



**School of Computer Engineering Faculty of Engineering
The University of New South Wales**

Ethical Approaches to Bias and Fairness in Artificial Intelligence

by

William Lazaris

Thesis submitted as a requirement for the degree of
Bachelor of Engineering in Computer Engineering (Hons)

Submitted: April 2023

Supervisor: Sebastian Sequoiah-Grayson

Student ID: z5309766

Abstract

In this thesis, I attempt to analyse the current state of research about the problems which bias, and fairness are introduced into machine learning algorithms. Throughout this research I have explored from the foundation of what bias and fairness is on a conceptual level to how it is involved in artificial intelligence, to potential tools and solutions that are available to combat these problems. As a byproduct of this research, I have created proof of concept educational resources which attempt to showcase a way to mitigate the risks of bias and provide more ethical artificial intelligence. With the field of artificial intelligence only going to expand and grow in the future, the significance of understanding this issue is critical to ensure that this technology is used appropriately.

Acknowledgements

I would like to thank my supervisor, Sebastian Sequoiah-Grayson, for allowing me to take up this thesis topic and for guiding me through all the valuable resources, whilst giving helpful feedback. I would also like to thank John Shepherd for assessing my work at late notice and for providing feedback on the seminar presentation.

This work has been inspired by the many academics within the field of AI and Machine Learning who have ultimately provided the foundation for this very young field.

Abbreviations

AI Artificial Intelligence

ML Machine Learning

Table of Contents

Chapter 1: Introduction.....	8
Chapter 2: Background.....	9
2.1 AI and Machine Learning.....	9
2.2 Main Ethical Frameworks.....	9
2.3 Ethical Principles Applied to AI.....	9
2.4 Emerging Themes	10
2.5 Tools for Assessing AI.....	10
2.6 Bias and Fairness in Machine Learning.....	11
2.7 Challenges of the Entire Domain	12
2.8 Possible Solutions to Bias.....	12
Chapter 3: Literature Review	13
3.1 Classification	13
3.2 A Survey on Bias and Fairness in Machine Learning	15
3.3 Putting AI Ethics to work: are the tools fit for purpose?	19
3.4 Tackling Bias within AI	20
3.5 Ethical AI.....	23
3.6 Thinking Responsibly During Development	26
3.7 Impact of AI and its Significance.....	28
Chapter 4: Project Plan & Description.....	29
4.1 Thesis C Term Plan	29
4.2 Expected Outcomes and Outputs	30
4.3 Project Motivations	31
4.4 Description of Project.....	32
Chapter 5: Project Results.....	38
5.1 Infographic	38

5.2 Digital Handbook	42
Chapter 6: Conclusion.....	53
6.1 Conclusion.....	53
6.2 Future Work	53
References	55

List of Figures

Figure 1 Google's AI Principles	10
Figure 2 Heatmap of previous work in fairness grouped by domain.	18
Figure 3 Final Version of Infographic	40
Figure 4 Version 1 of Infographic	41
Figure 5 Website Homepage	43
Figure 6 Understanding AI.....	45
Figure 7 Ethical Principles	47
Figure 8 News Page	49
Figure 9 Further Reading	50
Figure 10 About Us	51
Figure 11 Contact Us	52

List of Tables

Table 1 Thesis C Expected Outcomes and Outputs.....	30
Table 2 Existing Online Ethical Resources	34
Table 3 Criteria Justification	37

Chapter 1: Introduction

In this thesis, I aim to uncover the problems and issues that the field of artificial intelligence (AI) is currently facing regarding bias and fairness within machine learning algorithms. I have broken down and attempted to demystify this concept of bias and fairness within AI to reach the fundamental and key ideas needed to implement effective solutions. Initial thoughts regarding AI would seem to be that an algorithm wouldn't be capable of bias as it is just an entity consisting of code and executing an objective function. However, as more and more machine learning algorithms are implemented and applied in real world situations, it has become more apparent that such biases can emerge leading to outcomes which can be considered unfair. This project aims to understand what bias and fairness is and how they emerge within AI algorithms. Throughout this report a brief background regarding AI and ethics will be introduced as this will provide a basis for the literature that is examined within this report. Some background topics will cover the foundation of what bias and fairness is on a conceptual level, to how it is involved in artificial intelligence, to potential tools and solutions that are available to combat these problems.

The literature analysis will start with understanding classification within AI by analysing chapter 4 of the Atlas of AI by *Kate Crawford*. I will then move onto understanding what bias and fairness is through A survey on Bias and Fairness in ML by *Mehrabi et al.* Potential tools for assessing AI will be examined with the article Putting AI Ethics to Work by *Jacqui Ayling* and *Adriane Chapman*. Other ways of addressing the issue of bias within AI will be considered through the article Tackling bias in AI by *Silberg and Manyika*. This will be followed by a case study: Racism in Medicine by *Poppy Noor*. Finally ending with possible solutions to bias with the article Identifying Bias in AI using simulation by *McDuff, Roger Cheng, and Ashish Kapoor*. The literature review will also cover topic areas including thinking responsibly during development and the impact of AI and its significance within society.

With the field of artificial intelligence only going to expand and grow in the future, the significance of understanding this issue is critical to ensure that this technology is used appropriately, and I believe a greater sense of fairness of outcomes will come because of more people being aware of the risks and problems that arise with AI. To help create awareness, part of this thesis involved the creation of online educational resources designed to inform the community about AI and more specifically how to ensure it is ethical. The final goal of this report is to present an understanding of the current landscape of bias within AI and showcase the educational resources I have developed.

Chapter 2: Background

2.1 AI and Machine Learning

One of the new and exciting technologies within computer science is AI and machine learning (ML). It has rapidly developed over the last decade and has started having many applications from recommender systems to medical diagnosis as well as streamlining and automating many business processes. There are many types of ML algorithms ranging from the simplest, regression analysis, to the more complex and cutting edge, neural networks, and deep learning. The basis for this technology is data, as these algorithms require very large datasets to be effective. As big data becomes more prominent and grows over the coming years, so too will the applications and the impact that these algorithms will have on the world.

2.2 Main Ethical Frameworks

Part of this thesis requires having a foundational understanding of normative ethical theory. The three main branches consist of, utilitarianism, Kantian ethics, and lastly virtue ethics. These three ethical theories form the basis for almost all ethical arguments and will prove useful in providing insights to the ethical principles of AI.

2.3 Ethical Principles Applied to AI

When we apply ethical principles specifically within AI, some of the questions that are raised include:

- Is AI fair and does it provide the same benefit to everyone? Does it need to?
- Are there dangers with autonomy, and can it make the situation worst?
- Will AI remove any risks of human error or introduce new risks to society?
- Are the current systems and approaches transparent?
- Are the current regulations and standards suitable?
- Should we use AI at all, or should we avoid using it?
- Will machine learning help with this problem of bias or make it worse?
- Is the integrity of the methodology in design AI compromised?

For example, Google has their own set of principles which they have outlined in which they try to follow to ensure that the AI they develop is ethical. These principles can be seen below in figure 1.

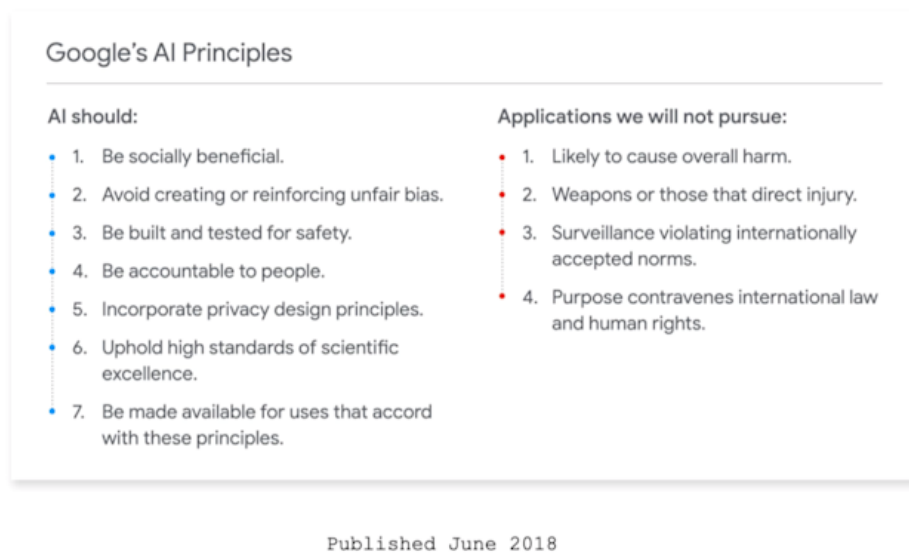


Figure 1 Google's AI Principles

2.4 Emerging Themes

Within *AI and Ethics* article by *Sebastian Sequoiah-Grayson and Toby Walsh*, some of the emerging themes within the ethics of AI are identified, these include:

- Ethics Washing, where large corporations try to justify their ethical standpoint in a way that favours the systems they are producing.
- Extractivism, in which through the use of acquiring some knowledge through data, an advantage can be made as a result of this knowledge.
- Power, in which a motivating factor for extracting this knowledge, is in order to gain power over another and benefit a personal or specific agenda.

2.5 Tools for Assessing AI

Since the middle of the 00's there have been three different waves of AI ethics, each taking a different approach and highlighting different concerns regarding the rapid growth within the field.

First wave of AI Ethics:

- Black box function is epistemic concern in which we don't fully understand the mechanisms.
- Fairness is normative concern.

Second wave of AI Ethics:

- How do you appropriately apply the mechanism or algorithm.
- Protocols and use of AI.
- Engineering design process and how it should be considered.

Third Wave of AI Ethics:

- Ethics washing.
- Reducing moral dilemmas to design issues.
- Avoidance of responsibility.

2.6 Bias and Fairness in Machine Learning

Bias in AI is simply a heuristic or assumption that is utilised to make processes easier and is relied on within algorithms to act efficiently. At first glance, bias does not seem like an issue at all, but the problem arises when one of these heuristics is used in a way that produces or reinforces an unwanted outcome, one in which we would not want within a society we are trying to build. Throughout this thesis I will refer to this bias as ‘negative bias’, in which a bias produces an unwanted outcome.

As a machine learning model is a product of the data it uses, the outcomes and results will ultimately reflect and showcase characteristics from the data set. In this case, data has inherent biases due to many factors including: how data is collected, as well as being a product of society in which humans themselves are inherently biased.

Fairness can be generally defined as the act of making a judgement or decision without favouritism or preference. Although a problem with defining fairness is that there is no singular concrete definition for the idea. Despite this, there are ways to ensure that ML algorithms are fair most of the time.

One way to ensure fairness is by determining whether any biases present in the output, are not negatively biased and will not produce a society that we would not want to live in. Some types of biases include data based bias, algorithmic based bias, and user generated bias. These three types of biases ultimately form a feedback loop in which any initial bias will be reinforced over time.

An example of a biased AI is the COMPAS algorithm. This was a machine learning algorithm designed to predict the likelihood of recidivism of a criminal. The results of the algorithm showed that there was a higher number of false positives for African American offenders. The algorithm did not even use a variable of race as a metric for prediction, however this skewed result was a byproduct of unwanted bias from the rest of the data set. It is clear that even if we specifically try to design an algorithm in a certain way, bias may still be introduced from the data. The algorithm was no better than that of human judgement and this poses the fundamental question of what are the underlying motivations for AI and decision making algorithms?

2.7 Challenges of the Entire Domain

Distinguishing between equal opportunity and equalised odds

- Equal opportunity is having access to the same opportunities.
- Equalised odds is the likelihood that it occurs or the rate of a positive/negative outcome.
- For example, everyone having access to a melanoma detecting AI would be equal opportunity
- That melanoma AI having the same success rate regardless of your characteristics is equalised odds.

Creating a concrete and universal definition of fairness that everyone agrees upon. Another challenge is a lack of definition and searching for unfairness to understand if an algorithm is also unfair and not just fair. AI also being a black box in which we know the inputs and outputs but don't fully understand the intermediate process that a ML algorithm uses to produce that output. As a result of this, we may face challenges and problems that are not yet known or fully understood. Although in contrast to this I would argue that humans are also a black box, yet we accept it and let ourselves and others make important decisions that could have a great impact on others. Perhaps the problem is not that AI is a black box but the danger of autonomy. I believe that the autonomy of decision-making AI will lead to a loss of accountability and ownership of mistakes, as people will seek to use AI for taking the blame for poor decisions. The motivating factor to use AI also poses the question of whether people care more about an outcome or the reason for a decision?

2.8 Possible Solutions to Bias

One possible solution is "Fairness through awareness" in which a greater awareness of the problems by society will lead to a better understanding of how we might tackle and solve the problem. It forefronts the idea that you first must be aware of the issue before any progress can be made fixing it,

Another outlook to consider is whether the solution to bias is inherently a technical problem or is there a bigger more fundamental issue at stake? Possibly the technical solutions are limited in capability and a more productive path would be understanding the impacts of these models and knowing how to best utilize them instead of relying on them blindly.

Some ways to maximise fairness and minimise bias could include:

- Establishing processes and practices to test for and mitigate negative bias.
- Fully explore how humans and machines can work best together.
- Invest more in bias research and adopt a multidisciplinary approach.
- Invest more in diversifying the AI field itself.

Chapter 3: Literature Review

3.1 Classification

Within *The Atlas of AI* by Kate Crawford, the issues regarding the use of machine learning in a social context are discussed, specifically concerning the technique known as classification. Classification is a machine learning tool in which the algorithm uses the data to determine whether something belongs to a specific group or category. Crawford outlines an example of classification from the 1800s in which over a thousand skulls were measured and observed and classified into one of five “races” of which was defined by the scientists at the time. The example of classifying skulls was used as way to demonstrate the potential dangers and bias within classification where the classifier, the scientists in this instance, will tend to categorise based on pre-existing notions of what those categories should be. Crawford states that “explorers used this as a justification for racist violence and dispossession” (2021, p.125). This presented the problem that the concept of classification is inherently dangerous, in that although classifying can be true it can also be unfairly presented and used, and even an objective method using mathematics and statistics will still have this fundamental problem.

We have now come to the root of the problem of classification, in which the machine learning algorithm being inadequate is not the only problem, but rather the underlying problem that classification can provide an unfair epistemic advantage. The ability to divide and name things such as in the case of the scientists classifying skulls into groups is not inherently oppressive but using this knowledge as a source of power as a way to segregate certain races is where the threat emerges. Crawford uses some modern examples including Amazon’s algorithm used to read job applications and determine a suitable candidate. In this case, the algorithm was heavily skewed towards selecting men over women as it had been trained using real world data with imperfections, including the fact that Amazon tended to hire more men than women in the past.

Although Amazon claimed that this was just a bug and was corrected it still poses the issue that no one is willing to discuss the problem of classification having the potential to be exploited for the gain of power. Major players in the tech industry including Amazon and Google, keep their algorithms very discrete, not allowing much space for criticism. Real world data in this instance is not reflective of fairness, however it still can be true data, this poses another greater issue of what we deem to be fair, and what companies deem to be fair when tuning their algorithms.

One of the final examples of classification given was ImageNet in which a dataset of images was classified into 9 major categories with thousands of sub-categories each being a subset of another category. The major take away from this example was that bias within data is determined at the stage of determining data types to

collect, in that at the point of collection someone still needed to determine the specific classes and inadvertently introduced bias through this sampling. Each data point is a discrete classification within itself, and the bias of each singular point eventually compounds to produce artificial biases politically and culturally. Adding onto this, *Crawford* mentions the linguist George Lakoff, in which he defined nouns being on a spectrum from concrete such as an apple or abstract such as health in which the latter is not a physically tangible item. However, this idea is lost when data is collected, as everything is forced into pre-built structures with little room for differences.

Crawford ends the section of classification by presenting some major questions including, what if anything is considered safe when classifying people? This idea homes in on *Crawford's* perspective that “classificatory strata in training data and technical systems are forms of power and politics represented as objective measurement” (2021, p.147). Although I agree that it is potentially hazardous to use classification in a social situation, the idea that classification is not objective and only used for power is only true in some instances. In order to know something, you need to make a judgement or classification at some point, and this is still objective despite it also having the potential to be used for power.

One of my major criticisms of this chapter is a lack of presenting any positives from which classification has emerged successfully. *Crawford* focuses on the social and political impacts of the classifying of people into specific groups which itself is a big issue outside of machine learning. Although I agree with a lot of the sentiment and issues that are raised including questioning the foundation of the use of this technology, I think more of an effort needed to be made to outline that this is only within social situations, such as hiring a new employee or labelling an image with abstract elements. Currently there are many positive and successful uses of classification, especially within the health information industry. It has been used for tumor detection for breast and brain cancers, along with early detection of Alzheimer's.

Despite these achievements within health, the future of health informatics will revolve around DNA sequencing and analysis. This itself holds both immense potential and the greatest risk for exploitation of the individual. Like the initial example of using skulls to determine race, the potential to use DNA to separate humans is an extremely likely outcome if not carefully considered and managed. In this instance the premise of the problems of social classification will eventually crossover into the world of health sciences.

Another criticism I have of *Crawford's* writing is the perception that the lack of public knowledge of multinational tech companies is necessarily a bad thing. Open-source software can be a viable option for developing new technologies in some instances as it promotes greater collaboration within the community, it can also be restricting in that many choose to copy rather than innovate. The view that large companies such as Amazon and Google are private with the intent to be malicious is not a viewpoint that I share, many of these large companies need to remain private for both security reasons and for having invested a lot of time

and money into developing these technologies. I believe the best course of action in trying to discuss these issues of classification is to not make these large companies out as villains, but to work collaboratively in any capacity while allowing them to remain private.

From a utilitarian perspective, it could be argued the gain they get from the use of these classifications has proved to be more valuable than the problems they have caused so far. Perhaps the ideal world is one that uses classification despite its shortcomings. Technology is still very young and, in its infancy, thus, to completely grasp how significant it will be in discrimination and bias against individuals, or how effective it will be in finding the ideal solution.

Overall, *Crawford* raises an essential question of what is at stake when we classify anything, whether it be through machine learning or through human interactions. These outcomes need to be considered as our past history has shown that if not used effectively it can lead to misguided outcomes despite the right intentions. “The Power to decide which differences make a difference” is a quote from *Crawford’s* chapter which I believe ties all the ideas nicely, being that we need to know what and why we use this technology before using it. *Crawford* pushes that power is used for negative consequences, but I believe that not all acts of power are an act of violence or misused.

3.2 A Survey on Bias and Fairness in Machine Learning

In this reading *Mehrabi et. al.* has listed many different sources of bias and have aimed to create a taxonomy for fairness definitions. One of the major examples that is recurring throughout this reading is the use of the COMPAS case example where the algorithm, designed for predicting repeat offenders amongst criminals, produced a higher number of false positives for African American offenders. Currently this algorithm is not any better than the judgment of a human using basic statistics at the moment.

In order for us to address bias we need to understand the root cause of it within a particular system, before being able to address it with a viable solution. In some cases, we have a duty to assess whether these tools are indeed worth using as they affect real world people, and this sentiment can be lost behind the numbers and statistics. *Mehrabi et. al.* provides a background regarding two toolsets, Aequitas and AI Fairness 360, that have both been developed as tools to help understand how fair a particular algorithmic model is respective to a specific population.

Mehrabi et. al. makes a clear point to identify and list the major sources of bias and types of bias within algorithms. They start by outlining that data is the foundation of AI and its applications, thus data that contains biases will also mean the algorithms will learn these biases. Apart from data itself, there are also design choices which can cause algorithm-based biases that are reinforced even if the data is unbiased. The last form

of bias is from user interaction in which it forms a cyclical loop with the former two types, being that user interaction will affect data collection, which in turn will affect algorithmic design and finally affect user interaction once again forming a feedback loop that perpetuates any initial biases. Some of the biases from each category are listed below.

Data based biases:

Measurement bias - Revolves around how we choose and measure features.

Omitted variable bias - critical variables which are left out.

Representation bias - process of population sampling, it's impossible to sample everything so specific design is used to collect.

Aggregation Bias - False conclusions are drawn due to not observing the entire population.

Sampling Bias - Nonrandom sampling.

Longitudinal Data fallacy - Cohorts overtime.

Linking Bias - misrepresentation due to interactions with other users such as on a social media platform.

Algorithm Based biases:

Algorithmic - Added purely by the algorithm.

User Interaction - presentation can lead users to select one thing over another, and ranking can misguide users to think the top ranked is the best result.

Popularity - If something is popular it is seen more often and reinforces popularity.

Emergent - Results when population or culture changes within society after the algorithm is completed.

Evaluation - Inappropriate benchmarks.

User based bias:

Historical - Pre-existing real-world bias.

Population - User characteristics are different from the original target population.

Self-selection - A variation of the sampling bias where the user has the control to influence the data such as an election poll on a news site.

Social - Other actions impede judgment.

Behavioral - Different user behavior across platforms.

Temporal - Change in population and behaviors over time.

Content Production - Difference use of language across different demographics.

All three types of biases, however, are very much intertwined. All leading into one another and forming a cyclical nature. *Mehrabi et. al.* also discusses another form of unfairness, that be being discrimination, which can be considered either explainable or unexplainable. The former is when “differences in treatment and outcomes for different groups can be justified via some attributes” (*Mehrabi et. al.*, 2021, pg10), whereas the

latter is “discrimination toward a group is unjustified and therefore considered illegal” (*Mehrabi et. al.*, 2021, pg10). Although *Mehrabi et. al.* provides these definitions to both types of discrimination, to justify discrimination as being explainable seems like wordy way of stepping around the problem of discrimination.

Mehrabi et. al. proceeds to discuss the different ways the concept of fairness has been operationalized and studied over time. The main idea is that there is no universal definition of fairness, as differences in culture and values lead to different ways of determining what is considered fair. Despite this there are 10 widely used definitions which are used. These include:

- Equalized odds.
- Equal Opportunity.
- Demographic parity.
- Fairness through awareness.
- Treatment equality.
- Test fairness.
- Counterfactual fairness.
- Fairness in relational domains.
- Conditional Statistical Parity.

There are also different categories for the above types of fairness, including: individual, group and subgroup fairness. *Mehrabi et. al.* goes on to explain that these definitions of fair can be almost impossible to always satisfy as many will not overlap.

Mehrabi et. al. spends a great deal of time explaining the many different machine learning approaches and how fairness manifests differently and ways the bias can be reduced. There are 3 main types of methods to target bias:

Pre-processing - In which the bias is managed prior to the algorithmic alterations.

In-Processing - In which the bias is managed during the algorithmic process.

Post-processing- In which the bias is managed by post algorithmic process.

Mehrabi et. al. concludes that the 10 types of algorithmic fairness is still a new area and requires a lot more research to determine the best way to remove bias from each algorithm and make it as fair as it can be. A lot of research is still needed, and some areas have been neglected. A figure used within their literature, seen below, shows the machine learning algorithms which are receiving the most amount of time and work put into finding solutions to fairness. Primarily Group classification is receiving the most, however this is also one the largest impact areas so that would be reasonable to expect this current outcome.

Going forward the goal would be to have bias reduction using different benchmarks such as using AI Fairness 360 by IBM which has implemented many of the explained methods already. Although AI360 exists to assess how bias a machine learning method is, it still does not target the challenges that overarch the entire domain including:

- Creating a concrete definition of fairness.
- Distinguishing equality from equity.
- Searching for unfairness to understand if an algorithm is also unfair and not just fair.

Perhaps, these problems will never actually be solved, and our time would be better suited finding more practical solutions such as AI360 which somewhat help manage the problem of bias. Perhaps various understandings of fairness need to be used and a somewhat average could be used as a practical approach. Regardless, a greater emphasis on considering these problems is essential to make any progress.

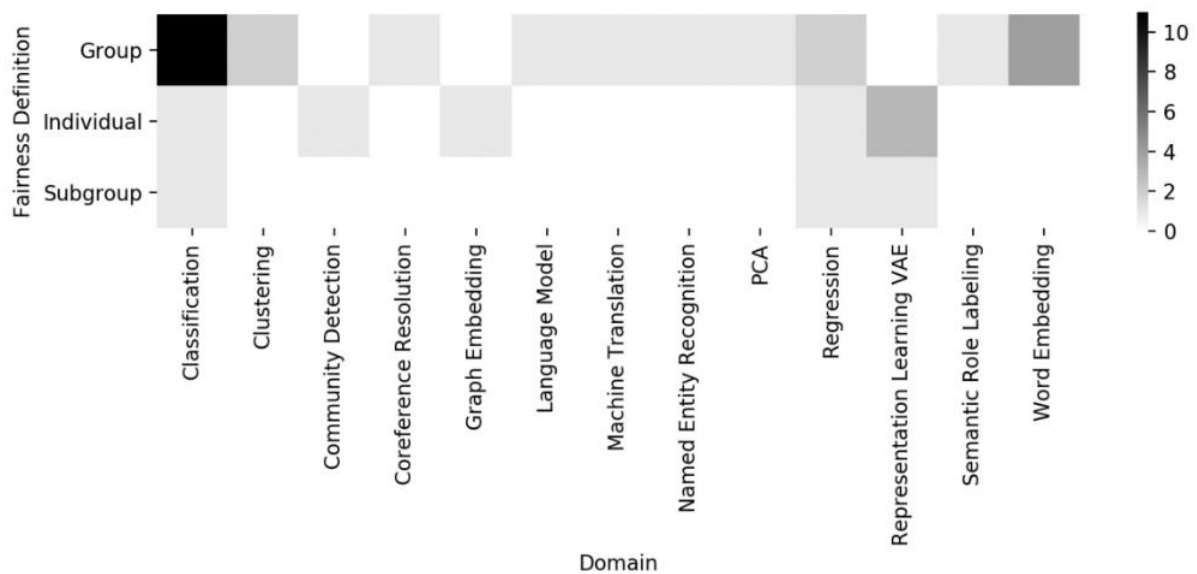


Figure 2 Heatmap of previous work in fairness grouped by domain.

3.3 Putting AI Ethics to work: are the tools fit for purpose?

Ayling and Chapman make a practical breakdown of ethics within AI and initially pose a major question of what does fairness and justice look like in real life? *Ayling and Chapman* propose that there have been two initial phases of ethics in AI with 2016-19 being the first phase, dealing with high level ethical principles focused on a purely philosophical approach as opposed to a legal or technical approach. The second phase focuses on a more technical approach, with this idea of “ethical by design” being a driving factor amongst engineers and scientists. The current phase is focused on governance mechanisms, regulations, impact assessment, auditing, and standardization. *Ayling and Chapman* also consider *Crawford's* view of “how power is wielded through technology” (2021, pg. 2).

Ayling and Chapman give some background into impact and audit practices, outlining the practices of nine different assessments.

These included:

- Impact assessment which is based on relevance, evidence and normative claims.
- Technology Assessment
- Environmental Impact assessment
- Social and human rights impact assessment
- Privacy and data protection impact assessment
- Audit
- Risk Assessment and techniques
- Stakeholder theory and participation
- Technical and design tools

The methodology for reviewing different documents was using a qualitative method of understanding key information on how each of these tools are being used. Some key findings were that the concept of big data was more prominent in earlier years of 2017 to focusing more on technical concerns including algorithms in the later years of 2020. The stakeholder types are directly related to the phase in which AI is being used, for example developers were distinctly related to the product development whereas the reporting tools were more in line with the decision makers using the tools. Another key takeaway was the exclusion of the customer within these tools in terms of assessment or auditing. This heavily leads to the finding that most tools were designed and used as a way of internal self-assessment rather than assessment to comply with other standards. As there is limited oversight from external factors, there was never any need to process the results of these tools to the wider public.

Three key areas were identified in where these tools are being developed, that being impact assessment, audit and technical tools. Overall *Ayling and Chapman* did a good job of creating a framework to assess the current status of AI and data ethical guidelines. The results showed that current guidelines are centered around the product development phase, primarily focusing on the development and delivery of an AI powered product. Another major development was the finding that some stakeholders were given little participation including the users and the voiceless. Almost every tool was for internal use, limiting accountability from outside standards, only needing to satisfy the IEEE. *Ayling and Chapman* end by noting that there is no official regulation regarding impact assessment or auditing, thus limiting the need of people and companies adopting these into their tools and applications.

3.4 Tackling Bias within AI

As the use of AI and machine learning increases within social areas including criminal justice, hiring and healthcare, there also needs to be a greater emphasis on addressing bias and fairness within these algorithms. *Silberg and Manyika* in their article *Tackling bias in AI*, define and address the many issues regarding bias and how we might be able to manage the problems that arise as a result. They start by questioning whether AI will be less biased than human judgement or if it will actually make the problem worse. *Silberg and Manyika* define bias and fairness as a form of preference that has unwanted and undesirable traits that can lead to a negative outcome for a specific individual or group. It is important to note that this definition of bias that *Silberg and Manyika* have used is a very colloquial and everyday use of bias. It should be differentiated from the technical use of the word bias, in which it is used as a heuristic and a metric that is tuned depending on the use of an algorithm.

AI can reduce bias but also scale it up, in that algorithms are very good at reducing the subjective influences and thoughts humans may have about someone that might not have any influence on the bearing of a decision, however as these algorithms use a lot of data from the real world, any societal and systematic biases that have been collected as a result will scale up far greater than if a human was to use that information. This links to the idea that the underlying data is a primary source of bias amongst machine learning algorithms, as the process of data collection is a form of classification itself performed by an individual with inherent biases.

Another major issue *Silberg and Manyika* bring to our attention is the problem of defining fairness. There are dozens of ways to define fairness each with their own metrics, so finding a universal one-size fits all models is somewhat problematic. Despite this problem it is one that needs to be addressed with a possible solution being those different cases having to define a relevant fairness metric for their own situation. Perhaps the best way to move forward is to use these tools as a way to augment our current decision-making methods and have them act as a source of recommendation rather than of an absolute decision. Human judgement will always be a critical skill that is required to make fair and informed decisions. With the introduction of these

algorithms that seek to replace our judgement, we must recognize that our efforts are still required but now shifted to making educated and informed decisions about the design and use of these algorithms. It is important not to rely on these algorithms alone, as they are still a product of personal human judgements and should probably be used to augment our current processes. With this in mind, perhaps a greater emphasis is needed on how humans currently make decisions and whether we need to hold ourselves more accountable and to a higher standard. *Silberg and Manyika* provide some suggestions on how to maximize fairness and minimize bias from AI including the following:

- Be aware of the contexts in which AI can help correct for bias as well as where there is a high risk that AI could exacerbate bias.
- Establish processes and practices to test for and mitigate bias in AI systems.
- Engage in fact-based conversations about potential biases in human decisions.
- Fully explore how humans and machines can work best together.
- Invest more in bias research, make more data available for research and adopt a multidisciplinary approach.
- Invest more in diversifying the AI field itself.

An example of the dangers with the use of AI in medicine is outlined in *Poppy Noor's article Racism in Medicine*. *Noor* outlines a healthcare inequality of melanoma detection and diagnosis amongst people of colour being less than that of white skin. *Noor* explains that a lot of mole checking guides and training programs predominantly use images of moles on that of white skinned subjects. Although using white skinned subjects is not inherently racist, it systematically makes it harder for people of colour to be diagnosed with melanoma as doctors are less trained for this situation. As a result, “black people are less likely to get melanoma but more likely to die” (*Noor*, 2020, p.1) as they are usually diagnosed later as doctors are less experienced with identifying moles on dark skin types. Due to the less emphasis on melanoma detection for people of non-white skin types it creates a systematic racial health inequality.

Machine learning is suggested to solve this issue, where “policy makers in the NHS see automated diagnosis as an easy fix” (*Noor*, 2020, p.1). Although *Noor* argues that this easy fix may not be as good as it sounds. Historically the NHS research has not been representative of a diverse group of people and there is no current requirement for data to be representative and no system for companies to upload data in a way to ensure diversity. With a lack of diversity in data, the problem of diagnosis of melanoma amongst people of colour is not solved as the machine learning models themselves will still be trained on images of white skinned subjects just as the real-life doctors were. This just scales up the problem and provides a false sense of security with these models being just as problematic for people of colour.

Another problem *Noor* outlines, is the inability to produce more data effectively to adhere to the scientific method of repetitions. As the NHS is somewhat profit driven, “the drive to make a profit outweighs the scientific method” (*Noor*, 2020, p.2) which is a very hazardous way to develop a reliable tool which is responsible for the health of an individual.

Another example of this inequality being overlooked is with an AI, trained to detect melanoma, created by scientists at Stanford university and having it praised without anyone questioning if there are any flaws for people of colour. Although the AI was successful for those of white skin types, it vastly overlooked any other cases, and it is concerning that this could be peer reviewed and then praised with there still being a fatal flaw being overlooked. This shows that these products could be misleading if data is withheld and may provide a false sense of security.

Some solutions proposed by *Noor* included:

- Ensuring that data is representative of multiple groups of people.
- Reports on data quality.
- Lack of diversity limits market scalability, there is a monetary incentive to fix this.

A possible solution to the problems seen above is using simulations to generate life-like data to test and train models. The advantages of using this method would be to find blind spots of these machine learning algorithms and be able to rectify them before they are able to become a problem within society.

The use of simulations also provides a way to test the current performance on whether it is effective or not without needing to use real world data. In the article *Identifying bias in AI using simulation*, the authors attempt to produce facial data to test facial recognition software and see how effective it is. The authors address this idea of “fairness through awareness as presented by Dwork et al (2012)”, in which in order for us to eliminate these biases from the algorithms and attempt to create it fairer, we need an effective way to first know how and why these biases are forming and what they are a result of. Essentially it is a diagnosis tool to see if there is anything that the model is missing or could be improved. This is essentially the process of engineering iterative design where people would identify what's wrong with a model and change it based on the results, except in this instance it is also using simulations to carry out this process.

Overall if we are to try and stop and solve the problem of bias and fairness within AI, we must be proactive in taking the necessary steps not to overlook the origins of these biases. Just like in the example of melanoma detection, it was a matter of recognizing a lack of representation while training both people and AI models to be able to attempt to solve this problem of health inequality. Our ability to assess and find the problems in our current methods is how we can progress and develop, regardless of whether it is human, or AI based methods.

3.5 Ethical AI

Vousinas et. al, do provide an overview of the current state of artificial intelligence and the major ethical dilemmas that both exist today and may come to fruition in the future. The article also attempts to assess the challenges brought upon by AI and ways that these issues can be managed and dealt with.

Vousinas et. al start drawing upon another author's definition of AI in which “AI is defined as a system's ability to interpret external data correctly, to learn from such data and to use those learnings to achieve certain goals and tasks through flexible adaptation” (*Kaplan and Haenlein*, 2019). Following this, the author gives their perspective on whether AI is beneficial or detrimental to society, in which they go on to consider that AI has both positive and negative consequences, which they address in further detail throughout their article. They summarize some examples of the benefits including applications within smart cities, learning environments, agriculture and transportation. They conclude their introductions by providing the main purpose of their article which is to provide a roadmap of ethical dilemmas concerning AI and promote discussion amongst the major problems.

Vousinas et. al move on to address the problem of automation and how it may “lead to high levels of unemployment” (*Vousinas et. al*, 2022). They address that many domains including finance and healthcare are at a high risk of having a lot of jobs within them be automated and replacing the human based workforce with a more autonomous workforce. Although this is a very big concern of AI, the article they quote by Frey and Osborne is one from 2013, almost 10 years ago, and since then the unemployment rate has not drastically increased, in fact the overall unemployment has slowly been declining in both Australia and United States (Although there was a large spike in 2020 due to COVID-19, however it has dropped back down since then). *Vousinas et. al* also mentions that Finland has implemented “AI trainings”, however no detail of what this entails is included and whether this is actually something that needs to be considered. It is worth noting that technology has always removed jobs but created new ones, electric trains for example saw many jobs lost but also many new jobs were created as a result.

From this the article then shifts its focus to discrimination issues and inequality of wealth distribution. *Vousinas et. al* mentions that a “manipulated AI-driven price may lead to higher paying rate for a specific group of people” (*Vousinas et. al*, 2022), however no example of what this means is given nor is there any explanation on what is meant by manipulating an AI. *Vousinas et. al* also mentions governing bodies that seek to address discrimination including the Council of Europe and the European Commission against Racism and Intolerance. Although the author mentions these bodies, no indication on how or what they do to mitigate discrimination is mentioned, neither are any specific examples are given. From this it makes me question whether these governing bodies are actually as effective as they hope to be or are they just ineffective systems.

The topic of manipulation of human judgment and behaviour is then considered, with the idea that AI can affect an individual's preferences regarding consumption and also when voting. They somewhat address the idea that social media is a system set up and designed to lead individuals in a way that will keep them using the specific app or service and as a result of this system, "hinders the development of personality and emotional intelligence" (Vousinas *et. al*, 2022).

Following on from the limited assessment of manipulation, roboethics is then focused on, with Vousinas *et. al* describing this potential risk of a dystopian future as a result of systems developing to a very high level of complexity and result in "widespread exploitation in society" (Vousinas *et. al*, 2022). Although I agree that AI is growing quickly and the complexity is advancing rapidly, this idea that we will end up in a dystopian future is unfounded and has no evidence to support this claim. Vousinas *et. al* does bring up an interesting point that there is a technological shift of having robots which are more specific and personalized approach with the scope of these autonomous systems. For example, autonomous vacuum cleaners are designed for a specific task rather than a multitude of tasks, and this allows the design to be more streamlined resulting in a product which is more economically viable. Another example of this specific and personalized AI is self-driving vehicles, in which the AI is trained specifically for the singular purpose of driving. Vousinas *et. al* mentions the dilemmas of autonomous vehicles, posing the problem that a vehicle will have to decide "which life to save" if the trolley problem was to come up in real life. Although this problem of which life to save is an ethical dilemma, there is an even greater ethical dilemma of whether self-driving cars despite some flaws would save more lives overall.

A section in which I think could have been expanded on was the section on the mistakes of AI bias. Vousinas *et. al* elaborates that as AI is a technology developed by us as humans, it is also vulnerable to some of the shortcomings that we possess. This is due to the fact that the foundation of an algorithm is its data, and if data collected is heavily skewed in favour of some demographics, then these indicators will then appear in the final model and result in a biased algorithm. Expanding on this idea of indicators, if an AI is dependent on a specific variable, then if that variable is taken away the system may fail, and this will show that the model may be over reliant on that specific piece of data. The example given is a neural network design to classify cows, however the algorithm became over reliant on associating cows with grass, that when there was no grass in an image, it failed to classify the animal as a cow.

Linking to the previous idea of data and its impact on AI, Vousinas *et. al* link to this idea of data privacy and protection. They claim that the privacy of individuals is increasingly violated due to the influence of technology as well as being forced to expose their personal data. I would have to disagree with this sentiment as although many social media and other tech companies collect a lot of data, every individual still has a choice to use a specific platform or technology, it is just that most people are willing to accept the idea that the use of a technology outweighs the negative impacts of data collection. They also then mention the General Data

Protection Regulation (GDPR), however no mention on whether this organization is a viable solution to the problem they have described.

Vousinas et. al also mentions this idea of being able to control AI, and that we are reaching a point, the singularity, that we will eventually no longer be able to control it or understand its goals leading to possible human extinction. Although this idea gets a lot of hype in the media, this is still just pure speculation as we have no idea what the uppermost limit of AI is or how it will act at the limit. I personally think the more dangerous aspect is how humans would exploit, AI to gain control over others, rather than how AI will gain control over all humans. Even someone with the best intentions may accidentally misuse or cause an error with AI leading to unforeseen dangers. *Vousinas et. al* does link to a similar problem of misinformation and fake news where humans are exploiting AI to gain advantages over others by deceitfully gaining people's trust and trying to remove the need for an individual's critical thinking skills. *Vousinas et. al* does provide an interesting insight, that the solution to misinformation and fake news is not a mathematical or engineering solution but rather one that requires attention from a philosophical standpoint.

Finally, the idea of ethics regarding the interaction of children and AI is one that also needs to be considered carefully. *Vousinas et. al* has some mixed ideas stating that children's data should also need to be carefully considered to prevent biased algorithms, however they state that there is also a lack of data protection and privacy rights regarding children's data collection. I think in this case, the responsibility falls partly on parents and guardians to ensure that children are only using online tools designed for children. Just as alcohol is only for adults, it's the guardian's responsibility to ensure that a child does not consume alcohol. That being said, the lines can become blurred online with what is a resource for children and what is not, and it is then in which we need to ensure that there are some forms of protection to ensure the safety of a child online. Even online tools in general and tools specifically made for children need to be carefully considered as there may be accidental unwanted breaches of data.

Vousinas et. al concludes by outline solutions for some of these problems mentioned above, including having diversity in datasets to manage bias and reduce its impact. Overall *Vousinas et. al* state that in order to address ethics within AI a holistic approach is needed, as the solutions are not 1 dimensional mathematical problem but rather requiring interdisciplinary expertise in order to provide a framework or global standard with how to deal with AI.

Continuing on from these ideas of ethical AI, an article by *Widder et. al*, addresses the limits and possibilities of open source AI, where the AI is transparent and part of an online community. Specifically, regarding deep fakes. One of the main ideas presented is that of how an open source community such as SkiKit or TensorFlow can suffer just as much from unethical practices as private companies due to the differing practices from person to person of how to approach an ethical challenge. *Widder et. al* also found that open source can allow for

greater ethical scrutiny as many would be scrutinized by fellow community members regarding an ethical approach. Technological neutrality, the idea that the choice of technology to achieve a specific task should not be limited or imposed by legislation or another organisation. The article found that technological neutrality is important to many members of the open source community regarding how an individual approaches an ethical idea. The belief that a policymaker or governing body should not dictate a clear winner of ethical approaches, however the best approach is the one that is naturally favoured by the community.

Widder et. al summarize their findings into 6 main ideas with suggestions for the open source community. The first idea being that each individual needs to stand against the misuse of technology and call others out for misusing as well. It is through the self-maintained standard that we can create an ethical community. The second idea was to ensure that you educate yourself and others working on a project as to potential ways a technology could be misused. This idea of being proactive rather than reactive is essential to prevent any unwanted consequences. The third idea is to implement technical restrictions to ensure an effective measure to prevent misuse. This can be effective as it makes it difficult for most users, but some leaders feared that it might not be a long lasting solution.

Another idea was to implement a somewhat reputation or status based reward system where repositories on GitHub could be rated based on its ethical standards. This could be quite an effective way as many people are concerned with their reputation as it sometimes has a great impact on job and career prospects. *Widder et. al* also suggested that platforms such as Google, GitHub and Discord should provide up to date policies that are enforced. The final suggestion was to publicize and study ethical source licenses as a way to “clarify misconceptions over the implications of license choice” (*Widder et. al* 2022).

Overall ethical AI has many potential implications in many domains as outlined by *Vousinas et. al* and it is important that we define ways to ensure that moving technology is created in an ethical manner. *Widder et. al* shows that within the open source community more progress is needed as there isn't a one size fits all standard of ethics when working with AI.

3.6 Thinking Responsibly During Development

Following on from the overview of ethical AI, an article regarding how we should think responsibly about AI and the potential dangers in which it can arise poses some interesting ideas about how we should act as a developer. The article by *Mikalef et. al*, explores AI from the perspective of trying to misuse the technology so that they can better understand possible ways someone who is malicious may try to exploit this technology. *Mikalef et. al*, takes this dark side perspective or “lens” in an attempt to better understand how someone may try to exploit AI and use the technology to their own benefit. Throughout the study it allowed them to understand where cases went wrong or when AI was not meeting standards. The importance of this

was that we now have identified a useful tool which can help aid us with analysing ethical AI and whether it is suitable or not. At the end of the paper, they summarise what can be done to ensure that AI is produced in an ethical manner.

What can be done:

1. Personally take a stand against unethical behaviour.
2. Educate and learn about project specific harms.
3. Consider technical restrictions on use of AI.
4. Leverage reputational incentives such as a GitHub rating.

The first two ideas take a more personal approach and hope that the individual who is developing AI will make a conscious effort to produce high quality work. A short coming of these methods include that many developers lack quality ethical understanding and training. The last two ideas focus on a more systematic approach in which they suggest something like a rating system on GitHub will lead to developers being more conscious and reinforcing the initial two ideas. Currently nothing like this has been implemented so we have no data to show if this will be effective.

In contrast to the previous article *Chubb et. al*, analyses a case study regarding conversational AI and its impact on children. It dives into how we must account for certain considerations when developing these new tools, and how such new technology will impact certain groups of people. It found that not all demographic groups are equal in the sense that we may have to be more sensitive and account for more sometimes. In the case of children, development needs to maintain higher standards by empathising more with the stakeholders. The stakeholders being both the children and their guardians as both will likely be interacting with the technology. Developers need to understand that the content that they personally may seem suitable for children may not be viewed the same way by their guardians and this will ultimately impact the overall effectiveness of a tool or AI.

Moving forward *Chubb et. al* suggest that more work needs to be done regarding situational ethics as this will have a large bearing on the impact of different individuals. Since we do not know the full extent at which AI will impact society, it is critical to understand the foundations now before any negative outcomes are produced.

3.7 Impact of AI and its Significance

An article by *Peeters et. al* analyses 3 major perspectives that concern the future impact of AI and addresses the current state regarding each perspective as well as a research design framework that addresses the potential dangers. The 3 major perspectives are technology-centric, human-centric, and collective-intelligence centric.

Technology-centric:

- The idea that AI will outperform humans.
- Major concern and threats are from humans becoming redundant.

Human-centric:

- That humans will remain superior in social and generalist aspects.
- Major impact will be in specialised situations with humans using these tools.

Collective intelligence-centric:

- Main threat for humans is not seeing the problems or being unaware of an unwanted result.
- Collectively underestimating AI and how it will impact society.

A second article by *Li and Huang* continues on with this idea of potential impact of AI by addressing the concern that has come from both academics and the general public. *Li and Huang* identify the major concerns that need to be addressed as well as understanding the shortcomings of current research and where more work needs to be done.

Eight AI anxiety factors identified:

1. Privacy violation.
2. Bias.
3. Job replacement.
4. Learning.
5. Existential risk.
6. Working against ethical standards.
7. Artificial consciousness.
8. Lack of transparency.

Chapter 4: Project Plan & Description

4.1 Thesis C Term Plan

Week 1:

- Review Summer Readings and create term plan

Week 2:

- Focus on achieving Goal 1
- Start Infographic with the target audience being the general population
- Revise Term Plan

Week 3:

- Finish Infographic
- Research into current resources regarding ethics of AI
- Create criteria for current resources and justification for each criterion.
- Analyse current resources found and find what the gaps that need to be filled are.

Week 4:

- Revise Infographic
- Review current limitations of assessing bias in AI
- Outline the types of ways to mitigate risks of negative bias and the pros and cons for each
- Based on these types, what are the overall limitations and shortcomings of each method

Week 5:

- Create initial concept for Digital Handbook
- Provide ethical approaches to handling bias in AI, in order to reduce negative bias
- Revise writings from week 4

Week 6:

- Flexi week
- Revise concept for digital handbook

Week 7:

- Review All completed content from Thesis A to C and Refine ideas that have been found
- Start on Presentation Slides

- Continue Refining both the Infographic and the digital handbook
- Send Seb copy of slides for feedback
- Give mock presentation to Seb and other Thesis students

Week 8:

- Give Presentation based on overall findings linking to the outcomes
- Continue Refining both the Infographic and the digital handbook

Week 9:

- Revise writings from Thesis A to C for final report
- Continue Refining both the Infographic and the digital handbook

Week 10:

- Finalise final report and ensure the desired outcomes have been met

4.2 Expected Outcomes and Outputs

Goal	Outcome	Output
1	Be able to present the problems of bias within AI, to the public in a non-technical manner	Infographic that explains basic concepts of AI and Bias
2	Be able to justify what the limitations of mitigating bias within AI	Research that helps the infographic The written thesis Digital Handbook
3	Assess the current possible solutions which mitigate these risks.	Research that helps the infographic The written thesis Digital Handbook
4	Justify what the ethical responsibilities are of an engineer in order to mitigate these risks.	Research that helps the infographic The written thesis Digital Handbook

Table 1 Thesis C Expected Outcomes and Outputs

Relation of outcomes and outputs to each other:

There is a priority and emphasis on goals 1 and 2, with 3 and 4 to be completed if time remained.

All the outputs directly achieve their relevant outcomes

Output 1 → Outcome 1

Output 2 → Outcome 2

Output 3 → Outcome 3

Output 4 → Outcome 4

Outputs 2,3,4 are the same output with each goal building on the previous knowledge.

Outputs 2,3,4 also help achieve Output 1 through better personal understanding of the topic area.

The output of writing the thesis helps my own understanding of the content which in turn allows me to better express and convey the ideas in a digital format.

4.3 Project Motivations

Maximising value of thesis:

- Engaging with the idea, learnt during Thesis A and B, of “fairness through awareness” by bringing the ideas I have learnt and developed to a more accessible platform.
- Goal 1 is targeted at bringing awareness to those from a non-technical background not involved in development of AI technology.
- Goals 2-4 will provide useful information to bring awareness to those of a more technical background and who are also involved in the development of AI but may not have had any formal ethics training.
- Bring awareness to developers while also not being just another survey article by providing my own ideas developed from my research.
- Another possible way to ensure that it is not just another survey article is to present the developed ideas in an interactive way such as a website, where the information is easily accessible.

How did I come about and why have I chosen to address the idea of “Fairness through awareness”:

- Through my literature research in Thesis A and B I became aware of this being a possible solution to the problem of bias and fairness in AI.
- I am also personally passionate about AI and believe others should also be well informed.
- I am also passionate about teaching and science communication.
- As a result of my passion for both AI and teaching, I believe I can maximise the value I add through this methodology .

Thesis C Targets:

- Creating a proof of concept with how this information can be communicated.
- Due to limited time and ethical restrictions, quantitative research will not be carried out.
- Although future work after Thesis C could include assessing the impact and effectiveness of these communication methods.

4.4 Description of Project**Who needs to be aware and why do they need to be aware?**

General public

- This technology is affecting their day to day lives and will so even more in the future.
Without an informed understanding of this technology, people may experience negative repercussions unknowingly.
- It is also important to note that with AI such as Chat GPT and others entering mainstream media, the public's current view of AI is important to consider when trying to communicate specific ideas.
- Specifically, we are trying to make the general public aware of the issues relating to bias and fairness in AI and how that might affect their own lives.

Policy makers

- Those in power making decisions need to be well informed as their decisions will ultimately affect the general public.
- Ethics washing is a common practice where ethics are communicated in a specific way to frame what is being done as appropriate for the gain of that specific organisation.
- The hope is that if someone is well informed they will know whether what is being done is ethical for society as a whole.

Developers

- Need to understand their own impact of how they may contribute to the problem and how they can reduce their impact.
- Some may never have had formal ethics training, so providing a guide can prove beneficial
- Also, important to consider is whether someone would actually take the time to learn about ethical AI if it is not mandatory for them to do so.
- With this as the case, how do we still ensure developers take an ethical approach.

Awareness ideas:**Infographic**

- Conveying the simple basic ideas of bias and fairness in AI
- Easy to understand
- Possibly interactive
- Aimed at general public and anyone completely new to AI

Digital Handbook/Website

- Dive deeper into the problems
- Dynamic and can be constantly updated
- Outline current limitations and solutions of mitigating bias
- Provide ethical approaches to handling these problems in order to reduce risks
- Provide pathways for people to carry out even more personal research if needed
- Aimed at developers and anyone else seeking to learn more about the topic

How does my concept provide something useful without rehashing what is already out there?

- By looking at what already exists online regarding communicating ideas, which are not technical articles or news articles, we can see where the gaps are.
- Currently only 3 resources were found that had similar intentions (Table 2 in red).
- Each 3 had positives and negatives, but none were able to completely fulfil the goal of communicating the ideas in my personal opinion.
- Purpose of a guide is to avoid the situation where people defer responsibility to the AI rather than themselves the developers and users of the AI.
- Educational resources with different target audiences and purposes depending on the medium.

What already exists online regarding ethical AI and effectively conveying the ideas?

Organisation	Summary	Link
<i>Australian Government</i>	<ul style="list-style-type: none"> • Outlines 8 ethical goals to ensure AI is safe • Voluntary framework • Testing conducted with other companies • How they developed the framework 	Australia's AI Ethics
CSIRO	<ul style="list-style-type: none"> • Discussion paper funded by Australian government • Technical document regarding Australia's Ethics Framework 	CSIRO Discussion Paper
<i>Gradient Institute</i>	<ul style="list-style-type: none"> • Non-profit website centred around responsible AI • They have news, case studies, research, training courses • Training courses are not online resources 	Gradient Institute
Ethical AI Advisory	<ul style="list-style-type: none"> • Links to the Gradient Institute 	Ethical AI Advisory
<i>Ethical AI</i>	<ul style="list-style-type: none"> • Provides a pdf guide to practical ethics of AI • Provides an example manifesto for ethical AI 	Ethical AI
Oxford	<ul style="list-style-type: none"> • Textbook from Oxford university regarding the Ethics of AI • Technical writing aimed at those in the industry 	Oxford Handbook of Ethics of AI
C4E Journal	<ul style="list-style-type: none"> • Supplementary content to the Oxford textbook • Annotated Bibliography 	C4E Journal
pwc	<ul style="list-style-type: none"> • Article • Outlines 10 principles for more ethical AI 	Ten Principles for Ethical AI
Built in	<ul style="list-style-type: none"> • Article • Guide to Ethical AI 	built in
Harvard Business Review	<ul style="list-style-type: none"> • Article • Practical Guide to building AI 	Harvard Business Review

Table 2 Existing Online Ethical Resources

Criteria for online resources that promote awareness of AI Ethics:

For a digital resource focusing on ethics, the two main factors for determining how successful it would be are content and presentation. Both content and presentation have sub criteria that can help evaluate how effective a resource is.

Content:

Features:

- Handbook/Guide.
- An about me page and who is involved.
- Contacts page.
- Further reading.
- Correct references.

Ethics

- Ethical principles.
- Applying principles.
- Testing principles.
- Developing the principles.
- Relevance and accuracy of content.

Objective of the website:

- Does it have a manifesto or vision page?
- Are the intentions clear and well stated?

Presentation:

Layout

- Ease of navigation.
- Visual design.
- Interactivity.

Features

- Multimedia (Videos, Images).
- Diagrams.

Accessibility

- Language availability.
- How easy it is to find online.

Justification for Criteria:

The goal of a digital resource is to be able to easily convey and teach an idea. A website for example should have all the necessary features in the one site to ensure that the answer or piece of knowledge that someone may be looking for, can be found.

In this situation, the criteria needs to assess the ability of a website to convey the necessary ethical ideas of ethical AI. For a website to achieve this goal it needs to be able to satisfy both the technical side of quality content and needs to be well presented so that the information can be found and interacted with easily.

Presentation could have even more detailed criteria, and using more specific human computer interaction concepts as well as implementing proper UX design methods could be very useful for making a design that achieves the user goals and provides an overall greater user experience.

Criteria	Justification
CONTENT	
Features: <ul style="list-style-type: none"> • Handbook/Guide • An about me page and who is involved. • Contacts page • Further reading • Correct references • Log of updates and news feed. 	<ul style="list-style-type: none"> • A guide is essential to convey the key content. • The About me page and contacts page is useful so people can learn about the background of the project being presented as well as interact and form a community. • Further reading and references are needed to provide a possible avenue for people to do even more learning if the website does not completely satisfy their own learning goals. • To ensure that someone using the website is aware of the rapid pace of development within the field of AI, and how ethics may be approached differently depending on context.
Ethics <ul style="list-style-type: none"> • Ethical principles • Applying principles • Testing principles • Developing the principles • Relevance and accuracy of content 	<ul style="list-style-type: none"> • Ethical principles relevant to the topic of bias in AI need to be outlined and discussed. • It needs to be well documented as to how such principles can be applied. • How the principles can be tested is important, in order to know that they are not completely useless principles. • A record of the motives and justifications behind the development of these principles is crucial, so that there is an understanding of why specific principles have been developed in the first place. • Relevance and accuracy is important to ensure that the information is up to date and still can be applied to today's world.

<p>Objective of the website:</p> <ul style="list-style-type: none"> Does it have a manifesto or vision page? Are the intentions clear and well stated? 	<ul style="list-style-type: none"> Having the object of a website is crucial as motives can influence how ethics is presented. By having a manifesto or intentions laid out, people can hold the ethics that have been presented accountable.
PRESENTATION	
<p>Layout</p> <ul style="list-style-type: none"> Ease of navigation Visual design Interactivity 	<ul style="list-style-type: none"> Navigation is essential to ensure a user knows how to find the relevant information to them. Visual design and interactivity contribute to the overall user experience of the website, and this can subconsciously determine how well a user learns what they need to know.
<p>Features</p> <ul style="list-style-type: none"> Multimedia (Videos, Images) Diagrams 	<ul style="list-style-type: none"> It is important to have multimedia, being a website, it is essential to have images, videos, and diagrams. Without these features, there may not be any incentive to use a website instead of writing an article.
<p>Accessibility</p> <ul style="list-style-type: none"> How easy it is to find online. Language availability High contrast and visibility features 	<ul style="list-style-type: none"> As we want to reach the largest audience possible, making a website as accessible as possible is important. The ability to actually find the resource online is a key factor to determining how well the website is achieving its goal. Having other features such as different languages and visibility features are important to have in order to allow more people to interact with the website.

Table 3 Criteria Justification

Chapter 5: Project Results

5.1 Infographic

The design process of the infographic was an iterative process. It started off with a draft version created with Canva as seen in Figure 4. Week by week, through the use of feedback from my supervisor and general discussion, the infographic was revised to the final version as seen in Figure 3. The final version has six introductory questions to AI and is designed to target the general population and anyone who does not know much about artificial intelligence and the risks involved. At the bottom of the infographic, there is a link to ethicalaiguide.com which is the digital handbook and second component of this thesis. The link is displayed to give those who are still interested in learning more, a path in which they can develop their knowledge. Adjacent to the report and digital version of the infographic, I was able to print and laminate, creating a hard copy A4 poster which could be used as a poster in a classroom for example. Below is the corresponding final text in the infographic.

What is AI and Machine Learning?

Artificial intelligence or AI for short, is a term used to describe a computer system that can perform complex tasks such as automation. Contemporary AI is often equated to machine learning, a method that processes data to make evaluations and carry out tasks.

Should we be using AI?

AI has many benefits to society including reducing human error and being more efficient than humans at repetitive tasks. There are also downsides to AI, such as negatively biased algorithms being used for decision making, and the inability of AI to adapt to completely new situations.

What are the risks of AI?

Automation is currently replacing humans in jobs such as customer service, in turn causing unemployment for those being replaced. Other risks include the loss of accountability and ownership of mistakes as people defer responsibility to AI in situations such as research analytics and decision-making tools.

Is AI negatively biased?

In AI, bias is an assumption used to simplify algorithms and improve efficiency. Negative bias is when an assumption forms a decision that would create a world in which we would not want to live. AI bias can emerge through machine learning as AI may use biased data.

Is AI fair?

Fairness has many definitions, often associated with the idea of equalised odds and opportunity. The challenge with AI fairness is locating and maintaining the balance between all fairness types. AI often struggles to distinguish between equality and equity, which can result in unfair outcomes.

Possible solutions to biased AI?

Fairness through awareness is crucial! If more people are aware of the risks of AI, then society will be more capable of preventing these problems from arising. Often such risks are less of a technical issue than they are a human issue. This needs to be considered carefully to ensure it is used ethically.

General notes for the infographic:**Purpose:**

Convey basic ideas of AI, as well as bias and fairness in AI.

Target Audience:

General Public, Policy Makers, Anyone new to AI

Context:

Due to the rapid development of AI and yet a lack of reputable sources that convey these ideas of AI, there is a gap of knowledge that needs to be filled in order for people to be able to learn about AI. Currently most people will only learn from the news media this is not always the most reliable source to trust.

ARTIFICIAL INTELLIGENCE IS IT FAIR?



What is Artificial Intelligence?

Artificial Intelligence (AI) is a term used to describe a computer system that is capable of performing complex tasks such as automation. Contemporary AI is often equated with machine learning, a method that processes data to make evaluations and carry out tasks.

Should we be using AI?

AI has many benefits to society including reducing human error and being more efficient than humans at repetitive tasks. There are also downsides to AI, such as negatively biased algorithms being used for decision making, and also the inability of AI to adapt to completely new situations.



What are the risks of AI?

Automation is currently replacing humans in jobs such as customer service, in turn causing unemployment for those being replaced. Other risks include the loss of accountability and ownership of mistakes as people defer responsibility to AI in situations such as research analytics and decision making tools.

Is AI negatively biased?

In AI, bias is an assumption used to simplify algorithms and improve efficiency. Negative bias is the situation that occurs when an assumption forms a decision that would create a world that we do not consider to be morally desirable. AI bias can emerge through machine learning as AI may use biased data.



Is AI fair?

Fairness has many definitions, often associated with the idea of equalised odds and opportunity. The challenge with AI fairness is locating and maintaining the balance between all fairness types. AI often struggles to distinguish between equality and equity, which can result in unfair outcomes.

Possible solutions to biased AI

Fairness through awareness is crucial! If more people are aware of the risks of AI, then society will be more capable of preventing these problems from arising. Often such risks are less of a technical issue than they are a human issue. AI needs to be considered carefully to ensure it is used ethically.



For more information, please visit
ethicalaiguide.com

Figure 3 Final Version of Infographic



Figure 4 Version 1 of Infographic

5.2 Digital Handbook

Website URL:

ethicalaiguide.com

Purpose:

The purpose of a guide is to avoid the situation where people defer responsibility to the AI rather than themselves the developers and users of the AI. It also provides more detailed information, specific to bias and fairness in AI, and provides further resources for people to continue their learning.

Target Audience:

Developers and engineers, anyone seeking to have a more in-depth understanding. Although the primary target is developers, the digital handbook is designed to be an online extension of the infographic and an interactive hub that is continually updated and adapting to the fast-paced landscape of AI.

Context:


Due to the rapid development of AI and yet a lack of reputable sources that convey these ideas of AI, there is a gap of knowledge that needs to be filled for people to be able to learn about AI. Currently most people will only learn from the news media this is not always the most reliable source to trust. Furthermore, many developers and engineers may be lacking understanding or training regarding ethical practices and this website is designed as an online handbook or guide to help with day-to-day decision making.

Home Page:

The homepage of the website, as seen in Figure 5, provides an overview of the ethical content. At the top of the page there is a button that says, “Learn More About AI” and this directs the user to a page that will help them learn more about AI. The other four sections are the ethical sections:

1. Ethical Principles
2. Applying Principles
3. Testing Principles
4. Developing Principles

Each section when clicked on will direct you to the respective page.



[Home](#)
[Guide](#)
[AI](#)
[News](#)
[Further Reading](#)
[About Us](#)
[Contact Us](#)


The Ethical AI Guide

Providing an ethical approach for building ethical AI.

[Learn More About AI](#)

ETHICS


Ethical Principles



A framework designed to help with the development of safe and ethical AI.

[Find out more](#)


Applying Principles



A guide to show how someone can apply these ethical principles in their workplace.

[Find out more](#)


Testing Principles



A review of how effective these ethical principles can be applied within industry.

[Find out more](#)

Developing Principles



An outline of the motivations and justifications for why the principles were created.

[Find out more](#)

Figure 5 Website Homepage

Understanding AI

The understanding AI page as seen in Figure 6 has a brief overview of the main ideas about artificial intelligence. Currently it requires further depth, but the following text was inputted just to show proof of concept.

AI and Machine Learning

One of the new and exciting technologies within computer science is AI and machine learning. It has developed rapidly over the last decade. Many applications range from recommender systems to medical diagnosis as well as streamlining and automating many business processes. There are many types of ML algorithms ranging from the simplest, regression analysis, to the more complex and cutting edge such as neural networks and deep learning algorithms. The underlying basis for this technology is data, as these algorithms require large datasets to be effective. As big data becomes more prominent and grows over the coming years, so too will the applications and the impact that these algorithms have on the world.

Ethical Principles Applied to AI

When we apply ethical principles specifically within AI, some of the questions that are raised include:

- Is AI fair?
- Are there dangers with autonomy, and can it make the situation worst?
- Are the current systems and approaches transparent?
- Are the current regulations and standards suitable?
- Should we use AI at all, or should we avoid using it?

Emerging themes

Overall, some of the emerging themes within the ethics of AI include:

- Ethics Washing, where large corporations try to justify their ethical standpoint in a way that favours the systems they are producing.
- Extractivism, in which through the use of acquiring some knowledge through data, an advantage can be made as a result of this knowledge.
- Power, in which a motivating factor for extracting this knowledge, is in order to gain power over another and benefit a personal or specific agenda.

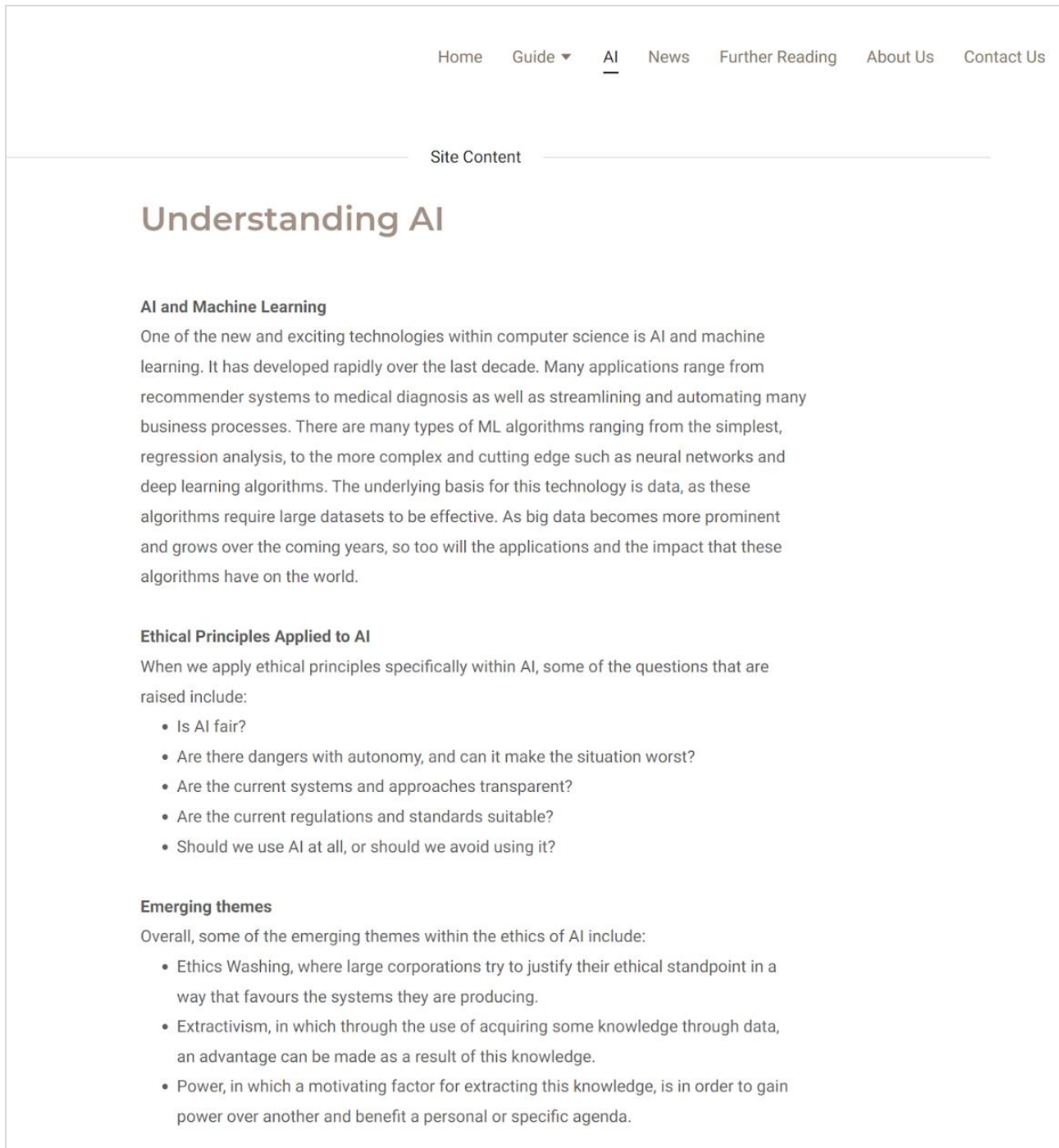


Figure 6 Understanding AI

Ethical Principles

The ethical principles page is the first ethical sub-page from the homepage. As seen in figure 7, the page describes some basic ideas of what ethical AI might entail and provides some info to the user about some principles they can use in their day-to-day work.

What does Ethical AI look like?

Rigorous: Is it built with transparency and is it well documented? Does it allow for adjustments to improve the model if needed?

Consentful: Is the data used acquired in an appropriate matter? Is the model used in a way that is known by those being affected by the outcomes of the algorithm?

Socially Conscious: Is the model designed to address an issue within society that will better the world in one way or another?

Sustainable: How resource intensive is the process of training the model and using the model. Is the process economically viable and does it fit within the realm of realistic constraints?


Inclusive: Is the model designed to be inclusive for multiple groups and does it consider edge cases. Is it truly representative of a society we want to live in or is it negatively biased?

Inquisitive: Do the developers consider the repercussions of the model and how it may impact society both for the better and for the worst.


What are the current limitations that prevent us from mitigating negative bias in AI?

Challenges of the entire domain

- Creating a concrete definition of fairness.
- Distinguishing equality from equity across diverse use cases.
- Searching for unfairness to understand if an algorithm is also unfair and not just fair.



HomeGuide ▼AINewsFurther ReadingAbout UsContact Us



Ethical Principles

What does Ethical AI look like?

Rigorous: Is it built with transparency and is it well documented? Does it allow for adjustments to improve the model if needed?

Consentful: Is the data used acquired in an appropriate matter? Is the model used in a way that is known by those being affected by the outcomes of the algorithm?

Socially Conscious: Is the model designed to address an issue within society that will better the world in one way or another?

Sustainable: How resource intensive is the process of training the model and using the model. Is the process economically viable and does it fit within the realm of realistic constraints?

Inclusive: Is the model designed to be inclusive for multiple groups and does it consider edge cases. Is it truly representative for a society we want to live in or is it negatively biased?

Inquisitive: Do the developers consider the repercussions of the model and how it may impact society both for the better and for the worst.

What are the current limitations that prevent us from mitigating negative bias in AI?

Challenges of the entire domain

- Creating a concrete definition of fairness.
- Distinguishing equality from equity across diverse use cases.
- Searching for unfairness to understand if an algorithm is also unfair and not just fair.

Ethical AI Guide

Figure 7 Ethical Principles

Some other ethical questions and topics to include in the future are:

- What are some current ways to mitigate risks of negative bias?
- Pros and Cons for each method of mitigating negative bias.
- Limitations and Shortcomings of each method of mitigation.

The other three subheadings from the homepage are not fully implemented at the moment, but this has not been considered a problem as the main goal for the website in this thesis was to provide a proof of concept. Moving forward however, the following is some of the content that would be valuable to include for each sub-page:

Applying Principles

Provide an ethical approach to handling negative bias in AI

- Include ideas from thesis readings and writings so far.
- Normative ethical principles.
- Methodology to help someone know what course of action to take.
- Perhaps the purpose of a guide is to avoid the situation where people defer responsibility to the AI rather than themselves, the developers, and users of the AI

Testing Principles

- Case Studies of where certain principles have helped or may have helped.
- Actual testing in the future would be better.
- Overall quality of ethical principles that have been shown on the website analysed.
- Testing is focused on ensuring that the principles in the other pages are thorough and to a global standard.

Developing Principles

- Reflection of the entire thesis itself essentially, in this instance.
- A way to help others understand the process in which the principles they are using are actually effective.

News

The news page is focused on staying up to date with the latest trends and innovations within the field of AI. As is rapidly changes, there needs to be a way to ensure that the website stays dynamic and relevant to the current time. If there was nothing changing then the website would become outdated very quickly. Some ways to ensure this is effective is by:

- Populating with some relevant news articles regarding ethical AI and bias in AI.
- Having a news feed that people can easily see what the major updates which they should be aware of.
- Having a log of updates which logs the updates to the website, so people have understanding regarding changes. The log should outline what the change is, when it was done, and why the change was necessary.

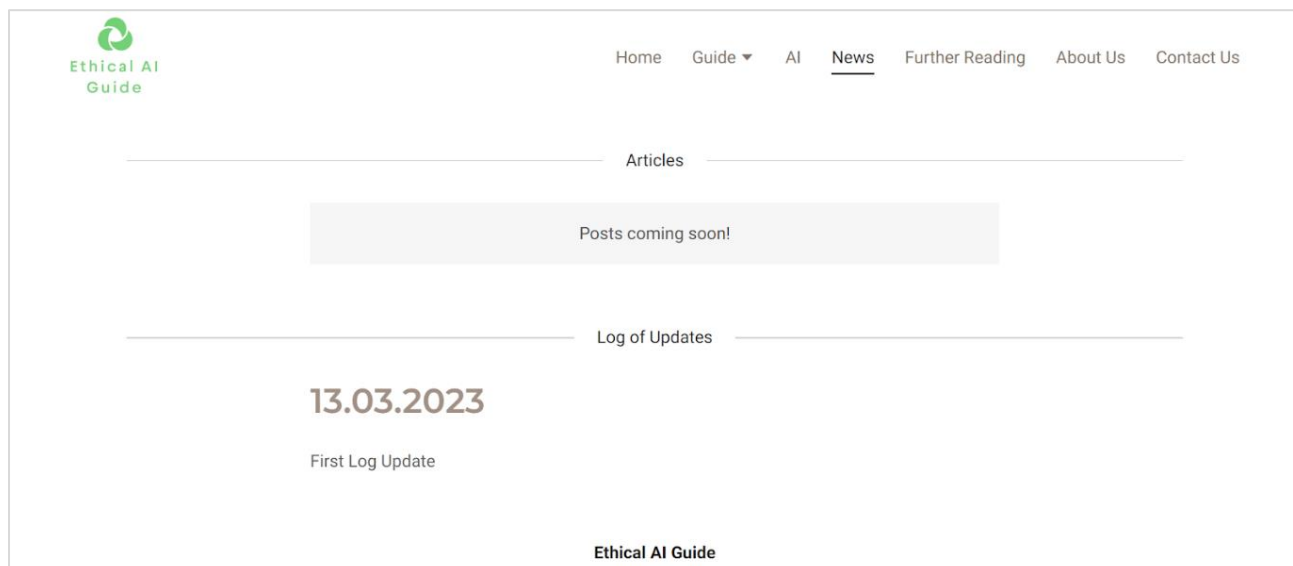


Figure 8 News Page

Further Reading

The further reading page is designed as a way to give those that want to learn even more a path they can follow. Some of the elements it contains include:

- Downloadable pdf of the infographic to allow for people to share and print as they wish.
- Populated case studies section with interesting real life examples of how ethical dilemmas have arisen with the use of AI.
- Case studies may provide benefit to those looking for a similar situation and wanting to know more to inform their own decision making process.
- Link and summary of case studies should be shown, in doing so creating a resource catalogue.
- Provide interesting technical papers that may be useful to anyone looking to find in-depth analysis.
- Like the case studies catalogue, the technical papers should also be catalogued.
- FAQs are a good feature to minimize unnecessary communication of issues.
- References to different resources used.

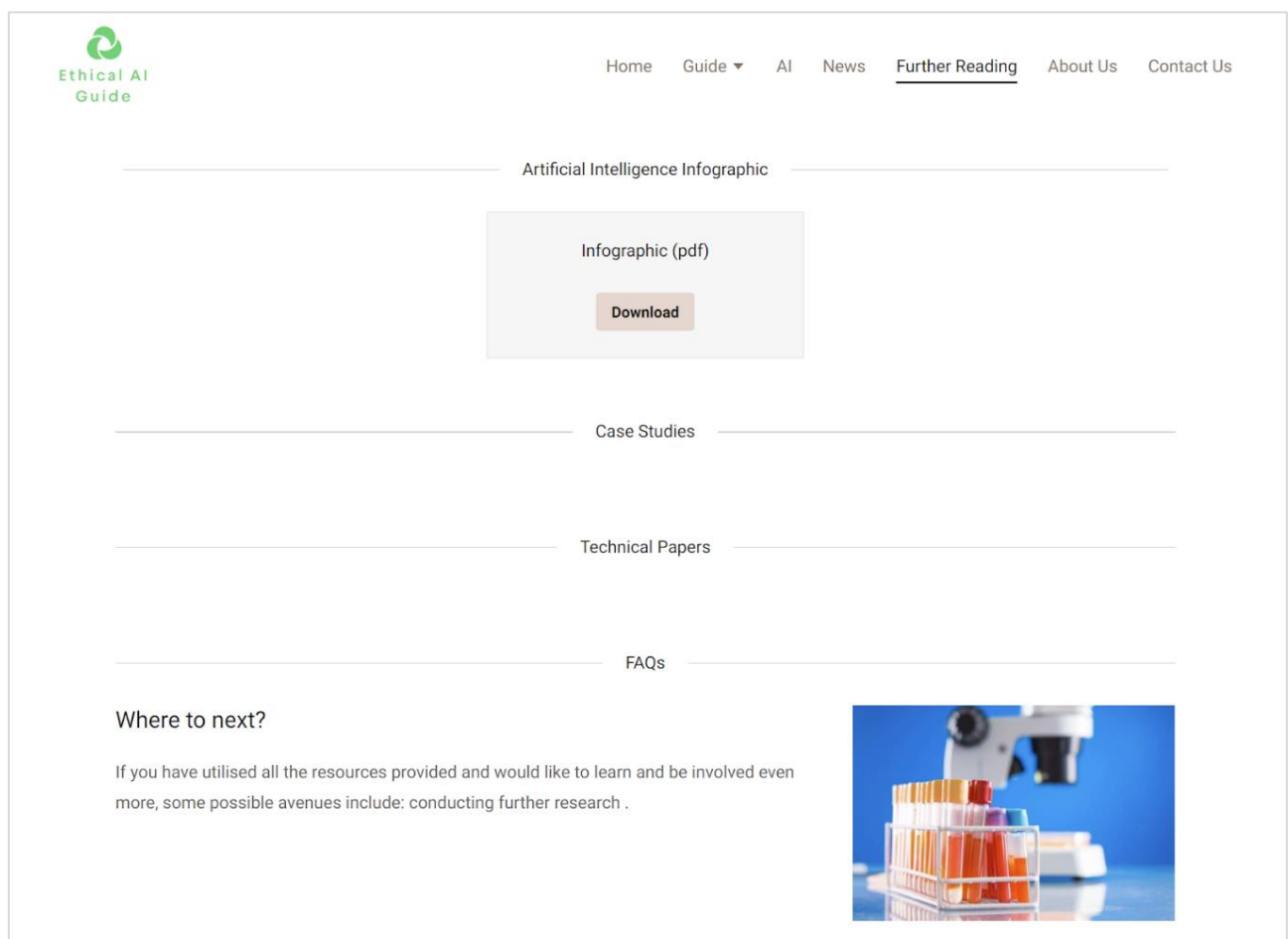


Figure 9 Further Reading

About Us

The about us page is designed to be a manifesto and commitments outline to ensure people understand the motivations behind the website. Transparency is critical for the website to succeed, and this is one way that will help. Clearly outlining the Mission, Vision, and Values will give most people a sufficient understanding of what the goal of the website is, that being to promote ethical AI. I have also included a bio of myself, to show who is involved. If the website was to grow and more people to become involved, it is important to include those who will significantly be impacting the website.

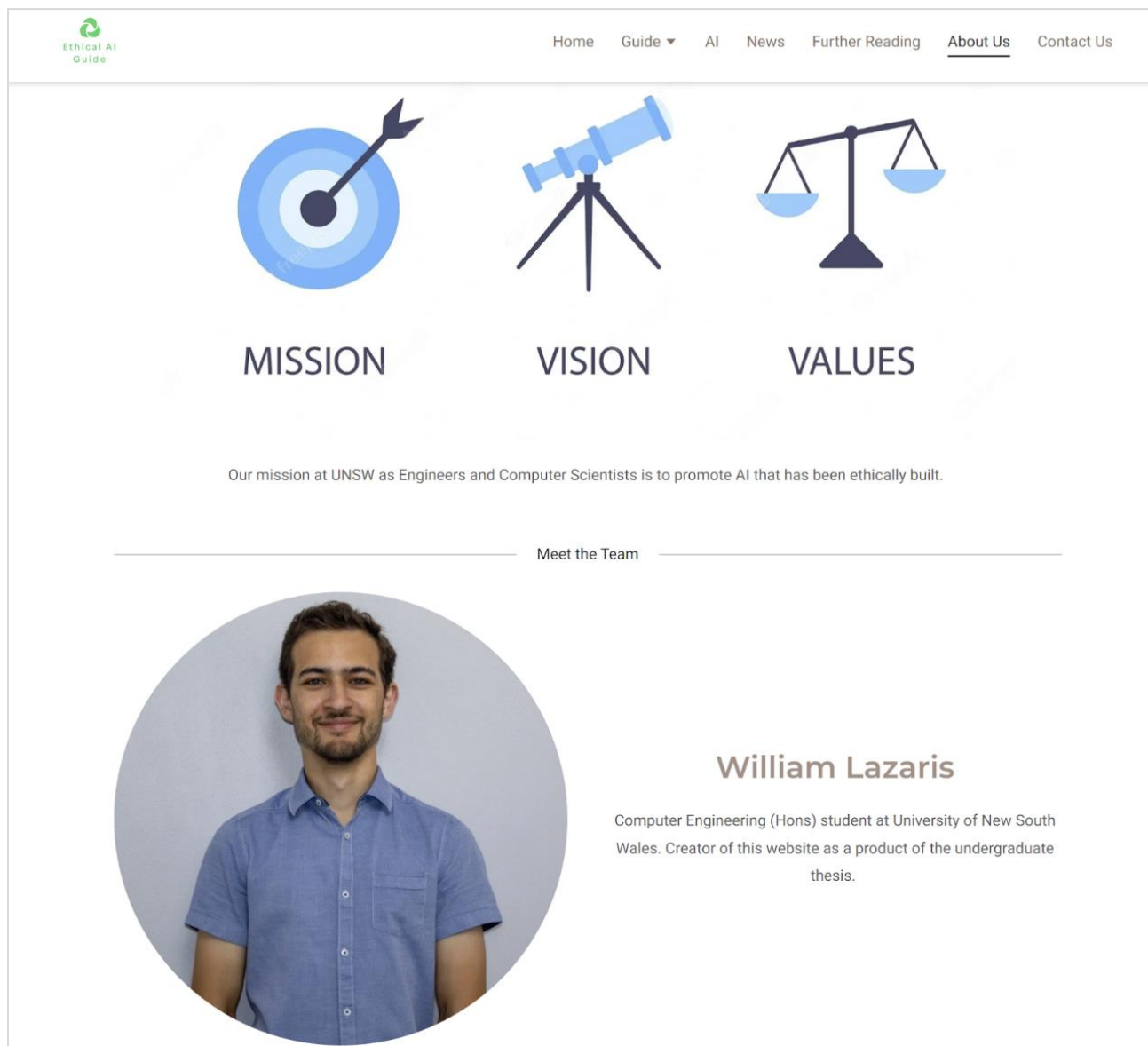


Figure 10 About Us

Contact Us

The contact us page is designed to be very standardized and easy to use. The goal in mind is that if someone has a query they can submit a form and I would be contacted via email. Below the form is a link to my LinkedIn, this is just to demonstrate that you can have links to different social platforms. If the website was to grow, then community interaction and image would be important so having something like a discord or Instagram account would be beneficial, and in that case you would be able to embed a direct link to those communities.

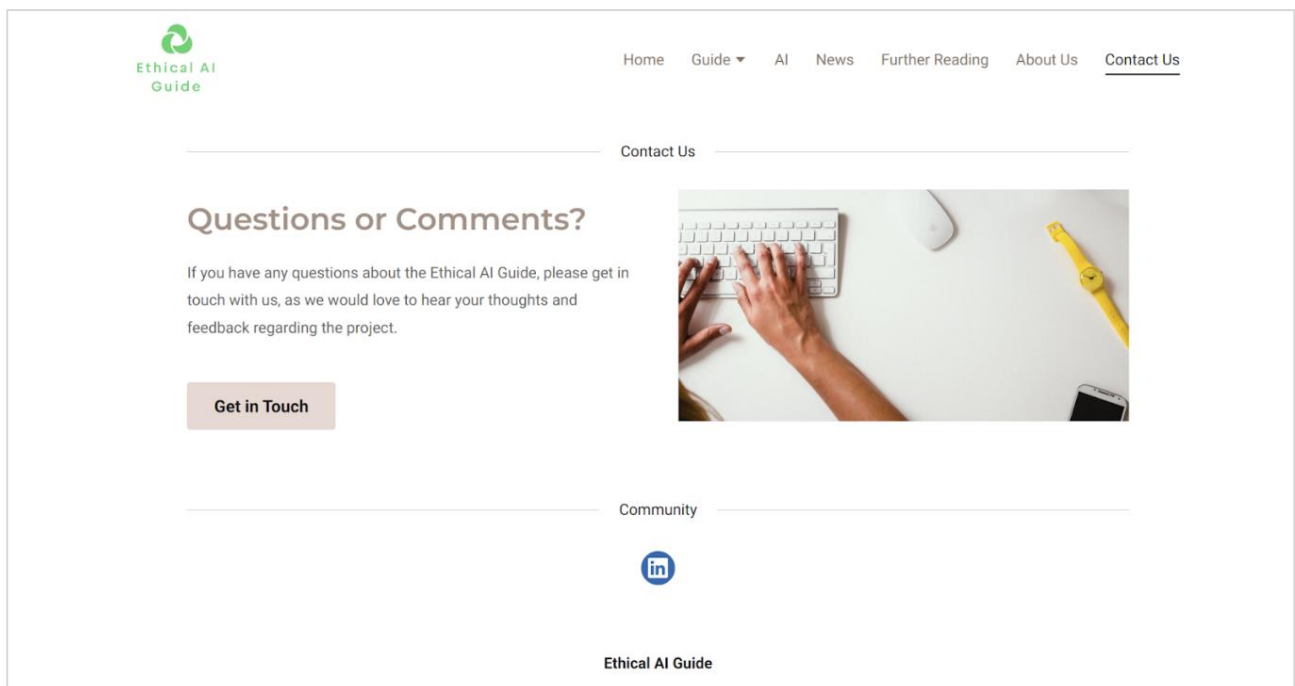


Figure 11 Contact Us

Chapter 6: Conclusion

6.1 Conclusion

Throughout thesis A and B, I have broken down and attempted to demystify this concept of bias and fairness within AI. Through my research I have been able to identify the areas in which the most research is currently been produced and this has allowed me to see what needed to be focused on and where to spend my time.

As a result, I was able to address the issue of bias and fairness in AI through the idea of “Fairness through awareness”. This led me to create educational resources about these issues but also more broadly ethical AI. Despite starting my thesis by researching the ethics of bias, I discovered that these digital resources apply universally to all ethics within AI and not just bias and fairness. In saying that, bias and fairness is a significant aspect of the domain.

Overall, the purpose of this thesis was to understand and applying the ethical principles of AI, and I believe through the use of my educational resources, I have been able to create something that would be of actual value to the world beyond just this report.

6.2 Future Work

Some future work could include the following:

1. Empirical Analysis with test groups, to identify any gaps or shortcomings of current iteration of project and improve criteria. This would require ethics approval and hence why it was not conducted during this project. Although if approval was achieved, the benefit of detailed feedback would be instrumental in creating a website that is truly effective.
2. Based on Human Computer Interaction principles, redesign the entire website using the UX design process. Possibly also get a professional graphic designer to help with this stage.
3. Include information regarding the economics of AI.
 - What is the economic cost benefit analysis of resource intensive AI?
 - The cost benefit of technology in general?
 - Economics is often a key deciding factor in decision making so this would be very useful.
4. Create a greater social presence through social media:
 - Instagram, Discord, YouTube, GitHub, Etc.

5. Training courses, consulting, and events are a possible avenue to explore for organisations that are seeking upskilling and learning.
6. Analyse what do the existing sites do well that we can learn from?

Why was the design of the website not a contrastive analysis?

In this thesis the design and criteria was not based on a contrastive analysis as I believed to ensure that the best product is achieved, a criterion that is derived from first principles would ultimately be better than just basing it off other websites.

After Thesis C

Moving forward, my goal is to keep working on developing the website and fully fleshing it out. I think that it can prove to be a productive resource within the community as the demand for ethical AI is increasing. Most recently this week, UNSECO made a press release on its global normative framework of its recommendation of ethical AI. The article calls on all governments and organizations to fully implement their framework. With this idea for ethical AI entering the mainstream media, the importance and impact of this website is relevant more than ever. I am also planning on reaching out to other existing organizations that have aligned goals who may be interested in joining up and working on this project.

References

- Sequoiah-Grayson, S. and Walsh, T., 2022. AI and Ethics.
- CRAWFORD, K., 2022. ATLAS OF AI. [S.l.]: YALE UNIVERSITY PRESS, pp.123-151.
- Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (July 2022), 35 pages. <https://doi.org/10.1145/3457607>
- Ayling, J. and Chapman, A., 2021. Putting AI ethics to work: are the tools fit for purpose?. *AI and Ethics*,.
- Noor, P., 2020. Can we trust AI not to further embed racial bias and prejudice?. *BMJ*, p.m363.
- Silberg, J. and Manyika, J., 2019. Tackling Bias in AI.
- McDuff, D., Cheng, R. and Kapoor, A., 2018. Identifying Bias in AI using Simulation. [online] Available at: <<https://doi.org/10.48550/arXiv.1810.00471>>
- Chubb, J., Missaoui, S., Concannon, S., Maloney, L. and Walker, J.A., 2022. Interactive storytelling for children: A case-study of design and development considerations for ethical conversational AI. *International Journal of Child-Computer Interaction*, 32, p.100403.
- Jaton, F., 2021. Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application. *Big Data & Society*, 8(1), p.20539517211013569.
- Li, J. and Huang, J.S., 2020. Dimensions of artificial intelligence anxiety based on the integrated fear acquisition theory. *Technology in Society*, 63, p.101410.
- Merhi, M.I., 2022. An Assessment of the Barriers Impacting Responsible Artificial Intelligence. *Information Systems Frontiers*, pp.1-14.
- Mikalef, P., Conboy, K., Lundström, J.E. and Popovič, A., 2022. Thinking responsibly about responsible AI and ‘the dark side’ of AI. *European Journal of Information Systems*, 31(3), pp.257-268.
- Peeters, M.M., van Diggelen, J., Van Den Bosch, K., Bronkhorst, A., Neerincx, M.A., Schraagen, J.M. and Raaijmakers, S., 2021. Hybrid collective intelligence in a human–AI society. *AI & society*, 36(1), pp.217-238.
- Tanweer, A., 2022. Tradeoffs all the way down: Ethical abduction as a decision-making process for data-intensive technology development. *Big Data & Society*, 9(1), p.20539517221101351.
- Tomalin, M., Byrne, B., Concannon, S., Saunders, D. and Ullmann, S., 2021. The practical ethics of bias reduction in machine translation: why domain adaptation is better than data debiasing. *Ethics and Information Technology*, 23(3), pp.419-433.
- Vousinas, G.L., Simitsi, I., Livieri, G., Gkouva, G.C. and Efthymiou, I.P., 2022. Mapping the Road of the Ethical Dilemmas Behind Artificial Intelligence. *Journal of Politics and Ethics in New Technologies and AI*, 1(1), pp.e31238-e31238.
- Widder, D.G., Nafus, D., Dabbish, L. and Herbsleb, J., 2022. Limits and Possibilities for “Ethical AI” in Open Source: A Study of Deepfakes.

Zajko, M., 2021. Conservative AI and social inequality: conceptualizing alternatives to bias through social theory. *AI & SOCIETY*, 36(3), pp.1047-1056.

Ai Ethics Framework (2018) CSIRO. Available at: <https://www.csiro.au/en/research/technology-space/ai/ai-ethics-framework> (Accessed: March 8, 2023).

Department of Industry, S.and R. (2023) Australia's Artificial Intelligence Ethics Framework, Department of Industry, Science and Resources. Available at: <https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework> (Accessed: March 8, 2023).

Responsible Artificial Intelligence (2023) Gradient Institute. Available at: <https://www.gradientinstitute.org/> (Accessed: March 8, 2023).

Ethical AI Advisory (2023) Gradient Institute. Available at: <https://www.ethicalai.ai/> (Accessed: March 8, 2023).

The Practical Guide to Ethical AI (2023) Ethical AI. Available at: <https://ai-ethical.com/en/practical-guide/> (Accessed: March 8, 2023).

The Oxford Handbook of Ethics of AI (2020) OXFORD Academic. Available at: <https://academic.oup.com/edited-volume/34287> (Accessed: March 8, 2023).

The Oxford Handbook of Ethics of AI: Online Supplement (2022) C4E Journal. Available at: <https://c4ejournal.net/the-oxford-handbook-of-ethics-of-ai-online-companion/> (Accessed: March 8, 2023).

A Practical Guide to Building Ethical AI (2020) Harvard Business Review. Available at: <https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai> (Accessed: March 8, 2023).

AI Ethics: A Guide to Ethical AI (2022) builtin. Available at: <https://builtin.com/artificial-intelligence/ai-ethics> (Accessed: March 8, 2023).

Ten principles for ethical AI (2022) pwc. Available at: <https://www.pwc.com.au/digitalpulse/ten-principles-ethical-ai.html> (Accessed: March 8, 2023).

Ethics and Society at Hugging Face (2023) Hugging Face. Available at: <https://huggingface.co/spaces/society-ethics/about> (Accessed: March 8, 2023).

Artificial Intelligence: UNESCO calls on all Governments to implement Global Ethical Framework without delay (2023) UNESCO. Available at: <https://www.unesco.org/en/articles/artificial-intelligence-unesco-calls-all-governments-implement-global-ethical-framework-without> (Accessed: April 25th, 2023)