Investigating Sources and Effects of Bias in AI-Based Systems – Results from an MLR

Caoimhe De Buitlear[1], Ailbhe Byrne[1], EricMcEvoy[1], Abasse Camara[1], Murat Yilmaz[2] [0000-0002-2446-3224], Andrew McCarren[1,3] [0000-0002-7297-0984], and Paul M. Clarke[1,4] [0000-0002-4487-627X].

[1]School of Computing, Dublin City University, Dublin, Ireland
{caoimhe[2].debuitlear4, ailbhe.byrne287, eric.mcevoy23, abasse.camara2}@mail.dcu.ie
[2]Department of Computer Engineering, Gazi University, Ankara, Turkey
my@gazi.edu.tr
[3]Insight, the Science Foundation Ireland Research Center for Data Analytics
andrew.mccarren@dcu.ie
[4]Lero, the Science Foundation Ireland Research Center for Software
paul.m.clarke@dcu.ie

Abstract. AI-based systems are becoming increasingly prominent in everyday life, from smart assistants like Amazon's Alexa to their use in the healthcare industry. With this rise, the evidence of bias in AI-based systems has also been witnessed. The effects of this bias on the groups of people targeted can range from inconvenient to life-threatening. As AI-based systems continue to be developed and used, it is important that this bias should be eliminated as much as possible. Through the findings of a multivocal literature review (MLR), we aim to understand what AI-based systems are, what bias is and the types of bias these systems have, the potential risks and effects of this bias, and how to reduce bias in AI-based systems. In conclusion, addressing and mitigating biases in AI-based systems is crucial for fostering equitable and trustworthy applications; by proactively identifying these biases and implementing strategies to counteract them, we can contribute to the development of more responsible and inclusive AI technologies that benefit all users.

Keywords: AI, bias, artificial intelligence, risks

1 Introduction

The concept of artificial intelligence (AI) has been around for a long time. The idea appeared c. 8th century BC when Homer wrote about the Gods being waited on at dinner by mechanical 'tripods'. It appears consistently throughout history from science

fiction writers who wrote about intelligent machines being a possibility, to their presence in religions such as Judaism, where the artificially created being known as a Golem appears [1]. An AI-based system is a "computer system able to perform tasks that ordinarily require human intelligence" [2]. Examples of this include systems that can play games like chess, draughts and checkers, filter spam emails and autocorrect text [3] [4].

Bias is defined as "the action of supporting or opposing a particular person or thing in an unfair way" which can be done by making an unreasoned judgement or allowing personal beliefs to influence a decision [5]. Bias may present in AI-based systems for various reasons, including how they were trained and the very data used to train them. This can lead to "algorithmic bias", which is when an AI-based system produces a result that is systematically incorrect [6].

These incorrect results can have varying effects on the groups of people the AI-based system is biased towards. An AI healthcare system was found to be racially biased, facial recognition AI systems are being used which have a lower accuracy on darker-skinned females, and search engines using AI were found to discriminate based on race and gender [7]. In this paper, we will examine what AI-based systems are, we will attempt to understand bias and the types of bias which appear in AI-based systems, we will study the potential risks and effects of bias in AI-based systems, and lastly, we will strive to identify techniques to reduce bias in AI-based systems.

The objectives of this study are to:

1.Elucidate the nature and functioning of AI-based systems and their increasing prevalence in various aspects of daily life.

2.Define and categorize the different types of bias present in AI-based systems.

3.Investigate the potential risks and consequences stemming from biased AI-based systems, and their impact on targeted populations.

4. Identify effective methodologies and strategies to minimize bias in AI-based systems.

5.Emphasize the significance of reducing bias in AI-based systems and advocate for the development of equitable and reliable technologies.

2 Research Methodology

2.1 Methodology

This research paper was created as part of a multivocal literature review. To fulfil the research objective, we adopted an MLR approach which included both academic/peer-reviewed (white) and non-academic/non-peer reviewed (grey) literature. Only a partial application of the MLR process [78] is implemented, whereby the concept of using predefined search strings is employed to evaluate white literature alongside non-traditional academic sources. Accordingly, we made use of Google, Google Scholar, ScienceDirect and other sources to conduct our research.

2.2 Search Queries

Initial high-level surveys of the related literature allowed us to break down the main topic into subtopics, for which associated research questions (RQs) were elaborated. Individual team members examined different sub-topics in detail. The search queries used included strings such as: “AI systems”, “bias in AI systems”, “what causes bias in AI”, “machine learning and bias”, “risks of bias in AI”. We made use of Google, Google Scholar, IEEE and other sources in order to conduct our research.

2.3 Inclusion/Exclusion

We limited our search space to papers that were written in English so that no translation was needed during our research. To determine which sources were appropriate to use during our research, we endeavored to only include literature from peer reviewed academic outlets and grey sources that offered robust levels of quality (e.g., moderated blogs).

2.4 Methodology Limitations

There are a number of significant methodological limitations in this work that are discussed in Section 4.

3 Literature Review

3.1 RQ1: What are AI-based Systems?

Although the concept of AI may have existed previously, it was only in 1955 when the term ‘artificial intelligence’ was coined by John McCarthy. McCarthy was the organizer of the first academic conference on AI, which was held in Dartmouth, New Hampshire [3]. In the years following this conference, computers advanced and became more readily available with larger available memory and greater processing power. This allowed the field of AI to progress [8].

There are many existing definitions for the field of AI, which is considered to be “intelligence demonstrated by machines”, as opposed to the natural intelligence shown by humans and animals [3]. In Alan Turing’s 1950 paper “Computing Machinery and Intelligence”, he asked the question “can machines think?” and created a test where a human evaluator must ask questions and determine which answers belong to a human and which belong to a machine. Although Turing created this test, he did not define what it meant for a machine to be intelligent [9] [10]. Later, Stuart Russell and Peter Norvig published a textbook on AI where they consider four potential definitions of AI to be:

- Systems that think like humans,

- Systems that act like humans,
- Systems that think rationally,
- Systems that act rationally [9].

In perhaps simpler terms, AI "is a field, which combines computer science and robust datasets, to enable problem-solving" [9]. In a general sense, AI systems work by analysing data to identify associations and patterns, later using this knowledge to make predictions about future events and states [11]. The data utilised in the analysis is sometimes referred to as training data [11].

AI-based systems may require large amounts of data during the training process, and therefore it is necessary to have access to large and reliable datasets. This access began in the 1960s with the arrival of the first data centres and as relational databases began to be developed. Following this in the 2000s was the rise of big data. Big data is "data that contains greater variety, arriving in increasing volumes and with more velocity" [12]. This rise was due to the exponential growth of the internet and the amount of internet users, from humans to objects and devices connected to the internet [12].

While training an AI-based system, there are different learning types that can be applied. These learning types are categories of machine learning. Machine learning is a subfield of AI and is "the field of study that gives computers the ability to learn without being explicitly programmed" [13]. It allows machines to learn from data by finding patterns, gaining insight and improving at the task they are being designed to do [14]. In supervised learning, a labelled dataset is used to allow an algorithm to learn. In unsupervised learning, the algorithm uses an unlabelled dataset to attempt to extract features and patterns. A combination of supervised and unsupervised learning can also be used [3].

Deep learning, a sub-field of machine learning, is considered *deep* because it is composed of a neural network containing more than 3 layers [9]. Neural networks comprise a series of algorithms that seek to identify relationships in a set of data by attempting to emulate human brain operation [13], using mathematical formulas and functions to make decisions [15]. Deep learning means that a system can learn very complicated behaviours without the need for a feature extractor to be designed through "careful engineering and domain expertise" [16]. This has allowed for noticeable advances in many problematic fields of AI. It has surpassed the ability of previous techniques in the fields of speech recognition, image recognition, use in healthcare and natural language understanding [16].

AI-based systems can be broken down into two main categories: weak/narrow AI systems and strong AI systems. Weak AI is "AI trained and focused to perform specific tasks" [9]. These systems are explicitly trained to carry out specialised tasks. Although the term 'weak' is used to describe them, these systems are very powerful at what they are trained to do [4]. These AI-based systems have been proven to be able to perfect

their task and outperform humans, dating back to 1996, when IBM's Deep Blue defeated chess grandmaster Garry Kasparov. This was possible as per second it was able to process 200 million positions [10]. This category of AI is prevalent in contemporary AI-based systems, for example in smart assistants like Siri and Alexa, Google maps, recommendation systems on websites like Spotify and Netflix and chatbots, which have replaced some customer service agents [4].

Strong AI, also known as artificial general intelligence (AGI), is "the hypothetical intelligence of a machine that has the capacity to understand or learn any intellectual task that a human being can understand or learn." [3]. To accomplish this, the system would need to be able to solve problems, learn new skills, apply knowledge to tasks, adapt to changes, communicate in natural language and plan for the future. It would need to be able to carry out all the functions human intelligence can achieve. Some researchers believe strong AI may be impossible to achieve, while others agree it may happen before 2045, as stated by Ray Kurzweil in 'The Singularity is Near' [3] [4].

Two subfields of AI involve natural language processing (NLP) and computer vision. NLP is the ability of a computer to understand and manipulate text and speech the same way that a human can. NLP does this by using a combination of computational linguistics and machine learning methods [17]. Computer vision is the ability of a computer to analyse images and videos, and identify components within them. It can then take action based on what it has analysed, powered by neural networks and deep learning [9].

AI is currently being used in many industries around the world. For example:

- Research and development - to model experimental scenarios and to automate testing,
- Healthcare - to help diagnose patients by using image analysis and to monitor patients via wearable sensors,
- Finance - to predict future outcomes and trends to help improve investing,
- Education - to automate administrative tasks such as marking exams,
- Retail - to eliminate the need for cashiers by using cameras and sensors to track items customers take and then charging their account once they leave the store,
- Media and entertainment - to moderate content by scanning and then removing unwanted content,
- Transportation - to manage traffic by collecting and analysing traffic data,
- Agriculture - to produce more accurate weather forecasts and to analyse soil quality [3].

AI is developing extremely quickly and is likely to have a significant impact on the future of many industries. There are advantages to the evolution of AI-based systems. It could lead to increased productivity, as the systems would not need to take breaks, and lower costs, as employees would not need to be paid. The systems may also be

more accurate than humans. The downside to this is that it could lead to unemployment in many areas, which could be highly disruptive. As AI-based systems progress, the ethics behind what they can do also needs to be taken into consideration, particularly in the areas of security, privacy and safety [10] [18].

3.2 RQ2: What are the Types of Bias in AI-based Systems?

Data bias occurs when we use biased data to train the algorithms. This can happen if we perform data generation or data collection that does not include disadvantaged groups in the data or where they are "wrongly depicted" in the data [20]. Data generation acquires and processes observations of the real world, and delivers the resulting data for learning [21]. This data is synthesised and is "generated from a model that fits to a real data set" [22]. There is ongoing research surrounding synthetic data generation and there have been some positive results in areas such as healthcare [23].

Data collection is needed in areas where there is insufficient data readily available to train an AI-based system. There are three main methods for data collection; data acquisition, measuring and analysing [24]. Data acquisition harvests data from a variety of sources such as surveys, feedback forms, websites, datasets etc. This data must be gathered in a meaningful way that makes sense for the AI-based system being trained. This data then needs to be measured so that the model can learn correctly from it. This can be done by labelling the raw data and/or adding meaningful tags to it [25]. This process can be done manually or automatically (i.e., by another AI-based system that is trained to do so). The data can then be analysed to extract "meaningful knowledge from the data and make it readable for a machine learning model" [25].

Institutional biases include discriminatory practices that arise at the institutional level of analysis and operate in mechanisms that go beyond prejudice and discrimination at the individual level [26]. This bias is not always a result of conscious discrimination, as the algorithms and data may appear unbiased, however their output reinforces societal bias [27]. Even if an individual's negative associations, stereotypes and prejudices against outgroups are removed, the discrimination still happens, even in an ideal environment with no shared biases or prejudices. Societal bias happens within AI-based systems because it relies heavily on data generated by humans or collected via systems created by humans [28], and therefore human assumptions or inequalities are reflected in data. This can happen due to certain expectations humans have or areas they are not informed about.

A global initiative "Correct the Internet" by DDB Group Aotearoa New Zealand is trying to change bias within the internet. This is due to the fact that upon using Google to search for "who has scored the most goals in international football?", the search engine returns "Cristiano Ronaldo", even though it is a fact that Christine Sinclair holds the record. This might suggest that the Google search engine has learnt to be gender-biased, returning results that might be considered incorrect [29]. There are certain common societal definitions that become intermingled in search engines, and they sustain

societal norms, for example “football” could also potentially mean not just soccer football but also rugby football. It is perhaps unavoidable that technology of this nature will sustain existing common language usage, even if it is considered by some parties to be biased in some way. This same impact can be seen in emerging AI technology such as chatGPT, where bias in training data for large language models may be carried forward into resulting AI applications [77].

Sampling bias can occur when a dataset is created by selecting particular types of instances more than others. Therefore, when the model is trained with this dataset it can result in a group being underrepresented. For example, Amazon created an AI recruiting tool with the aim of reviewing resumes for certain positions. However, in 2015, they discovered that for software developer roles, the system was biased towards men, due to the fact that the models were trained using resumes over the past 10 years when males dominated the tech industry [30]. It therefore seems crucial to keep humans integrated in the evaluation of AI models as it is important for accurate model performance [31]. This process is called “human-in-the-loop”. Humans can correct a machine's incorrect results using their own expertise, which can improve the performance of the machine by teaching it how to handle certain data [32].

However, human evaluation bias can affect the performance of an AI model, due to the fact that human evaluators need to validate the performance of an AI-based system. An example of this would be confirmation bias, which is the tendency for a human to look for, interpret, focus and remember information that supports their own prejudices. Labels can be assigned based on prejudices or prior beliefs rather than objective evaluation [33]. Data scientists should assess the data a system uses and make a judgement based on how representative it is. If there are biases identified then the correct adjustments can be made, which means that machine learning biases tend to decrease over time and will create much more fairness than harm [34].

There can also be a design-related bias in which the biases happen due to the limitations of an algorithm or constraints on the system such as computational power, for example in Spotify's shuffle algorithm [33]. How we as humans perceive randomness is not how it is perceived in computers, because computers essentially use pseudo-random numbers [35]. Due to this, Spotify users were complaining as they were getting the same song multiple times within a shuffle. This is because each song had an equal chance of being in the order. Spotify has since changed its algorithm to make it less random - essentially more biased - to certain songs so that it appears more random to the listener [36]. This is a particularly interesting observation, as although randomness is a property that may be associated with reduced bias, there are clear instances where randomness is not attractive to human agents.

It is important to also note that not every bias within algorithms leads to discrimination or less favourable treatment. For example, some algorithms may contain biases that are justifiable in job situations such as an age limit or certain qualifications [37]. This is why algorithms have to be assessed to determine any legal implications. For

example, New York City passed a measure that bans employers from using automated employment decision tools to screen applicants without having a bias audit performed on the tool in advance [38]. Laws like this are put in place for AI-based systems in order to identify and mitigate risks.

3.3 RQ3: What are the Potential Risks/Effects of Bias in AI-based Systems?

Recent scrutiny suggests a large number of cases of discrimination were caused by AI-based systems. A couple of fields that it seems to be heavily infiltrating are risk assessments for policing and credit scores [39]. To tackle the potential risks of bias in AI-based systems we first need to understand that there are many different perspectives that can be looked at when addressing bias. The risks within technical, legal, social and ethical AI-based systems will be the key aspects that are highlighted.

From a technical perspective there are two well established approaches used to measure bias. The procedural approach, which focuses on recognizing biases in the decision-making algorithms [40], and the relational approach, which focuses on preventing biased decisions in the algorithmic output. The potential risk in procedural approaches is that interventions can be complex and difficult to implement due to the AI algorithms being too sophisticated. This leads to major upbringing of bias in these AI-based systems. They are also trained with monolithic datasets and utilise unsupervised learning structures that might make bias difficult to comprehend. With further advancements in explainable AI, procedural approaches will become more beneficial [40]. However, declaring that an algorithm is free from bias does not ensure a nondiscriminatory algorithmic output. Discrimination can appear as a consequence of bias in training.

The metrics for measuring bias from a technical perspective are called statistical measures. Statistical measures focus on investigating relationships between the algorithms' predicted outcome from the different demographic distribution to the actual outcome achieved. This measure covers group fairness. As an illustration, if 7 out of 10 candidates were given a mortgage, the same ratio from the protected group should have the right to obtain a mortgage. Despite the demand in statistical metrics, a potential risk has appeared that statistical definitions are inadequate to estimate the absence of bias in algorithmic outcomes. This is because they already assume the accessibility of verified outcomes and may ignore other critical attributes of the classified subject rather than the sensitive ones [41].

During the rise of AI-based systems, it remains unclear as to whether we can render them immune to anti-discriminatory behaviours. Anti-discrimination laws vary across countries; there is no universal law for various actions. This raises complex challenges for those engineering AI-based software systems for global audiences. For instance, in the European Union, anti-discrimination legislation is classified in Directives. 2000/43/EC is a Directive against discrimination on the grounds of race and ethnic origin, or Chapter 3 of the EU Charter of fundamental rights [42]. Whereas in the US, anti-discrimination laws are described as the “Title VII of the Civil Rights Act of 1964”.

This states that it forbids discrimination in employment in the sense of race, sex, national origin and religion [43]. A major bias risk in AI-based systems in legal trials addressing discrimination concerns the discrimination measures that characterise underrepresented groups e.g. disparate impact or disparate treatment [44] [45]; and the relevant population that is affected by the case of discrimination [46]. These two risks run parallel to the problems explored in the technical perspectives introduced earlier. The Castadena Rule asserts that the specific number of people in the protected group from an applicable population is prohibited from being smaller than 3 standard deviations from the number expected in an arbitrary selection [44]. Although such laws can relieve a number of discriminatory issues, the risk presented is quite complex as completely different scenarios could arise that the AI-based system will fail to take account of, leading to bias.

Digital discrimination is prevalent in AI-based systems as it gives a set of individuals unfair treatment based on certain characteristics such as gender, ethnicity, income and education. When people think of digital discrimination, they think about it as a technical phenomenon regulated by law. However, it needs to be taken into account more from a sociocultural perspective to be heavily cognizant of. There are infinite possibilities of what can amount to bias in AI-based systems from a social perspective. A potential risk of bias in AI, highlighted from a social perspective, is the potential of digital discrimination to fortify existing social inequalities. This phenomenon is called intersectionality [47]. Ultimately, this is formed by the heterogeneous ways that gender and race link with class in the labour market. There is no set evaluation methodology that exists amongst AI researchers to ethically assess their bias classifications, as it can be dissimilar through different contexts, and assessed differently by different people [48]. The way a dataset is defined and maintained may incorporate assumptions and values of the creator(s) [49]. Hildebrand and Koops are examining the design of sociotechnical infrastructure that allows humans to anticipate and respond as to how they are profiled [50]. Morality and associated moral values – which are not universally agreed upon – must also somehow be considered in the design of AI based systems.

From an ethical standpoint, Tasioulas states that discrimination does not need to be unlawful to be unfair [51]. Some may say that Isaac Asimov’s Three laws of Robotics could help systems address bias. However, these principles have been criticised as being quite vague in a way that makes them not helpful. Frameworks for AI ethics to prevent bias have been proposed by the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [52]. Ethical questions geared towards AI usage have been arranged into three interconnected levels [51]. One of the levels is the social morality of AI. The risk of bias in this case is catastrophic due to the human's emotional response like anger, guilt or empathy possibly having an influence on the creation of AI-based systems. Another level includes people's interaction with AI. Citizens have a right to exercise their own moral judgement in relation to appropriate codes of practice, but from a technical point of view it is not yet very clear as to what is considered bias and what is not [51].

3.4 RQ4: How can Bias be Resolved/Prevented in AI-based Systems?

Evidence of unintentional algorithmic bias in AI based systems has been documented and recorded considerably in recent investigations [53]. Algorithmic discrimination can be introduced to a system during development stages [54]. Therefore, it is critical for corporations to detect and address the origin of such biases throughout the duration of the development process. One potential procedure to accomplish this is to attempt to inhibit the presence of bias through the incorporation of appropriate specialised capabilities in the system development process [55]. However, as a result of AI-based systems being designed by humans and are also established on input data supplied by humans who are inadvertently prone to bias as well as inaccuracies, AI-based systems accidentally obtain these human qualities which are then embedded within their system [56]. Since the ambition of AI-based systems is to conclude impartial, data-driven and neutral decision making to have a positive consequence on many people's lives, algorithmic discrimination has to be mostly nullified [54]. Subsequently, it is important to keep in mind, given the various complexities of impartiality and its circumstantial nature, it is typically not feasible to completely de-bias an AI-based system or to ensure its fairness [57] [58].

There are also moral requirements about algorithmic integrity procedures, given that these procedures are designed to produce predictions that are impartial, instead of sanctioning the impartial treatment of specific humans [59]. Moreover, modern resolutions to accomplish administrative necessities are directed on aggregate-level functionality, which can conceal stratification amongst subpopulations. The first and most important stage when it comes to naturally preventing biases contained within AI-based systems, is to establish methods for identifying them. There have been numerous research projects designed at identifying discrimination in AI-based systems. However, most of them necessitate comprehensive understanding of the internal mechanics of the algorithm and/or the dataset provided. For instance, McDuff [60] advocates a structure of "classifier interrogation" which necessitates characterised data to investigate the capacity domains that may result with a bias. Furthermore, procedures for identifying bias within AI-based systems can be to some extent job specific and complex to establish. A variation of the Implicit Association Task can be utilised in order to identify prejudice in word implants. Regardless of this being a legitimate function of the initial objective of Implicit Association Task, it is ambiguous how to progress above bias identification in independent settings [61].

Bias mitigation techniques are established in the position in which these algorithms can interfere within a determined AI-based system which is based on the distinction from Calmon et al [62], which delimit three scopes of interest. If an algorithm has the ability to alter the training data, then pre-processing could be applied [63]. If the algorithm is authorized to alter the learning mechanism for a model, then in-processing could be applied [64] [65]. If the algorithm can exclusively operate the concluded model as a black box with the absence of capabilities to alter the training data or learning algorithm, then post-processing can be applied [66]. The majority of this section

concentrates on pre-processing because it is most adaptive when it comes to the data science pipeline, it is self-sufficient in the modelling procedure and can also be unified with data delivery as well as publishing processes [62] [67].

Excluded variable bias expresses that disregarding a variable is an inadequate approach to prevent prejudice as any remaining variables which correspond to the absent variable still accommodate data about that variable [68]. Furthermore, modern research has identified that in order to establish that AI-based systems do not contradistinguish, it is essential that data with respect to the variable is utilized when conceiving the algorithm [69]. One such example of this approach is the ideology of dropping the gender variable which originates from Calmon et al [62], in which they use gender elimination as a benchmark model. The composers of this publication have also noted that this technique may not always be effective as other variables that are not eliminated could potentially be associated with the protected feature, which would still commission bias.

Measuring prejudice in AI-based systems can assist the elimination of bias from data sets that have been identified to be biased, unfinished or accommodate inequitable decisions, and accordingly encourage fairness in such systems [70]. In order to accomplish this, the algorithms calibrate non-discrimination procedures into restraints translatable by a machine, and subsequently models are developed fixed on the chartered restraints. Software toolkits are presently being established which comprise statistical procedures for calibrating and mitigating bias in AI-based systems. Although it is difficult to conclude how quickly these modern toolkits are being utilised in application, their accelerated adoption advocates an exceptional necessity in the private and public sectors. Examples of such software tools include IBM's 'AI Fairness 360 Open Source Toolkit' and Accenture's 'Fairness Tool'.

IBM's 'AI Fairness 360 Open Source Toolkit' has an ambition to facilitate developers to "examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle" [72] [73]. This instrument administers tests and algorithms to calibrate fairness and prevent discrimination in dataset and models. Accenture's 'Fairness Tool' has an ambition to detect bias and potential proxies for protected attributes contained inside datasets which are utilised by algorithmic systems [71]. This instrument can abolish correlations amidst sensitive constants as well as proxies which can conclude with biased outcomes.

Unfortunately, investigating the bias-accuracy tradeoff is an unfinished picture of impartiality of an algorithm. Such deficiencies have been analysed by Dwork et al. [41] with regard to how an adversary could accomplish statistical consistency whilst still addressing the preserved party unjustly. Fish et al. [64] exhibited these deficiencies in action even with the lack of antagonistic manipulation. Along with other procedures, they have also shown that adapting a classifier by arbitrarily flipping specific output characteristics with a specific probability already substantially exceeds the previous fairness literature in bias as well as accuracy.

An additional challenge with the resolutions of the equality with the discriminating factors of AI-based systems is whether they take into consideration factual recorded inconsistencies or under-representation, or in other words meaning not all unequal outcomes are unfair. One such instance of this can be the inconsistency between the amounts of male versus female CEOs within the business industry. This instance features the socio-technical view about data and prejudice where an abundance of concerns are social at first, and technical second [74] [75]. Simultaneous amendments to these complications will only result in aggravating current problems [76].

4 Limitations of Research

Research is a crucial instrument in furthering knowledge and understanding topics in various fields. Throughout our research, we encountered many limitations. One of the preeminent limitations was the timeframe available to undertake this MLR, which was conducted over a 6-week period in January and February 2023. The reason for the time limit arises from the nature of the original assignment: a four-person team research project as part of a final year undergraduate module in Software Engineering. It is also the case that the focus of this review has largely examined data driven AI based systems, but not the earlier knowledge-driven system that were more prevalent in the 1990s but which continue to be in us today.

Given the undergraduate status of the primary researchers, there was an absence of prior formal research training. However, all researchers received instruction on the MLR technique at research outset and were furthermore engaged on a weekly basis with a senior academic researcher to direct efforts. This training and direction helped to reduce the impact of core researcher inexperience. Guidance on writing academic papers was also provided so that the core researchers could strive towards high academic quality in their work products.

A final limitation emanates from the adoption of an MLR methodology. While this methodology permits the inclusion of non-peer-reviewed work which can have a positive impact on the research, it nevertheless may result in the inclusion of materials that are not of traditional scientific standing. For example, there are various arxiv.org papers included in this work but some of them may ultimately not be successful in scientific peer review and therefore of diminished scientific value.

5 Directions for Future Research

This research has identified various complex challenges and risks that arise in AI-based software systems. Future important research could integrate important ethical and moral knowledge into software engineering education and practice. This has clear resonance with societal and cultural values and expectations, both of which vary among and within populations. This may indicate that a bumpy road ahead might be expected for AI-based software systems that seek broader population effectiveness.

It is furthermore the case that evaluating the extent to which a software system might be biased is worthy of further research. Perhaps the design and deployment of AI-enabled systems can learn from safety-critical systems development where approaches such as Failure Mode and Effect Analysis (FMEA) help to build safer systems. Building less biased (or appropriately biased) software systems might incorporate an analysis of the possible sources of bias and their impact in production systems, this could become known as Bias Effect Analysis.

6 Conclusion

This research paper provides a review of AI-based systems by exploring their history, the techniques used to train them, the types of AI-based systems there are and the areas they are currently being utilised in around the world. It also investigates the various kinds of biases that may occur in AI-based systems by presenting findings on data bias, institutional bias, societal bias, sampling bias, evaluation bias and design-related bias. It highlights the challenges of what bias in AI-based systems can do today by looking at the different perspectives of technical, legal, social and ethical risks. It also examines the approaches used to measure bias and whether discrimination can be removed from AI-based systems. Finally, it touches on some best practices suggested by AI experts to help prevent this bias. It also examined how crucial it is to identify bias in order to prevent it, and the process of removing algorithmic bias by examining pre-processing, in-processing and post-processing techniques.

Interesting questions revolve around such areas as “What is bias?”, “Is bias bad?” and “Should we aspire to remove bias?” These questions are for the most part the prevail of academic participants outside computer science and software engineering. However, given the rise of AI, we find that these questions now take on much greater significance in day-to-day life, as AI based systems are increasingly supporting decisions that affect broader swathes of society. It therefore seems wise to integrate the accumulated learning of fields such as ethics into software engineering in the near term, and certainly well before the AI-enabled systems revolution takes full effect.

Acknowledgements. This research is supported in part by SFI, Science Foundation Ireland (https://www.sfi.ie/) grant No SFI 13/RC/2094_P2 to Lero - the Science Foundation Ireland Research Centre for Software. It is also supported in part by SFI, Science Foundation Ireland (https://www.sfi.ie/) grant No SFI 12/RC/2289_P2 to Insight - the Science Foundation Ireland Research Centre for Data Analytics.

References

1. Buchanan, B.G.: A (Very) Brief History of Artificial Intelligence. AI Magazine, 26(4), 53 (2005). https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1848 Accessed 15 Feb 2023
2. Schroer, A.: What is Artificial Intelligence? Built In (2022). https://builtin.com/artificial-intelligence Accessed 15 Feb 2023
3. Lowe, A., Lawless, S.: Artificial Intelligence Foundations. BCS, The Chartered Institute for IT, London, UK (2021).
4. Glover, E.: Strong AI vs Weak AI: What's the Difference? Built In (2022). https://builtin.com/artificial-intelligence/strong-ai-weak-ai Accessed 16 Feb 2023
5. Bias. Cambridge University Press dictionary, translations & thesaurus. https://dictionary.cambridge.org/ Accessed 10 Feb 2023
6. Nelson, G.S.: Bias in Artificial Intelligence. N C Med J, 80(4):220-222 (2019).
7. Leavy, S., O'Sullivan, B., Siapera, E.: Data, Power and Bias in Artificial Intelligence (2020). https://arxiv.org/pdf/2008.07341.pdf Accessed 16 Feb 2023
8. Anyoha, R.: The History of Artificial Intelligence. Harvard University (2017). https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/ Accessed 15 Feb 2023
9. What is Artificial Intelligence (AI)? IBM (n.d.). https://www.ibm.com/topics/artificial-intelligence Accessed 15 Feb 2023
10. Taulli, T.: Artificial Intelligence Basics: A Non-Technical Introduction. Apress, NYC, USA (2019).
11. Burns, E.: What is Artificial Intelligence (AI)? TechTarget (2023). https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence Accessed 15 Feb 2023
12. What is Big Data? Oracle (n.d.). https://www.oracle.com/ie/big-data/what-is-big-data/ Accessed 15 Feb 2023
13. Mahesh, B.: Machine Learning Algorithms - A Review. International Journal of Science and Research, 9(1) (2020).
14. How Does AI Actually Work? CSU Global (2021). ndyhttps://csuglobal.edu/blog/how-does-ai-actually-work Accessed 15 Feb 2023
15. What are Neural Networks? IBM (n.d.). https://www.ibm.com/topics/neural-networks Accessed 16 Feb 2023
16. LeCun, Y., Bengio, Y., Hinton, G.: Deep Learning. Nature 521:436-444 (2015). https://www.nature.com/articles/nature14539 Accessed 16 Feb 2023
17. Kurzweil, R.: The Singularity is Near: When Humans Transcend Biology. Duckworth Books, London, UK (2016).
18. What is Natural Language Processing (NLP)? IBM (n.d.). https://www.ibm.com/topics/natural-language-processing Accessed 16 Feb 2023
19. Bundy, A.: Preparing for the Future of Artificial Intelligence. AI & Society 32:285-287 (2016).
20. Lopez P: Bias does not equal bias: A socio-technical typology of bias in data-based Algorithmic Systems. Internet Policy Review. (2021)
21. Hellstrom T, Dignum V, Bensch S: Bias in machine learning - what is it good for? - arxiv. (2020) https://arxiv.org/pdf/2004.00686.pdf. Accessed 16 Feb 2023
22. Hernandez M, Epelde G, Alberdi A, et al: Synthetic Data Generation for Tabular Health Records: A systematic review. Science Direct. (2022)

23. Rankin D, Black M, Bond R, et al: Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for Data Sharing. JMIR Medical Informatics. (2020)
24. Creative AI: What is data collection? Creative AI. (2022) https://creative-ai.tech/en/what-is-data-collection/. Accessed 7 Feb 2023
25. Kniazieva Y: Data collection. High quality data annotation for Machine Learning, (2022) https://labelyourdata.com/articles/data-collection-methods-AI. Accessed 16 Feb 2023
26. Henry PJ: Institutional Bias. The Sage Handbook of Prejudice, Stereotyping and Discrimination. p.426 (2010)
27. Kulkarni A: Bias in AI and machine learning: Sources and solutions, Lexalytics. (2022) https://www.lexalytics.com/blog/bias-in-ai-machine-learning/. Accessed 2 Feb 2023
28. Ntoutsi E, Fafalios P, Gadiraju U, et al: Bias in data-driven Artificial Intelligence Systems—an introductory survey. WIREs Data Mining and Knowledge Discovery. p.3 (2019)
29. Mishra N: Break the bias, 'correct the internet' to make women in sports more visible, Campaign Asia. (2023) https://www.campaignasia.com/video/break-the-bias-correct-the-internet-to-make-women-in-sports-more-visible/483047. Accessed 10 Feb 2023
30. Dastin J: Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. (2018)
31. Wodecki B: Human evaluation of AI is key to success - but it's the least funded. AI Business. (2022)
32. Maadi M, Akbarzadeh Khorshidi H, Aickelin U: A review on human-ai interaction in machine learning and insights for medical applications. International journal of environmental research and public health. p.4 (2021)
33. Srinivasan R, Chander A. Biases in AI systems. Communications of the ACM. 2021 Jul 26; 64 (8): 44-9
34. Moschella D: Machines are less biassed than people. Verdict. (2019) https://www.verdict.co.uk/ai-and-bias/. Accessed 7 Feb 2023
35. Rubin JM: Can a computer generate a truly random number? Mit Engineering. (2011) https://engineering.mit.edu/engage/ask-an-engineer/can-a-computer-generate-a-truly-random-number/. Accessed 10 Feb 2023
36. Brocklehurst H: Ever feel like the Spotify Shuffle isn't actually random? Here's the algorithm explained. The tab, (2021) https://thetab.com/uk/2021/11/17/spotify-shuffle-explained-228639 Accessed 6 Feb 2023
37. Bias in Algorithms – Artificial Intelligence and Discrimination. European Union Agency For Fundamental Rights. (2022)
38. Mithal M, Wilson Sonsini Goodrich & Rosati: Legal requirements for mitigating bias in AI Systems. JD Supra. (2023) https://www.jdsupra.com/legalnews/legal-requirements-for-mitigating-bias-3221861/. Accessed 10 Feb 2023
39. Weapons of math destruction: How big data increases inequality and Threatens Democracy', Vikalpa: The Journal for Decision Makers, 44(2), pp. 97–98. https://journals.sagepub.com/doi/10.1177/0256090919853933. Accessed 21 Feb 2023
40. Mueller ST, Hoffman RR, Clancey W, et al. Explanation in human-AI systems: A literature meta-review, Synopsis of key ideas and publications, and bibliography for explainable AI. In: arXiv.org. https://arxiv.org/abs/1902.01876. (2019) Accessed 20 Feb 2023
41. Dwork C, Hardt M, Pitassi T, et al. Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. (2012) doi: 10.1145/2090236.2090255

42. Council directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin. Pharmaceuticals, Policy and Law 13:301–310. (2011)doi: 10.3233/ppl-2011-0332
43. Title VII of the Civil Rights Act of 1964. In: US EEOC. https://www.eeoc.gov/statutes/title-vii-civil-rights-act-1964. Accessed 22 Feb 2023
44. Barocas S, Selbst AD. Big Data's disparate impact. SSRN Electronic Journal. (2016) doi: 10.2139/ssrn.2477899
45. Feldman M, Friedler SA, Moeller J, et al. Certifying and removing disparate impact. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2015) doi: 10.1145/2783258.2783311
46. Romei A, Ruggieri S. A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review 29:582–638. (2013) doi: 10.1017/s0269888913000039
47. Walby S, Armstrong J, Strid S. Intersectionality: Multiple inequalities in social theory. Sociology 46:224–240. (2012) doi: 10.1177/0038038511416164
48. Miller T. Explanation in artificial intelligence: Insights from the Social Sciences. Artificial Intelligence 267:1–38. (2019) doi: 10.1016/j.artint.2018.07.007
49. Van Nuenen T, Ferrer X, Such JM, Cote M. Transparency for whom? assessing discriminatory artificial intelligence. Computer 53:36–44. (2020) doi: 10.1109/mc.2020.3002181
50. Hildebrandt M, Koops B-J. The challenges of ambient law and legal protection in the profiling era. Modern Law Review 73:428–460. (2010) doi: 10.1111/j.1468-2230.2010.00806.x
51. Tasioulas J. First steps towards an ethics of robots and Artificial Intelligence. SSRN Electronic Journal. (2018) doi: 10.2139/ssrn.3172840
52. The IEEE global initiative on ethics of autonomous and intelligent systems. In: IEEE Standards Association. https://standards.ieee.org/industry-connections/ec/autonomous-systems/. (2023) Accessed 20 Feb 2023
53. Edelman B, Luca M, Svirsky D. Racial discrimination in the sharing economy: Evidence from a field experiment. American Economic Journal: Applied Economics (2017) 9:1–22. doi: 10.1257/app.20160213
54. LEMONNE E. Ethics guidelines for Trustworthy Ai. In: FUTURIUM - European Commission. https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html. (2021) Accessed 22 Feb 2023
55. Dignum V. Ethics in artificial intelligence: Introduction to the special issue. Ethics and Information Technology 20:1–3. (2018) doi: 10.1007/s10676-018-9450-z
56. Kirkpatrick K. Battling algorithmic bias. Communications of the ACM 59:16–17. (2016) doi: 10.1145/2983270
57. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR). 2021 Jul 13; 54(6): 1-35
58. Selbst A, boyd danah, Friedler S, et al. Fairness and Abstraction in Sociotechnical Systems. In: Sorelle Friedler. http://sorelle.friedler.net/papers/sts_fat2019.pdf. (2019) Accessed 18 Feb 2023
59. Hughes S, Aiyegbusi OL, Lasserson D, et al. Patient-reported outcome measurement: a bridge between health and social care? In: WARWICK. http://wrap.warwick.ac.uk/151243/7/WRAP-Patient-reported-outcomes-measures-what-are-benefits-social-care-2021.pdf. (2021) Accessed 17 Feb 2023
60. McDuff D, Ma S, Song Y, Kapoor A. Characterizing Bias in Classifiers using Generative Models. In: arxiv. https://arxiv.org/pdf/1906.11891.pdf. (2019) Accessed 22 Feb 2023
61. Caliskan A, Bryson J, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. In: Science. https://www.science.org/doi/10.1126/science.aal4230. (2017) Accessed 9 Feb 2023

62. Calmon FP, Wei D, Vinzamuri B, et al. Optimized Pre-Processing for Discrimination Prevention. In: NeurIPS Proceedings. https://proceedings.neurips.cc/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf. (2017) Accessed 20 Feb 2023
63. Hajian S. Simultaneous Discrimination Prevention and Privacy Protection in Data Publishing and Mining. In: arxiv. https://arxiv.org/pdf/1306.6805.pdf. (2013) Accessed 22 Feb 2023
64. Fish B, Kun J, Lelkes ÁD. A confidence-based approach for balancing fairness and accuracy. Proceedings of the 2016 SIAM International Conference on Data Mining. (2016) doi: 10.1137/1.9781611974348.17
65. Zafar MB, Valera I, Rodriguez MG, Gummadi KP. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In: Proceedings of the 26th international conference on world wide web, pp. 1171-1180. 2017.
66. Hardt M, Price E, Srebro N. Equality of Opportunity in Supervised Learning. Advances in neural information processing systems (29). 2016.
67. Vigild DJ, Johansson L, Feragen A. Identifying and mitigating bias in machine learning models. Thesis, Technical University of Denmark (2021) 11-18
68. Clarke KA. The phantom menace: Omitted variable bias in econometric research. Conflict Management and Peace Science 22:341–352. (2005) doi: 10.1080/07388940500339183
69. Žliobaitė I, Custers B. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. Artificial Intelligence and Law 24:183–201. (2016) doi: 10.1007/s10506-016-9182-5
70. Žliobaitė I. Measuring discrimination in algorithmic decision making. Data Mining and Knowledge Discovery 31:1060–1089. (2017) doi: 10.1007/s10618-017-0506-1
71. Duggan J. Fairness you can bank on. In: Accenture. https://www.accenture.com/ie-en/case-studies/applied-intelligence/banking-aib. (2023) Accessed 15 Feb 2023
72. Varshney KR. Introducing AI Fairness 360. In: IBM. https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/. (2018) Accessed 15 Feb 2023
73. Bellamy RK, Dey K, Hind M, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development. (2019) doi: 10.1147/jrd.2019.2942287
74. Lloyd K. Bias amplification in artificial intelligence systems. In: arXiv.org. https://arxiv.org/abs/1809.07842. (2018) Accessed 14 Feb 2023
75. Oakley JG. Gender-based Barriers to Senior Management Positions: Understanding the Scarcity of Female CEOs. Journal of Business Ethics 27:321–334. (2000) doi: 10.1023/a:1006226129868
76. Domino. On ingesting Kate Crawford's "The trouble with bias". In: Domino Data Lab. https://www.dominodatalab.com/blog/ingesting-kate-crawfords-trouble-with-bias. (2022) Accessed 22 Feb 2023
77. Zhai, X., 2023. ChatGPT for next generation science learning. XRDS: Crossroads, The ACM Magazine for Students, 29(3), pp.42-46.
78. Garousi V, Felderer M, Mäntylä MV. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. Information and software technology. 2019 Feb 1; 106: 101-21.