Responsible Tech and AI Impact Assessment (IA) guide

Auror.

Responsible Tech and AI Impact Assessment (IA) guide

Overview

While technology and artificial intelligence (AI) can help empower our organisation, our customers and partners, the use of different technologies and AI also has the potential to introduce a range of negative impacts. Potential impacts include to our reputation, our customers, our customers' customers, already marginalised groups, and society at large.

This document is a guide for completing a Responsible Technology and AI Impact Assessment (IA) as part of a Responsible Technology and AI framework. An IA is an accountability tool to help assess possible societal impacts of technologies and AI systems before implementation. It is designed to help us identify, assess and consider how to mitigate any potential negative or unintended impacts of technologies, AI, and other algorithmic systems before deployment, helping to build trust in the use of such systems and maximise their potential for benefit.

The questions in the IA template are aligned with our Responsible Technology and AI Guiding Principles. This IA forms part of our Responsible Technology and AI framework, which aims to encourage both innovation and trust in the use of technologies by our organisation, our customers and partners.

Purpose

The IA template contains a series of questions to help you:

- provide high-level information about your project
- consider the possible consequences of your project, including any ethical and legal issues, potential harms or unintended consequences
- consider how any identified harms can be addressed.

The completed IA will serve as a record of impact assessment to support decision making, facilitate compliance with legal requirements, help us identify ways to make the project more effective, safe and successful and to facilitate a transparent approach to Technology and AI that will help build trust.

Instructions

- An IA should be completed in the Discovery phase or, at the latest, at the beginning of the design phase of a project
- It should be completed by the project lead/product manager with input from a diverse, multi-disciplinary team to get the best results.
- Keep the answers succinct and clear, using plain language. Avoid jargon and/or technical terms where possible and if they must be used, please provide a clear explanation of their meaning. Imagine you're explaining this to a 12 year old!

Definitions

AI/artificial intelligence: An umbrella term for a cluster of technologies and techniques that involve programming computer software to execute algorithms that can recognise patterns, reach conclusions, make informed judgments, optimise practices, predict future behaviour and automate repetitive functions. Examples of AI include:

- **machine learning/ML**, which is AI that learns from historical data and includes sub-fields like Natural Language Processing and artificial neural networks used in image recognition technology (such as facial recognition); and
- **machine perception** and motion manipulation, for instance, used in industrial robotics.

AI model: An AI model is a program or algorithm that has learnt from a set of data to recognise certain types of patterns. This allows it to reach a conclusion or make a prediction when provided with sufficient similar information.

AI system: Any AI-based component, software and/or hardware, often embedded as components of larger systems.

Algorithm: A sequence of instructions telling a computer what to do. Algorithms generally take in some form of input, perform a series of actions or calculations, and may return an output.

Customer: Any purchaser of your product, platform, software, etc.

Customer users: Any user of your product, platform, software, etc., typically the employees of our customers.

Data subjects: Individuals whose information is being collected and processed.

1. Project information

This section of the IA requests background information for readers who may not be familiar with your project . It includes high-level context questions to help inform our thinking on legal and ethical considerations and potential harms in later sections of the document.

2. Data

A significant number of the questions in the IA are about data. Data powers many technologies and AI systems so we need to understand what training and production data we will be using and how accurate and reliable it is. You can use the table below to collect information you'll then use when completing the IA.

	Training data	Production data
Where does the data come from (i.e. what is the source? Which country does it originate from?		
What and who is represented in the data? What and who is not represented?		
When was it collected?		
How was it collected?		
Who collected it?		
Why/for what purpose was it collected?		

When considering the quality of the data you will be using both for training and in production consider potential issues with the data in terms of any of the following:

- ☐ Completeness - Do we have enough data to meet our purposes? Are there any gaps?

- ☐ Timeliness - How recently was the data compiled or updated?

- ☐ Validity - Was the data captured correctly?

- ☐ Accuracy

- ☐ Credibility - Does the data come from a credible source?

- ☐ Consistency

- ☐ Integrity - How well was the data extracted, particularly where multiple datasets are involved?

- ☐ Representation - Is the data representative of the environment the system will be deployed in? Is it sufficiently diverse?

- ☐ Usability - How well is the dataset structured in a machine-understandable form?

- ☐ Organisation - Has the data been appropriately classified and labelled for the stated purpose?

3. Ethical considerations

⚖️ **Fairness and Equality**

When designing, developing and operating AI systems, we will respect:
- **Principles of equality and fairness** so that AI systems do not unjustly harm, exclude, disempower or discriminate against individuals or particular groups.
- **Applicable laws** in the relevant jurisdictions.
- **Human rights** recognised under domestic and international law and specific rights of indigenous people and cultures where we operate.

Transparency, explainability and accountability

We will be transparent about our AI activities, including how and when AI is used in the product/platform/system we build. The operation and outputs of our AI systems will be transparent, auditable and generally explainable in line with their use and potential risk profile so outcomes can be understood and interrogated. We will use our Responsible AI Framework to help ensure the proper functioning of any AI system we develop, operate or integrate with throughout its lifecycle.

Reliability, security and privacy

We will partner with our customers and partners to ensure AI systems and related data are reliable, accurate, and secure and that individuals' personal information and privacy is protected and any risks are identified and managed on an ongoing basis.

Human control and oversight

We will build an appropriate level of human oversight into our AI systems and their outputs as well as implementing safeguards to reduce and mitigate possible misuse of the products or platforms we build. We will use AI to augment, complement and empower our customers and partners, following human-centric design principles that provide meaningful opportunity for human choice.

Community benefit

We will ensure that any AI system that we design, develop and operate or that the product/platform/systems we build integrates with promotes and supports the wellbeing and safety of our customers, partners and the communities in which we operate.

4. Impact & harm analysis

Consider the best and worst-case scenarios that could arise from use of the proposed system after it has been deployed. Detail any expected hurdles or challenges to overcome.

When thinking about affected stakeholders, consider both direct stakeholders (such as customers, users, and partners) and indirect stakeholders (such as data subjects, marginalised groups, regulators, civil society and the general public).

Use the risk assessment table to indicate the expected likelihood and severity of each harm once the proposed mitigations are in place.

RESPONSIBLE AI RISK MATRIX

Impact (of harm to affected person/group)

		LOW Minor impact	**MED** Some impacts felt	**HIGH** Significant impact
Likelihood (of harm occurring)	**LOW** Unlikely to occur in most circumstances	LOW Unlikely to happen/ low frequency with minimal impact	LOW Unlikely to happen but moderate harm will occur	MED Low frequency but would cause significant harm
	MED Could occur in some circumstances	LOW Could happen but minimal impact	MED Could happen with moderate harm occurring	HIGH Could happen causing significant harm
	HIGH High chance of occurring in many circumstances	MED High frequency but low impact	HIGH High frequency with moderate harm occurring	HIGH High frequency causing significant harm

RISK DECISION-MAKING KEY

LOW Risk manageable through existing or proposed internal controls & mitigants	**MED** Requires additional risk treatment actions	**HIGH** Unacceptable
OK to proceed	Proceed only once appropriate mitigants are in place	Place project on hold pending investigation of further potential mitigants

5. Additional resources and reading

Many organisations are thinking about Responsible Technology and AI, and we can learn from and build on these in ways that help us build better products more ethically.

- [Data protection laws across the world](#)
- [Foundations of Humane Technology course](#)
- [Ethical OS - Risk mitigation checklist](#)
- [Ethical Explorer - tools to help navigate the future impact of today's technology](#)
- [Consequence Scanning - an agile practice for innovators](#)
- [New Zealand's Privacy Act 2020 and Privacy Principles](#)
- EU's [GDPR](#) data protection laws
- [Consentfultech.io](#)
- Book: [The Ethical Algorithm: The Science of Socially Aware Algorithm Design](#)
- Book: [Weapons of Math Destruction](#)
- Book: [Shouting Zeros and Ones: Digital Technology, Ethics, and Policy in New Zealand](#)
- Book: [Hello World: Being Human in the Age of Algorithms](#)
- Book: [The Alignment Problem: Machine Learning and Human Values](#)
- Book: [Invisible Women: Data Bias in a World Designed for Men](#)
- Book: [Queer Data: Using Gender, Sex and Sexuality Data for Action](#)
- Book: [Algorithms of Oppression: How Search Engines Reinforce Racism](#)
- Book: [Data Feminism](#)