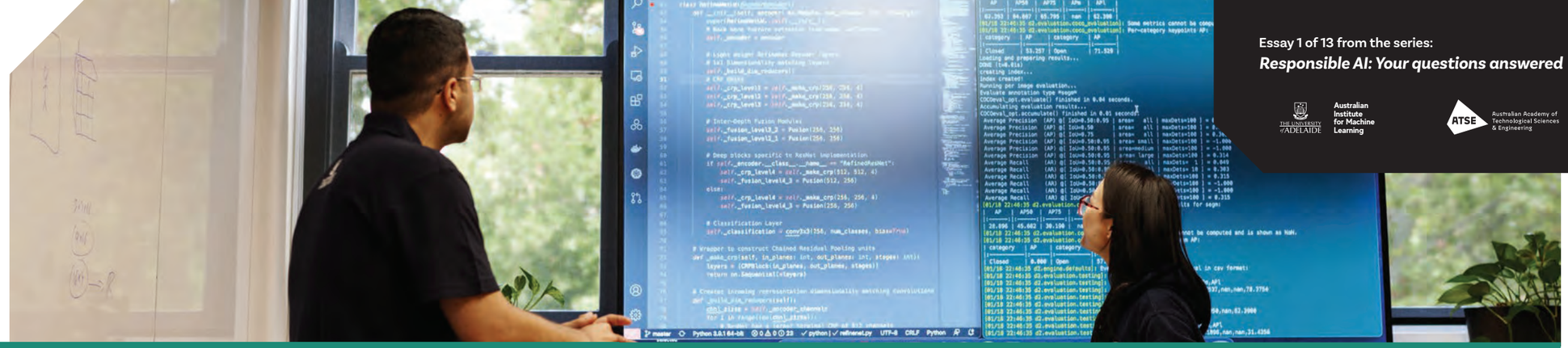

Essay 1 of 13 from the series:
Responsible AI: Your questions answered

The University of Adelaide
Australian Institute for Machine Learning
ATSE Australian Academy of Technological Sciences & Engineering

What is responsible AI anyway?

Professor Jon Whittle
Director, CSIRO's Data61

THE FIELD OF 'responsible AI', which has become increasingly popular in recent months, is intended to put the development and adoption of AI on a sound footing and to ensure that AI has net positive benefits for society.

This is a broad definition, however. This article delves into the details and refines what responsible AI means in practice. In our forthcoming book on the topic,[1] my co-authors and I came up with the following definition:

> Responsible AI is the practice of developing and using AI systems in a way that provides benefits to individuals, groups and the wider society while minimising the risk of negative consequences.

There are many benefits arising from AI. CSIRO's *Everyday AI* podcast[2] presents many examples where AI is benefiting society: Tennis Australia is using AI to assist blind spectators to watch tennis more easily; CSIRO and Google are using AI to help manage invasive species in the Great Barrier Reef; conservationists are using AI to help track biodiversity; at St Vincent's Hospital in Melbourne, AI is reducing the time it takes for women to get the results of breast cancer scans; and AI is also being used as a tool by artists such as the Grammy Award–winning pop duo Yacht, who use AI to help write their music.

This article could go on for pages about all the benefits, but the focus here is on the potential negative consequences of AI and how to minimise them. What are these negative consequences? Unfortunately, there are many.

AI risks can be divided into three categories: risks associated with the **use of the technology**, risks that arise from **how the technology is developed**, and broader **societal risks**.

Risks from AI use

Let's take ChatGPT as an example. Putting aside consideration of how this was developed and its many benefits, two of the well-known issues with ChatGPT are bias and hallucinations:

Bias
It's been known for years that AI systems can suffer from bias. One of the earliest public examples of this was the COMPAS system in the US, which was used to predict reoffender rates at parole boards but was found to discriminate against black people.[3] In the case of ChatGPT, because it is trained on a large proportion of the text available on the internet, and because a lot of that text is biased and discriminatory, ChatGPT can also be biased and discriminatory.
In recent times, significant effort has been put towards trying to reduce bias in AI systems — for example, there are guardrails in ChatGPT that avoid much of the potential bias and discrimination. So one might ask, is ChatGPT an example of responsible AI? It's hard to say — the time and effort put into developing guardrails is evidence of a responsible approach. On the other hand, ChatGPT still suffers from some bias issues, which doesn't look like responsible AI. This example illustrates that responsibility in relation to AI isn't necessarily based on a binary between entirely responsible and irresponsible. Rather, responsibility lies on a spectrum, much like AI itself.

Factual inaccuracies (or hallucinations)
AI systems are never 100% accurate. This is the nature of the technology. Data-driven AI, in particular, applies statistics to look for patterns in data. But statistics, by definition, are not 100% accurate — it's all about probability. So a self-driving car will never perfectly identify obstacles in its path. And an AI-driven movie recommendation system can't always recommend the movie you want to see right now. Of course, AI systems don't need to be 100% accurate to be useful. However, some AI systems need higher degrees of accuracy than others. In the ChatGPT world, these factual inaccuracies have been called 'hallucinations'. Examples range from making up fake citations in academic papers, to providing wrong answers to simple math puzzles, and faking people's bios. So are AI systems that hallucinate irresponsible? Again, it depends on what the AI system is used for. Context matters when it comes to responsible AI.

Risks in AI development

Responsible AI isn't only about the end product. It's also about the *way* the product is designed and developed.

Indeed, one clear principle behind responsible AI is that responsibility must be considered at all stages of a product's lifecycle — from initial concept, all the way through design and implementation, to adoption and use. Responsible AI is like a chain — a weak link means the whole system fails the responsibility test.

In our forthcoming book,[1] we think about this issue in the context of three dimensions: governance, process and product. Each of these are characteristics of *how* an AI system is developed, and the care and rigour applied in each dimension will affect the extent to which the system can be considered responsible.

Governance
If you are an organisation developing or using AI, what governance do you have in place to ensure the AI is responsible? Governance in AI is a huge topic, which can't be fully covered here. But some of the questions organisations should be addressing are: Where is AI used in my organisation? What data is used to train the AI and do we have the right to use that data? Has the data been properly curated to ensure it isn't inherently biased? Have we considered what can go wrong and put in place mitigation strategies, or have we just assumed the AI will work as intended and everything will be okay?

Process
The best governance framework in the world won't save you unless your organisation has rigorous processes in place to monitor whether things are being done responsibly. Are the right kind of people involved in the development of your AI system (i.e., end users, and any relevant external stakeholders)? Have you defined what 'responsible AI' means in your particular context, and is this definition specific enough to be falsifiable? Have you introduced any relevant training for your workforce? Does your culture support responsible AI — is there psychological safety so that people feel free to speak up if needed? Is there a "human in the loop" for any critical decisions that an AI might make?

Product
The final dimension is the AI system — or product — itself. Important considerations here relate to the detailed design of an AI system and ensuring that best design practices are incorporated to ensure responsibility. This could involve, for example, having redundant systems so that a critical system doesn't rely on AI alone, quarantining new AI features until they have been shown to work responsibly in the field, or undertaking continuous real-time testing of an AI product so that responsibility can be monitored.

Broader societal risks

Our first two categories of responsible AI are concerned with AI systems and how they need to be managed at an organisational level. Arguably more important than this, however, is to understand the broader societal impacts of an AI system and whether there are unknown, unanticipated or unintended negative system-level consequences. As a good example, even if ChatGPT were implemented without any biases or hallucinations, the computing machinery needed to train and run ChatGPT for millions of users can have significant negative environmental impacts. Another example of negative consequences concerns those critical minerals that need to be dug out of the ground to build the data centres and mobile phones required to run AI. A third example is the use of low-paid workers in Africa to train ChatGPT, who were asked to label text containing violent, sexist and racist remarks so that ChatGPT could avoid generating such text.[4]

Broader societal risks are the hardest risks to manage. They are often hidden and not talked about. Systems thinking is one way to help understand and manage these hidden risks. Another way is to engage with experts outside the technology disciplines, such as lawyers, social scientists and anthropologists. These experts will typically bring a different lens that can enable hard questions to be asked that otherwise might go unnoticed.

While it is difficult, if not impossible, to guard against all negative unintended consequences, a truly responsible approach to AI development and deployment will involve a rigorous attempt to understand not just the AI system's risks but also the risks arising from how that AI system is used in a broader environmental and social context.

PROFESSOR JON WHITTLE is Director of CSIRO's Data61, the digital and data science arm of Australia's national science agency.

With around 1000 staff and affiliates, Data61 is one of the largest collections of R&D expertise in artificial intelligence and data science in the world. Data61 partners with over 200 industry and government organisations, over 30 universities, and works across vertical sectors in manufacturing, health, agriculture, and the environment.

Prior to joining Data61, Jon was Dean of the Faculty of Information Technology at Monash University.

Essays

SECTION 1: INTRODUCTION

What is responsible AI anyway?
Professor Jon Whittle — Director, CSIRO's Data61

10 examples of AI that are here now and have been embraced by the general public
Stela Solar — Director, National Artificial Intelligence Centre

SECTION 2: WHAT DO WE NEED TO BE TALKING ABOUT?

A unique opportunity for Australia: bridging the divide between fundamental AI research and usable, embodied AI
Professor Michael Milford FTSE — ARC Laureate Fellow, Joint Director QUT Centre for Robotics

Responsible AI means keeping humans in the loop: what are other social implications of the mainstream adoption of this technology?
Associate Professor Carolyn Semmler School of Psychology, Faculty of Health and Medical Sciences, The University of Adelaide and Lana Tikhomirov — Australian Institute for Machine Learning (AIML), The University of Adelaide

AI is changing the way people work: how do we skill our future workforce to ensure these new jobs stay on shore?
Professor Katrina Falkner FTSE — Executive Dean of the Faculty of Sciences, Engineering and Technology, The University of Adelaide

Responsible data management: a precursor to responsible AI
Dr Rocky Chen, Associate Professor Gianluca Demartini, Professor Guido Zuccon, and Professor Shazia Sadiq FTSE — School of Computer Science and Electrical Engineering, The University of Queensland

Open the pod bay doors please, HAL
Andrew Dettmer — National President, Australian Manufacturing Workers Union

Innovation needs to create value: how do we tool universities to remain relevant to industry needs?
Professor Simon Lucey — Director, Australian Institute for Machine Learning, The University of Adelaide

An AI-literate community will be essential for the continuity of social democracy
Kylie Walker — Chief Executive Officer, Australian Academy of Technological Sciences and Engineering

SECTION 3: WHAT ARE THE NEXT STEPS?

What are the limits of current AI, and what opportunities does this create for Australian research?
Professor Anton van den Hengel FTSE — Director, Centre for Augmented Reasoning, Australian Institute for Machine Learning, The University of Adelaide

Australia's unfair advantage in the new global wave of AI innovation
Professor Mary-Anne Williams FTSE — Michael J Crouch, Chair for Innovation, UNSW Business School

The $1 billion dollar question: What should Australia's responsible AI future look like?
Kingston AI Group

What are we doing now to ensure that Australia is recognised as a global leader in responsible AI, and what else should we be doing now and into the future?
Dr Ian Opperman FTSE — NSW Government's Chief Data Scientist, Department of Customer Service

For acronyms, abbreviations and endnotes please see the composite document with all the essays.

Responsible AI

Your questions answered

ACKNOWLEDGEMENTS
The Australian Academy of Technological Sciences and Engineering (ATSE) and the Australian Institute for Machine Learning (AIML) acknowledge the Traditional Owners of the lands on which we meet and work and we pay our respects to Elders past and present. We recognise the deep knowledge and practices embedded in the oldest continuous culture on the planet. Australia's history of engineering, technology and applied science spans more than 60,000 years.

This artefact is produced by ATSE in partnership with AIML. We would like to thank all the experts for their contributions to this edition.

PROJECT TEAM
Eddie Major, Dr Kathy Nicholson, Peter Derbyshire and Suryodeep Mondal

DESIGN AND PRODUCTION
Elizabeth Geddes, Edwyn Shiell and Alexandra Horvat

SUGGESTED CITATION
Responsible AI: Your questions answered. Australian Academy of Technological Sciences and Engineering (ATSE), and the Australian Institute for Machine Learning (AIML) at The University of Adelaide. Canberra, Adelaide 2023

Cover image: An artist's illustration of artificial intelligence (AI). This image represents the boundaries set in place to secure safe, accountable biotechnology. It was created by artist Khyati Trehan as part of the Visualising AI project launched by Google DeepMind. Source: unsplash

Responsible AI

Your questions answered

Australian Academy of Technological Sciences & Engineering

THE UNIVERSITY of ADELAIDE

Australian Institute for Machine Learning