alone, and often interact with other systems. It is important to consider any interconnectedness in your ecosystem.

For example, determining whether an individual has been treated differently based on a protected attribute can be difficult where there is a combination of variables entered into an automated or semi-automated system using AI or algorithms.

If no explanation or reasons are produced, this can make it even more difficult, if not impossible, to determine whether unlawful discrimination has occurred.

Q2.29 Will the effect of the AI system outputs result in an individual or group being considered less favourably because of a protected attribute or a proxy for a protected attribute?

Q2.30 Will the effect of the AI system outputs result in an unreasonable requirement imposed, or likely to be imposed, on an individual or group because of a protected attribute or a proxy for a protected attribute?

5 Impact mitigation

The questions in this section concern impact mitigation and remedies. Where human rights impacts are identified, mitigation strategies and management need to be considered if an AI system is to go ahead.

Impact mitigation includes:

- measures to mitigate human rights impacts;
- transparency and right to plain English reasons (or explanation);
- accountability measures ensuring responsibility for use of AI systems, including Non-Executive Director responsibility and risk ownership by Executives; and
- human review to remedy problems and to monitor and evaluate the use of an AI system throughout its lifecycle.

5.1 Mitigation of human rights impacts

Impact mitigation should follow a hierarchy of (i) avoiding the impact; (ii) mitigating the impact; (iii) remedying the impact.

For example, it is preferable to avoid any discrimination on the grounds of sex by not using information about sex, or proxies for sex, in an AI system in the first place. If data about sex is to be used, the impact might be mitigated by ensuring the AI system's data set does not reflect any risk perpetuating, historical inequalities. If discrimination does occur, there should be a remedy for those affected.

Q3.1 Is mitigation of all identified human rights impacts addressed?

Q3.2 In addressing mitigation, are efforts made to first avoid the impact altogether, and if this is not possible, to mitigate and remediate the impact?

5.2 Mitigation of algorithmic bias

Acquiring more data about under-represented cohorts, for example, can help reduce the inequality between current and accurate data and an AI system's data set.

An AI system may be designed or modified to correct for existing societal inequalities, and other inaccuracies or issues in data sets causing algorithmic bias. Furthermore, an AI system may be deployed and operated in a test and learn context to identify opportunities to remedy societal inequalities.

Q3.3 What substantive measures can be taken to mitigate unfair outcomes for individuals due to algorithmic bias?

Q3.4 Can any mitigation strategies be used to remedy algorithmic bias?

5.3 Transparency and right to reasons

Banks should be transparent about the use of AI in decision making.

It is always good practice to produce reasons or an explanation for decisions. This analysis may form part of existing change risk assessments.

Q3.5 Are individuals affected notified about the use of AI in banking decision making where the decision affects their legal or similarly significant rights?

Q3.6 Does the bank provide reasons or an explanation for AI-informed decisions?

Q3.7 Is the use of 'black box' or opaque AI in decision making avoided?

'Opaque' decision making (formerly known as 'black box') describes where a person cannot determine the reasons or basis for the decision.

Q3.8 What is the appropriate level of human oversight and review for the AI system?

It is important to include human review to correct for errors and other problems in an AI system, and for humans to monitor and oversee the use of AI at the system level.

6 Access to remedy

The questions in this section concern access to remedies for individuals affected by an adverse decision made by an AI system.

Q4.1 Is it clear who is owns the risk for the use of an AI system?

While legal liability for decision making may be clear, there are some situations where this may need to be clearly stated in relation to an AI system.

Some complexities can arise where an AI system operates largely autonomously, or numerous parties are involved in designing, developing and using the system.

For example, if a bank is using a program provided by external companies, confusion might arise for the customer regarding responsibility for decisions produced by program, particularly if the program uses external branding.

Q4.2 Is there an internal bank complaints mechanism for decisions made using the AI system?

Q4.3 Does the complaints mechanism ensure that affected individuals are not denied access to external dispute resolution processes, including the courts?

Individuals affected should be given a practical means of appealing to a person or body that can review and remedy AI-informed decisions.

Such a person or body may be internal or external, legal or non-legal, as appropriate, and consistent with existing complaint-handling policies and procedures.

An internal complaints mechanism should meet the effectiveness criteria for non-judicial grievance mechanisms set out in the UNGPs.

Australian Human Rights Commission
www.humanrights.gov.au