UNIVERSITY OF BIRMINGHAM

How Does Bias Affect Users of Artificial Intelligence Systems?

By

Hebah Abdullah Bubakr.

A thesis submitted to
The University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

School of Engineering

Department of Electronic, Electrical and Systems Engineering

The University of Birmingham

September 2022

ABSTRACT

In large companies, artificial intelligence (AI) is being used to optimise workflow and ensure efficiency. An assumption is that the AI system should remain unaffected by bias or prejudices to contribute to providing fairer results. For example, in the recruitment process, AI ensures that each applicant is judged by the exact criteria in the job description. Our results suggest otherwise; therefore, we wondered whether the problem of bias extends from the training data (which, replicates existing inequalities in organisations) to the design of the AI systems themselves. These learning systems are dependent on knowledge elicited from human experts. However, if the systems are trained to perform and think in the same way as a human, most of the tools would use unacceptable criteria because people consider many personal parameters that a machine should not use. The question remains whether the potential impact of bias is considered in the design of an AI system.

In this thesis, several experiments are conducted to study unconscious bias in the application of AI with the aid of two qualitative frameworks and two quantitative questionnaires. We first explore the unconscious bias in user interface designs, then examine programmers' understanding of bias when creating a purposely biased machine using medical databases. A third study addresses the effect of AI recommendations on decision-making, and finally, we explore whether user acceptance is dependent on the type of AI recommendation, testing various suggestions.

This project raises awareness of how the developers of AI and machine learning might have a narrow perspective of 'bias' as a statistical problem rather than a social or ethical problem. This limitation is not because they are unaware of these wider concerns but because the requirements relating to the management of data and the implementation of algorithms might restrict their focus to technical challenges. Consequently, bias outcomes can be produced unconsciously because developers are simply not attending to these broader concerns. Creating accurate and effective models is important but so is ensuring that all races, ethnicities and socioeconomic levels are adequately represented in the data model (O'Neil, 2016).

ACKNOWLEDGEMENTS

To Prof. Chris Baber, my supervisor, for the amazing experience and invaluable guidance. I thank you from the bottom of my heart for your encouragement, patience and support.

Thanks to my family for the continued love, support and encouragement through the hard times. Without you, none of this would have been possible.

I am grateful to the King Faisal University and the Saudi Arabian Culture Bureau for funding this degree.

This thesis was copy edited for conventions of language, spelling and grammar by Cambridge Proofreading.

Contents

List of Tables

List of Figures

Conference proceedings and list of publications

- Presented my work at the World Congress on Artificial Intelligence and Machine Learning, 24–25 October 2019, Valencia, Spain, with the title 'Explore the Differences in Human and Automated Hiring Decisions'. In this paper I collected the data, did the analysis and draw the conclusion. Prof. Chris Baber helped with the structure, the organization of the paper and correct any language mistakes. World Congress on Artificial Intelligence and Machine ...

- Presented my paper 'Using Toulmin Model of Argumentation to Explore the Differences in Human and Automated Hiring Decisions' at VISIGRAPP 2020 (15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications), Valletta, Malta, 27–29 February 2020. I wrote this paper, did the analysis and presented the conclusion. Prof. Chris Baber provided a better organization of the paper and correct any language mistakes. . Using the Toulmin model of argumentation to explore the differences in human and automated hiring decisions

- Participated in the 'Virtual Ergonomics & Human Factors Conference', 27–28 April 2020, with a poster + recorded presentation 'How Sensemaking by People and Artificial Intelligence Might Involve Different Frames'. I wrote this paper, did the analysis and presented the conclusion. Prof. Chris Baber provided a better organization of the paper and correct the language mistakes. https://publications.ergonomics.org.uk/uploads/How-sensemaking-by-people-and-artificial-intelligence-might-involve-different-frames.pdf

- Presented my paper 'Conflict Between Frames When Using Machine Learning' at the Naturalistic Decision-making (NDM) conference 21–24 June 2021. In this paper I collected the data, did the analysis and presented the conclusion. Prof. Chris Baber helped with the structure, the organization of the paper and correct the language mistakes. **Competition** and Conflict Between Frames in Using Machine Learning

Chapter 1

Introduction and research questions

In this thesis, the following research questions are addressed.

- How do people use their own knowledge and experience to interpret the output of artificial intelligence/machine learning (AI/ML) algorithms?
- How do programmers creating AI/ML systems make sense of the values and beliefs encoded in their algorithms?
- How do AI system recommendations affect people's decisions?
- What factors are used to determine the acceptance of AI systems?

The reason for selecting these research questions is because while several research traditions focus on bias in human decision-making and social behaviour (which are reviewed in Chapter 2), the question of bias in computing is less agreed. In part, this conflict stems from whether the focus lies on a statistical definition of bias, that is, the manner in which data are manipulated, or a social definition describing the consequences of how the analysis is interpreted. AI models are complex as they embed assumptions that could be prejudicial. However, it is difficult to tell whether these assumptions were intended by unethical developers or simply arose from a systematic mistake. Therefore, we saw the need to ask questions related to how people interpret the output and how the programmers understand bias within their work. Companies such as Google, Amazon and Facebook precisely tailor their algorithms but do not normally allow these to be transparent (O'Neil, 2016). This obscurity makes it difficult to definitively answer questions such as whether the model works against the user's interest, is unfair, or even damages and destroys lives.

We focused on the relationship between these biased systems and people interpretation and acceptance to their outcomes. And while there were several studies address bias in AI (reviewed in Chapter 2) few of them mentioned the ethical consideration of the developments. An example that supports this statement is a recent study that address the process by which data scientists work with machine learning models (ML). The researchers conducted a semi-structured interviews asking the scientists about the process of creating the ML models, what difficulties they might face and other several questions related to the development of ML and AI models. The scientists' answers did not mention any ethical consideration of these model outcomes (Baweja et al.,2023). That is why this thesis investigates how bias can affect the development, use and acceptance of AI systems. A key aim is to highlight how people

understand and respond to ‘bias’ in AI systems. The definition of an AI system for this thesis extends from simply an algorithm to the way the algorithm is programmed and how people use the output of the algorithm. Previous papers have explored bias in automation, but few studies have examined the programmers’ influence on the design of AI systems. Although some prior studies have investigated users’ acceptance of automation, less research has been conducted on whether acceptance is moderated by the context in which AI systems provide recommendations. In this thesis, the first two works explore the relationship between the programmer, data and algorithms. The third and fourth studies highlight the interaction between the AI system and the user. The results of these studies lead to guidelines for the development of AI systems.

The definition of bias is controversial, in each domain the term is defined differently. In AI, as we stated, the conflict stems from whether the focus lies on a statistical definition of bias, or a social definition describing the consequences. For this thesis, bias is defined as *any recommendation or output from an AI system that can be considered socially unacceptable due to an ill-trained algorithm, unrepresentative data, or a consciously or implicitly biased programmer.*

We consider how the processes of sense-making might differ in the components of a human–AI system (which would include users, AI algorithms, and developers) and the relationships between them in terms of values, beliefs and expectations. In many scenarios, the decisions made by AI might not be acceptable according to the human values and beliefs affected by these decisions (O’Neil, 2016). This unacceptability might be due to failure of the algorithm but more likely arises from a selection of data that might inadvertently impose a ‘frame’ that creates a biased decision.

For example, an experimental hiring tool from the Amazon corporation showed gender bias due to male dominance in the training (Dastin, 2018), which is one of the causes of the issue. Using historical data, the AI system reflected the underlying biases and discrimination in society. For the Amazon hiring tool, female applications might be rejected automatically because their career might be interrupted by the duties of motherhood. Therefore, algorithmically, the system would regard female applicants as high risk. Other research has shown that people of colour might suffer adverse consequences in education, courts and hiring (Levinson & Young, 2009; Mitchell et al., 2005; O’Neil, 2016). In those case studies, the hiring system tools were designed to help companies and employers to select the right employees. However, the results were racist, sexist, prejudiced and discriminatory.

O'Neil (2016) suggests that the desire to use analytical science in human resources is to make selection fairer by removing the potential for human bias, yet the AI systems are increasingly being used to filter applicants. This raises the question of whether these systems are optimised for the objective of selection (to aid in picking the most appropriate person for the role) or rejecting applicants (to aid in filtering applications down to a manageable set). We propose that these objectives represent distinct forms of argumentation that emphasise different aspects of the resume and test scores.

The AI Committee of the British Parliament states that:

> the development of intelligible AI systems is a fundamental necessity if AI is to become an integral and trusted tool in our society. Whether this takes the form of technical transparency, explainability, or indeed both, will depend on the context and the stakes involved, but in most cases, we believe explainability will be a more useful approach for the citizen and the consumer.[1]

Recent developments in regulations in Europe and countries globally have emphasised the need for ethical approaches to the development of AI. The AI act of the European law on artificial intelligence[2] assigns applications of AI to three risk categories. First, systems that create an unacceptable risk**,** such as government-run social scoring of the type used in China, are banned. Second, high-risk applications, such as a CV-scanning tool that ranks job applicants, are subject to specific legal requirements. Lastly, applications not explicitly banned or listed as high-risk are largely left unregulated. The Act lays down some rules to deal with each type.

Moreover, many approaches can be related to the principles of fairness, accountability and transparency in machine learning (FATML),[3] which question the responsibility for algorithms. FATML queries examine sources of error and uncertainty that affect algorithms, how algorithms explain their decisions and whether a third party checks the behaviour of the

---

[1] 'AI in the UK: Ready, Willing and Able?' report, UK Parliament (House of Lords), Artificial Intelligence Committee, 16 April 2017. https:// publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002 .htm.

[2] 'https://artificialintelligenceact.eu/the-act/

[3] https://www.fatml.org/

algorithm. Moreover, whether the algorithm is guaranteed to prohibit bias against different demographics is examined.

Components of AI Systems

In this thesis, a systems perspective is adopted where the system is defined as including human and technical components. An overarching aim of this research is to understand the relationship between the components of AI/ML systems, namely programmer, algorithm, data and user. The focus is not only on the algorithm or data (although these are essential) but also on various roles humans have in the AI system, for instance, the programmers who develop and deploy AI systems and the users who interact with the systems.

Each component in an AI/ML system can influence the recommendations that are made. For example, a programmer or a user could exhibit biases in decision-making that can affect how an algorithm is programmed or a recommendation is interpreted. The algorithm could be tailored to suit particular measures of performance, and this could limit how well it applies to non-standard or edge cases. Moreover, the data might be collected from sets that reflect biases in society (e.g., conviction rates for particular types of crime might disproportionality include particular races) or organisations (e.g., a company might be dominated by male employees).

In many aspects of our lives, a data model or software can play a significant role in how we make decisions, either by selecting relevant information for us or by proposing a decision. For instance, search engines tailor the advertisements presented to their users based on information gleaned from individual online search histories (Woodson, 2018). Grocery stores use algorithms based on shopping habits obtained from loyalty cards to tempt consumers with coupons for future purchases (Woodson, 2018). Furthermore, data models can determine whether applicants are suitable for jobs or decide whether we are good citizens or even criminals (Dickson, 2018).

The danger of developing these models in domains such as health, justice and banking is that rather than providing intelligent aid and services, they can provide the exact opposite.

> Establishing broad norms…exert upon us something very close to the power of law. If a bank's model of a high-risk borrower, for example, is applied to you, the world will treat you as just that, a deadbeat – even if you're horribly misunderstood. And when that model scales, as the credit model has, it affects your whole life whether you can get an apartment or a job or a car…. (O'Neil, 2016, P33.)

To consider these issues further, we contemplate a media recommender system that suggests a TV series or a film to a user. The programmer develops an algorithm that relies on previous media viewed by the user to compare views with predefined categories. This creates a 'personal' (i.e., defined by viewer behaviour) model of 'interest' based on the individual's viewing history. The system might use parameters such as a film's genre, date of release, duration, title, actors' names or soundtrack, and combine all that information into clusters of these parameters. From these clusters, the model of a specific viewer can be constructed and used to make recommendations.

However, these suggestions are not completely accurate. Some suggestions might be rejected by the user, particularly when the user's interest is based on different criteria than the system uses. For instance, a parent of small children might have many children's films in the viewing history, but those are not defined as 'favourite' films of their own. When a recommendation is provided, one of two reactions is expected from users: (1) agree with the recommendation because it matches their interests, and the system will use the same category to build future suggestions; (2) reject the recommendation, where the system will adjust its model.

To better understand the process for recommending movies, consider that the programmer assumes that the behaviour of viewers of films and movies can be described by clusters of parameters that define the films. These clusters, however, might not regard aspects of the person's life, such as time of day or day of the week, which might indicate that 'children's films' are watched before 6 pm and other types of films are watched after 9 pm. Furthermore, the programmer assumes that a 'preference' (for type of movie) can be itemised in this manner, and that 'personalisation' (through these models) is desirable and useful to users. Likewise, the 'model' that the recommender system employs assumes that the clusters it uses represent a complete 'space' of combinations of parameters.

Finally, the user responds to the recommendations considering whether they are preferable. Acceptance or rejection is not simply a matter of the recommendation being true or false, but rather that the user can recognise the recommendation as being something that is or is not interesting. In some cases, we might expect the recommendation to offer a movie that the viewer has not heard of or one the customer might reject because the algorithm would offer little benefit if only currently known or familiar films were recommended. This suggests that the user expects the recommendation to offer novel or unfamiliar titles. Underlying these assumptions and expectations is a set of poorly defined values and beliefs that are not incorporated into either the model or the user's understanding of how to use the

recommendation system. That is, the 'system' provides a way of making choices (for what movie to watch) that relies on beliefs about what makes movies attractive (to that individual) and what makes one movie preferable to others.

While recommending or choosing movies might feel like a low-risk activity, there are other models which have more significant consequences (O'Neil, 2016). Such models carry assumptions and predictions or present results which can be flawed, resulting in faulty conclusions that affect society with many consequences. For these reasons, this research investigates how developers' values are input into these systems consciously or otherwise. Furthermore, this study examines whether the generated algorithms reflect the developers' own values and beliefs or produce socially acceptable results. AI algorithms use deep learning and neural networks to function; thus, we cannot always have an exact idea of how the results were produced. Even when the programmers of AI algorithms are asked to explain the logic behind their decisions, they often do not have good answers (Dickson, 2018).

Can automation reduce bias?

While the question of bias is central to this thesis and touches on the questions relating to responsibility, accuracy, and checking, it is also essential to consider explanations. According to Hoffman et al. (2018), explainability often focuses on *how* questions, that is, how does the machine do what it does? The essential aim of explanation is to provide a robust mental model of a system for its users and offer users reasons for the decisions made by an automated system.

> When AI is used as a decision aid, users would seek to use explanations to improve their decision-making. If the system behaved unexpectedly or erroneously, users would want explanations for suitability and debugging to be able to identify the offending fault and take control to make corrections. (Wang et al., 2019, P4.)

Therefore, according to the European Commission's High-Level Expert Group on Artificial Intelligence (2019), it is vital to ensure that 'the capabilities and purpose of AI systems are openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected'. An explanation is important to gain user trust, especially to explain why a model has generated a particular output. As Olhede and Wolfe (2017) remarked,

> Automation in decision-making seems attractive in its potential to remove the bias and idiosyncrasy of human decision-making. Yet anything automated must be designed,

and this design requires input data generated by real human interactions, and so is liable to reflect any existing inequities.

This inherent bias implies that developers must adopt an ethical perspective and not let their own personal values interfere as they supervise the automation of decision-making.

In automated support for recommendations, there is a need to know whether human bias decisions have been reduced with AI or simply camouflaged with technology (O'Neil 2016). In this thesis, the question relates to whether the result of the model was intended by the programmer, and whether this result is acceptable (or is at least comprehensible) to the user. The new AI models are complex, embedding and hosting various assumptions, some of which could be prejudicial. Some major companies, according to O'Neil (2016), tailor their algorithms and normally will not allow for transparency. This obscurity makes it challenging to definitively respond to assumptions such as whether the model purposely produced results against the subject's interest. Moreover, it is difficult to ascertain whether the AI was implemented to ensure efficiency and reduce costs without aiming to affect or damage people's lives. O'Neil (2016) recounts a study in 2013 conducted by the New York Civil Liberties Union which found that the percentage of Black and Latino males between the ages of 14 and 24 subject to police stop-and-frisk checks was 40.6% while they were only 4.7% of the city's population. Additionally, over 90% of those stopped were innocent. O'Neil (2016) said, 'Some of the others might have been drinking underage or carrying a joint. And unlike most rich kids, they got in trouble for it'. The problem is these biased data and percentages is what most AI models are learning from to produce predictions and recommendations, at least in some societies.

Another model described by O'Neil (2016) is the Level of Service Inventory–Revised (LSI–R). The model was created in the 1990s as a tool to provide the criminal justice system even-handedness and efficiency. The model was based on the answers provided by prisoners about their criminal history as well as personal information regarding their circumstances. Statisticians created the model in which prisoners' answers were highly correlated to convictions. The more information the prisoners provided, the heavier the weight was going to be and therefore, more points. After answering the questionnaire, convicts were categorised as high, medium and low risk on the basis of the number of points they accumulated. Sometimes the model was used to help convicts by recommending anti-recidivism programs for high-risk score prisoners. However, in some cases, the scores could be used by judges to assist in sentencing offenders.

According to O'Neil (2016), LSI–R could also help non-threatening criminals land lighter sentences. However, sentencing models that account for a person's circumstances help to create an environment that justifies assumptions because it is based on a questionnaire that judges the prisoner by details that would not be permitted in court. Thus, the model is unfair. While many might benefit from it, the model leads to suffering for others. A vital component of this suffering is the destructive feedback loop that emphasises biased elements. This destructive loop continues to circle, and in the process, becomes increasingly unfair (O'Neil, 2016).

Marsh (2019) provides another example of the unfair impact of these models in a piece in *The Guardian* '*Wed 10 Apr 2019'*. A policy and campaigns officer at Liberty said that when decisions were made on the basis of arrest data, it was 'already imbued with discrimination and bias from the way people policed in the past' and was 'entrenched by algorithms'. The officer added, 'One of the key risks with that is that it adds a technological veneer to biased policing practices. People think computer programs are neutral, but they are just entrenching the pre-existing biases that the police have always shown'. Using freedom of information data, the report finds that at least 14 forces in the UK are using algorithm programs for policing, have previously done so or conducted research and trials into them. 'This means the public can't hold the programs to account – or properly challenge the predictions they make about us or our communities,' said the officer. 'This is exacerbated by the fact that the police are not open and transparent about their use.'

Thesis organisation

In this thesis, both qualitative and quantitative methods are used to answer the research questions. Qualitative approaches include the critical decision method (CDM) interviewing technique, which is interpreted in two frameworks for sensemaking: Toulmin argumentation diagrams and the data frame model (DFM). These methods are used to explore how people interpret AI systems. Quantitative approaches involve statistical analyses of survey data from two questionnaire studies that explore perceptions of bias and user acceptance of AI systems.

According to Kahneman (2011), decision-makers usually make a better choice when they expect their decision to be judged by how it was made, not only by how it turned out. When people make a decision or solve a problem there are distinctive patterns of thinking that could lead them into making errors, potentially due to biases. Thus, our concern in this research is with the relationship between algorithms, data and output during the development of AI/ML systems. We explore whether the type and reliability of recommendations matter.

Furthermore, we assess whether programmers understand not only automation biases and systematic errors but also social biases they have encoded into their developing systems.

Chapter 2 contains a review of the literature on bias in different sectors, decision-making processes and the hiring process prior and after implementation of AI algorithms. In Chapter 3, a description and explanation of the frameworks used in this thesis are offered. To visualise complex concepts, a motivation example is shown to demonstrate how models function and present data.

Chapter 4 presents research where participants are tasked with creating a job application for a fictional position. The study examines the process in which developers create the system and select its parameters, which can lead to biased output. The study explores developers' understanding of their own selection and the consequences that would make sense to them. Chapter 5 describes a study in which a group of students was asked to design and build a 'bias machine' capable of producing biased recommendations and decisions from medical data. This study explores the factors and values used by the developers to define bias in AI using DFM.

In Chapter 6, a study is presented on people's ability to select a candidate correctly based on the candidate's qualifications. The study examines how confident are they with their answers, particularly when the AI system offers wrong recommendations for some candidates. Chapter 7 describes a study examining users' acceptance of AI recommendations in various scenarios. We suggest several factors that might help people to accept the AI systems in various applications. Finally, Chapter 8 offers conclusions of the studies, possible limitations of the thesis and suggested future work. Guidelines for AI/ML system developers are offered arising from this research.

Chapter 2

Literature Review and Related Work

This chapter provides several definitions of bias in various domains, namely bias in psychology, bias in automation and computation, bias in decision-making and bias in the hiring process. The primary use-case for the thesis involves the role AI systems play in personnel selection. It is well known that bias has long plagued the personnel selection domain, and recent technological developments to support personnel selection have been proposed to reduce such bias. Therefore, it is reasonable to test how AI systems might remove or retain bias.

Human biases

In this section, we present bias according to human understanding and interpretation. When people interact with AI or any kind of automation, they will understand its results, accept its products and use its output based on the knowledge and experience they had prior to the interaction. Therefore, we found it critical to investigate and understand human biases to address the first question of the thesis, 'How do people use their own knowledge and experience to interpret the output of artificial intelligence/machine learning (AI/ML) algorithms?' The study of human decision-making has often sought to explore the degree to which decisions are rational (Hilbert, 2012; Reynolds, 2012; Tversky & Kahneman, 1974) and the extent to which these decisions exhibit cognitive bias (Haselton, 2015; Hilbert, 2012). Kahneman's (2011) popularised summary of research on decision-making proposed two systems of human thinking:

- System 1 operates automatically and quickly, with little or no effort and no sense of voluntary control.
- System 2 allocates attention to the effortful mental activities that demand it, including complex computations. The operations of System 2 are often associated with the subjective experience of agency, choice, and concentration. (Kahneman, 2011, p. 22)

While System 1 enables people to develop rapid rules (or heuristics) for decisions in familiar situations, these could offer a potential for bias. For example, we might believe that we are making a choice after reasonable consideration (using System 2). However, we tend to respond more quickly toward familiar situations and act unconsciously (using System 1) to

produce a response that most of the time is appropriate but could be inappropriate for the given situation. Thus, System 1 thinking – without conscious checks and balances – has the potential to produce biases, that is, systematic errors, in decision-making. This situation suggests one form of bias in human decision-making: a heuristic that has proved reliable in the past is applied inappropriately. Such biases have been studied extensively in cognitive psychology, typically through laboratory experiments in which participants are provided with information and asked to make a decision. On occasion, the participant might misconstrue the information and make a wrong decision.

For Tversky and Kahneman (1974), wrong decisions indicate that people 'frame' the information that they are provided, and such framing could lead to bias in decision-making. As the frames (reflecting either heuristics or the ways in which people interpret the information provided) tend to produce the same errors, this bias is regarded as systematic. Following the seminal research, cognitive psychology has studied a wide range of biases, for example, cognitive bias, which is defined as 'the mental behaviours that prejudice decision quality in a significant number of decisions for a significant number of people' (Arnott, 2006). The behaviours usually are imminent in human thinking and reasoning. Bias has also been described as a variation from critical judgement leading to irrational conclusions. A rational decision is made by the decision-maker's current assessments and the possible consequences of the choice (Hastie & Dawes, 2009). Human decision-making is vulnerable to cognitive biases that can often negatively affect decision quality. Social bias, also known as an 'attributional error, occurs when we unwittingly or deliberately give preference to (or look negatively upon) certain individuals, groups, or races due to the systemic errors that arise when people try to develop a reason for the behaviour of certain social groups' (Haggarty-Weir, 2013, Part 2). Many of these social biases can occur naturally and not out of harmful intent. Examples of biases include:

- 'Cultural bias' occurs when we tend to judge other aspects based on our own culture or by the standards of a particular culture. The term 'has been used generically to pertain to a variety of sociodemographic comparison groups, including racial, ethnic-cultural, and socioeconomic groups, whose essences are presumed to be misrepresented by standard tests used for assessment or for decision-making purposes' (Helms, 2010).
- 'Racial bias' is the disparate treatment minorities of ethnic groups (coloured people) received by white people in court, decision-making and social judgment. This unfair

treatment can affect the distribution of data when creating an AI model (Levinson & Young, 2009; Mitchell et al., 2005).

- ‘Gender bias’ includes prejudiced acts or ideas based on the gender-based perception that women and men are not equal. This term is also interpreted as differentiating people as male and female based on gender or gender-based functions, treating them uniquely in the matter of social functions or treating them unjustly in the distribution or burdens and benefits in society (Mukherjee, 2015).

In many AI/ML research papers, bias refers to the prejudice or preference towards some things, people, or groups over others (AI, 2019; Arnott, 2006; McDuff et al., 2018). These biases can affect the collection and interpretation of data, the design of a system and how users interact with a system. Biases in AI can encompass several definitions and be caused by many reasons. Note that AI/ML are simply tools for continuing the patterns they learn from – their outputs are related to their inputs.

Other forms of AI and design bias include ‘automation bias’, which occurs when a decision-maker prefers the choices that are made by automated systems even when these systems make errors compared to other information provided by non-automated agents. In particular, automation bias is mostly defined as the scenario where a human interacts with an automated system that is providing false or incorrect advice but the decision-maker (human) does not question the suggested decision or seek additional information (Adensamer et al., 2021; Parasuraman & Manzey, 2010). ‘Algorithmic bias’ occurs when a model algorithm creates systematic and repeatable errors that can lead to unfair results, such as prioritising or selecting one arbitrary group of data over others. Biased algorithms can occur due to a falsely design of the algorithm or the unintended use of decisions relating to the way data is collected, coded, selected to train the algorithm. (Cramer et al., 2018; Kordzadeh & Ghasemaghaei, 2022).

‘Implicit bias’ happens when someone makes an assumption based on memories or mental ideas (Dietz & Hamilton, 2008; Elsbach & Stigliani, 2019). Implicit bias affects how machine learning systems are designed and how the classification of the data is made. ‘Representativeness bias’ is applied when a person might assume that the most common aspects of a situation generalise to all situations (Cherry, 2021), such as assuming that men would be better at computer programming because all the company’s software engineers are men. ‘Confirmation bias’ is a form of implicit bias, which occurs when ML developers collect or select data that confirms their beliefs. For example, developers might favour a specific dataset that influences the outcome of their system. Confirmation bias can be made

accidentally (Dardenne,1995). An issue with confirmation bias is that people sometimes interpret it as a social skill because interviewers tend to ask questions that seem biased to confirm their presumptions about a person, idea or group (Dardenne,1995). Confirmation bias also was defined by Nickerson (1998) as when a person only looks for information that supports their initial decision, such as deciding that a candidate for a job is unsuitable because of what they are wearing, then seeking information in the candidate's replies to questions, to confirm this decision.

'Anchoring' occurs when a higher degree of importance than necessary is placed on a single piece of information when making decisions (Haggarty-Weir, 2013, Part 1). In anchoring, a person concentrates on one or two pieces of information to the exclusion of others, such as only asking the candidate about one aspect of their previous work and not others. According to Taylor and Doria (1981), another form of bias is known as in-group and out-group biases. 'In-group bias' occurs when special treatment is given to people in one's own group. For instance, when a developer uses his group of friends and family to create a model dataset. The real problem is the possibility of showing preferences that affect the results later on. 'Out-group bias' happens when developers tend to evaluate all the people outside their circle with one biased point of view compared to the people they interact with regularly (Haggarty-Weir, 2013). For example, out-group bias occurs when the characteristics and values from all people outside one's own group are seen as alike. An example of this bias is when developers gather information by asking people to provide characteristics about other groups. Those characteristics may be less accurate and more stereotyped than characteristics that participants list for people in their in-group.

Not all biases are social – some types are systematic errors introduced by the data in the system or through people who participate in the training process when creating the model. Forms of these biases include 'systematic bias', which is a broad, general term as most biases are considered systematic. Nevertheless, a definition provided by Tripepi et al. (2010) describes systematic bias as 'any error resulting from methods used by the investigator to recruit individuals for the study, from factors affecting the study participation (selection bias) or from systematic distortions when collecting information (information bias)'. 'Confounding bias' is seen when a relationship or a connection is assumed to exist between the exposure and the outcome when there is none. This bias occurs when a factor is autonomously associated with either the exposure or the outcome. According to Aronson et al. (2018):

> Randomization is the best way to reduce the risk of confounding. However, it may not be enough, particularly when it is anticipated that imbalances in prognostic factors may

occur despite randomization, or when imbalances occur by chance. Stratification and statistical adjustment can reduce the risk of confounding in such cases.

'Selection bias' is defined as any errors resulting from the sampled data due to a selection process. This type of bias usually occurs when the researcher attempts to recruit individuals for the study and generates systematic differences between the sampled data and reality (Tripepi et al., 2010). 'Reporting bias' occurs when individuals misreport information related to the study. Reporting bias can affect the data that ML systems learn from, which influences the outcome of the model. As explained below, various definitions of reporting biases have been proposed:

- The Dictionary of Epidemiology defines reporting bias as the 'selective revelation or suppression of information (e.g., about past medical history, smoking, sexual experiences) or of study results' (Canoy, 2015).
- The Cochrane Handbook defines reporting bias as 'when the dissemination of research findings is influenced by the nature and direction of results' (Bradley et al., 2020).
- The James Lind Library states that 'biased reporting of research occurs when the direction or statistical significance of results influence whether and how research is reported' (Bradley et al., 2020).

A summary of bias definitions and their consequences on decision-making are listed in the table below.

TABLE 2. 1: BIASES AND THEIR CONSEQUENCES

Type of Bias	Consequences
Cognitive bias	Serious errors of judgement in strategic decisions.
Social bias	Affects social treatment, causes discrimination and creates unfair systems.
Cultural bias	Narrow point of view of the design modelling.
Racial bias	Unfair social treatment, causes discrimination and creates prejudicial systems.
Gender bias	Unfair social treatment, causes discrimination and creates prejudicial systems.
Automation bias	lookCan lead to the wrong decision by falsely appear to be beneficial.

Algorithmic bias	Provides a wrong prediction of the model and incorrect results.
Confirmation bias	Incorrect results, unrealistic outcome of the system and unreliable study findings.
Implicit bias	Sometimes happens by accident, can have unwanted effects on the data or the design.
Representativeness bias	Unfair social treatment, incorrect results leading to a false outcome.
In-group bias	Invalid product testing or unrealistic dataset.
Out-group bias	Would create an unbalanced and stereotyped dataset.
Systematic bias	Errors in the model depending on the type of systematic error.
Confounding bias	Leads to a false accusation between the variables, which affects the validity of the results. Unreliable system outcomes.
Selection bias	Can have varying effects.
Coverage bias	Unrealistic result for the machine learning model because the dataset does not represent the target.
Sampling bias	Unrealistic result because the dataset does not represent the target.
Reporting bias	Nonrepresentation of the truth. Deceiving people of interest and does not illustrate the model work.

All these types of bias can be categorised into three types of mistakes regarding their consequences. Some of these biases will lead to wrong answers (errors in decision-making that occur because of mistakes either in the data or rules), unfair answers (errors regarding social discrimination and treatment) or oblivious answers (problems regarding trusting the answer without questioning or not realising there are mistakes in the design).

TABLE 2. 2: TYPES OF BIAS MISTAKES REGARDING BIAS CONSEQUENCES

Type of mistake	Type of bias
Wrong	Confirmation, In-group, Out-group, Confounding, Reporting
Unfair	Social, Cultural, Racial, Gender, Non-response
Oblivious	Algorithmic, Automation, Cognitive, Implicit

Given these examples, it is plausible to assume that decisions in personnel selection might exhibit some form of bias. Recognising these possibilities, personnel selection procedures are designed to minimise the possibility of overt biases. Nevertheless, covert or unconscious biases remain, which have been well documented in social decision-making (Bodenhausen, 1988). Examples include racial (Knotek, 2003), cultural (Helms, 2010; Pedersen, 1987) and gender biases (Mukherjee, 2015).

It is well accepted by most people that when meeting face-to-face with candidates, interviewers make decisions as to the candidate's suitability well before the end of the interview. Our social judgments tend to be constructed from the limited information we have of the situation combined with our expectations. In interviews, social events or simply meeting new people, we tend to form 'first impressions', that is, 'one's initial perception of another person, typically involving a positive or negative evaluation' (American Psychological Association, 2018).

In first impressions, people are judged by the slightest things. For instance, the implicit association test (Greenwald et al., 1998)[4] rests on the assumption that we might have specific expectations that reflect gender or racial stereotypes of particular activities or professions. While this test is not without its critics (Oswald et al., 2013), it underpins a popular belief that these attributes can be influenced by culture, gender or race (Ross, 2014). These subjective judgements have the potential to create bias, namely 'a particular preference or an inclination, especially one that inhibits impartial judgment' (Ross, 2014). Given the potential for humans to exhibit biases in their decision-making and social interactions, it is unsurprising that technology has been mooted as a means of debiasing personnel selection.

Biases in the recruitment process

According to Torrington and Hall (1998), personnel and line management use various imperfect methods to aid the task of predicting which applicants will be most suitable for the job. While one might assume that 'first impressions' can immediately influence the decisions made by interviewers, Frieder et al. (2016) found that only a small number of interviewers, 4.9%, reported making a decision within the first minute, which seems to contradict some of the more extreme claims on interviewer bias. However, in their study, around 60% of the interviewers had made a decision within 15 minutes of meeting a candidate. Using cognitive

[4] https://implicit.harvard.edu/implicit/takeatest.html

load theory, Frieder et al. (2016) argue that ‘interviewers manage cognitively intensive information processing demands by employing cognitive schema that render decision-making automatic’ (p. 241). While this automatic processing does not explicitly cover bias, it does imply that decisions are made in terms of factors that are beyond the semantic content of an interviewee’s responses or the complexity of the questions being asked.

This process reflects what Humphrey (1976) termed ‘social problem-solving’ (p. 316) and hints that this social orientation of thinking toward solving social problems also affects other fundamental computational problems that can cause social biases within the technological world. Frieder et al. (2016) further suggest that these interviewer decisions are influenced by the experience of the interviewer, performance of the interviewee and the sequence in which interviewees were interviewed. A slight increase in decision time occurs over the first two or three interviewees with a decline in time beyond that. Interviewer decisions are not influenced by the sophistication of the questions, which implies that the depth of discussion in the interview (which might align with System 2 thinking) seems to be less significant than other factors.

While the interview studies suggest that interviewers might be applying heuristics to help manage cognitive load in interviews (which does not imply that these heuristics reflect biased or prejudiced thinking), there is some evidence that bias might arise in other parts of the personnel selection process. Bertrand and Mullainathan (2004) sent 5,000 fictional résumés for various job ads that were offered by the *Boston Globe* and the *Chicago Tribune* newspapers. The curricula vitae (CVs) were sent in applications to various occupational categories. In each category, the quality of the CV was either high or low. Identity was assigned to each CV using a personal name to suggest the race of the applicant, for example, Emily, Anne and Brad suggest White names, while Kenya, Lakisha and Jamal suggest African American names. A statistically significant difference was reported in callbacks (Bertrand & Mullainathan, 2004): CVs for White-named applicants received 50% more callbacks than those with African American names. Furthermore, the callback rate for White applicants with a high-quality résumé was 27% more than for White applicants with low quality résumés. In contrast, callbacks for African American names with high-quality CVs was only 8% higher than African American names with low quality CVs. This indicates that high-quality CVs for African American names received less attention than high-quality CVs for White names (Bertrand & Mullainathan, 2004).

Additionally, biases against women, people of colour or other underrepresented groups have long plagued hiring. Many studies have covered and discussed biases in hiring (Bogen &

Rieke, 2018; Chichester et al., 2019; Dalton, 2018; Krishnakumar, 2019; Pena et al., 2020; Peng, 2019). The term 'bias' is often used to refer to interpersonal bias: prejudices held by individual people, whether implicitly or explicitly (Bogen & Rieke, 2018). To this day, many hiring managers evaluate candidates in ways that contribute to disparate hiring outcomes, leading to underrepresentation and pay disparities in roles across industries (Quillian et al., 2017).

Uses of automation in personnel selection

Automation has played a number of roles in hiring, for instance, automating the job application and CV processing or conducting psychometric testing and analyses (Ahmed et al., 2015). Decision support systems provide an automated shortlist of potential employees during recruitment in an organisation (Samad, 2012), and automated systems can be used to screen candidates (Singh et al., 2010). In this section, we present several studies exploring the use of AI in personnel selection to address our investigation of 'How do AI system recommendations affect people's decisions?' Some studies suggest using automation only as part of the process, such as during interviews (Langer, 2020). Other studies discuss the benefits of using these systems in the context of hiring people (Buckley et al., 2004; Dunnette & Borman, 1979; Jackson, 2016; Lodato, 2011; Schmid, 2018).

O'Neil (2016) remarked that the Kronos Company brought science to human resources departments to make the selection fairer. Over time, more software tools used for hiring decisions were developed to optimise, eliminate and judge employees. Many of these tools are used in the hiring system, included automated programs or personality tests. Although these systems were originally developed for fairness and efficiency, the hiring system in some cases used unfair criteria to filter applicants (an example is provided in the next chapter). This raises the question of whether the tools were optimised for filtering or for selecting. If it is the former, is there potential for the underlying algorithms to exhibit bias?

As O'Neil (2016) suggests, the desire to bring analytical science to human resources is to make selection fairer by removing the potential for human bias. Increasingly, however, such systems are being used to filter applicants. This raises the question of whether they are optimised for the objective of selection (i.e., to aid in picking the most appropriate person for the role) or rejecting applicants (i.e., to aid in filtering applications down to a manageable set). We propose that these objectives represent distinct forms of argumentation that place emphasis on different aspects of the resume and test scores.

Nevertheless, with more deliberation, transparency and oversight, some new hiring technologies might be poised to help improve this poor baseline. The use of hiring aids is common to assist the recruitment process. These tools are referred to as ‘hiring algorithms’ or ‘artificial intelligence’, and their recommendations aim to forecast outcomes by analysing existing employee data (Bogen & Rieke, 2018). Hiring technology builds predictive models to reflect employee characteristics throughout the hiring process (Walsh & Volini, 2017). These tools rely on ML, where computers detect patterns in existing data (e.g., the profiles of current employees. The systems build models that make recommendations about prospective employees in the form of scores and rankings (Bogen & Rieke, 2018) or suggestions for a specific action, such as an invitation for an interview (Ricci et al., 2011). Moreover, employers turn to hiring technology to increase efficiency. Employers often adopt these tools in an attempt to reduce the time and cost of hiring or to maximise the quality of the hiring process (Ahmed et al., 2015; Bogen & Rieke, 2018; Buckley et al., 2004; Pena et al., 2020).

Moving back from the interview, we should note that the recruitment process starts when a company decides to hire employees to meet its business objectives. The employers (or the human resources department) in the company creates a job profile that specifies the role, job category, essential skills, location of the opening and a brief job description detailing the nature of the work. Applicant profiles might be established regarding the prospective employee’s work experience, required qualifications or the desired skills applicants should possess. The job openings are advertised through multiple channels such as online job websites, advertisements, newspapers and other avenues. Candidates who are interested in applying for the job opening upload their profile through a designated portal. The portal typically provides an online form where the candidate enters details about their application such as personal information, education and experience details, skills and more data. The candidates can also upload their résumés and photographs. Following this, a review of applications is performed to shortlist candidates who will undergo further evaluation in interviews, written tests or group discussions. It is during the application review that most hiring aids are used.

Bias in computing

Recent studies have focused on bias in AI (Bahner et al., 2008; Barocas & Selbst, 2016; Caliskan, 2017; Franke, 2022; Goddard et al., 2014; Helms et al., 2011; Kliegr et al., 2021; Morar & Baber, 2017; O’Neil, 2016; Serna et al., 2019; Stanovich, 2003; Weber, 2019). Many of these studies discuss discrimination within the data or the data sources used to train the AI, embedded cognitive biases included in the algorithms and quantitative methods to remove statistical biases. As an effort to address the second thesis question, ‘How do programmers

creating AI/ML systems comprehend the values and beliefs encoded in their algorithms?', the following section explains how bias can be manifested in AI system design and outputs and how people interpret bias based on their perspectives.

Many variations of AI are found. The concept was defined by Russell and Norvig (1995) as intelligent systems with the ability to reason and learn. AI systems embody a heterogeneous collection of techniques, tools and algorithms (Jarrahi, 2018). AI is also defined as algorithm-based and data-driven computer systems that provide machines and people with digital capabilities such as perception, reasoning, learning or even autonomous decision-making (Tito, 2017). AI algorithms interpret a vast amount of information (data), draw conclusions, learn, adapt or adjust parameters accordingly, supporting human decision-making or automated actions (Misuraca et al., 2020; Sun et al., 2019). Human interactions with these systems have been well documented in various studies (Bock, 2020; Guszcza & Evans-Greenwood, 2017; Huang et al., 2018; Misuraca et al., 2020; Parasuraman & Riley, 1997; Parasuraman et al., 2000).

Several studies have examined the process of designing AI systems and investigated to what extent the process should be automated (Parasuraman, 2000). Some systems have focused on specific type of users – customers, for instance – asking how AI may influence marketing strategies and customer behaviours (Davenport et al., 2020). However, our focus is within the preferences, recommendations and outcomes of AI systems. Logg et al. (2019) discussed people's preferences in using AI tools over human judgment. In healthcare, Pezzo and Beckstead (2020) stated that patients prefer AI to a human provider, and Longoni et al. (2020) mentioned that people do not always reject healthcare provided by AI.

Tripepi et al. (2010) discussed selection bias and information bias in clinical research. To understand bias in the context of an AI system and automation, a definition should be provided. Several attempts have been offered to define bias in automation systems in general, for example, Parasuraman (2010) defined automation bias as 'the results in making both omission and commission errors when decision aids are imperfect'. Automation bias, Parasuraman believed, cannot be prevented by training or instructing the data and will affect the decision-making process. Peng et al. (2019) defined a biased decision 'to mean any difference in the system outcome such that one gender is favoured over another in a manner that does not correspond to the gender distribution of the candidate slate as fed into the system'. Meanwhile, Bahner et al. (2008) described automation bias as the use of automated decision aids that involve 'commission errors' that occur when people (operators) follow the recommendation of an automated aid system even though the recommendation is wrong.

The main purpose of using an automatic system is to ensure efficiency and fairness, and to cut the thousand applications to a reasonable number for HR employees to take to the interviewing process. Moreover, the machine ought to remain unaffected by bias or prejudices, and each applicant should be judged by exact criteria. However, most AI systems are trained using thousands of records of employment data from many years. CVs are traditionally composed of structured data, including name, position, age, gender, experience and education. Using these data, the AI systems compare the structured information to a 'model' applicant. This model could be defined in terms of the applicant profile, but it could equally be defined according to the company's definition of 'successful CVs', that is, the CVs of people whom the company has employed in the past. This reliance on past data could mean that bias has been institutionalised in favour of specific employee groups.

Gender bias is an issue in jobs that reflect a male dominance such as the tech business, for example. Amazon developed an experimental hiring tool using AI to review job applicants' résumés with the goal of automating the search for the best talent (Dastin, 2018). This system would give candidates a ranking from one to five stars. However, the team found that their software was biased against women because the models were trained on résumés submitted to the company for the previous 10 years, and most of these were submitted by men. As a result, the model learned that males were preferable and excluded résumés that carried the word 'woman'. The result not only was biased against women but also could result in rejecting plausible candidates.

CVs might also include unstructured data such as a face photo or a short biography. A face image is rich in unstructured information such as identity, gender, ethnicity or age (Gonzalez et al., 2018; Pena et al., 2020; Ranjan et al., 2018). That information can be recognised in the image, but it requires a cognitive or automatic process trained previously for that task. The text is also rich in unstructured data. The language and the way we use that language determine attributes related to one's nationality, age or gender. Both images and text represent two domains that have gained interest in many AI studies in recent years.

McDuff et al. (2018) suggest that many of the biases in AI exist because of data homogeneity in the field. This means that the variability one might normally assume within a population is not reflected in the datasets used to train AI. Many learned models exhibit bias as training datasets are limited in size and diversity (Tommasi et al., 2017; Torralba & Efros, 2011) or reflect inherent human biases (Caliskan et al., 2017). For instance, datasets used to train facial recognition systems might be trained from data collected from young to middle-aged

white men who work in universities or tech companies. From this, one might expect the algorithm to perform better with images of white men than women or ethnic minorities. This outcome is partly because the 'landmarks' used to define facial features might differ between genders and racial groups. Moreover, the effect of ambient light on different skin tones will have an impact on recognition performance (Nagpal et al., 2019).

As an aside, a recent survey in *Nature* highlights concerns over the potential application of facial recognition software and problems relating to biased sampling in the dataset and using approaches that violate other aspects of research ethics, such as the provision of informed consent prior to collecting images, particularly in public settings (Van Noorden, 2020). The societal problem of ML systems for facial recognition is that these automated decision-making tools impact people's lives (e.g., in justice, health and education). This impact might mean that recognition errors could disproportionately affect some groups more than others, for example, leading to a higher likelihood that a black person could be misrecognised. In some countries and some US states, these technologies have either been banned or severely licensed.

Understanding bias and fairness and their impact on AI is necessary to advance the field as a whole and make better products and algorithms. This understanding raises several questions, not least of which is how creators of these systems can ensure that their work is ethical. Most importantly, are programmers aware of the consequences of their systems' decisions and recommendations? To explore these questions, we applied various frameworks, one of which is the data/frame model (DFM), to understand which decisions are made by developers of AI/ML. Another framework, the Toulmin model of argumentation, was employed to understand how the system works and clarify the beliefs, values and expectations of stakeholders.

In summary, AI and ML bias is driven from aspects of human culture known to lead to harmful behaviour. Recent reviews point out some of the risk factors that contribute to bias, including the following.

- Human biases. Even if they are implicit, human biases reflect the assumptions of the developer. The problem may still exist when algorithms contain the same biases that humans have without our explicit knowledge (Caliskan et al., 2017).
- Algorithmic biases. Recognition algorithms, particularly for recognising human faces, can either misrepresent particular facial features or overlearn unrelated aspects of a dataset, leading to greater misidentification of faces that are not white (Torralba & Efros, 2011).

- Biases in datasets that train the AI: Torralba and Efros (2011) identify a range of biases that relate to data:
    - Selection bias in the content of available datasets.
    - Capture bias in the elements used from datasets, for instance, subsets of elements rather than full sets.
    - Labelling bias in poorly defined data structures.

Trust and system reliability

If automation can exhibit bias, how can users trust it? Addressing this last thesis research question, we investigate trust and AI acceptance in this section. Human–human trust first begins with a basis in performance, then progresses and finally evolves into faith (Rempel et al., 1985). In an opposite development pattern, trust in automation begins with faith, which plays a significant role in the initial stages. In the research community, several studies address AI trust and acceptance (Del Giudice et al. 2021; Gursoy et al. 2019; Kim et al. 2021; Talley, 2020; Young & Monroe, 2019).

Studies have highlighted similarities and differences between human–human and human–automation trust (Madhavan & Wiegmann, 2007). Wickens et al. (2000) presented a view of the development of trust in a system given its specific level of reliability (depending on the prediction). Appropriate levels of trust in automation (i.e., a decision support system [DSS]) require proper standardisation between the individual's perceived reliability of the aid and its actual reliability. Starke and Baber (2020) presented the effect of known decision support reliability on outcome quality and visual information foraging in joint decision-making. Mathieson (1991) predicted users' intentions and compared his model of technology acceptance with the theory. The notion of system trust refers to an 'operator's belief that a system is operating in a predictable manner and keeping with expectations' (Jian et al., 2000).

Many works on trust development have discussed the human tendency to underestimate the actual reliability of recommendation aids. Nevertheless, the development of trust is likely to be strongly influenced by several DSS characteristics. Only recently have researchers attempted to directly test this assumption by comparing the effects that varying levels of recommendation

aid reliability have on users' trust levels.[5] Indeed, Madhavan and Wiegmann (2007) stated that most studies have examined how trust develops when interacting with the automation of a single reliability level or how a solitary automation failure can affect users' trust in a system that has been entirely reliable before the failure.

'Transparency' is a defining factor of a 'good' recommender system (Morar et al., 2019; Tintarev & Masthoff, 2012), that is, the extent to which the computational process behind the recommendation is visible and apparent to the human. It has been shown that increasing the transparency of recommender systems and making explanations available to the user improves decision performance (Morar et al., 2019; Skitka et al., 2000). Some studies show that in high-reliability conditions, participants will follow the advice of automation even when it is wrong. This finding could be interpreted as automation bias, where humans are complacent and do not check the automation output against other information sources (Parasuraman et al., 1997, 2010), that is, the human merely accepts the automation's output and does not contribute to the decision-making.

However, with these novel technologies, individuals can experience discomfort and uncertainty. Developing trust in AI plays a pivotal role in its adoption and acceptance, especially when human intuition and emotional intelligence are challenging and change rapidly (Misuraca et al., 2020). In addition, a vital research question is to understand if the level of risk of the recommendation affects how users accept AI-generated content and information. Researchers have identified potential issues and areas to develop to gain a deeper understanding in the past decade. For example, Bock et al. (2020) explored the internal and external values of incorporating AI technologies in service environments. AI and ML technologies are critical agents in value creation, engagement and experience (Kaartemo & Helkkula, 2018). Individuals' motivations, expectations, and concerns shape technological developments, and marketers must comprehend better how they can impact and alter consumer behaviours and experiences (Kim & Duhachek, 2020; Lu et al., 2020). Previous research has investigated several task-type factors that cause people to accept or avoid AI-generated information. One crucial aspect is the task accomplished by the AI systems.

[5] See Morar et al. (2019) for a review as they proposed that automation confidence interacts with user decision activity.

When a task is subjective, people tend to avoid AI (Castelo, 2019). This assumption relies on the belief that AI lacks sympathy. In addition, consumers avoid AI when the task domain involves instinctive (vs analytical) aspects of decision-making (Jarrahi, 2018) or 'common-sense' decisions (Brynjolfsson & McAfee, 2012; Guszcza et al., 2017). According to Castelo (2019), this tendency is related to the general belief that although AI shows superb performance and skills in analytical and mechanical tasks, AI would be less functional in performing intuitive and empathetic tasks (Huang & Rust, 2018). The critical advantage of AI is its ability to process large amounts of data, identify data-driven patterns, and therefore, produce sensible and scientific recommendations (Kim et al., 2021). Similarly, people tend not to associate AI applications with independent goals (Kim & Duhacheck, 2020) and will be less responsive to accept AI for more consequential and risky tasks, which we tested in chapter 7.

Longoni et al. (2020) claimed that people usually prefer a human to an artificially intelligent medical provider. However, Pezzo and Beckstead (2020) showed that this was only the case when the historical performance of the human and AI providers was equal. When the AI is known to outperform humans, their data showed a clear preference for the automated provider. Other studies suggest that when the AI systems provide some explanation to the users, it helps gain their trust and acceptance. However, users interpret the quality of the explanation based on their level of understandability (Samek et al., 2016).

In the process of interaction, issues of fairness play roles as heuristic cues, triggering user trust. In addition, with the rise of automated algorithms in every sector of our lives, fairness is becoming a key concept in AI (Shin & Park, 2019). 'There are increasing concerns about the use of data, which may be shared illegally or abused by others for the sake of content automation. Automated data decisions may be incorrect, unfair, non-transparent or unaccountable' (Shin, 2021). Recommending content or items requires a more complex user engagement to understand their needs. However, the system sometimes asks and uses information more than it needs for that specific recommendation to customise outcomes, resulting in a final biased product. Fairness and bias issues bring up critical considerations in designing and developing algorithm services (Shin et al., 2020).

Algorithms are usually designed to produce the most accurate predictions. However, unanswered questions remain regarding how the processes are conducted, whether the results actually reflect user preferences and whether the outcomes are reasonably accountable. Previous studies have exhibited that fairness, transparency and accountability define user trust and subsequent attitudes and behaviours (Shin, 2021; Shin et al., 2020).

Users' understanding and awareness of how and why a particular AI recommendation is generated and how their input impacts the outcome have been found to be significant in their acceptance of the technology.

Kizilcec's (2016) work illustrates that using explanations can enhance positive attitudes and satisfaction with the recommendation system, encouraging users to accept its outcome. It can be inferred that explainable artificial intelligence (XAI) would help users understand the process and thus increase their faith in the system. Recent studies have focused on explainable AI (Anjomshoae et al., 2019; Baber et al., 2020, 2021; Borgo et al., 2018; Shin, 2021), in which people interpret the output of the AI. Moreover, people are taught to use explainable systems because they want to understand how data are collected and processed and thus how recommendations are produced (Rai, 2020).

Users increase their input data to improve recommendation outcomes through a transparent mechanism. Algorithm users can also interpret the logic of a recommendation system (Shin, 2021). Developers of algorithms are urged to ensure the accuracy and lawfulness of results to increase user trust. How trust is shaped and grows during an interaction may provide valuable clues when designing and developing AI. This is significant because now, more than ever, people realise that algorithms are not neutral and may harbour human prejudices. People would like to understand how algorithms work, how their data are analysed and to what extent their results are fair (PR Newswire, 2018).

Studies suggest that intentions and wishes usually influence individuals' actual behaviours (Cronan et al., 2018). Research on customers' willingness to adopt AI recommendations reported many influencing factors. For example, Chi et al. (2022), Gursoy et al. (2019) and Niemelä et al. (2017) found that 'hedonic motivation' is the main factor influencing customers' intention to use and accept AI. Song (2017) found that the most significant determinants of service robot adoption are perceived usefulness, social capability and device appearance. Meanwhile, Stock and Merkle (2017) found that customers' intention to use AI services is influenced by their expectations of AI systems and the service's quality in a general service setting. However, studies that examined customers' willingness to adopt AI systems in service transactions were either explanatory or operated on the assumptions of technology acceptance models, which is argued to be inadequate for examining the AI systems' adoption (Lu et al., 2019; Song, 2017).

The customer's final decision to accept the use of AI devices during service encounters is also likely to be determined by their emotions generated through a complex multistage appraisal

process (Kuo & Wu, 2012). In addition, most studies have addressed AI acceptance or attitudes toward technology in a single system type in which people either accept or reject the use of AI in that context. Research showed that acceptance of the technology is most likely dependent upon human feelings and trust. However, interacting with AI systems is not subject to the user's preferences or biases alone but also can be dependent on the AI/ML systems and the type of service they provide. In this study, we explored several factors that can be implemented by AI that can help increase acceptance and user trust. In Chapter 7, we investigate users' acceptance and reaction to several types of recommendations and test whether users' attitudes would change with the context of the recommendation.

Conclusions

In this chapter, we discussed the literature on human biases and rationalities, human interaction with decision-making systems, the recruitment process and the long-embedded cognitive biases within the field. Next, we discussed the current widespread practice of using AI automated aid systems and DSS during the hiring process to save costs and time. However, we reviewed several studies that have proven disadvantages when using these systems, especially for users who are not well represented socially. The user's acceptance and their trust toward AI systems are dependent upon whether the results are socially acceptable and expected. We have introduced many studies demonstrating discrimination in AI. These studies highlighted specific type of bias (gender, culture, racial) as well as focused on the final outcome and how to prevent systematic bias error.

Questions remain, for example, as AI systems help people by recommending a decision, do they affect users' judgments and conclusions? What if the system is wrong or biased – will it guide the decision-maker to agree and make the same mistake? Did the discriminatory systems revealed in the literature intentionally produce unfair outputs, or was it coincidence? Do programmers understand bias?

As for understanding bias, In the table below we summarised the studies that involve bias definition in human interaction with computers, hiring process and decision making that we reviewed in this chapter.

TABLE 2. 3: SUMMARISING BIAS DEFINITIONS

The definitions of bias in previous studies in this chapter	**Source Reference**
Cognitive and Social Bias	
The mental behaviours that prejudice decision quality in a significant number of decisions for a significant number of people	Arnott, 2006;Hastie & Dawes, 2009. Used in DSS
Bias refers to the prejudice or preference towards some things, people, or groups over others	Many AI/ML research papers, AI, 2019; Arnott, 2006; McDuff et al., 2018
Happens when someone makes an assumption based on memories or mental ideas	Implicit bias, (Dietz & Hamilton, 2008; Elsbach & Stigliani, 2019
Is a form of implicit bias, which occurs when ML developers collect or select data that confirms their beliefs. For example, developers might favour a specific dataset that influences the outcome of their system. Confirmation bias can be made accidentally	Confirmation bias, (Dardenne,1995) Nickerson (1998)
Occurs when a higher degree of importance than necessary is placed on a single piece of information when making decisions	Occurs when a higher degree of importance than necessary is placed on a single piece of information when making decisions
Automation Bias	
Automation bias as the use of automated decision aids that involve 'commission errors' that occur when people (operators) follow the recommendation of an automated aid system even though the recommendation is wrong.	Automation bias, Bahner et al. (2008)
Occurs when a model algorithm creates systematic and repeatable errors that can lead	Algorithmic bias, Cramer et al.,2018; Kordzadeh & Ghasemaghaei, 2022

to unfair results, such as prioritising or selecting one arbitrary group of data over others	
Discuss discrimination within the data or the data sources used to train the AI, embedded cognitive biases included in the algorithms and quantitative methods to remove statistical biases	Bias in AI, Bahner et al., 2008; Barocas & Selbst, 2016; Caliskan, 2017; Franke, 2022; Goddard et al., 2014; Helms et al., 2011; Kliegr et al., 2021; Morar & Baber, 2017; O'Neil, 2016; Serna et al., 2019; Stanovich, 2003; Weber, 2019
Defined automation bias as 'the results in making both omission and commission errors when decision aids are imperfect'	Automation bias, Parasuraman (2010)
Bias in Hiring	
Interpersonal bias: prejudices held by individual people, whether implicitly or explicitly	Biases in hiring, Bogen & Rieke, 2018; Chichester et al., 2019; Dalton, 2018; Krishnakumar, 2019; Pena et al., 2020; Peng, 2019).
5,000 fictional résumés example.	Bias might arise in other parts of the personnel selection process, Bertrand and Mullainathan (2004)
Forecast outcomes by analysing existing employee data	Hiring algorithms' or 'artificial intelligence', Bogen & Rieke, 2018
Hiring technology builds predictive models to reflect employee characteristics throughout the hiring process	Hiring technology, Walsh & Volini, 2017
Bias in decision making	
Any error resulting from methods used by the investigator to recruit individuals for the study, from factors affecting the study participation (selection bias) or from systematic distortions when collecting information (information bias)'	Systematic bias, Tripepi et al. (2010)

Occurs when a decision-maker prefers the choices that are made by automated systems even when these systems make errors compared to other information provided by non-automated agents	Adensamer et al.,2021; Parasuraman & Manzey, 2010
The scenario where a human interacts with an automated system that is providing false or incorrect advice but the decision-maker (human) does not question the suggested decision or seek additional information	Adensamer et al.,2021; Parasuraman & Manzey, 2010
Defined a biased decision 'to mean any difference in the system outcome such that one gender is favoured over another in a manner that does not correspond to the gender distribution of the candidate slate as fed into the system'	biased decision, Peng et al. (2019)

In this table we divided the definition of bias in four sections (Social Bias, Automation bias, Bias in hiring and bias in decision making**).** As we can see in the table the meaning of bias is different between these studies which suggest that the term is yet to be agreed on. Especially in computing, our concern is when AI systems reflect personal values and preferences (biases) of their developers. Questions were addressed regarding social acceptability of the results and whether programmers understand biases as a social problem within the system because the social orientation of thinking toward solving social problems also affects other fundamental computational problems that can cause social biases within the technological world. This research explains how bias can be manifested in AI system design and outputs and how people interpret bias based on their perspectives. also explores user reactions and acceptance of different AI system outputs. We addressed these biases in the next chapters, for instance, we tested the social bias (racial and gender biases) in Chapter 6. Bias in hiring can be seen in both Chapter 4 and 6. We presented bias in decision making in Chapter 6 and 7. Bias in Automation is presented in Chapter 5 and 6.

Chapter 3

Introducing the Data Frame Model and Toulmin Model of Argumentation Frame Works

Part of this chapter was based on the paper 'Using Toulmin Model of Argumentation to Explore the Differences in Human and Automated Hiring Decisions' at VISIGRAPP 2020 (15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications), Valletta, Malta, 27–29 February 2020.

Other parts were presented at the 'Virtual Ergonomics & Human Factors Conference', 27–28 April 2020 in the paper 'How Sensemaking by People and Artificial Intelligence Might Involve Different Frames'.

Introduction

In the previous chapter, we provided various definitions of bias and several studies showing how different types of bias (gender, racial, social, systematic) have an effect in many domains. In most of these studies, bias causes problems for users, companies and society – especially when it exists and can clearly be seen in an AI application. When interacting with automation, people expect sufficient results and fair outcomes.

One of the main concerns in this research is to understand why a machine produces such unacceptable results. Whether the problem was an algorithmic mistake, biased dataset or unethical developer, it will be always be unacceptable to reject applicants for these reasons. Thus, people's beliefs and cultural values affect how we interact with an AI system and accept its results. Users of the system will have expectations based on their own knowledge. This knowledge is formed by the beliefs and the values of the users. For instance, the HR department will expect the résumé reader to choose applicants with higher education and sufficient years of experience for administrative positions. Former managers with higher degrees would also be expected to be called for an interview. Therefore, when a hiring system eliminates suitable candidates because of their name, gender or neighbour, the results will be prejudicial because the outputs conflict with people's ethical values and do not match the expectation they had of the system. This thesis seeks to understand how such a system works and what the beliefs, values and expectations of people are who are engaged in the process of sensemaking.

Despite rapid improvements in the underlying capabilities of intelligent systems and ML, increasing their acceptance by users has remained a formidable challenge. User acceptance depends on vital elements such as people's beliefs, emotions, trust and cultural values as well as whether the output of these systems are agreeable to their users. For this thesis, sensemaking is used to consider explainability and transparency. That is, for a system to be explainable or transparent, it must somehow make sense to its user. 'Sensemaking allows humans to comprehend complex information, assimilate it, create order from it, and develop a mental model of the situation as a precursor to responding to the situation' (Anderson, 1983). According to Klein et al. (2006a), 'sensemaking' is the understanding of complex events. Sensemaking also has been defined as 'how people make sense out of their experience in the world' (Duffy, 1995). By sensemaking, modern researchers seem to mean something like creativity, comprehension, curiosity, mental modelling, explanation, or situational awareness, although all these factors or phenomena can be involved in or related to sensemaking. Therefore, the process of sensemaking seems to be essential to deal with understanding complex concepts such as biases within AI systems.

By presenting two frameworks, we seek to understand the relationship between the components of AI systems and how this interaction may result in bias. We start by presenting the frameworks: the data-frame model (DFM) and the Toulmin model of argumentation. We then provided a fictional example as a case study that we apply both frames to and represented results. we also compared similarities and differences between the frame works. These frames works can be later seen in this thesis as we implement the Toulmin Model of Argumentation on the study we presented in Chapter 4, and DFM in Chapter 5.

Sense-making and the data-frame model

Before explaining the framework, we need to define what we mean by data in our DFM, even though the 'distinction between data and information has been discussed within the database and information systems communities for many years' (Aamodt, 1995). Most studies look at the two from a productive perspective, where data is the raw material and information the product. This perspective follows the waterfall of data-information-knowledge-understanding that insist on the fact that data must be treated as raw and undefined material. Ackoff (1989) also defined data as 'symbols that represent the properties of objects and events. Information consists of processed data, the processing directed at increasing its usefulness'. He also compares them saying 'Like data, information also represents the properties of objects and events, but it does so more compactly and usefully than data. The difference between data and information is functional, not structural'. However, this waterfall or hierarchy of data-to-

knowledge conflict with the data-frame theory of sensemaking that uses data or simple cues directly from the environment. 'Naive information-processing accounts assume that primitive data or isolated cues are successively massaged by inferential operations until they emerge from the other end as knowledge or wisdom' (Klein et al., 2006 a). Klein and his colleagues argue that the data-to-knowledge process is misleading for several reasons. First, when we apply sensemaking we do not have a specific point to start so this conflict with the data-to-knowledge process that emphasises the beginning with raw material (data) and ending with (knowledge). Second, in the DFM of sensemaking, we can collect any unstructured data or cues and apply it to the frame as long as it fits. Data are the elements present in a situation that contribute to finding and building an applicable frame (Zimmerman, 2008). These data help to create the frame are known as cues and anchors. The identification of data elements is crucial, especially not to treat each data element as an anchor. According to Klein et al. (2007), experts have the advantage of a much stronger understanding of the situations than novices. Additionally, experts tend to have more ways of achieving objectives, which extends the variety of frames they can pick. However, most experts are not reliable at describing events or proving the wanted details based on their memory, a special method is needed to extract the tacit knowledge and help order the chain of events.

So, people who engage with sensemaking are using the frame as visual representations of the data that cannot be understood as the frame provides an explanation. According to Thomas et al. (1993) each element in the sense-making process forms some relationship with others, using anchors, cues and any relevant data found in that environment. Thus, 'when people try to make sense of events, they begin with some perspective, viewpoint, or framework' Klein et al. (2006b). The primary function of frames in DFM, according to Minsky (1975), 'is recognition, to guide attention to fill in missing parts of the frame, to test a frame by searching for diagnostic information'. As we use the DFM, another important point to emphasise is that frames change as we add more data in the model. Therefore, frames shape and define the data and data affect how frames are, because of this relationship, frames are tools that you not only think with but also items you think about.

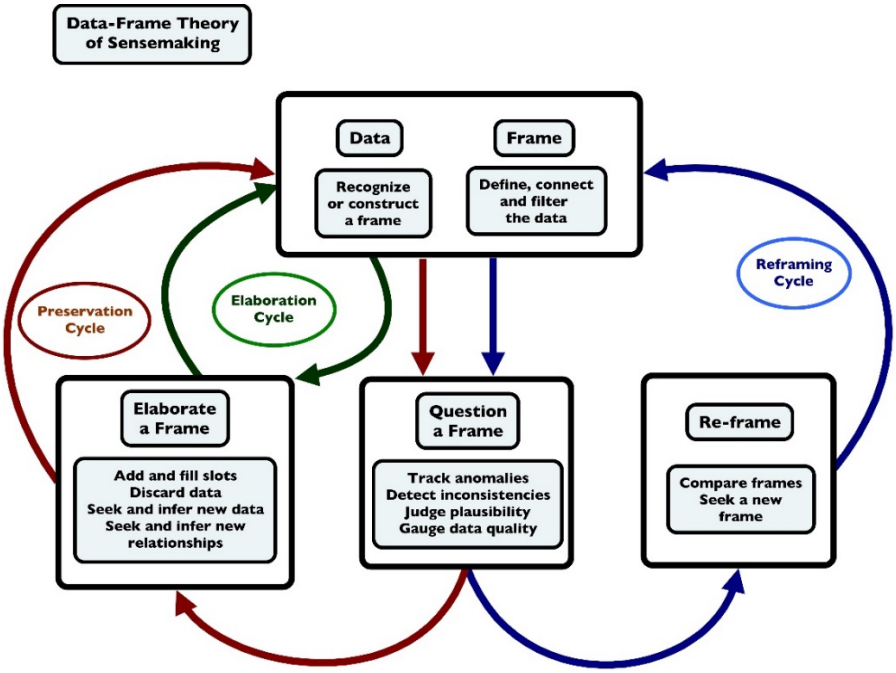


FIGURE 3. 1: DATA-FRAME THEORY OF SENSEMAKING (KLEIN ET AL., 2007)

To start building a frame, the situation must first be understood. Frames are used in this context to help shape and organise incoming data. As shown in Figure 3.1, the first action is represented in the higher rectangle. Certain cues or anchors within the scenario are recognised by the individual to elicit a frame. This frame may represent a typical single diagram or multiple fragments to construct the frames to adjust to the new scenario. Alternative frames might also be compared to determine which seems most accurate. This way, a strategy might be constructed on which data would be most useful in the future. Second, an individual will usually choose a frame based on three or four cues. More data might be searched to find cues that can be used to construct a suitable frame. Certain factors help in selecting the frame, such as a person's beliefs, society or the environment where the frame must be constructed and the availability of the data. Once the frame is selected, information can continue to be added if it fits.

In addition to the DFM, this research employs the Toulmin model of argumentation (Toulmin,1958). In this context, an 'argument' has a 'scheme', which relates to a specific form of reasoning, often inspired by human behaviour. The scheme has a structure for how information leads to a conclusion. A 'format' encodes the information and a 'representation' visualises the scheme, for which diagrams are commonly used (Reed et al., 2007).

Toulmin model of argumentation

Argumentation technology can help AI agents enhance human reasoning (Lawrence et al., 2017; Modgil & Prakken, 2013; Oguego et al., 2018). According to Freeman (1991), information in an argument diagram can be either a proposition or an inference. Together, these create the structure for the argument (Bench-Capon & Dunne, 2007). A widely known example of argument, 'Toulmin diagrams have proven a flexible approach in AI treatments of argumentation' (Bench-Capon & Dunne, 2007, p. 625). In this chapter, Toulmin diagrams represent arguments to consider how bias is revealed for jobs that have many applicants. An initial sift of CVs could be performed to reduce the total number of applications to consider. However, there is a risk that such a shift could be subject to bias. Inferring from the Bertrand and Mullainathan (2004) examples, this could mean that bias has been institutionalised in favour of specific employee groups.

Gender bias is an issue in jobs that reflect a male dominance, such as the tech business in the Amazon example. The main concern in this chapter is to understand where the problem of bias lies and how people (users) view and absorb bias because people's beliefs and cultural values affect how we interact with AI systems and accept its results. Users of the system will have expectations based on their own knowledge. This knowledge is formed by the beliefs and values of the users. We seek to understand how the system works and what the beliefs, values and expectations of stakeholders are by applying the Toulmin model of argumentation.

This chapter uses Toulmin's model of argumentation (Toulmin, 1958). Toulmin provides an argumentation technique to show that there are other arguments than formal ones, which offers more logic and further explanation. His argumentation model distinguished six different kinds of elements: data, claim, qualifier, warrant, backing and rebuttal (Verheij, 2009). The example below illustrates Toulmin's classic example of Harry, who may or may not be a British subject.

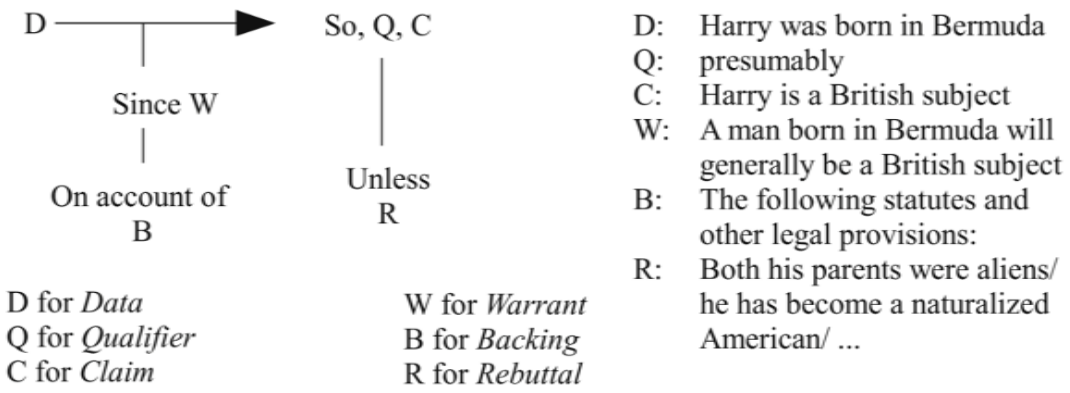


FIGURE 3. 2: TOULMIN'S LAYOUT OF ARGUMENTS WITH AN EXAMPLE (TOULMIN 1958, PP. 104–105).

First, we have the Datum: 'The Datum provides the ground to support the claim' (Toulmin 2003, p. 90). In Figure 3.2, 'D' represents the datum 'Harry was born in Bermuda', which is held to be the foundation of the claim 'Harry is a British subject'. 'The Claim is the assertion that we are committed to and must prove when challenged' (Toulmin 2003, p. 90). 'It is where the argument starts' (Verheij, 2009). What does this claim stand on? We must use a warrant. In our example, this is 'A man born in Bermuda will generally be a British subject'.

Note that a warrant is not universal. The claim is that a man born in Bermuda will *more likely* be a British subject. 'The Warrant provides the connection between datum and claim' (Verheij, 2009). 'Warrants are 'general, hypothetical statements, which can act as bridges, and authorise the sort of step to which our particular argument commits' (Toulmin, 2003, p. 91). Moreover, Toulmin (2003) declares that warrants are 'rules, principles, inference-licences or what you will, instead of additional items of information' (p. 91). A warrant can be considered as representing the reasons behind the inference and can be subject to defeat. As a result, a claim should be qualified (Verheij, 2009).

The 'qualifier' represents 'the degree of strength which our data confer on our claim in virtue of our warrant' (Toulmin, 2003, p. 91). In our example, the qualifier is to presume that Harry is a British subject. As the elements in this argument became more explicit, the warrant needs further support so it can provide a connection between the data and the claim. In Toulmin's example, this is called the 'backing' and refers to statutes and other legal provisions without the need to define them. 'The backing provides reasons why warrants themselves would' (Toulmin, 2003, p. 96). 'Backing occurs when not a particular claim is challenged, but the range of arguments legitimized by a warrant' (Verheij, 2009). This idea of backing helps illustrate the hidden values and beliefs that might inform models.

The final element in Toulmin's example is the rebuttal, which indicates any arguments against the claim or any exceptions to it. According to Toulmin (2003), the rebuttal indicates 'circumstances in which the general authority of the warrant would have to be set aside' or 'exceptional circumstances which might be capable of defeating or rebutting the warranted conclusion'.

Toulmin diagrams represent arguments, while DFM represents a visual framing to consider how bias for jobs that have many applicants could be performed to reduce the total number of applications to consider. Taking job applications as a case study, we consider how AI and humans might frame the decision to select an applicant in various ways. The idea is to develop an approach that could ultimately serve as a 'pre-mortem' on decision models prior to their application. However, there is a risk that such a selection could be subject to bias. We begin by describing a motivating example.

MOTIVATING EXAMPLE

This motivating example was developed from an article in the Wall Street Journal, *'Are Workplace Personality Tests Fair?' by Weber and Dwoskin (2014), with slight changes and different names.*

Kareem Bader, a student who was seeking a minimum-wage job, applied for a part-time position at Kroger after his friend recommended him. Kareem had a history of having bipolar disorder, but at the time of sending the application, he was a productive, high-achieving student and healthy enough to practise any type of work. However, Kareem was not called for an interview. When he asked, he was told that he failed the personality test taken during the application. According to Lundgren et al. (2017), personality tests investigate individual motivations, preferences and the differences between people. Some researchers argue that if these tests are used before hiring, it can lead to better hiring decisions. However, this argument depends on the stability of the test-taker preferences over time and in different contexts (Barrick et al., 2001).

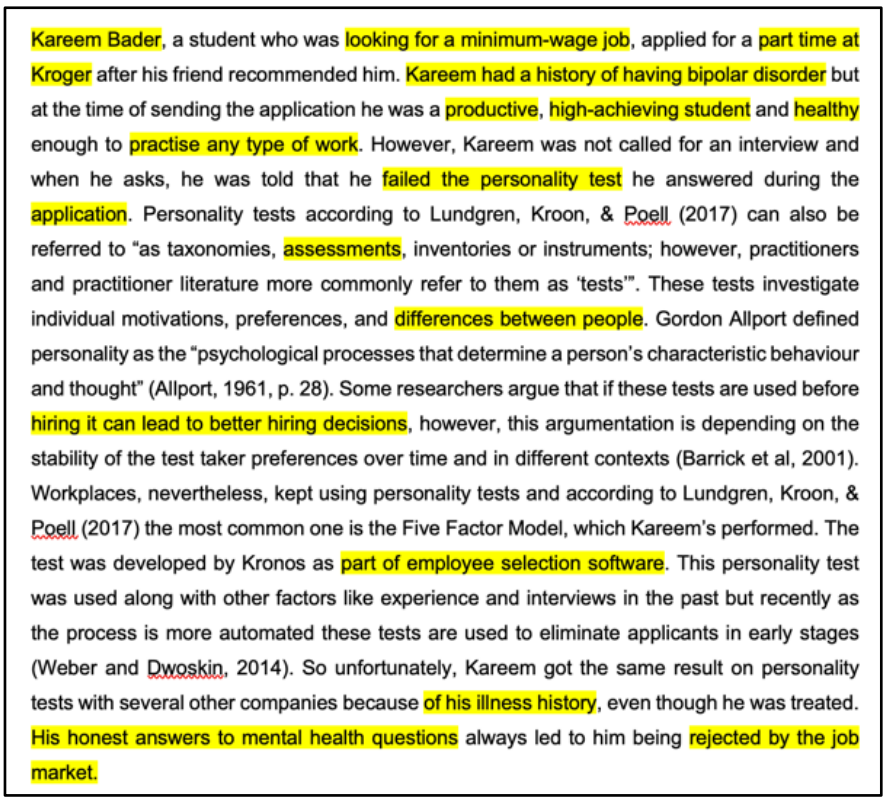
Nevertheless, workplaces kept using personality tests. According to Lundgren et al. (2017), the most common test is the five factor model, which Kareem performed. The test was developed by Kronos as part of employee selection software. This personality test was used along with other factors such as experience and interviews in the past. Recently, however, as the process has become more automated, these tests are used to eliminate applicants in early stages (Weber & Dwoskin, 2014). Unfortunately, Kareem received the same result on personality tests with several other companies because of his illness history, even though he

had been treated. His honest answers to mental health questions always led to him being rejected by the job market.

Applying the data-frame model to Kareem's application.

To provide a reasonable explanation to Kareem's case, DFM divides the sense-making process into seven elements: mapping data and create an initial frame, elaborating a frame, questioning a frame, preserving a frame, comparing frames, reframing and finding or seeking a frame. Therefore, the flexibility this model provides helps with complex data and an ambiguous environment. Using DFM, we provide an explanation for the hiring system and Kareem's application process by explaining the relationships between the data in that environment and what Kareem believes and expects from the system.

To analyse the motivating example to fit into an understandable frame, we created a concept map based on keywords highlighted from the text (see Figure 3.3) and then a concept map (Figure 3.4)



Kareem Bader, a student who was looking for a minimum-wage job, applied for a part time at Kroger after his friend recommended him. Kareem had a history of having bipolar disorder but at the time of sending the application he was a productive, high-achieving student and healthy enough to practise any type of work. However, Kareem was not called for an interview and when he asks, he was told that he failed the personality test he answered during the application. Personality tests according to Lundgren, Kroon, & Poell (2017) can also be referred to "as taxonomies, assessments, inventories or instruments; however, practitioners and practitioner literature more commonly refer to them as 'tests'". These tests investigate individual motivations, preferences, and differences between people. Gordon Allport defined personality as the "psychological processes that determine a person's characteristic behaviour and thought" (Allport, 1961, p. 28). Some researchers argue that if these tests are used before hiring it can lead to better hiring decisions, however, this argumentation is depending on the stability of the test taker preferences over time and in different contexts (Barrick et al, 2001). Workplaces, nevertheless, kept using personality tests and according to Lundgren, Kroon, & Poell (2017) the most common one is the Five Factor Model, which Kareem's performed. The test was developed by Kronos as part of employee selection software. This personality test was used along with other factors like experience and interviews in the past but recently as the process is more automated these tests are used to eliminate applicants in early stages (Weber and Dwoskin, 2014). So unfortunately, Kareem got the same result on personality tests with several other companies because of his illness history, even though he was treated. His honest answers to mental health questions always led to him being rejected by the job market.

FIGURE 3. 3: HIGHLIGHTED TEXT FROM THE MOTIVATED EXAMPLE

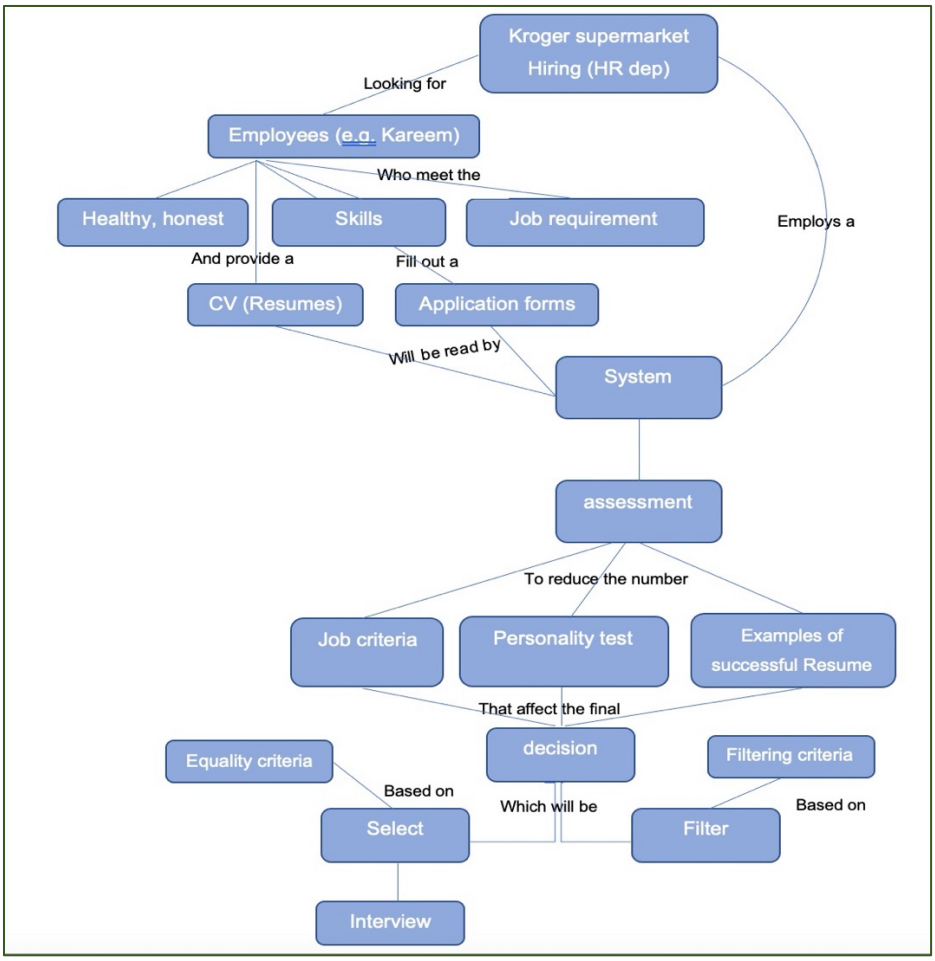


FIGURE 3. 4: CONCEPT MAP FOR THE MOTIVATED EXAMPLE

The initial stage of the DFM is to create a frame based on the data available in the situation. In our case. the concept map describes the hiring process in Kareem's case, which was extracted from the highlighted text. The concept map represents the big picture of the situation, which can represent a giant frame of the hiring process (Figure 3.5). To understand the process, however, the frame must be divided into multiple subframes. Kareem or applicants that interact with the system are the ones who will provide the data (inputs) which the system will use to filter or select the user (applicant), the departments that used the system (HR) and the system (Hiring algorithms) as we can see in hiring system frame, these main elements were created as individual frames because they represent an individual party in the process. There are some data that contribute to build only one frame and there are others in which build multiple ones at the same time. Each data element is treated differently in each frame, for instance the element 'résumés' in the HR frame process is presented differently in the system frame.

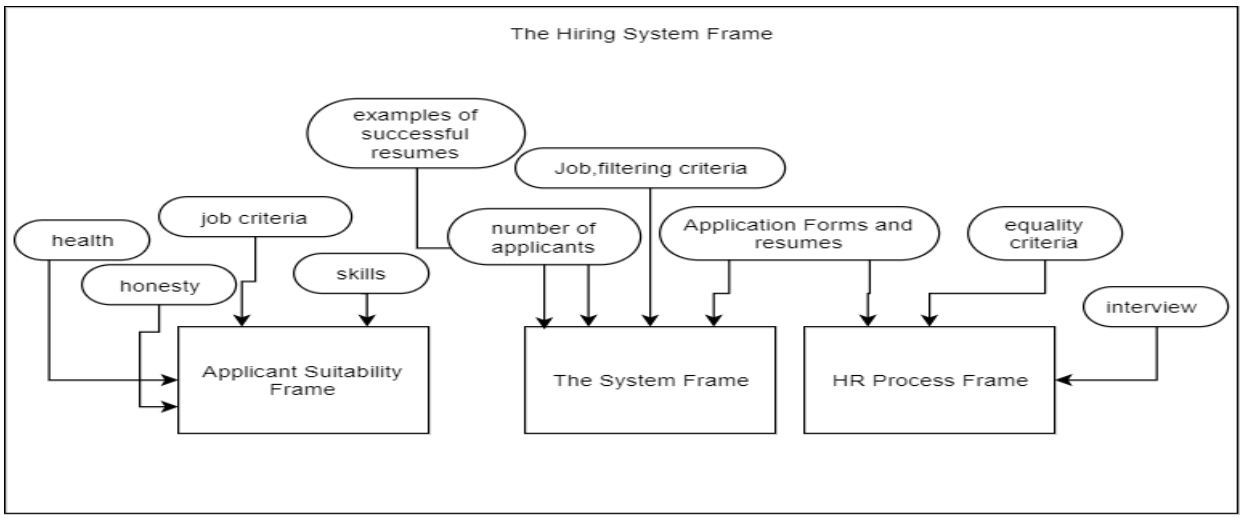


FIGURE 3. 5: HIRING SYSTEM FRAME

An important individual from both the concept map and the hiring frame (Figure 3.5) is the applicant (Kareem in our example). The applicant is the jobseeker who applies for a specific job and interacts with the process by filling out forms, answering test questions and submitting résumés. Therefore, we present the applicant suitability frame as the first subframe extracted from Figure 3.5. The frame explains the applicant's(Kareem) point of view of their suitability as a candidate for this job. Therefore, it is essential to understand what are the data that people used as inputs which can be used later against them. The applicant suitability frame will address the user's (applicant's) suitability for the job based on various elements such as job criteria, the skills they need to perform the job efficiently, their honest answers in the application and their health condition (see Figure 3.6).

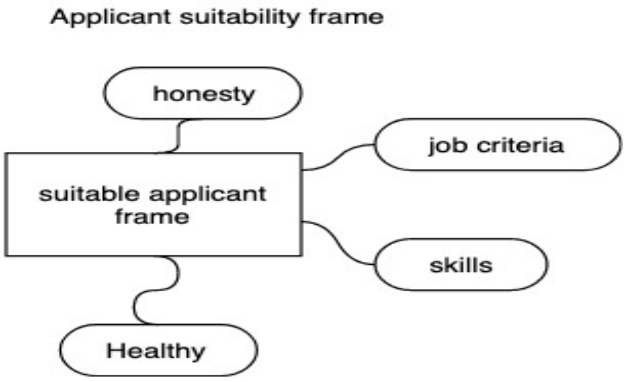


FIGURE 3. 6: APPLICANT SUITABILITY FRAME

Applicants without school-leaver certificates will not be accepted in managerial positions because of their lack of skills to perform the job and not meeting the needed job criteria. Applicants will have expectations of their suitability for the job based on their beliefs. If all the

data were met, the user would expect to be accepted for the job or at least called for an interview and not eliminated completely in early stages. Hence, when Kareem was rejected, he did not accept the output of the system because in his suitability frame, he believed he had all the required elements for that specific job. The system suitability frame was inferred from the concept map where the system attempted to assess the applicants based on various criteria to reduce the number of people who applied for the job. Applicants could revise, change and compare their data elements and the relationships between them to influence the suitability frame.

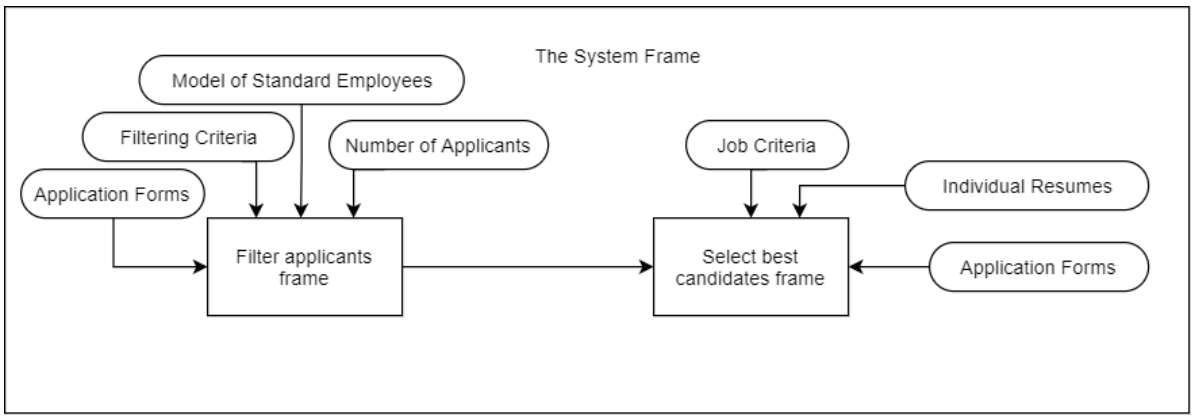


FIGURE 3. 7: SYSTEM SUITABILITY SYSTEM FRAME

To understand why the system rejected Kareem (and why other systems might eliminate applicants with specific race, gender, or type of disability), we must examine the system suitability frame and the data that affect its decision. Unlike the applicant's single frame, the system which is usually used by the company for its efficiency to reduce time and money, most likely will use the assessment results to filter/select applicants. In the motivating example Kareem had to do several assessment tests like the personality test which rejected him. We also presented the assessment in the concept map which helps to reduce the number of applicants and eventually aid the selection of the candidates. In Figure 3.7.

The system works as a bridge between the applicants and the HR people. It will help and aid the HR to select the right employee by recommending the most suitable applicants based on the qualifications the job require. The first subframe is filtering the applicants, when the number of applicants exceeded the required, it is important for the system to filter and reduce the number as possible based on a specific criterion (that should be agreed on before running the system). The filtering frame will build decisions based on the data, which are the model of standard employees, the number of applicants, the filtering criteria and of course the application forms. However, these inputs open a wide option for filtering, especially when the

number of applicants is massive. Additionally, any new applicants who do not resemble the standard employee of the company will be at risk of rejection because their information does not match the frame. This pattern explains Kareem's rejection because his mental illness made him an unwanted example compared to thousands of other employees who had clean mental illness records.

Moreover, in this frame, when the number is considered reasonable, the system will select the best candidates for review by the HR department. The selecting frame will compare the application forms and the resumes to the job criteria that were transferred from the filtering frame and then rank the candidates based on their qualifications. Now by applying these two subframes, the system believes that it fulfils the whole assessment part by providing the best recommendations for the HR department. However, the main problem here is the system frame of suitability does not match the applicant's suitability.

The system believes that Kareem is unsuitable because he failed to meet a specific requirement (the scores from the assessment tests), which conflicts with the applicant's frame of suitability (Kareem's point view of suitability). Moreover, when the system also rejected individuals because of their skin colour, foreign name, gender or other criteria unrelated to the job criteria, the system believes it did a good job by selecting only the good applicants. However, these outcomes are socially unacceptable. The problem in this specific frame can be the biased examples of successful data that trained the system or false algorithms that handled individual forms unethically but mainly the criterion of judging, is also a problem. The final results did not match any ethical expectation.

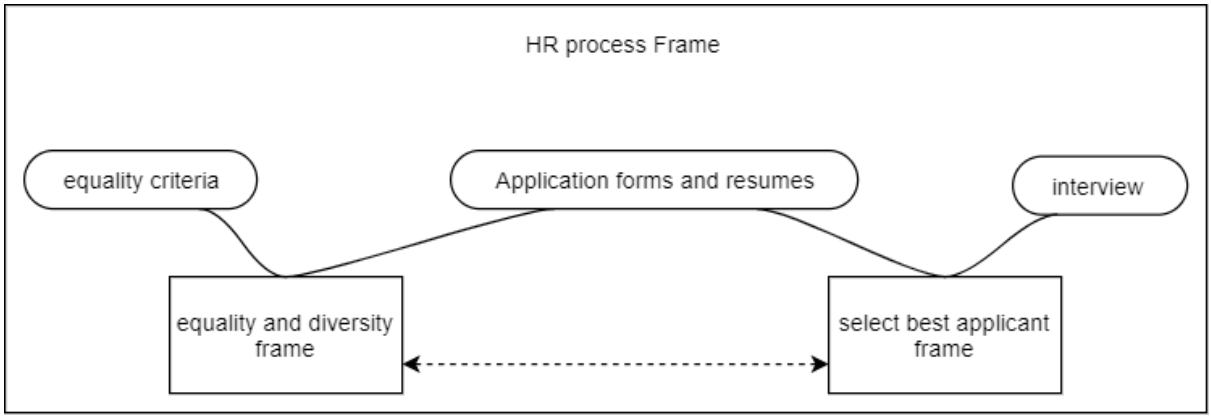


FIGURE 3. 8: HR SUITABILITY FRAME

Similar to the system suitability frame, the HR frame has multiple subframes (see Figure 3.8*)*. Using this frame, the HR can select the best employee for the offered job. By applying the 'equality and diversity' subframe, HR ensures that they treat all applicants the same this data

element was brought to this frame based on the assumption that companies and HRs are using AI aid system and testing applicants for that purpose. The data elements that create the 'equality' subframe are 'equality criteria' and 'application forms and résumés' that were recommended from the system's suitability frame (see Figure 3.7). After ensuring the equality, the HR frame will employ interviews and evaluates individuals to select the best applicant. However, this frame is dependent on the results of the previous frame, so we are unsure whether this frame will lead the HR department to the right selection or even the ethical one. If the system suitability frame provides a biased outcome, it is more likely that the HR suitability frame will perform the same biases. The reason is because the information or the input the HR will use to perform the decisions is depending on the outputs from the system.

Conclusion of applying DFM on Kareem's case

The obvious problem in these frames is that they can only make the change locally. Applying various DFM activities such as preserving, questioning, changing, comparing and revising will only be within the frame itself. Although they all are small frames creating one large frame (the system hiring frame; see Figure 3.5) they have no authorisation to interfere with each other. 'Application forms' for example, is a data element that was used differently in two different frames. None of the frames will influence the data outside its local environment. In the system's frame, the data element was scanned for specific information and skills to serve the purpose of 'filtering' any unqualified applicants.

On the contrary, the HR frame thoroughly used this data element and reviewed each one to 'select' the right applicant. Therefore, the result of Kareem's rejection or acceptance, for instance, has different point of view in each frame. In the applicant frame, he was eligible. In the system frame he was not eligible because he failed the personality test in his application which is an important data element that preforming the frame. In HR frame he might be eligible, but his application was filtered and rejected by the system, so he did not have a chance to be considered or interviewed by the HR department. To make the relationship between the second and third frames clear, they do overlap and share some elements but do not communicate directly. The HR department will only have the result of the system's frame as an input, without the ability to influence the creation of that result. Each frame is limited to its own data elements and their interaction only.

Applying the Toulmin model of argumentation to Kareem's application.

Applying this model to the processing of hiring (Kareem's example), which includes selecting and filtering applicants depending on their résumés and test results, we can construct the argumentation scheme shown below. First, to apply Toulmin model of argumentation on the motivating example, the terms must be interpreted. We converted the texts from the example into Toulmin element's terms, as shown in Table 3.1.

TABLE 3.1: TRANSLATING THE EXAMPLE TEXT TO TOULMIN ELEMENTS

Toulmin element	**Text from motivating example**
Datum	Kareem gave honest answers to mental health questions. Productive, high-achieving student. Healthy enough to practise any type of work.
Claim	Kareem Bader is suitable for this job.
Warrant	Productive, high-achieving, healthy, was recommended.
Backing	Minimum-wage job, which is part-time at a supermarket. Usually not hard to meet criteria.
Rebuttal	Unless he is not suitable.

Using Table 3.1, our first claim in the hiring process is in the applicant's argumentation. When a job seeker applies for a specific job, they will have some sort of expectation regarding their suitability. When an applicant claims that he or she is suitable for the job, we ought to seek why. Their answers will be based on their educational background, skills, heath condition and various other information that supports their claim in which they strongly believe. In the Toulmin model of argumentation, that supportive information from the text represents the data. However, to claim suitability for a job, providing simple facts is not enough. A warrant in the Toulmin model presents a general rule that offers sense and an explanation why a man born in Bermuda, for example, would *more likely* be a British subject (see Figure 3.2). In Figure 3.9, the warrant states that if applicant met the requirement, they would probably be suitable. That rule is supported by the backing, which is the job criteria that was generated from the employers who offered the job. A rebuttal occurs when the rule is not met or does not apply

to the data. Thus, when Kareem claimed he was suitable for the job, he did not understand why his suitability was challenged.

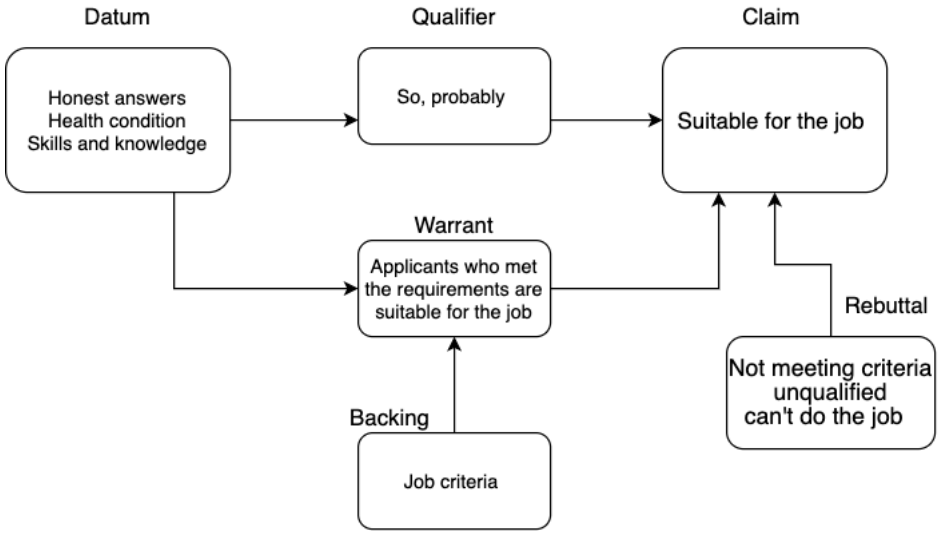


FIGURE 3. 9: APPLICANT'S ARGUMENTATION

Automated systems help HR department hire, fire, and promote employees efficiently. The hiring system might be used when there are too many people applying for the job, and the HR department can only interview a certain number of applicants. Hence, when the number of applicants is too high (claim) the system must reduce the number by filtering application which is the assessment the system performed on applicant to examine who is best fit. Similar to the DFM, the assessment used many tests and examples from the companies hiring history. The warrant is the limited time for interviewing. This rule is supported by backing, which states that this filtering process will save the HR time and effort. However, in the filtering process the system might use a model based on current profiles of employees, job criteria, to judge the applicants' resumes and forms. An exception to this claim is when the system filters good and qualified people. Going back to the motivating example, Kareem was rejected because of his scores on the personality test where he was honest about his illness. This problem is presented in the filtering argumentation when the system compared thousands of records of healthy employees to his and found no match. Unfortunately, anything the applicants provided in their forms could be used to exclude them if the AI was not fair or the people who built it were not careful about creating a debiased system because it is more likely than we think that the AI will produce unexpected results.

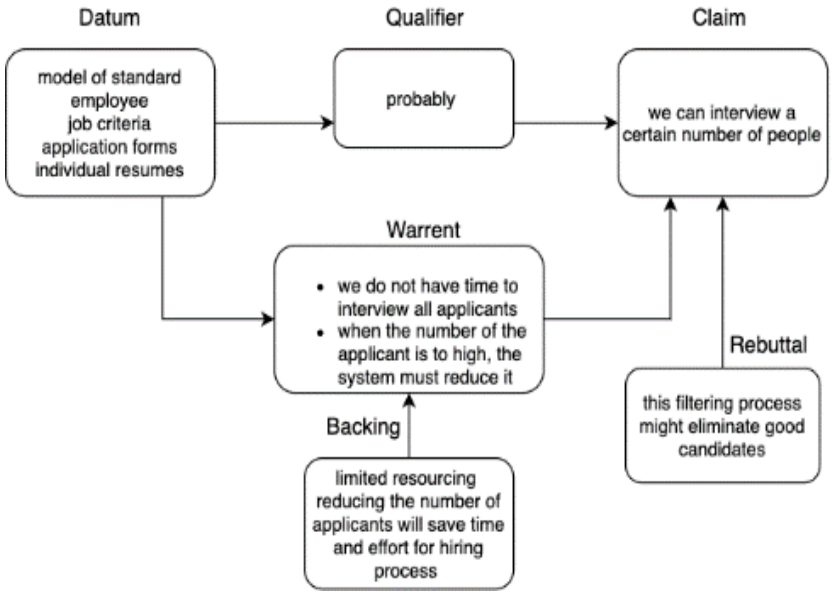


FIGURE 3. 10: FILTERING ARGUMENTATION.

After reducing the number of applicants to a reasonable amount, the system could select a set of best candidates among the applicants by matching the job criteria with the individual resumes and forms for the applicants then ranking them based on their match to the 'model of employees', see Figure 3.11.

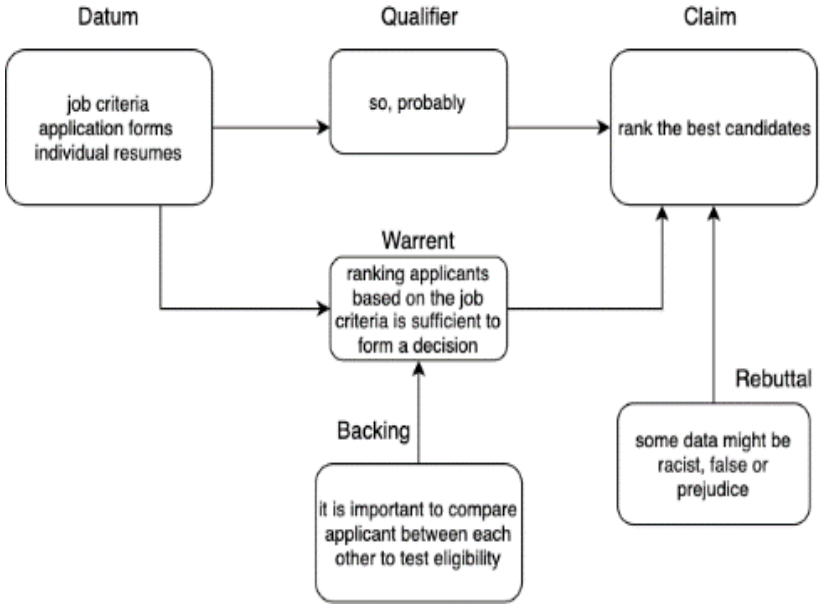


FIGURE 3. 11: SELECTION ARGUMENTATION

The last stage in the hiring process is in the HR department's hands. As Figure 3.11 shows, the system, which the HR uses to ensure the filtering efficiency, will provide a number of 'best candidates' for the HR to select from. The HR will define the best applicant after reviewing each CV or resume the system recommends. This information is supported by the belief that the system is reliable enough to trust its results. Note that the HR department assumes that this process ensures equality and diversity in its hiring process, depending on data from the hiring system and specific equality criteria applied to all candidates. However, if the model is inherently biased, for example, because it is derived from a homogeneous employee population that does not reflect diversity, then the process will have prejudice embedded in it.

Conclusion of applying the Toulmin Model of argumentation on Kareem's case

Conflicts arise between different claims in distinct situations from various stakeholders. Each claim represents the values and beliefs of its claimer. Moreover, each claim was formed based on the available data in that situation and has no authorisation to interfere with other claims, similar to DFM. Although the filtering and selection processes are dependent on each other, each had their own procedure and criteria. For instance, the above hiring process argumentation about Kareem being rejected has multiple claims. In the applicant's suitability diagram (see Figure 3.9), Kareem claimed that he was suitable for the job, which was supported by the data and the warrant that he was good enough for the job.

In both filtering and selecting the argumentation's diagrams (see Figures 3.10 and 3.11), the system claimed that Kareem was unsuitable because he failed one of the essential tests on his application form. Therefore, his application was not transferred to the selection process nor to the HR department who claimed nothing about Kareem's stability, except that his name was not listed as a possible candidate. Thus, based on the data, warrant and backing, Kareem was not claimed as a suitable employee. Whether Kareem was filtered from the start or not selected at the end, the system clearly did not see him as fit for the position, which proves that there were values and belief conflicts between the applicant and the system's suitability. This example illustrates that each argumentation has its own elements that do not interact with any other elements in other argumentations. Therefore, the claim that was formed by the system could not use the data from the applicant's argumentation or the HR arguments. Moreover, the same data was treated differently in the filtering and selection stages. Therefore, the Toulmin model of argumentation is a way of illustrating how different criteria supporting an argument can led to a biased decision.

Similarities and differences between the two frameworks

TABLE 3.2: COMPARING DATA-FRAME AND TOULMIN MODEL ELEMENTS

Data-Frame	Toulmin
Frame	Claim
Data	Datum
Seek	
Preserve, Elaborate	Warrant
Reframe	Rebuttal
Compare	
Question	Qualifier
Reframe	

Toulmin's argument model analysis and the data-frame model can create compatible frameworks for the process of sensemaking as they provide an explanation and logic for the data. The data-frame model provides assumption, explanation and representation for the input data. In the Toulmin model, these data represent the claim, which holds the strongest point of argumentation. Furthermore, both models start with data or inputs that needed to be supported or may be argued. For instance, to build a frame, it is necessary to have initial data (cues) and the basis of any claim in Toulmin is the data (datum). Moreover, elaborating and preserving the frame in DFM is not quite the same but similar to the warrant and backing the warrant in the Toulmin model. Both elements try to support the original statement by keeping, adding or filling slots.

Questioning the frame in the DFM is necessary to challenge the accuracy and the validation of the data, Toulmin also discussed challenging the warrant to see if it fits in the argument context. Hence, other similarities can easily be found between the reframing in the data frame model and the rebuttal in Toulmin model as these element's purposes are to refuse or change the original frame (statement). Although both models are proving an evidence and explanation for complex and hard to understand events, they use different words and were developed for different purposes. Therefore, each one of them was used in a different study. One of the main differences between them is, in the Toulmin model we have the claim, but we need the data that can add support and solid base to the claim. Thus, in the next chapter (Chapter 4) a study was conducted to explore bias in user interface design. We interview participants who had

experience using designing tools in the past and wanted to explore their understanding of bias by comparing their claims about their designs. We chose to use the Toulmin model as its elements will help us in comparing and understanding each participant's point view of biased design. In the Toulmin diagrams, we present data, claims, warrants, backing and rebuttals in one direction as we extracted these elements from the participants' interviews and designs. The claims were clear for each participant, but we tried to understand what data led them to this claim.

In Chapter 5, the DFM framework is used to represent 'the bias machine' project. Unlike the Toulmin model of argumentation, DFM is more flexible to deal with unclear situations. The project aims to create a problem and investigate the causes. Therefore, DFM was found to be much more convenient for performing different activity from the frame itself (e.g., seeking, preserving and reframing).

Chapter 4

Exploring Unconscious Bias in User Interface Designers: A Case Study of A Job Application System

Introduction

In large companies, artificial intelligence (AI) is being used to optimise the recruitment process to ensure efficiency and fairness as well as to cut the thousands of applications down to a reasonable number for HR employees to take to the interviewing process. An assumption is that the AI system ought to remain unaffected by bias or prejudices, and that each applicant should be judged by the exact criteria in the job description. We wondered whether the problem of bias extends from the training data (which, after all, replicates existing inequalities in organisations) to the design of the AI systems themselves. That is, do people consider the potential impact of bias in the design of a 'hiring system'?

In traditional approaches to software engineering, the programmer specifies the rules that define the behaviour of a system. In contrast, unsupervised learning algorithms develop their behaviour by discovering patterns in the data provided to them. Although traditional software can sometimes have bugs or errors, a developer can go back and try to fix these mistakes or at least understand where and what went wrong. However, this scenario cannot be applied to unsupervised learning algorithms.

> In classical computer programming, you have precision with the algorithm. You know exactly in mathematical terms what you are doing. With deep learning, the behaviour is data driven. You are not prescribing behaviour to the system. You are saying, "Here's the data, figure out what the behaviour is." That is an inherently fuzzy and statistical approach. (Fernandez, as reported by Dickson, 2018)

When you let unsupervised learning algorithms develop their own model, you are losing visibility of their reasoning processes (Dickson, 2018).

In this chapter, we present an experiment to study the unconscious bias in user interface design for systems to support personnel selection with AI support. Investigating the first research question of the thesis, 'How do people use their own knowledge and experience to interpret the output of Artificial Intelligence/Machine Learning (AI/ML) algorithms?' The aim of

conducting this experiment is to explore the programmer parameters (choices) when developing an AI system because these parameters are used by the algorithms in the filtering and selection processes, which could lead to socially unpleasant outcomes. Implementing an AI system with what was believed to be the default of a typical user interface (UI) is not necessarily correct. Many job applications request considerable personal information that could be misused. This experiment also explores the programmer's understanding and awareness of encoding such biases within their systems.

Participants (developers) were shown a fictional job description for the position of 'IT team manager', with specific qualifications and experience requirements. To create the system, participants were asked to first design a conceptual sketches of graphical user interfaces (GUI) for the applicants to register their information. The purpose of this step was to understand the designer's choice of system inputs. Then, participants were asked how the data would be translated into the AI system. For this, they drew a data flow diagram sketches to explain how the system reads, stores and uses the data in order to select a suitable employee.

This chapter uses a combination of a UI design exercise and a critical decision method interview to reveal potential for unintentionally introducing bias in the design of job application systems. This experiment highlights the need for designers to appreciate the ways in which information provided by job applicants could be used by AI/ML in ways that emphasise 'unconscious bias' in the personnel selection process.

Method

Our assumptions in this experiment are: first, programmers and designers focus on the technical side of the system. This could mean that they overlook ethical implications of their work (or, if they recognise ethical issues, they might assume that these will be addressed during the deployment of the technology). Second, most of the biases influenced by the designer are unconscious. This could mean that these could be addressed if they are pointed out to the designer, that is, in a sort of pre-mortem review of the design choices that are being made. Finally, after highlighting bias, designers should modify their designs.

Participants

This study involved 10 participants (7 female; 3 male). Four were postgraduate students from a school of computer science (CS), and six were information systems graduates who currently work in the field. All participants had experience using designing tools in either school projects

or work. We focused on the majority of the participants as our experience required understanding how the UI was created. The study employed an in-depth qualitative investigation involving analysis of the designs and interviews with participants. Our view was that the sample would be appropriate to capture broad differences in outlook and approach. The design of the study was approved by the University Ethics Review Board. Prospective participants were approached on the basis of their experience in designing computer systems and software engineering, and the 10 who agreed to participate did so on a voluntary basis.

Procedure

We asked participants to produce a conceptual design in the form of sketches of GUIs and Data Flow Diagrams for an AI system that could aid in personnel selection. Each participant was provided with the following scenario:

The hiring process takes place when a company settles on the number of employees they are looking to hire and the skill sets they require of these employees. Employers mostly attempt to reach candidates through job postings, job referrals, advertisements, college campus recruitment, etc. Candidates who respond to these measures come in for assessments. Hence, using an AI tool can help save time and money. Information from applicants is uploaded and organised in a database, making it easily accessible and searchable for HR professionals because the information is collected and organised digitally and automatically. In this experiment, the subjects were asked to design an AI to select the best candidates to interview. Ideally, the AI would filter out unqualified applicants and produce a shortlist of candidates to take to interview. The client company offers the position of IT team manager for their next project. The candidate must have a hybrid skill set (soft and hard skills that are a combination of technical and non-technical skills), which include:

- A bachelor's degree in one of the following majors: computer science, computer engineering, electronic engineering or information systems;
- Computer skills (e.g., document management, Microsoft Office, running office machines and coding);
- At least two years of experience in IT work,
- At least 5 skills from the list below:
    - Supervising
    - Teamwork
    - Office coordination
    - Problem-solving

- Leadership
- Information management
- Scheduling
- Goal-oriented
- Digital communications
- Manage remote working teams
- Multitasking
- Meeting deadlines
- Critical thinking

Once participants had understood the scenario (with any clarifications addressed by the experimenter), they were asked to produce conceptual designs through the following.

- Design (conceptual sketch) the GUI for of a hiring system where applicants can upload a CV online and provide information about themselves through an application form found in their system profile. A GUI is an interactive visual component for computer software, displaying objects that convey information and representing actions that can be taken by the user. The GUI is the gate where the user enters the system. Thus, by asking participants to design the UI, we were asking what inputs as designers they believed should be entered and processed.

- Design (conceptual sketch) a data flow diagram (DFD) to show the system architecture and database. The reason for the DFD is because it was a familiar graph for most CS students. DFD also graphically represents the functions or processes that capture, manipulate, store and distribute data between a system and its environment, as well as providing a logical information flow for the system.

Participants were free to add all the features they saw fit and suitable in their sketches and designs. We wanted to explore what their own choices were and if that would lead to a biased outcome. In both sketches, we examined the choices the participants made, for instance, whether the GUI inputs had any personal information that could lead to biased results. Moreover, we examined how participants created the sequence in the DFD and interpreted how they understood the data and selected applicants. Letting the participants sketch first without interruption provided good reflections on what they believed was important in the recruitment process. Mainly, we were looking for any biased choices in the sketches. For example, if the participant chose to put in the application any information that had to do with

gender, age, race, nationality or religion, as these inputs not only were not required to select an employee but also consider illegal to enquire about in some jurisdictions.

In this experiment, we used the critical decision method (CDM) technique to ask participants to explain their designs after they completed each sketch. 'The CDM is a retrospective interview strategy that applies a set of cognitive probes to actual no routine incidents that required expert judgment or decision-making' (Klein et al. 1989, p. 464). This methodology is helpful for analysing the activities participants used to formulate their decisions and demonstrate how more and less experienced participants interpreted a situation. Consequently, CDM allows researchers to obtain information from a variety of decision-makers about the same situation and gather as much data and cues as possible. Hence, participants were invited to express their thinking while explaining the sketches. Therefore, using CDM probe-questions represented in Table 4.1, we asked participants about their sketches. The idea was not to ask the question explicitly, but let the interviewee talk and answer the main question through their true beliefs.

TABLE 4.1: CDM QUESTIONS

Probe type	Probe content (the interview questions)
Cues	Talk about your design? Why do you think this will help the recruitment process? What was the most difficult thing about it?
Information	Why do you think this would be a good design? What are the specific features?
Analogs	Can you think of other designs or features that will help the recruitment process?
Standard Operating Procedures	What are the rules of understanding or using this GUI? What cues are leading to your answer?
Goals	What do you think is the main purpose of using such a design?
Options	Are there different conclusion you can make?
Experience	Do you have previous experience with this process or design?
Assessment	How easy is it to create? Can it be improved? Or adjusted
Mental Model	How can the data be collected from the system? Can you explain?

The interviews were recorded with the participant's consent and transcribed within a few days. After transcription, the recording was deleted.

Following the initial design, participants were debriefed, and issues and consequences of bias were explained to them. Subjects had the opportunity to revise or review their sketches. They were provided with or asked to use different pens to highlight any changes they were going to

make. The aim for the second interview was to expose the designers to their previous designs and examine whether their choices were intentional or unconscious.

Finally, Toulmin diagrams (Toulmin, 1958) were created based on the participants' answers to the CDM questions. Toulmin provides an argumentation technique to show that there are other arguments than formal ones, which offers more logic and further explanation. This argumentation model distinguishes six different elements: Data, Claim, Qualifier, Warrant, Backing, and Rebuttal (Verheij, 2009). We code each statement from a participant into Toulmin elements. This was done by categorising the content of participants' sketches and interviews into data, claims, warrants, backing, rebuttals and qualifiers. The purpose of the qualifier is to define the special cases where the data led to a claim. If there is no qualifier, we assumed that in all cases the data led to the claim. From the sketches, we defined data elements that were essentially the inputs in the GUI. From the interviews, the data elements could be linked to the claims and connected to a warrant. Each warrant has backing that supports it. Finally, the rebuttal comprises any statements that conflict with one or more claims.

Analysis and Results

In this section, we present an example of the concept sketches of graphical user interfaces produced by participants (Figures 4.1 and 4.2) and the concept sketches of data flow diagrams for some participant (Figures 4.3 and 4.4).

The participants' concept sketches of the GUI

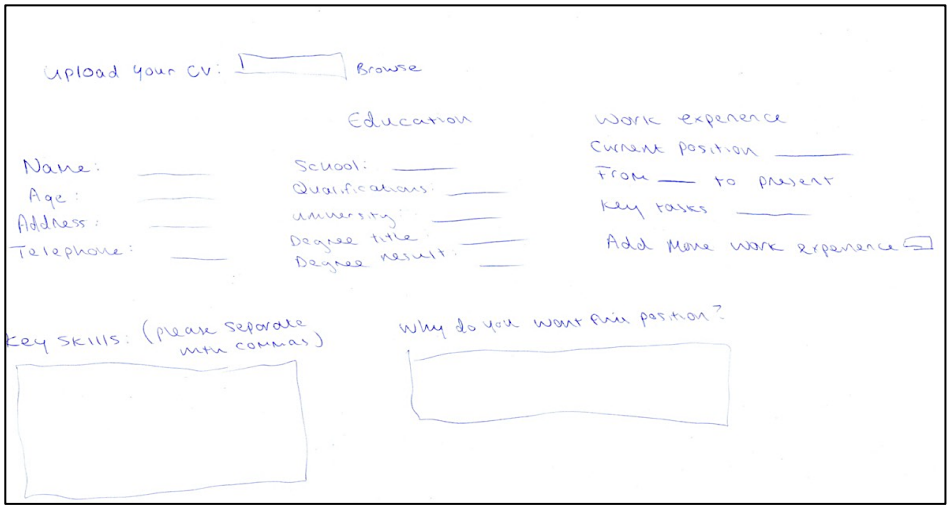


FIGURE 4.1: EXAMPLE OF THE CONCEPT SKETCH OF GUI (P2)

Figure 4.1 present an example of the participants' concept sketches of the GUI design. The figure shows how P2 created the sketch and the choices of selecting what the application should contain. For example, P2 chose to ask applicants to upload a CV or to fill in the form by asking for personal information on the left, educational information in the middle and lastly, on the right the work experience. P2 also added a text box for the skills and another one for the question "Why do you want this position?".

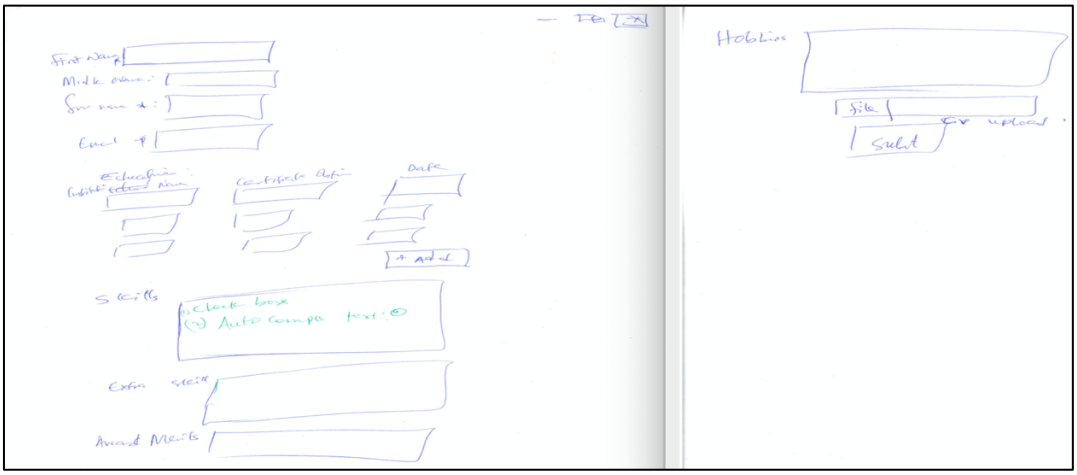


FIGURE 4.2: EXAMPLE OF THE CONCEPT SKETCH OF GUI (P9)

Another example of a concept sketches of the GUI is shown in Figure 4.2, where P9 has selected a different set of inputs to present in the design of the application. P9 application contained less personal detailed compared to P2's application and have other sections added like 'hobbies' for instance. The rest of the concept sketches of the GUI can be seen in the Appendix.

The participants' GUI concept designs contain information that job applicants are expected to provide in order for their application to get processed. Most participants created simple (and quite similar) application forms, with many sections. We suggest that these forms represent a sort of 'standard' form, that is, one that the participants had experienced themselves. The forms contain personal details, information about education, skills, experience and some added the option of uploading other information.

Despite similarities, we found some unique features that reflected the participants' individual choices in their design decisions (see table 4.2). For example, P1 provide a job list and description at the top of the application form. P2 gave the applicant the choice to upload a CV (or résumé) at the beginning. P6 suggested creating an authentication portal to log into the CV system then filling all the required information. P6 provide an option of connecting the

application with the social media platform LinkedIn. Most participants asked for unnecessary personal information like age (P1, P2, P5, P6 and P7), and address (P2, P3, P4).

Participants P9 and P10 explicitly stated that they were considering the potential for bias in their design. However, they created GUIs sketches that were not so different to the others (see Figure 4.2). Both P9 and P10 had detailed GUIs that included personal information, educational, skills, and other information. The main point was P9 and P10 both argued that the system should generate an applicant ID number. P9 stated that the system will use the ID number so that the system will not use name or gender. A summary of participants' inputs is presented below in table 4.2.

TABLE 4.2: SIMILARITIES AND DIFFERENCES IN PARTICIPANTS' DESIGNS

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Similarities	Name, education, Skills, Experience									
Differences	Job list description	Age address	address	address	age	Authentication portal Social Media age	age	–	ID	ID

The participants' concept sketches of Data Flow Diagram (DFD)

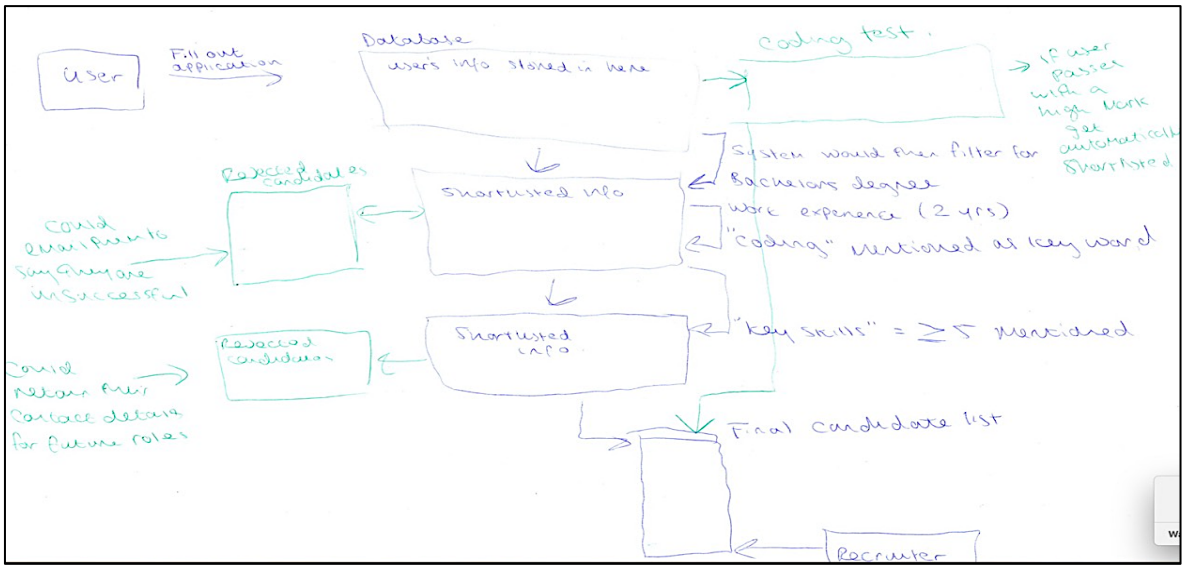


FIGURE 4.3: EXAMPLE OF THE CONCEPT SKETCH OF DFD (P2)

Figure 4.3 showed the DFD for P2 and the way they think the AI would process the inputs. As the difference of colours shows that p2 change some of the options after the second interview (the green is the added after confirming the possibility of bias in their original design). In the original concept of the DFD the P2 focused on shortened the list of candidates and filter out un-qualified applicants. Then after confirming bias the participants added more test and assessment as P2 believed that would make the process of filtering is more fare.

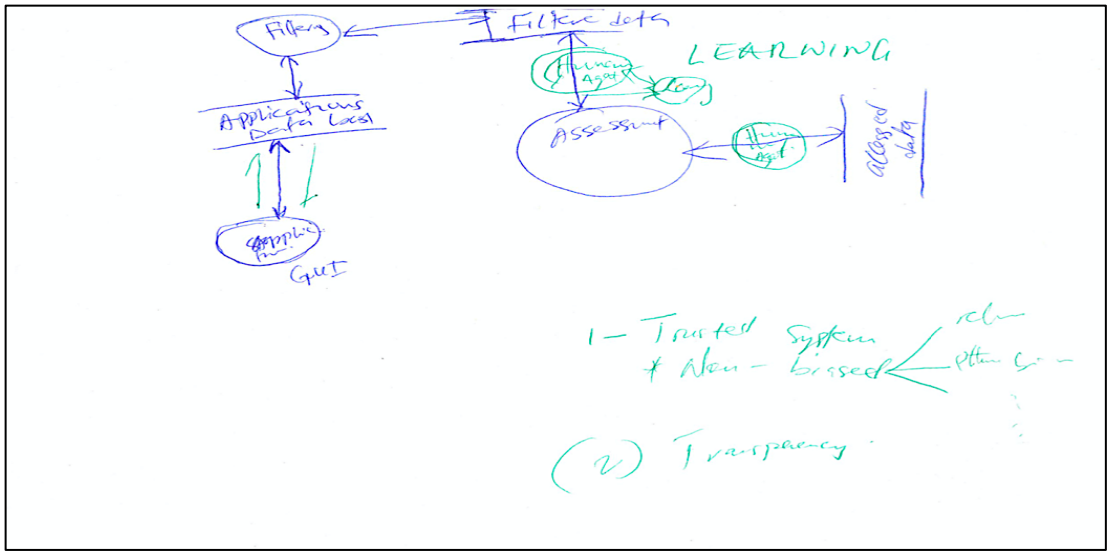


FIGURE 4.4: EXAMPLE OF THE CONCEPT SKETCH OF DFD (P9)

It can be seen in figure 4.4 that p9 also had some changes done after the second interview. The concept sketch of DFD shows that applications go to the system (AI) to be filtered then stored so the system (AI) will make several assessments. P9 added (in green) the learning phase of the neural network and human intervention to make sure results make sense. The rest of the DFDs can be seen in the Appendix.

The DFD concepts describe how information from the GUI concepts will be used by the system (AI) to create a shortlist of candidates. Most participants, drawing on information in the 'person specification' in the experiment scenario, focussed on the job requirements. They had almost the same process where the system read the application searching for all the criteria (work experience, degree, skills, additional skills) then, if these were satisfied the job criteria, the system transferred the applicant to a longlist. However, if one of the criteria was false or missing, the system rejected the applicant immediately. Some participants passed data to databases before the selection process, others suggested automatically reading all information as soon as the application is submitted.

Participants P1 to P8 and P10 had no specification of how the AI system will read the CV. Some participants said the system would scan the database for the job requirements information but did not specify how it will deal with other information in the CV (personal information, for instance).

Figure 4.4 illustrates the DFD concept sketch for P9. P9 stated that the system will have a learning phase and human intervention, so if the system produced any unwanted result, the human will correct it so that the algorithm will learn how to avoid similar problems in the future.

Summary of Features from Sketches

The concept sketches show the participants' ideas and imply the choices they have made. One issue here is that some participants did not consider the personal information as potential bias input. As seen in Figure 4.1, some participants asked for personal information which could be used to identify an individual. While none of the GUI concept sketches contained fields for gender or race, the Bertrand, and Mullainathan (2004) study shows how race can easily be inferred from a person's name (one would assume that similar inferences could be made in terms of gender). In this respect, while the GUI concepts might not be designed to collect information that is intended to produce a biased decision, the personal details *could* produce a bias outcome. For example, the Amazon selection process (in which male applicants were favoured over female applicants) could apply in this case. Table 4.3 shows the result of a summary analysis of the content of the GUI concept sketches that showed what variables and inputs the participants were focused on.

TABLE 4.3: PARTICIPANTS' INITIAL ANSWERS IN THE FIRST INTERVIEW

Participant / **Elements**	**P8**	**P3**	**P4**	**P5**	**P7**	**P1**	**P2**	**P6**	**P9**	**P10**	**TOTAL**
Name	1	1	1	1	1	1	1	1	1	1	10
Education	1	1	1	1	1	1	1	1	1	1	10
Experience	1	1	1	1	1	1	1	1	1	1	10
Skills	1	1	1	1	1	1	1	1	1	1	10
Age	0	0	0	1	1	1	1	1	0	1	6
Address	0	1	1	0	0	0	1	0	0	0	3

ID number	0	0	0	0	0	0	0	1	1	1	3
Gender	0	0	0	0	0	1	0	0	0	0	1
Human intervention	0	0	0	0	0	0	0	0	1	0	1
TOTAL	4	5	5	5	5	6	6	6	6	6	

The DFD concept sketches of the participants show that they followed the job description. Unlike the GUI, the DFD provides more individuality among the designers and shows more reflections of their own beliefs regarding the AI system's functionality. Table 4.3 presents elements that were included by participants while creating their designs. There were 10 participants (columns) and nine elements that were extracted from the concept diagrams (rows). The number (1) in Table 4.3 indicates implementation of the element (the participant used that element in the design, while (0) represent no implementation. We calculated the totals in which the column presents the highest elements that were used in participants' designs, based on this arrangement (name, education, experience and skills) got 10 which means all participants include these elements in their designs. The row (total) presents each participants' usage of the element, and we can divide the participants into two groups based on the total points they have, one group who has 6 entries, which are 1, 2, 6, 9 and 10, this group used much personal information and the designs for participants in this group were most likely to produce a biased outcome. The other group of participants with scores less than 6 entries were 8, 3, 4, 5 and 7. These participants used less personal information comparing to the first, which means they showed more consideration of their design's ethical outcome.

Table 4.3 indicates that the participants' priorities for job qualification-related elements (education, experience, skills) and one personal element (name) all received a 10, which means all participants used them. However, we still argue that some other personal information which should not be used to evaluated applicants can be inferred by the name. Age also received a high point of 6, which is more than half of participants considering that age is important in selecting a candidate for a job. Three participants chose to add address in their application designs. The system's algorithms could use applicants' names, age, address or gender when it reads the CV, and there is a possibility of making one of these parameters critical for filtering or selecting applicants. Only three participants thought about adding IDs to identify applicants instead of using personal information. In general, even though none of the participants wanted a biased outcome, we saw some aspects of the designs that could lead

to bias. Some participants were thinking of bias during the design process and tried to avoid it but made several decisions that could led to biased results.

TABLE 4.4: PARTICIPANTS AFTER THE SECOND INTERVIEW

Participant / Elements	**P2**	**P3**	**P7**	**P8**	**P4**	**P5**	**P1**	**P6**	**P9**	**P10**	**TOTAL**
Experience	1	1	1	1	1	1	1	1	1	1	10
Skills	1	1	1	1	1	1	1	1	1	1	10
Education	0	1	1	1	1	1	1	1	1	1	9
Name	1	0	0	1	1	1	0	1	1	1	7
ID number	0	1	1	1	0	0	0	1	1	1	6
Age	0	0	0	0	0	1	0	1	0	1	3
Address	1	0	0	0	1	0	1	0	0	0	3
Gender	0	0	0	0	0	0	1	0	0	0	1
Human intervention	0	0	0	0	0	0	0	0	1	0	1
TOTAL	4	4	4	5	5	5	5	6	6	6	
Total number of changes	11										

Following the debrief interview, some participants changed their design choices. Table 4.4 shows the changes the participants made. In the light of emphasising the unconscious bias, we re-interviewed the participants to ask further question about their design choices. Specifically, about whether their designs could lead to biased outcomes and what could they do to prevent that. Surprisingly, the majority of participants did not realise that their designs could result in bias, and they were willing to adjust their design decisions. In the sketches the different colours indicate changes. As shown in Table 4.3 P1, P2, P3, P7and P8 change their entries. P1, see Table 4.4, agreed that name and age should not be used in the application. However, P1 kept gender and added a possibly biased option like the country (nationality). P2 removed age as well as information about schools, universities and qualifications to prevent discrimination. P3 cross out the personal information and replaced it with an ID number also,

removed the address. However, some other participants were not sure how their choices would lead to biased results. P4, P5, P6, P9 and P10 kept their entries the same. After confirming the participants of these possibilities most of them agreed that their designs can possibly lead to biased design.

Analysis of CDM Interviews

The aim of using CDM probe-questions (see Table 4.1) is to create a semi-structured interview to let the participants talk freely about their sketches and explain their design choices.

First, we wanted to know if participants had an experience with personnel selection. Most participants had not designed recruiting process before. However, P1, P3, P4, P6, P7 and P8 state that the process was not difficult to understand, especially designing the concept sketches of GUI even without prior experience. P2, P5, P9 mentioned having experience filling in these applications as a user, not a designer, and they agreed with the rest of the participants in having no difficulty designing the GUI. P10, on the other hand, had previous experience designing a GUI for a similar process. P1 said ‘what I chose to add the basic requirement for any job application’, P5 also said, ‘this is a straightforward process where an applicant will fill in the boxes. In terms of the DFD process, the participants: P1, P3, P5, P8, P10 prioritised the education qualification in their sketches, that is, academic degree, as the first criteria to select the applicants. So, the process begins with checking if the applicant had a degree, if yes then will check other requirements, if no it will reject the applicant. P2, P6, P7, P9 processed all the requirements as one step in the DFD. So according to P6, the list of applicants will be stored in a database after submitting the GUI. Then the system will fetch the data and search for candidates who met the requirements. After that another list of candidates will be created, the system will rank the best applicants and will create fewer list of candidates until the number of applicants in the list reaches a reasonable number. P4 started the selection process by selecting applicants who have experience, then this experience should be more than two years, then at least five skills that related to computer science and finally the education or the degree.

In term of design, the GUI concept sketches were similar yet executed differently. All participants included personal information, education, experience and skills in their designs. However, there were many details that reflect a person’s choice. Asking questions about the sketches, P1 answered by saying ‘I created an application that contains a job listing, so applicants would choose the job they want to apply to first then they will fill the CV and other information related to the position. I believe that the design is simple and easy to understand’.

P1's design contains five main pages (or sections) job list, introduction about the job, personal information, skills and experience, self-introduction and assessment test. P2 explained the design by saying 'I gave the chance to upload a cv and I got very standard fields here like contact information, education, work experience and maybe these can be automatically populated from your CV. I also added key skills they should fill separated by commas and finally a space to write why are they suitable for the job'. P2 expressed what was good about the design by saying 'it asked all the right questions, and I liked the idea of auto-filling I think it is good functionality of any application'. P3, P4 and P5 said similar comments about their design. However, P6 had a different design. According to P6, 'the first thing in my application is the logo and the registration, so each applicant should have their own portal page. The user will have the chance to upload the CV from LinkedIn or create the CV by filling the personal information. The next pages will have the education, experience, skills and finally the certificates in case of upload'. P2, P3, P8 and P10 also were in favour of using auto filling tools. The job description requires the applicant to have five skills, and P5, P6, P7, P8, P9, P10 chose to make this field multiple choice or dropdown menu. The other participants chose a text field. These choices reflect what the designers believe is best for their models, and by asking them and letting them speak, we can understand their personal values.

An essential goal of using CDM is to understand the implicit values of the designers. The sketches in this experiment show potential for biased decisions (based on the data the GUI would collect or the DFD would process), so we did a second interview asking participants about bias in their designs. We asked, 'do you think your design will produce biased or unfair results?' and 'what are your recommendation to produce an unbiased system?'. Some participants agreed and understood the bias in their designs. P2 said ' I think you are right, you know I do not have my age in my CV because I am worried people might be biased against me so I think you are right most of the personal information is not important but not the name. We need the name to contact people yet again it could cause bias… um, this can be a tricky one'. P2 stated that all the information in the application could lead to bias not only the personal but also the educational, for instance, the name of the school because you might be judged by the schools you attended. So, P2 deleted some information in the application and added an ID number for each applicant. P3 shared the same concerns and deleted the personal data and added an ID number. P6 said 'I get it will be biased here because it depends on the requirements and we put degree first, so if we got people with high GPA it will choose them and ignore people who have more experience or more skills, but it hard to deal with a

system that was trained on biased data. Even if you removed the values that could produce bias, the system was based and build as a bias' and when asked P6 what could be changed to make the design less bias the answer was 'Now I believe I should remove some inputs, but also I know the recruiter would want to know the name, address and the name of the school, it is really confusing'. P5 and P8 agreed that some inputs could lead to bias but also believe that their system is not necessary bias. P5 said 'I think it depends on the input most than anything, but my DFD sketch doesn't have any personal information. So, I believe it is enough to specify to the machine what inputs should be used in selecting and filtering people to produce the results you want, bias or not. I can't say if training data was bias, so I don't know'.

On the other hand, P1, P4 and P7 did not agree or understand the biased results and believed that they did what it takes to make the system fair. P1 said 'ok I think the name and the age should not be used but gender and degree are a must, you know some jobs require male or female and other jobs require higher degrees than others.' P1 also added other fields in the application like 'nationality' and when we asked 'don't you think selection people based on their nationality or race is not fair?' P1 said 'Maybe selection people based on their colour is not fair, but nationality is important. Performing a job can differ between a native and international, for example, if a UK company chose to hire a UK resident, it won't need to pay for the visa and a company got to think about its profit more than anything'. P4 was certain that using a system (ML) would reduce all the biases that a human would. P4 said, 'If you read my DFD, you would notice I did ignore personal information, so why would the algorithm use them?' In addition, 'When I think of it as a coding process, choosing skills or the experience to select people is a clear order for a ML, it won't go for the name and I think this is one of most advantages for ML because when I tell the system to look at this label, it will only look at that label nothing else, so I don't understand where the bias here'. P7 said 'I think my system is fair enough and the criteria is clear so each applicant would have an equal chance'.

P9 and P10 designs showed some ethical consideration. P10 made sure to use the right algorithm and test the system before any real implementation. Also, both generated an ID number as a primary key in the database. However, they did not delete any field from the GUI sketches. P9 added when asked about bias and said 'the bias can exist because of the algorithms that implement the system not only because of the inputs or the data. Another real-life problem is sometimes you have many candidates that have met the same criteria, in this case first come first serve rule will apply, at least this is me, and I know I will be biased toward time, but people who apply first will get a higher chance, and this is not always fair'.

Using CDM to interview participants in this experiment provided a broad conclusion is that designing the application (i.e., user interface and dataflow) focused the attention of the participants on functionality and not on the consequences of using the application in terms of potential for biased outcomes. However, our results suggested that there was a slightly different when participants were aware of bias implementation in system design and try to avoid it. Providing more visual understanding of the participants beliefs and points of view on how their designs reflect bias, we are implementing Toulmin Model of argumentation.

Toulmin Model of Argumentation

An argumentation scheme was developed from the interviews to represent the primary decisions being made in the design process to enable the system to select the best candidates. In general, the decisions compared the application forms and the résumés to the job criteria then ranked the candidates based on their qualifications to provide the best recommendations. However, the Toulmin models indicate differences in the structure of argument between participants. This could produce conflicting designs which could cause the system to reject individuals for non-related job criteria.

In this section, we present some examples of Toulmin models for several participants (Figures 4.5 to 4.7), then we present the remainder in Table 4.4. The aim of using this model is to show which participant outcome was negative or positive to bias.

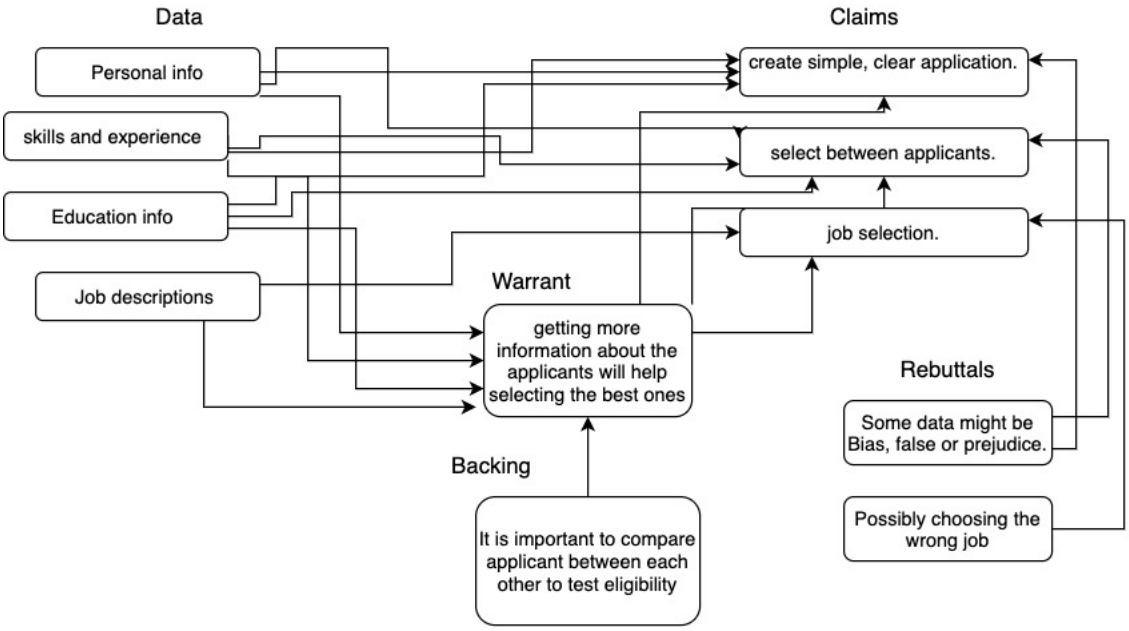


FIGURE 4.5: P1 TOULMIN ARGUMENTATION DIAGRAM

P1's (Figure 4.5) first claim is that a simple, and clear application was created. The second claim is the model will select between applicants. The third claim is that an applicant can chose the job offer from the system. As we can see three data elements connect to the first two claims while the fourth element connects to the third claim. The warrant and the backing in Figure 6 suggests that the more information obtained from the applicants, the better to make a selection. However, the rebuttal statements are saying the if we get more data there is a chance some data might be biased. Also, there is a possibility that applicants will choose the wrong job.

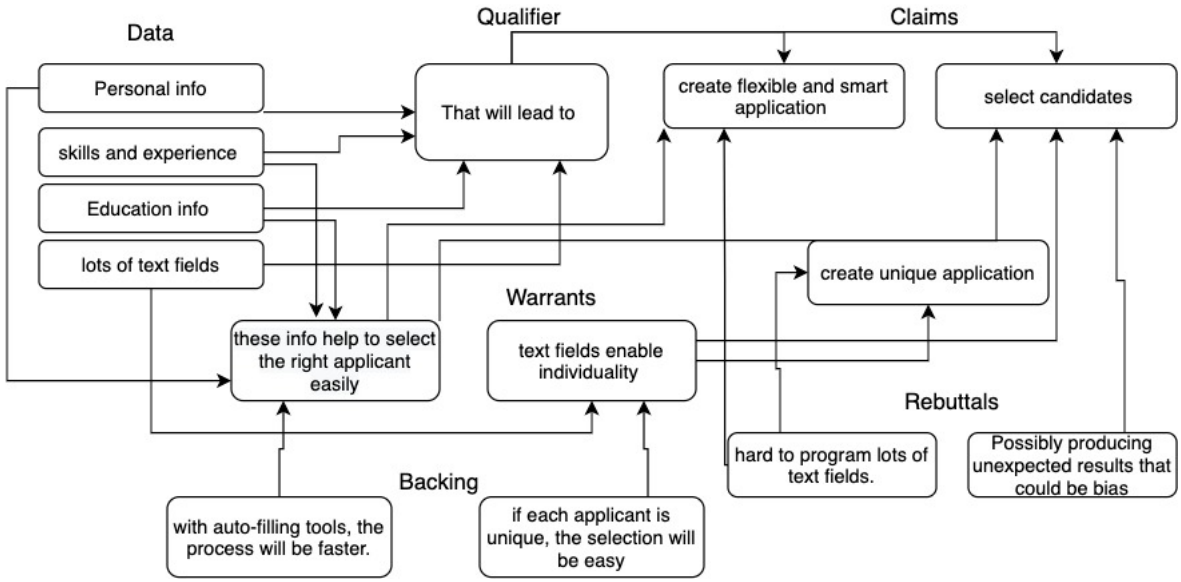


FIGURE 4.6: P2 TOULMIN ARGUMENTATION DIAGRAM

P2 (Figure 4.6) has four data elements (personal info, skills and experience, education info and many text fields were used in the design). P2 connects all the data to a qualifier which directly connect to three claims (create flexible and smart application, select candidates and create unique application. P2 had two warrants which support the relationship between the data to make the claims. All the data elements except the last one lead to the first warrant (select the applicant easily) which support the two claims (create flexible and smart application and select candidates) the last data element (many text fields) lead to the second warrant

(enable individuality) which support the claim (create unique application) and each warrant has a backing to support it. The rebuttal has two elements that conflict with each claim statement. This suggests that having a unique application will probably be hard to program. It also suggests bias to the claims about creating application and selecting applicants.

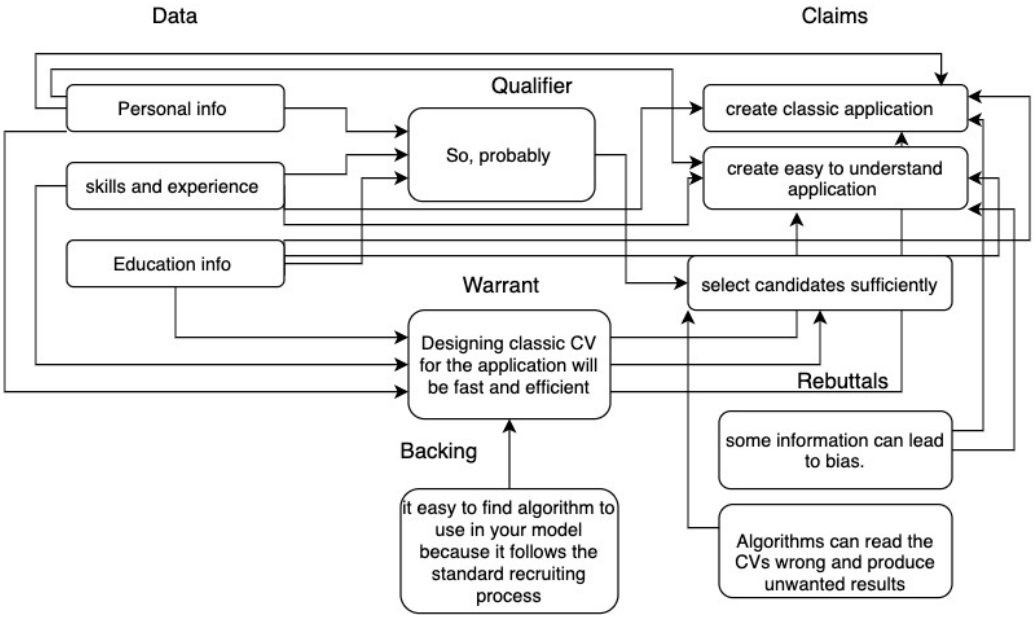


FIGURE 4.7: P3 TOULMIN ARGUMENTATION DIAGRAM

P3 (Figure 4.7) has three data inputs, the first one leads to the first two claims directly while connect to the last claim through a qualifier. The same goes for the other two data elements. All the data elements lead to one warrant and this warrant is supported by one backing and lead to all claims. The rebuttal contrasts with the claims and suggests bias and possibly unwanted results.

The remainder of the Toulmin diagrams for participants P4–P10 can be found in the Appendix. A summary of the elements is presented below to highlight similarities and differences on what participants believed was important in their designs.

TABLE 4.5: A SUMMARY OF TOULMIN MODEL ELEMENTS FOR P4-P10

	Data	Claim	Warrant	Backing	Qualifier	Rebuttal
P1	Personal info Skills, experience Education info	Simple and clear application Select between applicants.	Getting more info about the applicants will help selecting the best ones	It is important to compare applicants between each other to test eligibility		Some data might be biased, false or prejudiced

	Job description	Job selection				Possibly choosing the wrong job
P2	Personal info Skills, experience Education info Many text fields	Select candidates Create unique application	Text fields enable individuality This info help to select applicants easily	If each applicant is unique the selection will be easy With auto-filling tools the process will be faster	That will lead to	Possibly producing unexpected results that could be biased Hard to program many text fields
P3	Personal info Skills and experience Education info	Create classic application Create easy-to-understand application Select candidates	Designing classic CV for the application will be fast and efficient	It is easy to use the algorithm for your model (classic process)	Probably	Some info could be biased Algorithms can read the CVs wrong and produce unwanted results
P4	Personal info Skills, experience Education info	Detailed application is best to select the right candidate	More info will mean easy to prioritise	Using ML info will be organised and filtered fast and neat		Some info can lead to bias Many entries mean more data for the algorithms to read and use
P5	Personal info Skills, experience Education info	Simple design Interactive design	Easy application means an easy process	Good algorithms will work better if the model follows standard design	Usually	Some info can lead to bias Algorithms may read CV wrong
P6	Personal info Skills, experience Education info Registration Upload from other application	Smart, efficient application Good security Select applicant	New methods to collect applicants' info are efficient	Authentic registration is best for security. Other apps will provide more valuable info	That will lead to	Linking the application to other apps can lead the ML to use unwanted info that causes biased outcomes Algorithms may prioritise the wrong entries

P7	Personal info Skills, experience Education info Job description Contact info	Simple and easy-to-understand design List of the best	By asking the right questions, we provide good structure to select from	Asking enough questions will give ML more to compare applicants and select the best	Probably	Some inputs can lead to bias
P8	Personal info Skills, experience Education info Upload resume	Create easy-to-understand but smart application Select candidates	Adding new features like autofill will save time	Saving time means select candidates faster and save money.	Probably	Some inputs can lead to bias
P9	ID number Name email Education info Skills, experience	Create easy-to-understand app to select applicants Unbiased design	IDs will be used to identify applicants	If all criteria are met, the ML will apply first come first serve.	Will lead to	First come, first served can lead to time bias
P10	ID number Personal info Education info Skills, experience Test for IT skills	Create easy-to-understand app to select applicants Unbiased design	It is essential to compare applicants anonymously to ensure fair selection	Testing to IT skills will help select only qualified applicants.	Most likely	Some information can lead to bias

Most of the participants' argumentation diagrams and Table 4.5 show that participants were positive to bias outcome (meaning that their Toulmin diagrams showed, based on their data and claims, that there will be a potential biased output). In a previous section in this chapter, we presented analysis of CDM interviews and described the participants' designs and how they view bias and ethical selection. The diagrams presented what they believed their application design represents and the main purpose of their model. The data the participants used to create the claim were almost identical with slight difference in some participants. Yet their claims were different, some focused on creating flexible and a unique application (P2), some secure one (P6) but the most claimed they wanted an easy to understand application (P1, P3, P5, P7, P8, P9, P10). Warrants and Backing were also varied between participants

which means each one has his/her own way of achieving the design's purpose. Their points of view were mostly oblivious of bias outcomes.

Conclusion

The study of algorithmic decision-making in hiring has been introduced in various studies (see Chapter 2). While there has been a long history of studying discrimination in hiring and social psychology, there has been little work done on the interaction between the algorithmic systems' outputs and the developer's understanding of the ethical consequences. especially across a wide variation of different professions. The result of our experiment suggest bias in the participants' sketches and there was an obvious lack of ethical consideration. This type of bias can be categorised as bias in hiring as designing a job application is part of the hiring process (showed at the end of Chapter 2, table 2.3). It also fits well with our own definition of bias in this thesis '*any recommendation or output from an AI system that can be considered socially unacceptable due to an ill-trained algorithm, unrepresentative data, or a consciously or implicitly biased programmer.'* In this experiment, we believe that participants might produce unacceptable results because of the lack of the ethical awareness that should be highlighted in every AI development. Although we provided a scenario with specific description and goal, when we translated the CDM into an argumentation diagram, we noticed that each participant had a focused goal of their own.

Most of the claims in the Toulmin diagrams focus on the functional side and the representation of the design, not the ethical implication of the decision, because the claims focused on how the system should look and how it works rather than the decision it produces and its effect on people. Even when we told them that there might be a biased outcome, they still did not change their mind. Moreover, in the GUI sketches, most participants included personal details, which could introduce bias to the decision process (some of these details are proscribed in employment legislation in many countries because they are recognised as potential contributors to biased decision-making). In addition, participants specified the design based on technical requirements of the algorithm rather than on an appreciation of likely consequences of applying this algorithm. This point confirms the first hypothesis of this experiment because participants did overlook ethical implications of their work. Furthermore, by creating Toulmin model of argumentation to present the participants' points of view on how to create GUI and DFD to the recruiting process we notice that the participants were contradicting themselves because most of them denied bias in their designs or stated that they did what it takes to prevent bias, which is proof that most biases influenced by the

designer are unconscious. When we asked them about the possibility of bias because of their entries' selection, most of them (not all) tried to modify their designs.

Chapter 5

Building a Bias Machine

Part of this chapter was presented in the paper ‘Conflict Between Frames When Using Machine Learning’ at the Naturalistic Decision-Making (NDM) conference 21–24 June 2021.

Introduction

In the previous chapter, we presented an experiment where we asked the participants to illustrate an AI system. Participants selected an applicant for a specific job by drawing sketches of how they view the process based on their knowledge and experience and how they believe that the AI system will most likely process the applicants. A critical step we examined was the type of inputs the participant requested from the applicants in their designs. We discovered that most participants included unnecessarily personal inputs, which could be used to filter out applicants and lead to biased decisions against the applicants. In this chapter, we improved the experiment to investigate bias within AI/ ML applications by expanding the experiment to the implementation level. We asked the participants to create an AI system instead of doing the sketches.

We were also interested in how well designers of an AI system could recognise its potential to produce biased output. Therefore, rather than asking participants to produce a ‘perfect’ AI system, we asked them to create a ‘bias machine’, which they could use to produce outputs that reflect various forms of bias. This approach allowed us to explore how they define bias, which investigated the second research question in this thesis, ‘How do programmers creating AI/ML systems comprehend the values and beliefs encoded in their algorithms?’ This step aims to create the problem and then analyse what factors led to building that type of system.

In Chapter 2, we reviewed literature on how AI/ML systems promise to improve organisational decision-making by avoiding bias because the machine ought to remain unaffected by moods, prejudices or personal opinions when interpreting the data. However, this promise rests on the tools being independent of the biases of their developers. The purpose of this experiment is to investigate what bias means to developers of AI/ML systems and how they interpret bias through the results of the system. Our concern is the relationship between developer, algorithms, data and output. In Chapter 3, we introduced the data-frame model and its potential to provide explanations and sense making to complex situations. In this chapter, we applied the framework to understand what decisions are made by developers of AI/ML and

how they view bias through their work. We have discussed different types of bias in previous chapters, and this study explores what type of bias our developers might understand. Based on the results of the last chapter, our hypothesis is that programmers usually interpret bias as a technical rather than ethical problem.

Method

To understand how AI/ML developers work with their 'frames', we need to understand the process of creating an AI system. Interviews were conducted with computer science undergraduates who completed a project implementing ML on medical datasets. To encourage them to focus on the potential issues relating to 'bias', we devised a project which we called 'the bias machine'. The goal of the project was to deliberately produce biased output from datasets that could allow 'bias' to be dialled up or down, that is, from no bias to high levels of bias that would disadvantage particular groups of people. Using open access databases, participants selected two datasets to (1) make predictions about future smoking habits of different user groups based on racial profiles, and (2) classify severe visual impairment based on age, gender and race. From these datasets, participants would adjust datasets to inflict bias and implement the algorithms to produce biased results. Following this, we conducted critical decision method (CDM) interviews with the participants, focusing on three high-level questions:

- What frame do developers apply when they select a dataset?
- What frame do developers apply when they select an algorithm?
- What frame do developers apply when they are interpreting the results?

Participants

This study involved four undergraduate students from the School of Computer Science, University of Birmingham in which three of them agreed to participate. We accept that this might be seen as a limitation in the study but argue in our defence that many implementations of 'routine' AI/ML will be performed by junior employees in companies, that is, recent graduates. The knowledge of methods in junior employees could be similar to that of our student participants. The study employed in-depth qualitative investigation involving interviews with three participants. The design of the study was approved by the University Ethics Review Board. Prospective participants were approached based on their experience in computer systems and software engineering, and the three who agreed to participate did so voluntarily.

The Project (Procedure)

Rather than simply ask participants to analyse datasets, we set them the challenge of manipulating the data to produce different levels of 'bias' in the output. In this way, they were working on a project that aimed to produce a 'bias machine' that could dial 'bias' up or down (from no bias to high levels of bias). This project ran for the summer vacation as a team exercise. In that the participants chose to address the health care system, working with medical datasets, 'bias' could be defined in terms of 'social' factors (age, gender, race, social class, level of healthcare, etc.) or 'medical' factors (e.g., disease prevalence, predictive validity).

The activity began with a search for suitable datasets. The definition of suitability included factors such as size, representativeness and coverage of data, type of questions that could be addressed using these data, and other factors. Most of the medical datasets have missing or incomplete data on patients, or certain patient groups or medical conditions could be over- or underrepresented. Often this will require the analysts to modify the dataset, for instance, through normalisation, consistency of labelling, or ensuring a balanced distribution of factors in the data. These processes could introduce distortions to the original dataset. Furthermore, processes of attribute sampling (in which elements in the dataset are selected on the basis of their relevance or expected contribution to the research question) can also introduce bias into the choice of data. In this project, two datasets were selected on the basis of record sampling, that is, the project would use the complete dataset and there was little need for cleaning of the data (e.g., in terms of dealing with missing values):

- The tobacco dataset to make predictions about future smoking behaviours relative to age and race.
- The vision and eye health dataset showed patients who were severely visually impaired, or blind based on a person's age, gender and race.

The students found and defined the datasets from a US website that provides access to a vast amount of large data from different disciplines. Since the project focused on the medical field, the participants selected two databases to test bias using different AI/ML algorithms. The selected datasets were chosen for their size and variation. At first, participants kept 100% of the data so that they could demonstrate how selective filtering by the algorithms could force

bias. Unfortunately, datasets were still subject to problems in sampling, labelling and representativeness. For example, certain race, gender or age groups were disproportionately represented which reflect the demographics of the source of the data but could have an impact on the quality of the output from the algorithms. However, decreasing the quantity of data for a particular group (perhaps to provide 'balance' in the dataset) by 6% was enough to produce a marked change in the results.

Having selected datasets, participants then explored different algorithms. Three different algorithms were applied (neural network, *k*-nearest neighbours [*k*-NN] classifier and linear regression). The participants explained that the Linear regression uses data points to identify a trend and then makes a prediction by extrapolating the trend. They state that the *k*-NN classifier predicts classes based on the percentage of given parameters, while neural networks classify data based on two-thirds of the dataset for training and one-third for testing.

Participants applied the selected algorithms to the datasets, modifying parameters in the algorithms or datasets to introduce different outcomes. The aim was to identify parameters that would reliably produce 'biased' results (and to confirm that these parameters could also be specified to minimise bias). This objective meant some fields were removed, which affected the data. For instance, the year end was removed (from the tobacco dataset), so data collected at the end of a year (overlapping into the following year) would be treated the same as data collected at the start of the year. Moreover, location would likely have a large effect on data. Poorer areas with worse healthcare would have quite different results compared to richer areas. The sample size was also ignored or in one dataset, was seen as an 'equal' spread but was not representative of the population distribution of the United States, where the dataset was collected. Therefore, the results of this exercise formed the basis of group presentations of the design and development activity. The choice of algorithm is considered in the analysis and results section.

Interviews

Three participants were interviewed on an individual basis to explore their beliefs and understanding of bias in the study. These interviews were conducted using the Critical Decision Method, with the probes shown in Table 5.1. Each interview took approximately 45–60 minutes.

TABLE 5.1: CDM INTERVIEW QUESTIONS

Probe type	Probe content (the interview question)
Cues	Think about the last time you worked on machine learning (ML) project? What was about? Can you please talk in general about your project on Bias in Machine learning? What was the most difficult thing about the project? What were the feature that you were looking for in a dataset? What made a good dataset? What were the feature of the algorithms you apply? what made a good algorithm?
Information	Why you think that the feature you defined first, were the best feature you would use? Why do you think this dataset is the best choice for your model? Can you explain the dataset? Do you think the data can be improved? How where the data collected? What are the reasons for selecting these specific datasets? Who would benefit from this?
Analogues & Experience	What do you think is the main purpose of machine learning? Did you have previous experience with other algorithms? What about that previous experience did it seemed relevant to this project? Were you reminded of any previous experience using this method? What specific training or experience was necessary or helpful in making this project?

Standard Operating Procedures	What is the process that you use to make these algorithms? (what data sets, where from, how validated, how sampled) Can you draw a flowchart or timeline for doing this? what decisions happen at each point… What is the process that you use to make the final decision? Can you draw a flowchart or timeline for doing this?
Goals	Is there different output you can or want to make? Do you think the model can present more biased results? How? What do you think is the main purpose of presenting bias? Did the model present the specific goals you planed?
Assessment	How do you know that you produced a biased result? What would make a good bias result? What would make a good unbiased result?
Mental model	How do you train the models? How do you tune the algorithms? What model fitting do you do?
Decision-making & Options	Which algorithms do you think would be most useful for this application? Which ones did you use? Why did you select these? Which other (data) method do you think would be most useful for this project? Why did you select these methods? What other courses of action were considered or available to you?

Analysis and Results

Interviewing the participants in this study was conducted after their team had presented the results of their project (implementing bias Machine Learning algorithms on medical datasets). We started the interviews by asking the CDM questions (Table 5.1) in order, encouraging the

participant to express themselves as much as possible. However, in this section, we are highlighting only the most salient questions and answers.

Why do the project?

We asked the participants 'What do you think is the main purpose of presenting bias?' to understand their motivation and how they viewed bias. P1 said, 'The main purpose was essentially an educational tool to show how easy it is to have biased algorithms and more like how hard it is to have unbiased algorithms to the point of almost impossibility'. P2 said the project was aimed at showing that not every algorithm is suitable for any dataset – people can misuse them. You will only get a biased output if your inputs or processes are biased. On the other hand, for P3 it was much harder to define bias. P3 said 'We did not know what unbiased would be. I suppose that is one of the challenges because it was hard to tell'.

How is bias defined?

Answers to the initial question offered insight into how participants understood bias and how this was related to some incorrect and questionable mathematical results. Participants linked bias to how well or poorly the numbers appeared in the datasets and how many error values the algorithms produced. This point of view was confirmed when we asked participants to describe the features that made a good dataset. P1 stated:

> For datasets, we were looking for kind of equal spread attributes, so we had a dataset which has race, gender and where people lived and we wanted it to be equal in terms of spread, so we wanted an equal amount of every race to be included to make sure it was not biased in any way in those kinds of sets.

Although, P1 described a sample bias, P1 also believed that if the data is statistically balanced, then it is 'good'. However, sometimes 'balanced' data is not representative of the real-world population. Moreover, as P3 said,

> If [the data] had one group that was more likely to smoke and we knew for a fact that was not true, that would be wrong. For example, we had a high percentage of the population smoking, and the algorithm assumed it was from the American Indian/Alaska native racial categories. This prediction is totally unfair…because, in fact, American Indian/Alaska Native were hugely overrepresented in the dataset while in reality, this group only makes up about 0.8% of the total population.

This is not only a bias issue. With a certain part of the population overrepresented in the dataset, the result of the model or the prediction will probably be wrong. P2 said,

> It is a hard question, but you can tell bias when you see wrong values, or the results do not make sense, and there were overlapping [data] that caused bias because it is just not working and will not give the right predictions.

Reason for the choice of algorithm

The teams used three algorithms to test bias (neural network, *k*-NN classifier and linear regression) the reasons for selecting these were, because they had some experience with these algorithms, they were among the popular algorithms in ML and these algorithms gave a clear biased result according to the group. The participants accepted that the selection of the algorithm was not based on absolute principles, but more like testing and exploring what works. P2 stated that the algorithms they used seemed to be very popular and P1 said 'Honestly, we struggle to find a good algorithm, and the ones we tested were very unreliable, and most of them had so many big issues that we really could not see them being used in the light of the application we tested', while P3 said 'we tried the linear regression, *k*-NN classifier and the neural network, and they were different from each other. Not so sure, but they were good enough'. What was much more difficult for participants was to provide a coherent definition of what they meant by 'good'.

Results

Outcome of the algorithm

We asked participants whether the model produced results that fit the goals of their project (which is to present bias). Two of three agreed that they had the result they worked to achieve. P1 stated 'Yes, it showed how easy bias is added even if it not intended because before we started adding bias to things, we also tried to make an unbiased version of the algorithms and that was still biased'. P2 also confirmed this point by saying,

> Yes, this project can raise awareness about how people can be wrong in using certain algorithms with certain datasets. Just make people aware that high-end tools are not necessarily clever. If you get bad data, you will get the bad stuff out. It depends on what you feed your algorithm; also, what algorithm you are using as well.

Meanwhile, P3 was expecting more obviously biased results and addressed some of the limitations that led to the output:

> The results were unexpected. I think we were imagining something that would have been achieved if we had a bigger dataset. So, I guess that is one of the main limitations [having small datasets] Well, it was the largest we can find, but still, the model needed to be bigger.

P3 continued by remarking that

> We limited ourselves [to] the medical field but we could have much larger datasets, which would have been useful. Mainly the biases we were finding were more down to the very limited size of the dataset rather than any other limitation regarding the data.

Conscious or unconscious bias?

Our last point to investigate was the possibility of producing the same result unconsciously. P2 said

> I think our results can be produced unconsciously because some people are just not being aware that bias can be produced and then being naïve in which algorithm you chose for a certain dataset. I mean, we did not even have to try to get this result. We all tried not to be biased in a way because everything was just failing.

Where does bias lie?

To highlight the reason for the bias or where it comes from, we asked participants what caused the bias, algorithms or data? P1 said

> In terms of who is most at fault, I think the model itself and the algorithms we used are very biased. They put their own biases on top of the dataset. I do not think the datasets are perfect, but they are considered not too bad, either. We did not find a lot of extra bias, but the algorithms would take this small amount of bias in the datasets and extrapolated it. In other words, the dataset is what causes the bias, but the algorithm allows it to make a big issue.

P2 also agreed, saying,

> It is both. If you have a certain dataset, you have to study it, and you did some tailoring to select what algorithm you should use and vice versa. If you have an algorithm, you have to tailor the dataset, so it worked the algorithm.

However, P3 had a strong opinion that datasets are the cause of bias,

> Definitely, it all depends on the data. If the dataset is not representative, then the output is not representative either. I think people produce these outputs because of a biased dataset. It is difficult for the algorithms to be biased compared to the actual data you put in.

Frames

Based on the interviews and participants' answers, we propose that participants work with three distinct frames in which they explore bias:

1. Modellers must define a suitable dataset that will answer specific questions and be amenable to analysis. We term this the 'dataset frame', which includes factors such as size, representativeness and coverage, and the type of questions that could be addressed using these data.
2. Having selected datasets, participants then explored different algorithms to test the selected datasets. We term this the 'algorithm frame'.
3. Once the algorithm produces answers, they are reviewed. We term this the 'interpretation frame', which includes judgement on the algorithm's performance (so it overlaps with the 'algorithm frame'), judgement on the interpretation of the output to the original questions and judgement of the implications of this interpretation.

From our interviews, an initial observation is that the 'data' that forms the dataset for AI/ML is both 'data' and 'frame' for the developers in that the choice of dataset (in terms of what it contains and the relationships it defines) creates the definition of the 'frame' (in terms of which algorithm to select, what questions could be asked, the dataset balances between factors, etc.). To develop this further, similar to the way we created DFM in Chapter 3, we selected the elements of the frame based on the participants' answers to the detailed questions about the datasets and the algorithms. In this content analysis, we only used elements that were mentioned by two or more of the interviewees.

The first frame was the dataset frame, showing what were the main features of selecting the dataset from the participant's perspective as a team. As Figure 5.1 shows, five features led participants to select the tobacco and the vision datasets. First, data had to be from the medical field and should be large enough for the algorithm to execute, test and learn. Furthermore, datasets must be diverse (with different entries to explore bias). It is also better to have an equal spread of attributes because there is no point exploring bias with a highly biased dataset. Finally, data should contain a large amount of personal information because

personal information – even if not explicitly used by the algorithm – affects the outcome. Another reason for including personal information is that many real-world companies use this information in their processes. Therefore, selecting datasets that have all these characteristics in their beliefs will affect the fairness of the model. According to P1, data is a critical explanation for any biased system even if it is not the only reason for bias.

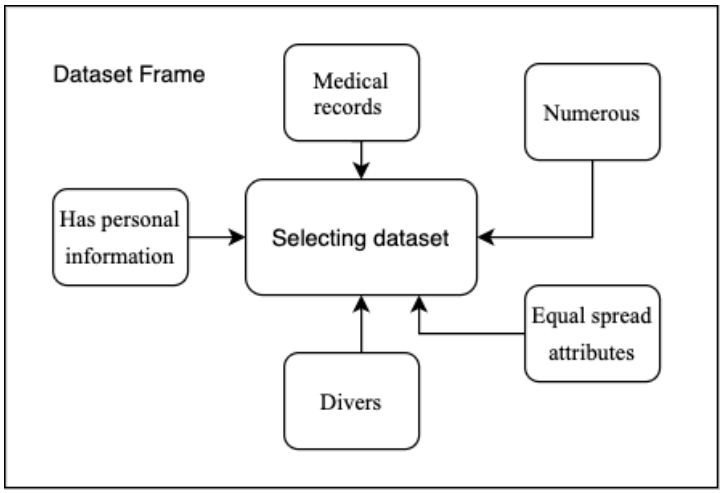


FIGURE 5. 1: DATASET FRAME

The algorithm frame presents features of what makes a good algorithm and what the reasons are for selecting the three algorithms used in the project. Selection was most likely a straightforward process where significant exploration was conducted to see what fits best. As Figure 5.2 shows, algorithms must be testable, used previously in ML and reliable. Moreover, some members of the team (participants) must have had experience using the algorithm before its implementation. First, when the team decided to select an algorithm, they did not intend for a biased one, but the idea was to find a good algorithm to explore with the data they had. Unfortunately, even when not aiming for bias, the results suggest biased execution from the algorithms' unreliability. This issue was described earlier by the team and how they struggled to find a suitable algorithm, and how unreliable their algorithms were.

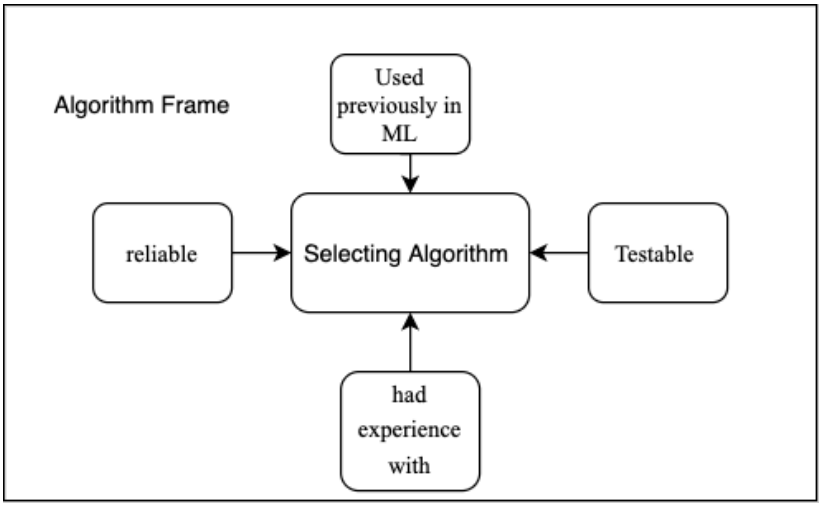


FIGURE 5. 2: ALGORITHMS' FRAME

Figure 5.3 presents the initial goal of the model and how participants viewed bias. Figure 5.4, in comparison, presents the final version of how participants interpreted their results. These figures show different types of bias. Our argumentation in the comparison between these frames is that developers might not know the differences. This lack of awareness is because while working on the selection of the dataset and the algorithms, the questions that address the project shift when interacting with various processes of cleaning data and testing algorithms. Developers often focus on statistical, sampling and coverage bias (because they are using frames that address datasets and algorithms) at the expense of frames which address epistemological bias.

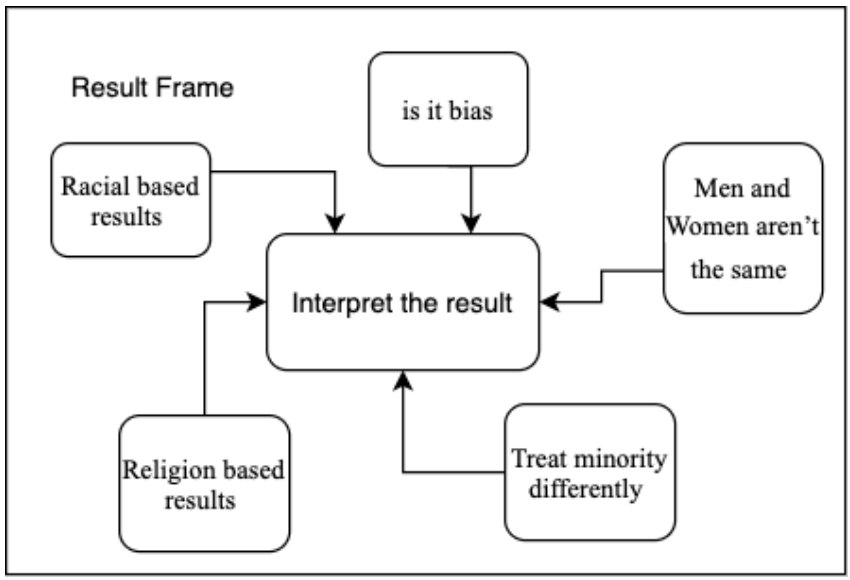


FIGURE 5. 3: INITIAL RESULT FRAME

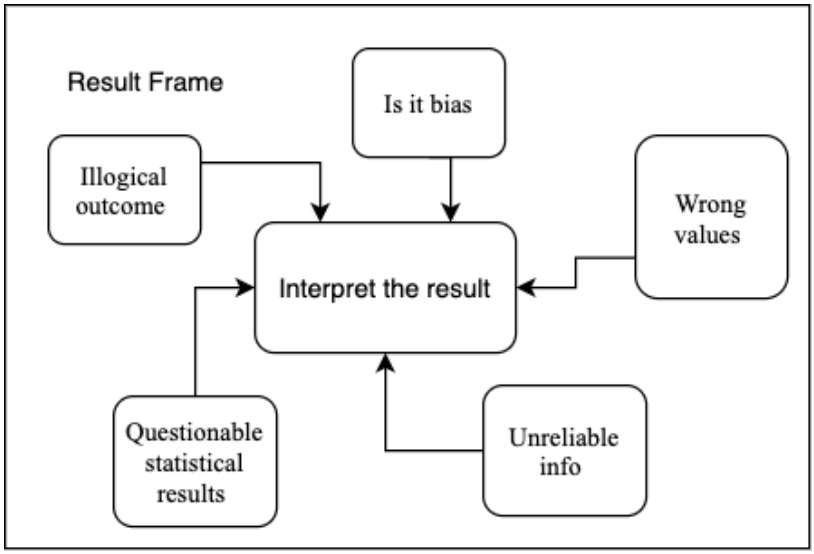


FIGURE 5. 4: FINAL RESULT FRAME 2

Conclusion

We identified three frames, (namely the dataset frame, algorithm frame and the resulting frame) Regarding the identification of bias in these frames, we believe that the data and the algorithms are usually considered by the developer, but it is unclear who would be responsible for the bias in the interpretation frame. Our results suggest that people change the initial version of the 'result frame' in light of the available datasets and algorithms without realising it. This shift explains why we have two subframes for interpretation. The first subframe presents the 'expected result' at the beginning of the project, which aims to demonstrate the ethical bias, and the second is a final result frame that presents the actual outcome. This situation occurs when for example, the data need to be cleaned, adjusted and modified, resulting in differences between the original and prepared versions of the datasets.

Moreover, the algorithm may need some data for training and other data for testing, which requires partitioning the dataset. In other words, we end up with different conflicting interpretation frames. Thus, there is the potential to introduce epistemological bias because the project changes slowly from project X to project Y, where the original interpretation frame may have morphed into one that can be addressed using the revised dataset. Instead of project X, developers have project Y, which has a nicely balanced dataset that executes well with the algorithm, but which might not reflect the real-world population and the results did not address project X's problem. This conclusion is supported by recent study that address how scientists work with machine learning models. The researchers conducted a semi-structured interviews asking the scientists about the process of creating the ML, and other several

questions related to the development of ML and AI models. The scientists' answers were mostly about how hard it was to clean and adjust the data, errors related the execution of the algorithms and the experience they have. They did not mention any ethical consideration of these model outcomes (Baweja et al.,2023). Bias in Automation was provided in different studies in chapter 2 and a summarisation table at the end of the chapter presented definitions that related to this experiment (you can see the definitions in table 2.3). This experiment also agrees with our own definition of bias '*any recommendation or output from an AI system that can be considered socially unacceptable due to an ill-trained algorithm, unrepresentative data, or a consciously or implicitly biased programmer* ' as the participants states that biased outcomes can easily be produced unconsciously and algorithms are tricky to train sometimes, especially with the lack of experience.

This experiment presented frames that revealed three important components of any AI system: data, algorithms, and developers. In Figures 5.1–5.4 the components seem to be different processes – they compete and complete each other to create a system. 'Compete', means that other components are modified to fit one of them. For instance, when changing the original data to fit the algorithm, the algorithm has an important role in the process. Conversely, when the data are the critical information, every process to create the system will change according to the dataset.

This chapter highlights bias in AI from the developer's perspective. We present an experiment in the next chapter investigating the effect AI bias might present from the user's point of view. We explore whether automation bias can affect people who use AI to make decisions. Moreover, we aim to determine what people would do if the system that aids them in making decisions offers incorrect or false values and whether that influences people to make the same mistakes.

Chapter 6

The Effect of AI Recommendations on Human Decision-Making

Introduction

In the last chapter, we presented the project of creating a 'biased machine'. We concluded that AI system developers might not fully understand the ethical biases within these systems and define bias based on errors relating to the datasets or the algorithms of the systems, especially when using the wrong algorithms or non-representative datasets. But what about the users of such systems, decisions makers? Could these biased outputs influence people to produce biased decisions? In Chapter 2, we discussed that employers have long used digital technology to manage their hiring decisions, and many are turning to predictive hiring tools to advise each step of their hiring process. Therefore, employers like Apple, Amazon, Ikea, and major staffing agencies, are testing and adopting data-driven, predictive tools. With increasing public attention on Artificial Intelligence (AI) and the emerging popularity of the technology in the employment context, these tools are simultaneously touted for their potential to reduce bias in hiring and vigorously derided for their capacity to exacerbate it. This advancement in technology can improve the job recruiting process to some extent. However, predictive models can reflect biases in other subtle and powerful ways, which can be challenging to detect and correct. Importantly, removing or obscuring sensitive factors like gender and race might not prevent predictive models from reflecting patterns of bias.

The third thesis question investigated 'How do AI system recommendations affect people's decisions?' This chapter tested the effect of hiring tools on people's decisions, especially when the system gives wrong recommendations or provides unreliable information about the candidates. In our experiment, we present a variety of candidates from different racial groups and different gender to test bias. To test the theory, we distributed a questionnaire containing 20 CVs. Each CV has a recommendation based on a (simulated) AI evaluation. The participant should review the recommendation to decide whether to accept or reject the candidate.

Method

We conducted an experiment to test whether the recommendation from an AI system would affect the judgment of a human (decision-maker). We ask participants to fictionally be in a position of a HR decision-maker who is supposed to select a candidate for the job. We assume that the AI system would be used by people at different grades in companies, but most likely

for people in junior roles who could be responsible for the initial screening of applicants, particularly when a job has received many applicants. To this extent, we did not feel that prior experience in recruitment would be necessary. We also believed we could use participants from a variety of backgrounds.

Participants

A total of 53 participants were involved in the study. The questionnaire was distributed through several online portals (emails and social media). No personal information was collected. All participants were shown a consent form at the beginning of the questionnaire and only if they agreed on the terms, the questionnaire would be proceeded to the questions.

Procedure

We distributed a questionnaire where we gave each participant a job description and showed them a set of CVs (20), where each CV had a recommendation from the system. The recommendation is either high or low.

The CVs are two different types:

one with no personal information and the other with personal information.

- CVs without personal information have only qualification information.
- CVs with personal information have personal and qualification data. In CVs with personal information, we created four groups of people to test bias selection based on race (black vs white) and gender (male vs female)

The CVs appeared randomly for each participant. Each CV holds a recommendation from the AI system which presented in Table 6.1. Four recommendation conditions are provided for each type of CV: correct recommendation, incorrect recommendation, recommendation that matches the percentage the system has and a recommendation that does not match the percentage.

TABLE 6.1: CONDITIONS OF THE AI RECOMMENDATIONS

	Match (M)	**Does not Match (M^)**
Correct (C)	1	2
Incorrect (I)	3	4

1. *Correct and match* recommendation means the system gave a correct recommendation based on the applicant's qualification and the percentage matches the system evaluation that provided in the lower lift of the CV. (See Figures 6.3 for system evaluation)
2. *Correct and not match* recommendation means the system gave a correct recommendation based on the applicant's qualification but the percentage does not match the system evaluation that is provided on the lower left of the CV.
3. *Incorrect and match* recommendation means the system gave an incorrect recommendation that does not fit the applicant's qualification, but the percentage matches the system evaluation that provided in the lower lift of the CV.
4. *Incorrect and not match* recommendation means the system gave an incorrect recommendation that does not fit the applicant's qualification and the percentage does not match the system evaluation that provided on the lower left of the CV.

Answering the questionnaire, the participants will review 20 CV applications for people who applied for interior designer position, the job was offered by X company and the job requirements are:

- The applicant should have a bachelor's degree or higher in interior design
- The applicant should have prior work experience of at least one year
- The applicant should have more to offer, for example, having some extra skills that help improve the work environment or the relationship with the customers
- The applicant should have clean driving license records and no history of any criminal offense

All the applicants applied through the company's system portal and went through an initial stage evaluation by an AI algorithm.

After reviewing the job description, participants examined 20 CVs and were asked two questions:

- Would you recommend this person?
- How confident are you with your answer?

The aim of the first question is to see if the participant would give a correct recommendation in the presence of the system recommendation. For the second question, the aim is to see if the system recommendation would affect the participants' confidence while making their decision.

The CVs in Figures 6.1 and 6.2 are examples of the candidates' CVs. Both types have work experience sections, education, skills, and system evaluations at the bottom left. The CVs differ in content regarding personal information in the questionnaire. All CVs were presented to the participants randomly. We first, tested gender and racial biases (white vs black) and (men vs women). Then we tested the effect of the system recommendation on people correct answers (they recommend the suitable candidate based on the information provided in the CV) and their confidence.

The user interface.

Twenty CV profiles were created using Balsamiq software. All the CVs represent applicants applied for a job. Each CV contains structured data (personal information, qualification, skills, and education) and unstructured data (photo, contact information). The UI also presents a system evaluation of the applicants based on four criteria (based on match to job description and company hiring history, driving license, and overall evaluation) and presents a percentage recommendation score at the top of the CV. Examples of both CV types can be seen in Figures 6.1 and 6.2. The remainder of the CVs can be found in the Appendix

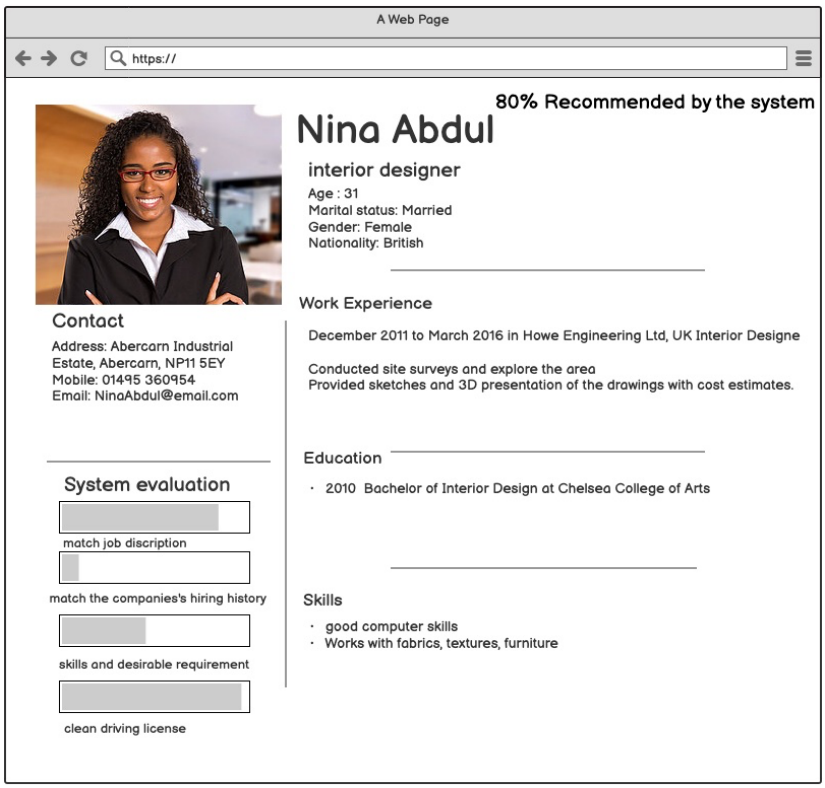


FIGURE 6.1: CV WITH PERSONAL INFORMATION EXAMPLE

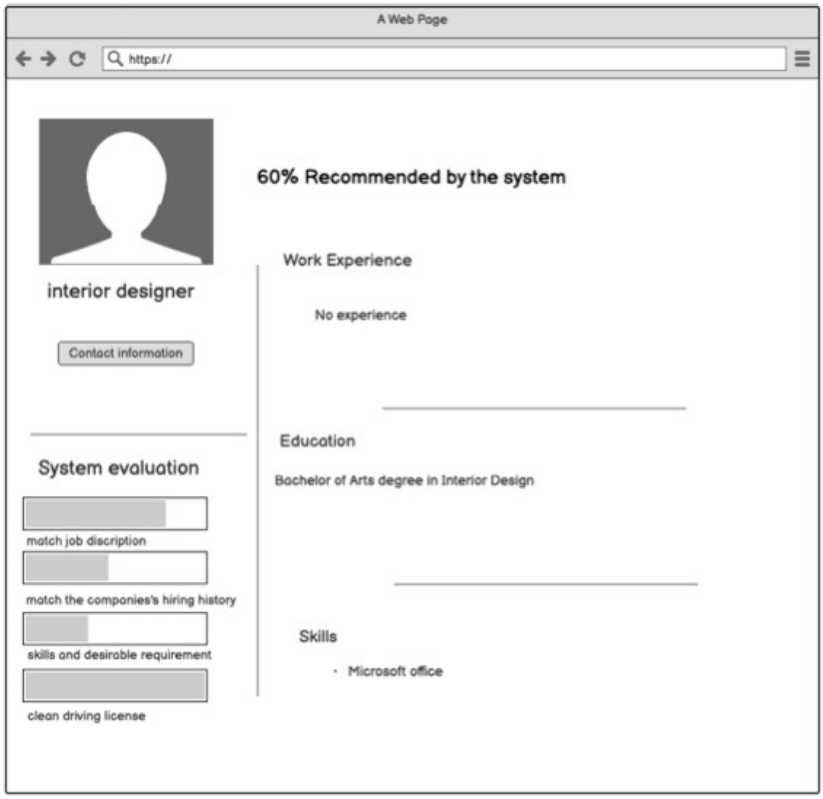


FIGURE 6.2: CV WITHOUT PERSONAL INFORMATION EXAMPLE

How to calculate if the percentage matched or not:

- The system makes a recommendation based on four criteria that were provided in each CV, and the algorithm gave each criteria a weight:
    - Matches job description (0.4)
    - Matches the companies' hiring history (0.2)
    - Skills and desirable requirement (0.3)
    - Clean driving license (0.1)

To demonstrate the evaluation (recommendation), the algorithm multiplies the weight of the participant's achievement in each criterion:

Each criterion will be demonstrated with a rectangular bar showing how the applicant scored in these specific criteria. To calculate the overall percentage, we divided each bar into four sections. For example, the bar above achieved 3 out of 4 sections.

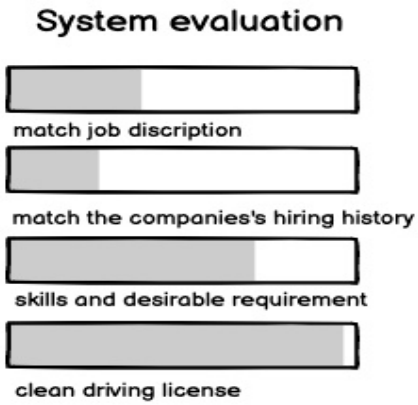


FIGURE 6.3: EXAMPLE OF A SYSTEM EVALUATION

For this example, the process will be:

(0.4*2) + (0.2*1) + (0.3*3) + (0.1*4) = 0.8 + 0.2 + 0.9 + 0.4 = 2.3/4 = 57.5, The system will give this applicant a (57.5%) = recommendation. Thus, for Nina, the system should give 57.5% based on system evaluation. Nina's example is (correct recommendation, no match) because based on the qualification she presented in her CV, she can be selected as a candidate and the system gave a correct recommendation as 80% but does not match the system evaluation presented in the CV. On the other hand, Figure 6.2 showed a CV with no personal information that presents (incorrect/match recommendation) the system giving the applicant a 60%

acceptance, which is incorrect based on the qualifications presented. However, the percentage matches the system evaluation when we calculated using the same formula presented above. For the last example, if the decision-maker would evaluate the applicants based on the system recommendation without any personal consideration, this means an unqualified applicant was chosen as a candidate. The purpose of these measures is to test the system's influence on the participants' decisions to see whether the participants would notice if the system gave them a correct recommendation based on the system evaluation.

TABLE 6.2: SYSTEM RECOMMENDATION AND THE TRUE RECOMMENDATION FOR EACH CV CANDIDATE

sys	CORRECT/MATCH	rec		sys	INCORRECT/MATCH	rec
Y	85%	Y		Y	60%	N
Y	Sara	Y		N	Clair	Y
N	Susan	N		N	Sofia	Y
Y	Omar	Y		Y	Bruno	N
Y	Tony	Y		Y	Theo	N
	CORRECT/NO MATCH				**INCORRECT/NO MATCH**	
N	30%	N		N	10%	Y
N	Mary	N		Y	Chloe	N
Y	Nina	Y		N	Lily	Y
N	Rafael	N		N	Arthur	Y
N	Joseph	N		N	Craig	Y

SYS= system recommendation in each CV

REC= the true recommendation the person deserve

Y= recommended

N= not recommended

To make the table easier to read, it was divided into four sections (CORRECT/MATCH, INCORRECT/MATCH, CORRECT/NO MATCH, INCORRECT/NO MATCH). Under each category, the names of applicants we presented in the CVs, the percentage represents the anonymised applicant with no name in our questionnaire. The percentage is simply the recommendation the system gave to that individual applicant, which might be correct based on the category in the table (6.2).

For example, a CV without personal information was presented with a system recommendation of 10%, the applicant was incorrectly recommended by the system (N) when the applicant deserved to be recommended (Y) based on the qualifications on the CV. We wanted to clarify that the correct recommendation sometimes is (Y) and sometimes is (N). Thus, when we conducted the analysis in the next section, we calculated the correct answers. For example, Bruno recommendation by the system is recommended (Y) but the correct answer should be (N). Therefore, when we calculated the participants' answers, we only considered people who did not recommended him, selecting (N) because it is the correct answer.

Analysis:

The responses collected from the participants were recorded on a Microsoft Excel sheet. Following this, the responses were binarised, namely yes = 1 and no = 0. To check for participant bias, we examined the data based on gender and racial categorisation to ensure that these factors did not have undue influence on decision-making. Following this, data were collapsed to the four recommendation conditions: YY (correct AI recommendation and match between recommendation and system evaluation), YN (correct recommendation; no match) NY (incorrect recommendation; match) and NN (incorrect recommendation; no match). The analysis was conducted for each question, that is, the correct answers the participants had to explore the effect of the system recommendation (because in our experiment the system gave a wrong recommendation in some CVs), and the confidence participants had when making their decisions to also explore the system recommendation effect on people (did they answer with less confidence when the system gave wrong recommendation?)

We carried out the analysis on the correct answers and the confidence ratings in R. We assume that the data will be analysed using non-parametric methods, given the ordinal scale on which they were collected. First, we apply a Friedman non-parametric Analysis of Variance,

and, if there was a main effect, a post-hoc Mann-Whitney test was applied for pairwise comparison.

The Friedman nonparametric test assesses differences between groups (three or more paired groups). The Friedman test also is preferred when compared to other non-parametric tests in a situation where same parameter has been measured under different conditions on the same subject. Additionally, the Mann–Whitney test is used to compare whether there is a difference in the dependent variable for two independent groups. Mann–Whitney compares whether the distribution of the dependent variable is the same for the two groups and therefore, from the same population.

Results:

To test for gender or racial bias in our participants, we compared the correct responses against the Control condition. As there were 53 entries for the Control condition, to produce a balanced analysis we compare race for each gender, and gender for each race. For Black and White candidates against the Control condition for each gender, the Friedman analysis showed no effect for race for Male candidates [$x^2(2) = 4.171$, $p = 0.136$] or for Female candidates [$x^2(2) = 2$, $p = 0.360$]. We therefore assume that our participants were not exhibiting racial bias. Comparing gender for each race, the Friedman analysis showed no effect for gender for White candidates [$x^2(2) = 0.725$, $p = 0.724$] or for Black candidates [$x^2(2) = 0.731$, $p=0.464$]. This means that we do not believe our participants were exhibiting bias for race or gender in their responses. From this, we pooled the data from all responses in our subsequent analyses

In terms of correct responses, Table 6.3 shows the total and percentage correct responses from all conditions. This suggests that there might be an influence of match for correct recommendations (i.e., YY vs YN). So, when the recommendation was correct and the information matched (YY), participants were most likely to produce a correct answer. When the recommendation was correct but the information did not match (YN), this seemed to reduce participants' ability to produce a correct answer. It might be that the lack of match was confusing. There is less of an effect for incorrect recommendations (NY vs NN), with both producing an average of around two-thirds answers being correct, that is, 67%. In this case, it might be that providing an incorrect recommendation was sufficient to confuse participants. What is striking is that the correct response is affected by incorrect recommendations, that is, when the recommendations are incorrect, correct answers reduces. This suggests that participants were paying attention to (rather than ignoring) the incorrect recommendations. In the absence of other information, this might be a reasonable strategy. But there was sufficient

information for participants to decide that the recommendation was not correct and to form their own opinion.

TABLE 6.3: TOTAL AND PERCENTAGE CORRECT ANSWERS.

	YY	YN	NY	NN	Mean
Total correct	239	211	180	179	
percentage	90	79.6	67.9	67.5	76.3

A Friedman Test shows the main effect of a condition (Friedman x^2 (3) = 50.175, p < 0.0001). Following this, pairwise comparisons highlighted differences between conditions as shown in Table 6.4.

TABLE 6.4: COMPARISON OF CONDITIONS. * = 9P<0.05; ** P<0.001

	YY	YN	NY	NN
YY		Z = −3.2**	Z = −5.8**	Z= −6.0**
YN			Z = -3.4**	Z = −3.1*
NY				Z = n.s.
NN				

The pairwise Mann–Whitney test showed significant differences when the system was correct and the percentage it gave matched the system evaluation. Table 6.3 shows that the YY condition received the highest correct answers (239) and the highest percentage (90%). On the other hand, the NN condition had the lowest correct answers from the participants in total (179) and as a percentage of 67%. Therefore, based on our result, participants' ability to answer our questions correctly was highest when we provided a correct recommendation and lowest when they received an incorrect recommendation. We also show that a match between the recommendation and the system evaluation can influence a response, but only when the recommendation is correct.

Confidence

A confidence Table 6.5 was conducted to test if the system recommendation affects people's confident when selecting a candidate. Participants were more confident in the YY condition than the other conditions.

TABLE 6.5: CONFIDENT ANSWERS

	YY	YN	NY	NN	Mean
Total confidence	230	205	185	194	
percentage	86.7	77.3	69.8	73.2	76.7

For confidence, the Friedman Test shows a main effect of condition (x^2 (3) = 26.238, p-value < 0.0001). Post-hoc Wilcoxon tests were applied, showing significant differences between conditions (YY x YN), (YY x NY), (YN x NY) and (YY x NN) (see Table 6.6).

TABLE 6.6: COMPARISON OF CONDITIONS. * = 9P<0.05; ** P<0.001

	YY	YN	NY	NN
YY		Z = −2.9**	Z = −4.6**	Z= −4.0**
YN			Z = −2.1**	Z = −1.1**
NY				Z = n.s.
NN				

The pairwise Mann–Whitney tests showed significant differences between the pairs, YY/NY p < 0.0001, which means that the participants' confident was highest in YY (when the system gave correct recommendation and match information) condition and the lowest in the NY (when the system got incorrect recommendation but match percentage). The NY condition seems to have confused the participants and shaken their confidence. It might also suggest that some participants did not fully understand where the percentage came from or associated the percentage with the system evaluation, which could also have caused this reduction in confidence. Additionally, the YY condition was significantly different to all other conditions which tells us that when the recommendation and the percentage where correct and matching participants' confidence was highest, compared to other conditions. On the other hand, when comparing the conditions NN/NY, the comparison was not significant, and confidence was similar to when dealing with such scenarios. Additionally, the total percentage mean of confidence showed that 76.7% of participants were confident with their choices which also state that 23.3% where not confident and confused because of the system recommendation which implies that AI recommendation's affect decision-making. In particular, when the AI system provides incorrect or mismatched information, participants are less likely to give a correct response and less confident in their responses. The first point might seem obvious (if you are given incorrect advice, this might reduce the likelihood of you giving a correct answer). But the second suggests that participants can perceive problems with the AI system (which

affects their confidence) even if they are unable to challenge this and provide an alternative (correct) response.

Discussion and Conclusion

The results show main effects on correct answer and confidence. This shows a statistically significant effect across the four conditions. Thus, we can conclude that the system evaluation and recommendation did affect the process of decision-making to those who participated in our study. Because the conditions presented four different scenarios where the system recommendation is different in each one and our results suggest that people's choices and confidence were mainly significant between them.

In this experiment we tested the ability to provide a correct recommendation and people's confidence while answering in four scenarios where AI system recommendation was provided to explore its effectiveness on people's decision-making. Our results showed that there was an effect on both the correctness and the confidence. People answered 90% correctly when the system provide a correct recommendation, there was over 30% wrong recommendation when the system provided an incorrect one. Additionally, confidence was at the lowest when the system gave an incorrect recommendation which is evidence that people put the system recommendation at consideration when making their decision because according to the results of the confidence table analysis, when the system provided a correct recommendation, their confidence was higher than when the system provide an incorrect one. Most of the biases we presented in this experiment was mentioned in chapter 2 (table 2.3) where we defined different types of bias and summarised the studies that support that specific type. For instance, this experiment examined social and racial biases, automation bias (when we gave wrong AI recommendation), bias in hiring (selecting candidates) and bias in decision making. And while our result showed no main effect of bias when we tested explicitly gender and racial data, the system can affect implicitly people decision-making and manipulate their choices and even can discriminate people using more subtle variable, in our design of CV we provided a system evaluation that contained information about (match the companies' hiring history) which can contain many variables that algorithms should not use. For instance, if the hiring history of the organisation happened to have more of a specific gender that the other, or the majority of the employees where at a certain age or from local neighbourhood, all these variables can be considered when screening the data for hiring history and while training the algorithm which should not. Therefore, AI hiring tools can reflect institutional and systemic biases, and removing sensitive characteristics is not always the solution. Predictions based on past hiring

decisions and evaluations can reveal and reproduce inequity patterns at all stages of the hiring process, even when tools explicitly ignore race, gender, age, and other protected attributes.

This study examined correct answer and confidence while providing different recommendation (correct/ incorrect and reliable/unreliable). An expectation was that participants would interact differently with these conditions because otherwise means that people follow blindly the system suggestion without any self-judgment and if this happens no matter how wrong, unfair or biased the system is (for many reasons, e.g., unrepresented database) decision-makers will follow and apply the same biases on people (usually groups that were not represented well in the society). Luckily, participants did interact differently. However, this study represented a small part of the population that used these systems and the percentages showed 30% of our participants provided wrong recommendation when the system encourage them to do so, in a major population the same percentage can be discriminatory against underrepresented groups, especially when these systems are trained on real world data. Moreover, hiring processes were usually unique to the specific organisation, society or recruiter when searching for candidates. However, algorithmic screening nowadays impacts individuals on a larger scale which was impossible to see before.

Currently, platforms such as HireVue, and LinkedIn Recruiter provide one centralised screening database for many fields and professions. This consolidation means that the impact of a biased AI system can affect people globally. We examined the effect of AI recommendations on the recruitment field, which is a domain whose biases have been well known for a decent time before the existence of AI. The use of AI remains a controversial subject for many people, many prefer the human interaction and refused to be assessed by a computer. Job seekers continue not to trust or accept such software. On the other hand, there are some fields where AI systems have thrived and flourished, including e-commerce, retail and streaming services and production. Consequently, we wondered whether people or specifically the users of AI systems preferred some interplay with automation for some fields more than others. Does the trust and acceptability of users depend on the naturalness of the service the AI provides? Or it will treat all the services and the recommendations from AI in the same manner? We investigate these issues in the next chapter.

Chapter 7

Exploring The User Reaction and Acceptance to Different Recommendation Systems

Introduction

In previous chapters, we investigated bias in UI design and asked people who were involved in developing AI systems to explain bias within their designs and explore whether they understood subtle and cognitive biases. We then implemented a biased machine to understand what factors led to biased output and whether programmers interpret their results correctly. Based on these first two experiments, we found that bias can exist in many AI systems, even if not as an intentional act of their developers. However, we asked whether systems with wrong and biased values would affect people who interact with them. We conducted an experiment to investigate the effect of a faulty recommender system on people's decision-making, concluding that 30% of the participants gave a wrong recommendation when the system encouraged them to do so.

The system in the previous chapter was a hiring recommender system. We wondered if the same effect could be found in different types of recommendations. In this chapter, we present our last experiment, which investigated the last research question of the thesis, 'What factors are used to determine user acceptance?' We presented various recommendation types and tested users' reactions and acceptance of each. Results showed that the effect and interactions would increase or decrease depending on how much the recommendation influenced people's lives.

Recommendation systems are efficient for corporations and organisations to suggest services and products to people based on data analysis of their previous buying history and personal behaviours. Many systems use AI-based recommendation agents and algorithms to predict the user's need for a service or a product (Davenport et al., 2020). One of the biggest strengths of AI is 'the ability to acquire information and solve problems based on deep learning and knowledge creation and transfer' (De Bruyn et al., 2020). With the capacity of AI to collect large amounts of data from individuals' history and interactions, AI recommendation engines can produce fast and relevant recommendations to reflect each person's preferences and wants. AI can transform those inputs into insights and improve personalised recommendations (Pardes, 2019).

The increasing adoption of AI technology suggests improving performance and that consumers acknowledge and appreciate its unique capabilities. For example, Netflix displays a limited selection of movies and TV shows that users are expected to enjoy rather than presenting their entire gallery to each viewer. This contributes to consumers' overall experience and results in considerable gains. In addition, companies like Amazon.com integrate recommendations into the purchasing process, and these systems have expanded into several other domains such as retail, travel, medical, and finance (Shin, 2021).

Despite the expanding adoption of AI and increasing use of AI-based recommendation agents, our understanding of the user reactions and acceptance of these emerging technologies is still limited. For example, a lack of sufficient data and information can lead to inaccurate recommendations from the AI recommendation system or agent. Despite the superior performance and accuracy of judgments made by statistical models, marketing and psychology research documents people's scepticism toward accepting AI recommendations (Shin, 2021). As is often the case with new and emerging technologies, individuals continue to have essential reservations due to the significant levels of risk and uncertainty associated with those novelties (Seegebarth et al., 2019).

Furthermore, many research work examine people acceptance to how well people understand AI (Del Giudice et al., 2021). Other look at how trust affects people's acceptance to AI (Choung et al.,2022; Kim et al,2020). These studies which look at attitudes to technology or technology acceptance seems to assume a stable trait for attitude (acceptance or rejection) that is mostly related to humans. For instance, Young et al. (2021) investigated patients' attitudes toward AI in which they concluded patients and the public conveyed positive attitudes toward AI. Another research in healthcare and customer support has shown that individuals may experience a significant level of discomfort and perceived risk when dealing with AI robots and recommendation systems (Huang & Rust, 2018). Lichtenthaler (2019) article describes employees' preference to collaborate with real humans rather than having virtual colleagues showed that the same persons may have positive or negative attitudes to AI, depending on the particular situation. These human related factors like preferences, emotions and trust are likely to play a role in accepting the system's recommendation, (Gursoy et al., 2019; Chi et al., 2022) found that social influence and hedonic motivation positively influence people acceptance toward AI. However, there has been less consideration of whether the acceptance is also associated with the type of recommendations the AI system provides (technology related factors). In this chapter, we assume that people might have a different reaction to accept different recommender systems based on the level of risk of the recommendation. We

provided four different scenarios to explore the user reaction and acceptance. For that, we distributed a survey and implemented a quantitative analysis.

Questions of the experiment

Q1: What factors are used to determine the acceptability of AI systems?

Q2: Do the factors that people use to define acceptability of AI systems change with AI systems for applications with different levels of impact on their daily life?

Method

We created a questionnaire exploring user interactions with recommender systems in various scenarios. The questionnaire aims to understand user reactions to different statements related to each scenario to understand the user's expectations of recommender systems.

The questionnaire contained four different scenarios of a recommender system and asked the participants to indicate the extent to which they agreed or disagreed with that statement.

The scenarios vary in terms of risk; the first scenario is about the movie recommender system, which supposedly indicates low risk. On the other hand, the fourth scenario is about a medical recommender system that recommends a patient procedure, which indicates high risk. Our hypothesis suggests that people will think differently when dealing with different recommender systems.

Participants

The questionnaire was distributed using a range of online platforms and social media applications, for example, email, WhatsApp or Snapchat groups. This delivery meant that participants were from different age ranges and different countries, religions, races and education levels. We decided not to collect demographic data from participants because the aim of the experiment was not to explore stratified sampling in this manner, and because asking for such data could be considered intrusive in terms of ethics for the study design. We assumed that all understood English as the questionnaire was written in English. Initial screening of responses, prior to analysis, was undertaken to ensure there were no inconsistent or clearly erroneous response. The questionnaire took 5–10 minutes to complete, and we collected 113 responses. We felt the number was adequate because initially we were

looking for over a hundred responses, and after waiting for approximately six weeks the last two weeks had no increase in the number of responses. Thus, we considered that 113 responses were reasonable.

Procedure

To build the questionnaire, first we created a table containing 120 words related to AI, ML automation and systems, which were collected from the literature review and pilot studies we conducted. We highlighted words, concepts and definitions relating to bias in AI (e.g., social acceptance, fairness, minorities and unconscious) or words related to AI (e.g., hyper-parameter, analysis, automatic and recommendation). In that our focus was on recruitment, we also included words such as filtering, selection, hiring and candidates. Finally, as bias is related to psychology, we added words including beliefs, reasoning, problem, consequences and epistemological (the table can be seen in the appendix).

We desired to capture all the concepts and terms related to the AI system and create the questionnaire based on logical grounds. Furthermore, we wanted to explore how well-connected and related the words were, so we created Table 7.1. The word ‘bias’ was sometimes mentioned when people discussed ‘social acceptance’, ‘hiring’, and ‘fairness’, for instance. This process was run by a group of PhD students, classmates, and staff, who we asked to act as judges to point out the relationship between the words, so if there was connection, we put ‘1’, as shown in Table 7.1. This also can be seen in the Appendix.

TABLE 7. 1: PCA INITIAL TABLE SHOWING THE RELATIONSHIPS BETWEEN WORDS

	Bias	Social acceptance	Hiring	Automatic	Representation	Name	Fairness	Personnel Selection	Professions	Machine Learning
Bias		1	1				1	1		
Social acceptance	1						1			
Hiring	1							1	1	
Automatic										1
Representation										
Name										
Fairness	1	1						1		
Personnel Selection	1		1				1		1	
Professions			1					1		
Machine Learning				1						
Distributions										
Race	1					1	1			
Prejudice	1	1					1			
Computer Science				1						1
Recruiter			1					1	1	
Artificial Intelligence				1						1
Outcomes		1			1					1
Gender	1					1	1			
Minorities	1						1			
Decision-making			1	1	1			1	1	1
Programmers									1	1
Algorithms				1						1
Design				1	1					1
Disability	1									
Criteria	1	1	1				1	1		
Sense-making									1	
Users								1		
Tools			1	1	1			1		1
Data				1	1			1		1
Religion	1						1			

Having defined a set of data, principal components analysis (PCA) was used to reduce the table. Essentially, PCA takes each column and compares it to the other column to conduct a correlation and does that with every possible combination of words. The process of conducting the PCA is: an Initial analysis with all terms was performed with no rotation of the PCA matrix. Although this resulted in 27 components which explained 80.3% of the variance, the matrix was not positive definite. See Table 7.2. Then, there were several terms that were cross loaded. Cross-loaded means that a term is correlated with more than one component which could distort the analysis and the convention is to remove them. After two passes of removing cross-loaded words, 23 components explained 77.9% of the variance. However, this had a Kaiser-Meyer-Olkin score of 0.4 which is very poor. Inspection of the Component Matrix indicated that some of the words made no contribution to any factor. These non-contributing words were removed. A final model was produced in which 20 components explained 76.7% of the variance. This had a Kaiser-Meyer-Olkin score of 0.6 which is modest but was felt to be acceptable. Full access of the PCA Matrix can be found in the Appendix

TABLE 7. 2 : COMPONENT MATRIX FOR PCA

Component Matrix[a]

	Component																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Bias	-0.272	0.360	0.543	-0.040	0.266	-0.093	-0.011	0.077	0.129	0.007	0.076	0.026	-0.110	-0.042	-0.240	0.188	0.141	-0.014	0.064	-0.124	0.043
Social_acceptance	-0.166	0.394	0.417	0.005	0.248	-0.453	-0.050	0.058	0.011	-0.053	0.064	0.081	0.052	-0.060	-0.253	-0.091	-0.067	0.094	-0.070	-0.280	-0.074
Hiring	-0.126	0.243	0.086	0.726	-0.079	-0.033	0.068	-0.056	0.133	0.013	0.073	0.149	0.001	-0.106	-0.066	-0.206	0.062	-0.004	-0.123	0.136	-0.042
Automatic	0.587	-0.428	0.106	-0.046	0.165	0.124	0.203	-0.155	0.150	0.113	0.042	0.150	-0.015	-0.008	0.213	-0.022	0.066	-0.072	0.081	0.065	0.087
Representation	0.060	-0.143	-0.163	0.067	0.313	0.232	0.427	0.014	-0.231	-0.041	-0.007	-0.045	0.262	-0.114	-0.234	-0.071	-0.096	-0.150	0.077	-0.018	0.137
Name	-0.313	-0.030	0.175	0.023	-0.042	0.347	0.066	0.267	-0.290	0.265	0.026	0.150	-0.175	-0.296	0.095	0.294	0.055	0.007	0.201	-0.085	-0.137
Fairness	-0.301	0.299	0.610	0.024	0.232	-0.140	0.054	-0.046	-0.024	-0.056	-0.145	-0.043	-0.072	-0.146	-0.185	0.195	0.020	-0.035	-0.106	-0.095	-0.001
Personnel_Selection	-0.227	0.174	0.209	0.689	-0.061	0.060	0.111	-0.115	0.080	-0.051	0.018	0.003	-0.057	-0.145	-0.113	-0.242	-0.004	0.047	-0.052	0.155	-0.046
Professions	-0.284	0.059	0.008	0.602	-0.163	0.249	-0.008	-0.167	-0.145	0.137	0.111	0.082	0.083	0.172	0.075	0.057	-0.181	0.076	-0.003	0.026	0.105
Machine_Learning	0.460	-0.265	0.072	-0.170	0.221	-0.057	0.408	-0.359	0.224	0.088	0.055	0.173	-0.076	-0.083	0.055	-0.033	0.043	-0.115	0.192	0.236	0.056
Distributions	0.005	-0.073	0.374	0.118	0.072	-0.138	0.470	0.112	0.125	-0.118	0.086	-0.140	-0.021	0.022	0.087	0.250	0.182	0.173	-0.206	0.002	-0.136
Race	-0.390	0.100	0.544	0.194	0.090	0.245	0.165	0.162	-0.228	0.121	0.110	0.041	0.020	0.037	0.315	0.021	0.021	-0.021	-0.089	0.024	0.169
Prejudice	-0.067	0.368	0.635	-0.106	0.158	0.011	-0.199	0.036	0.071	0.025	-0.139	0.057	0.001	-0.139	-0.194	0.018	0.119	-0.003	0.050	0.036	0.158
Computer_Science	0.439	-0.590	0.023	0.018	0.342	0.061	0.070	-0.158	0.033	0.055	0.027	0.091	-0.023	0.006	0.040	-0.028	-0.073	0.078	0.000	0.012	0.053
Recruiter	-0.240	0.089	0.047	0.605	-0.217	0.136	0.028	0.048	-0.027	0.226	0.172	0.149	-0.293	-0.087	-0.214	0.025	-0.084	-0.065	0.122	-0.097	-0.209
Artificial_Intelligence	0.408	-0.309	0.090	-0.106	0.361	-0.136	0.377	-0.366	0.157	0.026	0.064	0.216	-0.002	-0.061	0.066	-0.003	0.066	-0.015	0.127	0.073	-0.062
Outcomes	0.236	0.153	0.172	0.224	0.292	-0.020	0.205	-0.098	-0.378	-0.055	-0.076	0.006	0.323	-0.151	-0.260	0.112	0.091	-0.022	-0.169	0.153	0.010
Gender	-0.384	0.080	0.447	0.314	0.059	0.293	0.161	0.179	-0.204	0.088	0.170	0.105	0.051	0.059	0.287	0.141	0.066	0.047	0.005	0.000	0.146
Minorities	-0.368	0.226	0.616	-0.039	0.171	0.126	0.166	0.128	-0.212	0.066	0.056	-0.037	-0.146	-0.129	0.166	0.048	0.032	-0.036	0.138	0.037	-0.056
Decision_making	0.178	0.509	-0.114	0.197	0.153	0.062	0.094	-0.083	0.008	-0.264	-0.127	0.131	0.005	-0.093	-0.065	0.092	0.111	0.313	-0.202	0.025	-0.014
Programmers	0.241	-0.425	0.081	0.289	0.415	0.209	-0.009	-0.158	0.094	0.032	-0.012	0.024	-0.167	0.077	0.145	-0.051	-0.088	0.116	-0.096	-0.009	0.089
Algorithms	0.528	-0.359	0.115	-0.097	0.125	0.131	0.223	-0.310	0.226	-0.008	-0.062	0.289	0.053	-0.148	-0.044	0.040	0.059	-0.038	0.051	0.018	0.142
Design	0.373	-0.265	-0.170	0.206	0.559	0.044	-0.158	0.291	0.055	-0.003	-0.097	-0.135	-0.015	-0.048	-0.070	-0.088	-0.072	-0.103	-0.117	0.146	-0.076
Disability	-0.352	0.133	0.490	0.030	0.103	0.352	0.065	0.316	-0.179	0.205	-0.019	0.177	-0.096	-0.148	0.043	0.017	0.170	0.120	0.108	0.100	-0.039
Criteria	0.187	0.259	0.176	0.279	0.092	-0.130	0.495	0.046	-0.050	-0.288	-0.093	0.078	-0.016	-0.003	-0.068	0.140	-0.104	0.060	-0.042	-0.125	-0.001
Sensemaking	0.047	0.511	-0.246	0.083	0.276	0.257	0.212	0.001	-0.068	0.047	0.174	0.208	0.185	0.170	-0.045	-0.036	0.087	0.161	-0.035	0.001	0.043
Users	-0.204	0.092	-0.181	0.450	0.291	0.317	-0.181	0.051	0.081	-0.349	0.111	0.105	0.215	0.113	-0.181	0.133	0.076	0.041	0.219	0.084	0.034

The scree plot (Figure 7.1) presents the component number on the horizontal axis and the distance between these vertical axes. The figure shows that from component 1 to 5 we had a distance of 4, when we had 10 components the distance was 2, after that the distance is like a straight line. The graph suggests that around 10 components reflect much of the variance (9 components explain 61% of the variance which was felt to be acceptable).

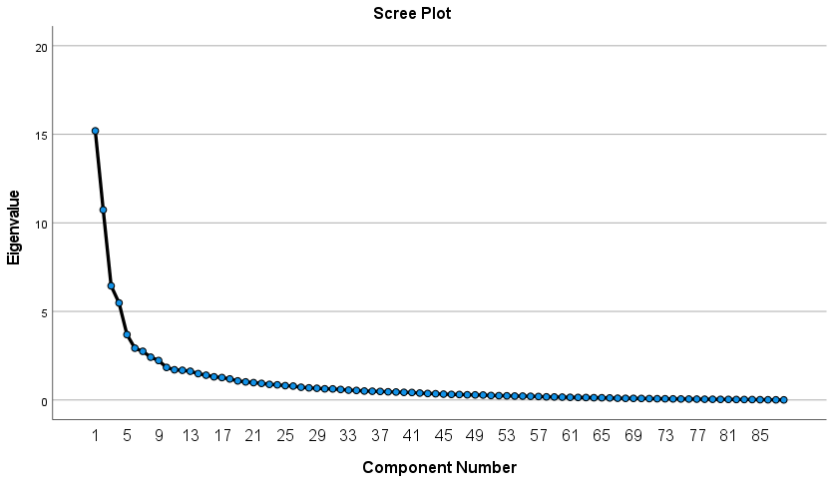


FIGURE 7.1: PCA ANALYSIS SCREE PLOT

Inclusion criterion was set at 0.4 and this produced the following components (three of the contributing components had no words, with correlations of at least 0.4):

TABLE 7. 3: COMPONENTS FROM THE PCA ANALYSIS

1	2	3	4	5	9	10
Automatic	Decision-making	Bias	Hiring	Representation	Participants	Study
Computer Science	Sensemaking	Social Acceptance	Personnel Selection	Programmers	Situation	
Algorithms	Problem	Fairness	Professions	Tools		
Hyperparameter	Psychology	Race	Recruiter	Diagrams		
Input	Argumentation	Prejudice	Users	Graphical User Interface		
Online platforms	Unconscious bias	Gender	Resume			
Statistical	Cognitive biases	Minorities	Application			
Dataset	Values	Disability	Candidates			
Reliable	Beliefs	Religion	Occupation			
Implementation	Reasoning	Illegal	Qualification			
Testing	Recommendation	Ethics	Experience			
Mathematical	Thinking	Culture				
Factors	Selection					
Symbolic	Critical Decision Method					
Model	Epistemological					
Subset	Concept					
Precision	Judgement					
Measurement	Debate					
Weighing	Inference					
Function	Critical					
Analysis	Elaborate					
Regression	Justifications					
Classification	Choice					
Clusters	Goal					
Defining	Consequences					

The Components are given descriptive labels which are (automation, logic and decision-making, bias, experience, and representation) because the words that each component had were most related to each other. The first component had words related to automation and technology, the second component has words related to logic like decision-making, thinking,

choice, select and so on, the third component has words related to bias, the fourth component had words related the hiring process (experience) and the fifth component contained concepts about GUI and representations. We decided to remove components 9 and 10 from the list (because these related to the design of experimental studies rather than the deployment of systems). This resulted in a set of statements that reflected the final components. The purpose of the statements is to discover which aspects of the system people view as critical while interacting with systems depending on the presented scenario. The participant would select if the statement made sense to them and agree with it or disagree. The last statement was created to define the level of security the system in a specific scenario might require from the participant point of view.

The statements used in the questionnaire were:

- In this scenario, it will be important to explain the algorithm used to make this recommendation.
- In this scenario, it will be important to protect personal data.
- In this scenario, it will be acceptable for the system to use the user's personal information to provide personalised services and recommendations.
- In this scenario, it will be important to capture where the system gives wrong recommendations so these can be improved.
- In this scenario, it will be important for users to have recommendations tailored to their social and cultural demographic.
- In this scenario, it will be important for the system to have a clear ethics policy.
- In this scenario, it will be important that users of system are expert in their domain.
- In this scenario, it will be important for the system to have an attractive user interface.
- In this scenario, it will be important for the system to explain its decision.
- In a scale from 1 to 10 how would you rank the level of security this system in this scenario needs (as 1 not very sensitive data and needs low security and 10 require high security procedure)

Scenarios:

The questionnaire had different scenarios that differ in their risks and sensitivity. The main reason for the variety of scenarios is to see if the participant will answer the statements' questions differently if the scenario addresses a recommendation that will affect their healthcare vs a recommendation that suggests which movie they should watch. All the

scenarios in the questionnaire are about a recommender system that offers a service to the customers (the participants). Scenarios will go from as simple as recommending a movie to as high risk as recommending an operation to a patient. Scenarios used in the questionnaire are presented below:

- Scenario 1: You are designing a system that recommends what movie someone might like to watch based on their viewing history. The reason for selecting this recommendation is because movie recommendation applications are very popular and participants most likely experience these systems in their daily life, we chose this as number one recommendation as the result of taking or rejecting the recommendation has no effect on people's life.
- Scenario 2: You are designing a system that recommends colours for decorating your house. As we tried to provide more examples of low-risk recommendation system, we thought decoration recommendation won't put your life at danger but if the recommendation was bad, it might affect your mood.
- Scenario 3: You are designing a system that advises a person on a weight loss programme. Trying to provide recommendation that can have an effect on people, we thought weight loss programs has a bit more effect on people than the previous two scenarios, but it can't put people's life at risk.
- Scenario 4: You are designing a system that advises where to travel based on your budget and interests. The reason for this recommendation is that we thought recommendation that related to money and spending can affect people's finance.
- Scenario 5: You are designing a system that advises a research methodology for research students based on their data. As this scenario should provide moderate but not risk-free recommendation. We suggest a research methodology as a wrong recommendation can cost time and money to students (who usually lack of these resources)
- Scenario 6: You are designing a system that gives suggestions about universities that might accept you based on your GPA. Also, this scenario was felt important in which can affect people selection about their life choices.
- Scenario 7: You are designing a system that advises a company on what people to hire for a specific job. This scenario has an effect on people's life and the way they are making money. People are very cautious about hiring recommendation.
- Scenario 8: You are designing a system that connects all the governmental services to a person's national ID. This scenario was felt risky because it deals with legal documents and information about people usually supervised by the governments.

- Scenario 9: You are designing a system that recommend infertility treatment to a couple. This is one of the riskiest recommendations because infertility treatments are hard to get and very expensive. Giving the wrong recommendation will not only cost money, but couples chance to make a family.
- Scenario 10: You are designing a system that recommends a medical procedure to a patient who needs an operation. We presented this scenario as the riskiest because it is a life and death matter. Wrong recommendations can end people lives.

Pilot Study

The experiment was first run as a pilot study where we used all the scenarios (ten) and the statements. Four people participated in this pilot study (1:M and 3: F), all between 20–35 years old, work in different domains, and voluntarily agreed to participate. It took each participant 15-18 minutes to complete the questionnaire.

After collecting the answers from the participant, they were transferred to an Excel sheet table, which made it easier to process and analyse. After distributing the questionnaire between participants, an Excel table was created to collect the answers. Next, a median table that listed all the participants' values was produced. To this median table, a correlation analysis was done to explore the relationships between the statements which the results showed different relationships between the statements in the scenarios. We assessed how well the statements were connected (the values 0.6 and above indicate strong relationships in the correlation analysis). Out of 10 scenarios, four gave well-connected statements, (Scenarios 1 and 3) which their recommendation can be consider not that risky, and the other two were from the risky recommended systems (Scenarios 8 and 10). Therefore, it was decided to distribute the questionnaire to a broader population using these four scenarios and the 10 statements.

Main Study

Following the pilot study results, our main study was conducted on four selected scenarios (1, 3, 8 and10) which provided well connected statements in the pilot study, and the 10 statements which exported from the PCA. With the same purpose of exploring people's reactions and acceptance in various recommendation systems that vary in degree of risk and how participants answer statements in each scenario. Would there be a difference, or would they be treated the same? We started the questionnaire by asking participants about their knowledge and experience in developing and understanding ML. The purpose of these questions was to know our participants and their level of expertise. Looking at people's

experience using and understanding ML and AI provides an idea about the type of participants answering our questionnaire and their level of knowledge they have interpreting the technology. Next, we pooled all the data together and conduct the analysis based on the questions for each scenario. Participants viewed each scenario in a section that contain 10 statements, the participant had to choose whether they: strongly disagreed/disagreed/neutral/agreed/strongly agreed with each statement in every scenario. After collecting the answers, we tested the internal consistency of the questionnaire items using Cronbach's Alpha for each Scenario. Then A Friedman rank-sum test (non-parametric equivalent to repeated measures analysis of variance [ANOVA]) was conducted on each question as the questionnaire contains mainly ranked data. If this analysis produced a main effect, then Nemenyi tests were used for pairwise comparisons.

Results

Results for the computer experience-related questions are shown in the graphs below.

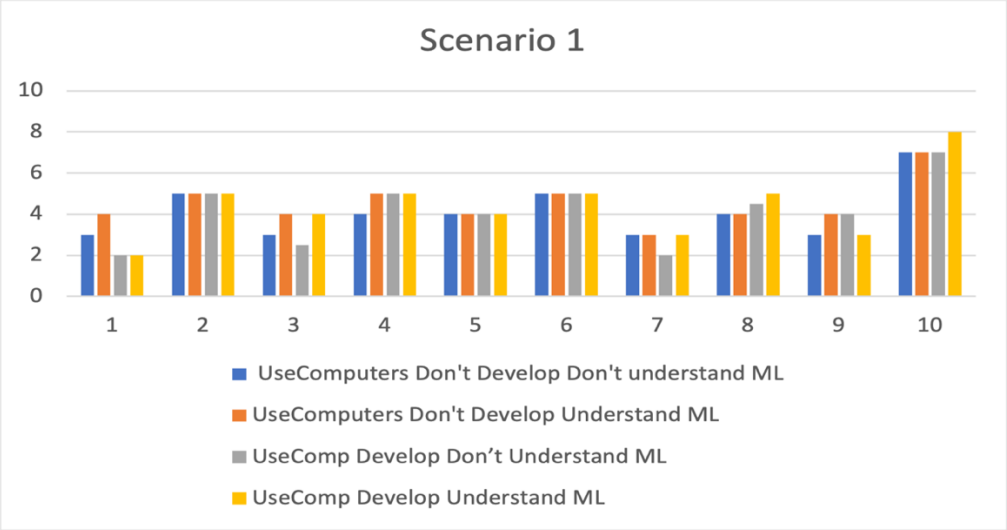


FIGURE 7.2: PARTICIPANTS' ANSWERS TO SCENARIO 1 BASED ON THEIR EXPERIENCE IN DEVELOPING AND UNDERSTANDING TECHNOLOGY

This graph (7.2) presents the participants' answers to the first scenarios based on their knowledge and experience in developing and understanding technology (AI/ML). the first scenario is the movie recommender system, as the graph shows questions (1,3,7,8,9 and 10) had some differences between the groups. The average answer for the first question has a little bit of variation between people who do not develop (3 for people who do not understand ML and 4 for people who do) which is higher than the average answer for people who develop

(both groups have average answer of 2). This means that participants who develop find it less important for the system to explain its algorithms if the recommendation was simple, such as recommending a move.

The third question (regarding the use of the user's personal information to provide personalised services and recommendations) has some differences in average values between people who understand ML (higher value of 4) and people who do not understand ML (do not develop 3 and develop 2.5). Question seven asks participants about their experience, and results showed that all groups find it slightly more important except for the developers but not for the understanding ML group. Question eight, which is about providing an attractive UI, got higher average answers for the developing groups, which could be because they are more aware of the design implementation of the system that they do not develop groups. The ninth question asked the participants if the system should provide an explanation behind its decision, the answers vary but according to the graph, people who develop and understand ML and do not develop and do not understand got higher average answers than the other two. The tenth question had high average answers in all the groups but the people who develop and understand ML received the highest, which might be due to their expertise as developers, requiring them to emphasise the importance of security. Other questions received fairly similar answers in this scenario.

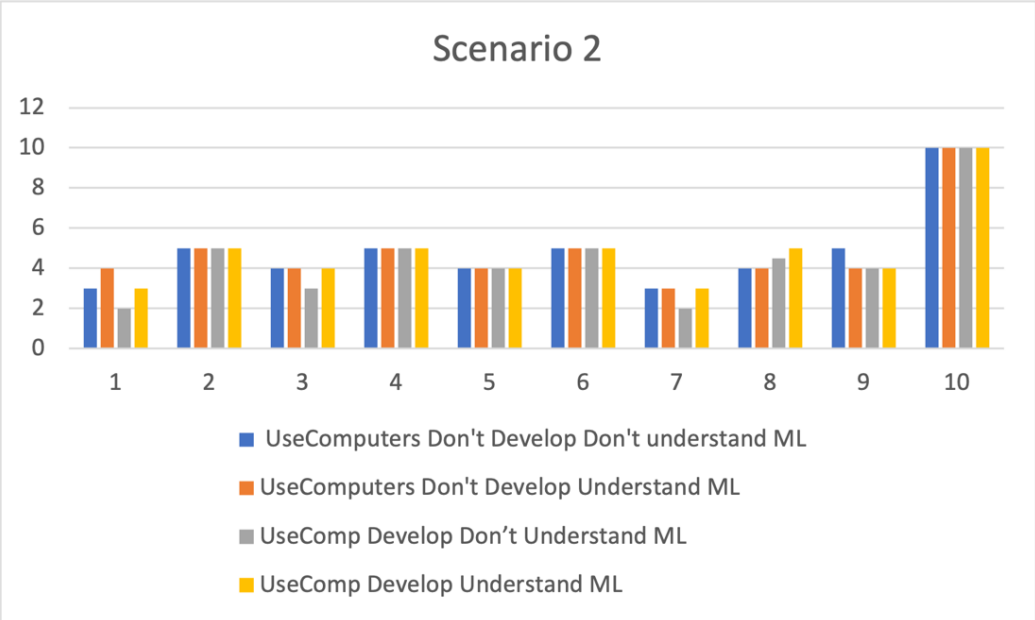


FIGURE 7 3: PARTICIPANTS' ANSWERS TO SCENARIO 2 BASED ON THEIR EXPERIENCE IN DEVELOPING AND UNDERSTANDING TECHNOLOGY

This graph presents the differentiation between participants' answers based on their level of knowledge and experience in developing and understanding ML in the second scenario (weight loss recommender system). The graph showed that questions (1, 3, 7 and 8) had variation of average answers. The first question gave a variation of the average value between the groups in question number1, where people who do not develop and understand ML got higher average of 4, people who develop but do not understand ML got the lowest of average of 2. This variation means that people who do not develop were more concerned about understanding the algorithms that contributed to provide the recommendation. In question 3, people who develop but do not understand ML got the lowest average answer of 3, which means this group of participants was less concern about the usage of personal information if the system will provide customised services. Similar to question 3, question 7 the same group (develop but do not understand ML) got lowest average answer. Question 7 address the user expertise while using the system. In question 8, the groups who develop got higher average answers than the other two.

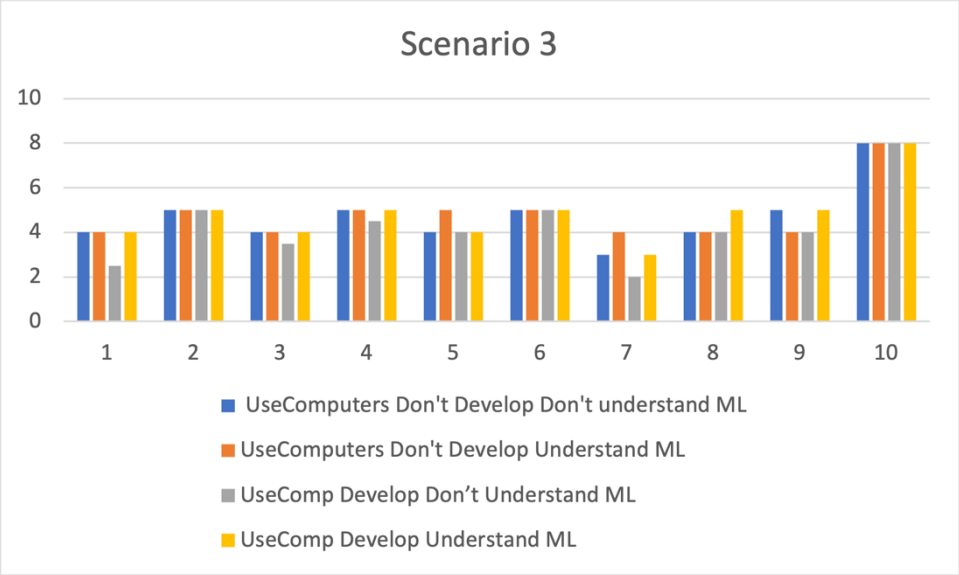


FIGURE 7.4: PARTICIPANTS' ANSWERS TO SCENARIO 3 BASED ON THEIR EXPERIENCE IN DEVELOPING AND UNDERSTANDING TECHNOLOGY

In Scenario 3 (which addressed the ID government service recommender system), we do not see much variation between the groups answering the questions, except for 1, 7 and 9. In question one the group who develop but do not understand ML got the lowest, but the other groups had similar average answers. The seventh question the highest answer was provided by the group who did not develop but understand ML, the group who do not develop but do

not understand ML and the group who develop and understand got similar results. however, the group who develop but do not understand ML got the lowest. In the ninth question, the first and the last groups (people who do not develop and do not understand and people who develop and understand got the highest average answers of 5 while the other two got similar average answer of 4. Questions 3, 4 and 5 had slight differences between the groups, and questions 2, 6 and 10 obtained the same results.

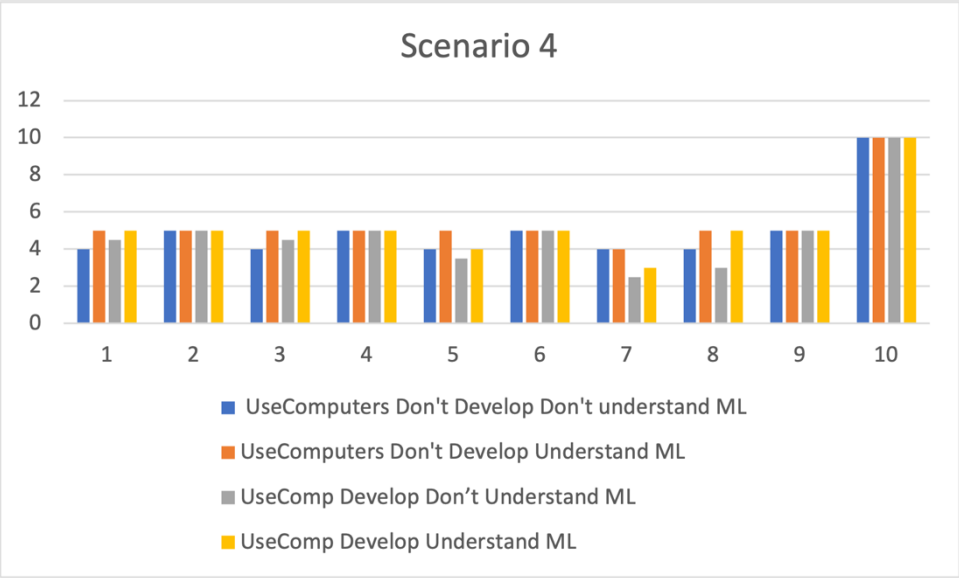


FIGURE 7.5: PARTICIPANTS' ANSWERS TO SCENARIO 4 BASED ON THEIR EXPERIENCE IN DEVELOPING AND UNDERSTANDING TECHNOLOGY

This graph (7.5) presents the fourth scenario (recommends a medical procedure to a patient who needs an operation), which showed not much variation in the average answers between the groups in questions 2, 4, 6, 9 and 10. However, questions 1, 3, 5, 7 and 8 had different answers between the groups. Starting with the 1, 3 and 8 questions, the average answers were quite similar. The groups who understand ML had a higher result of 5 than the other two groups who do not understand ML. For the fifth question, the group who do not develop but understand ML had the highest average answer. This group found it more important for the recommendation to provide services that reflect social and cultural demographics more than the rest of the groups. In the seventh question, a variation was found between the groups who develop and the groups who do not. This means that people who do not develop think the riskier the system is, the more expertise the user should have, which is the general idea. People who do develop might believe that technology should be accessible and easy to use by everyone.

The analysis for the survey questions in the different four scenarios based on people's experiences using and developing software systems showed that there was a difference in the point of views between people who have different experience using technology. We divided the answers based on people experience to software development and their understanding of ML for each scenario. However, we decided to combine all the data and conduct the analysis based on the questions for each scenario, meaning how much of a difference people (all groups) answer each question between the scenarios. This provides concentration on how people answer each question depending on the scenario which serves our hypotheses of exploring the acceptability of user to AI systems that change with different levels of impact on their daily life.

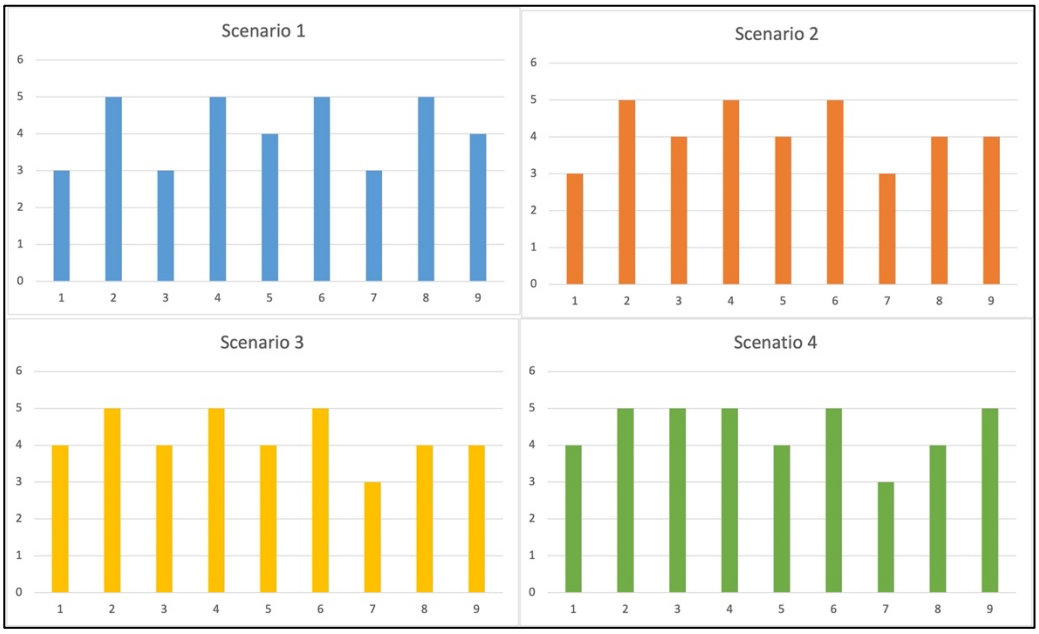


FIGURE 7.6: PARTICIPANTS' MEDIAN ANSWERS TO THE FOUR SCENARIOS

Figure 7.6 presents the median values of the answers to all four scenario questions. As this chart illustrates, some questions got the same responses from the participants in all the scenarios, others got differences, and we are interested in exploring those differences between the scenarios. We are going to name the scenarios (scenario 1:A, scenario 2:B, scenario 3:C, scenario 4:D).

First, we tested the internal consistency of the questionnaire items using Cronbach's Alpha for each Scenario. The assumption is that a high value of Cronbach's alpha would indicate that the different questions were measuring the same concept, for example, the perceived risk associated with the use of machine learning. Typically, the rule of thumb is that a value of $\alpha > 0.7$ indicates a reasonable consistency across items on a questionnaire, that is, that the

questions measure a single construct. We could assume that where α < 0.7, the questions could be less well related, perhaps because they reflect different constructs.

TABLE 7. 4: CRONBACH'S ALPHA

Scenario	Description	Cronbach's Alpha	Comment
1 (A)	Movie recommender	0.6	Low
2 (B)	Weight loss advisor	0.6	Low
3 (C)	National id system	0.7	Acceptable
4 (D)	Medical procedure	0.73	Acceptable

The scenarios scored between 0.6 and 0.73. Scenarios A and B scored less than 0.7. We assume those results are because people have less consistency in their answers due to divergent views of how significant bias in AI is to those situations. For the movie recommender, people do not care about issues of data security, discrimination, and others. If they do care, they believe it is not worth concerning themselves. However, in Scenario D about the medical procedure, people showed more concern about bias or mistakes, suggesting different attitudes that explain the higher value. Scenario C scored around 0.7, which also suggests some consistency in response.

A Friedman rank-sum test (non-parametric equivalent to repeated measures analysis of variance [ANOVA]) was conducted on each question as the questionnaire contains mainly ranked data, and non-parametric statistics compare the ranking of scores. If this analysis produced a main effect, then Nemenyi tests were used for pairwise comparisons. The mean rating was used to interpret differences produced from these pairwise comparisons. The choice of the mean (rather than the median) was simply to capture a higher granularity in the scores being compared. We report our results with the mean. However, the screenplot showed the median values, and we present both to show the subtle differences they both represent.

Answers to Individual Questions

- Q1 (You are designing a system that …, it will be important to explain the algorithm used to make this recommendation.) showed a significant main effect (x^2 (3) = 34.679, p < 0.001). There were significant differences, p < 0.05, between Scenario A (mean 3.044) and Scenario C (mean 3.531), between Scenario A and Scenario D (mean 3.832) and between Scenario B (mean 3.133) and Scenario D.

The first question asked the participants their opinion on how important understanding the algorithms in each scenario. Their answers suggest that Scenario A (movie recommender

system) was not that important, with a mean of 3.044. Comparing this mean value to the other scenarios, we obtained Scenario C (national ID system), with a mean of 3.531, and scenario D (medical procedure), mean 3.832. Therefore, people tend to give the algorithms of the system importance if the system recommendation was about something that touches a sensitive subject in their lives. For instance, in Scenarios C and D, the participants had a need to understand what the algorithm was doing.

- Q2 (You are designing a system that…, it will be important to protect personal data) showed no main effect.

The second question did not achieve a statistically significant value between the scenarios using the Friedman test. However, the median and the mean for this question suggest that people gave 'protecting personal information' high value (Median:5.000, Median:5.000, Median:5.000, Median:5.000) whichever was the scenario, so we did not see any significant difference because participants valued their personal data in all the scenarios. The mean shows some variance but not so much, which shows that participants thought it is essential to protect your personal data no matter what your system is.

- Q3 (You are designing a system that …, it will be acceptable for the system to use the user's personal information to provide personalised services and recommendations). Showed That there is a significant main effect (x^2 (3) = 38.969, $p < 0.001$). There are significant differences, $p<0.05$, between Scenario A (mean 3.195) and Scenario B (mean 3.673), between Scenario A and Scenario C (mean 3.726), and between Scenario A and Scenario D (mean 4.053).

The results in the third question suggest that the higher risk scenario requires much more personalised services and recommendations, with a mean (4.053) compared to a less or lower risk system like the movie recommender (lower than the rest with a mean of 3.195). This variation between the first scenario and the others could be when we address weight loss. People need to have the recommendation exactly to suit their personal needs, and the same goes for when they deal with their national ID system and other governmental services. With a higher median and mean, the fourth scenario (which is about medical procedure median: 5.000, mean:4.053) suggest that participants believe when the system recommend a medical procedure, it is essential to have the system personalise the services to fit their health requirements and need.

- Q4 (You are designing a system that …, it will be important to capture where the system gives wrong recommendations so these can be improved). Showed that there is a

significant main effect (x^2 (3) = 14.144, p < 0.005). However, there are no significant differences, p < 0.05, between pairs of Scenarios.

Analysis of the fourth question suggests that people gave attention and importance to how recommender systems must improve and fix their mistakes; however, the data suggest no significant difference in how people view this point through the different scenarios. Nemenyi test analysis suggests only a slightly apparent difference between scenarios A and D (0.08), which are still not more significant than 0.05. Moreover, the mean values indicate no difference or a minimal difference in scenario A, which the participants thinks is less important than the other but with a statistically significant value. The median suggests that this is one of the questions where participants believed it is essential and rated it with a higher score no matter what the system was recommending.

- Q5 (You are designing a system that …, it will be important for users to have recommendations tailored to their social and cultural demographic). Results showed no significant main effect.

The fifth question provides no significant statistical value, indicating that participants view the social and cultural demographic as somewhat important but with no difference between the scenarios. This finding can reveal how people view bias within these recommender systems. Thus, according to the median (4 in all four scenarios) and the mean (mean: 3.752, mean:3.814, mean:3.761, mean:3.619), all scenarios have similar average values. This question is critical to discovering how people view AI's treatment of them based on their culture or race. However, most answers gave identical values, for which we understood that people consider this point an important question and did not distinguish the scenarios from each other. This feature of the recommender system would treat the users based on personal information. Based on our results, participants seemed less concerned about the naturalness of the recommendation.

- Q6 You are designing a system that …, it will be important for the system to have a clear ethics policy. Showed that there is a significant main effect (x^2 (3) = 15.624, p < 0.005). However, there are no significant differences, p < 0.05, between pairs of scenarios.

The sixth question had a statistically significant effect; thus, people believed that this question was highly critical. Still, there was no differentiation on how important this question was among the scenarios, indicating that participants think having an ethical policy is a must no matter what the recommender system is. Looking closely at the data (median:5.000, median:5.000,

median:5.000, median:5.000) and (mean:4.442, mean:4.442, mean:4.708, mean:4.611) there are slight difference between the first two scenarios A and B and last ones, C and D, but no significant difference. These results statistically support that participants believed ethical policies are essential no matter what the system is.

- Q7 You are designing a system that …, it will be important that users of system are expert in their domain. Showed that there is a significant main effect (x^2 (3) = 28.252, p < 0.0001). There are significant differences, p < 0.05, between Scenario A (mean rating 2.788) and Scenario D (mean rating 3.398), and between Scenario B (mean rating 2.85) and Scenario D.

The seventh question produced a great effect according to the p-value, which indicates that being an expert in using the system is a significant point. An 'expert' can have different connotations, and we are unsure whether people define this word in a similar manner. Does 'expert' mean knowing how to use the system, or does it mean occupationally expert in your domain? We suggest that the general meaning everyone understood was knowledgeable and informed of how to use the system (i.e., having enough knowledge about the domain to understand what the recommendation is). Nevertheless, participants agreed that there should be a difference between the first two scenarios (less risky ones) and the fourth scenario (most risky) with a higher mean value to the last scenario. Participants believed that one should be an expert when dealing with a recommender system such as the medical-related ones and less importance was shown when considering a movie recommender (median :3.000 median :3.00 median :3.000 median :3.000) and (mean:2.788, mean:2.85, mean:3.212, mean:3.398).

- Q8 (You are designing a system that …, it will be important for the system to have an attractive user interface). There is a significant main effect (x^2 (3) = 25.901, p < 0.0001). There are significant differences, p < 0.05, between Scenario A (mean rating 4.23) and Scenario D (mean rating 3.832), and between Scenario B (mean rating 4.257) and Scenario D.

The eighth question, which asks the participant how important it is to have an attractive UI, obtained a significant main effect as a question. Participants offered different answers between the scenarios. According to the data, as opposed to most questions, this query received the highest median value. Also, there was a significant difference between the first scenario (high average) and the fourth scenario (lowest average). That disparity indicates that people care about attractiveness when the system is about fun, entertaining subjects and less about serious issues such as medical producers or governmental processes. Examining the

means (mean:4.23, mean:4.257, mean:3.973, mean:3.832), it is the opposite of the other conditions, which means for the movie recommender it is better if the UI looked nice but for the medical system, it was unimportant.

- Q9 (You are designing a system that …, it will be important for the system to explain its decision). There is a significant main effect (x^2 (3) = 37.768, p < 0.0001). There are significant differences, p < 0.05, between Scenario A (mean rating 3.584) and Scenario C (mean rating 4.088) and between Scenario A and Scenario D (mean rating 4.3725).

The ninth question asks the participants how critical explaining the system's decisions is to participants. This question also revealed significant main effects and different answers among the scenarios. People's answers to this question suggest that there was a great difference between the first scenario (the movie recommender system) and both (the weight loss system and the medical procured system) which we believe the last two systems were about the person's own health and that require the system to be extra careful when providing a recommendation. While it was primarily acceptable for a movie recommender system to provide a suggestion that does not make sense to its user, it is not acceptable to provide a questionable recommendation related to users' health. We have some apparent differences in the mean scores (mean:3.584, mean:4.009, mean:4.088, mean:4.372). D is much higher than the others.

- Q10 (You are designing a system that …, in a scale from 1 to 10 how would you rank the level of security this system in this scenario needs). There is a significant main effect (x^2 (3) = 145.41, p < 0.0001). There are significant differences, p <0.05, between Scenario A (mean rating 6.973) and Scenario B (mean rating 9.204), between Scenario A and Scenario D (mean rating 8.894), between Scenario B and Scenario C (mean rating 7.6110, and between Scenario C and Scenario D.

The last question asks the participant to rate the security each scenario needs. There was a main effect which people consider as an important question and also according to the data they treat scenarios differently. In this question, it was much clear that people treat the health-related system with extra care, both the weight loss system and recommending a procedure system got the highest median (10) and mean (9.204, 8.894), which point it is essential to have good security on your data in these systems. People probably care less if their TV viewing history were viewed by other parties, for commercial purposes perhaps but for the weight loss system and the medical one, people were more worried about their personal information and preferred not to share it with others.

A summarisation of the results can be seen in table 7.5

TABLE 7. 5: SUMMARISING THE ANALYSIS RESULTS FOR EACH QUESTION

The Question	The P-Value	The Mean
Q1	x^2 (3) = 34.679, p < 0.001	A: 3.044, B: 3.133, C: 3.531, D: 3.832
Q2	No significant value	
Q3	x^2 (3) = 38.969, p < 0.001	A: 3.195, B: 3.673, C: 3.726, D: 4.053
Q4	x^2 (3) = 14.144, p < 0.005	A:4.239, B:4.416, C:4.46, D:4.584
Q5	No significant value	
Q6	x^2 (3) = 15.624, p < 0.005	A:4.442, B:4.442, C:4.708, D:4.611
Q7	x^2 (3) = 28.252, p < 0.0001	A:2.788, B:2.85, C:3.212, D:3.398
Q8	x^2 (3) = 25.901, p < 0.0001	A:4.23, B:4.257, C:3.973, D:3.832
Q9	x^2 (3) = 37.768, p < 0.0001	A:3.584, B:4.009, C:4.088, D:4.372
Q10	x^2 (3) = 145.41, p < 0.0001	A:6.973, B:9.204, C:7.611, D: 8.894

Discussion and conclusion

In our questionnaire, we provided four scenarios. In each one, we asked participants 10 questions or (statements), in which they should choose to agree or disagree. Our results show that eight questions out of ten got statistically significant value which represent a variation between the scenarios.

Our experiment shows that people interact and interpret recommender systems differently when there are clear differences in 'risk' associated with the recommendation. How people reacted was most likely dependent on the nature of the question. When experimenting with different types of recommender systems, we proposed four different scenarios from the least critical to the most (A- movie recommender, B- weight loss program recommender, C- governmental national ID recommender and D- medical procedure recommender).

It was not always that the simple recommender systems (A or B) obtained the lowest scores, but more likely that each question had its own interpretation, which we found reasonable. For example, many statements considered system B to be more critical than C, and people showed more interest and careful choices because the questions concern their health. In question 8, we asked participants about the UI, and our results showed that it was more important for scenarios A and B than C or D. On the other hand, question 3 asked the participants about using users' personal information to provide personalised services and recommendations. Their answers showed that they were more interested in having this service (recommendation) for scenario D and less for A. Moreover, in question 10, people were more conservative and required more security levels on the system that address user health (B and D) than other systems. Additionally, our results showed that people have biases; their acceptance of the system recommendation was dependent on how they feel about the type of the services that was provided to them. We present bias in decision making in chapter 2 where we provided several studies that support the result of this experiment and defined different other types of bias (see table 2.3 in chapter 2).

To answer our main question in this experiment, there were several factors which determine the user acceptance to AI systems:

- First, explaining the algorithms used to make the recommendation.
- Second, providing personalised services, especially for systems that provide sensitive services.
- Third, consistently maintaining and updating the system.
- Fourth, providing a clear ethical policy.
- Fifth, for the users of the system to have a knowledge of what should be done when dealing with different recommendations.
- Sixth, have an attractive UI even though this factor was not much considered when dealing with critical types of services.
- Seventh, explainablity, for the system to explain its decision and for the users to understand where this recommendation came from is essential factor. We discussed this point in our literature review in Chapter 2, where we presented it as 'transparency' is a defining factor of a 'good' recommender system (Morar et al., 2019; Tintarev & Masthoff, 2012), that is, the extent to which the computational process behind the recommendation is visible and apparent to the human. It has been shown that increasing the transparency of recommender systems and making explanations available.

- Eight, the level of security the system provides is particularly important. Our data showed that the more sensitive the recommendation is, the higher security they expect from the system.

In this chapter, we tested 10 factors that might affect users' interactions with AI systems in which eight of them showed significant results and importance. These factors vary in terms of their worth. Some factors were more critical depending on the risk of services the system provided. When the system provided less risky recommendation (movie), factors such as UI attractiveness and improving performance were more considered. However, for AI recommendations related to people's health, which were believed to have a critical impact, participants were much more concerned about factors such as security, transparency and explainability, understanding algorithms and personalised services. In previous literature, we presented Castelo's (2019) statement, which is that people avoid AI when the task is subjective to them or impacts sensitive parts of their lives. This observation is in line with Kim et al. (2021) because AI is perceived to lack affective capabilities or empathy. However, this experiment suggested that when providing the factors we discussed in this chapter, we need to careful as these factors change depending on the context AI recommendation. it is shown that perception of AI system and trust depends on the level of risk the service provided by the AI systems and the extent to which the recommendation of these systems affects people's lives. Understanding users' needs is a critical part of human–automation interaction.

Chapter 8

Conclusion

In this thesis, we addressed people's understanding of different form of bias in AI systems. Our studies showed that people involved in the AI development process do not distinguish the difference. Moreover, acceptance of the outcome of AI systems depends on many factors, mostly personal. Bias can be represented by many names and can mean different things depending on the circumstances. People view bias differently. Professionals bring additional divergent views to the concept; for instance, researchers may not interpret bias the same as lawyers. Attorneys might see bias as an illegal act, while academics use it to investigate various possibilities and to test different scientific theories (Reynolds 2000; Reynolds & Brown 1984).

Bias remained within the domain of science until the 1970s, then later it became a major social issue and a subject of debate (Brooks, 1997). In general, bias is a lack of objectivity in a person's thinking brought by a reflection of his or her life (Reynolds and Suzuki, 2012). However, bias is sometimes not a merely in the mind but also an action that invites a notion of prejudicial attitudes. A person seen as prejudiced may be told, 'You are biased against something'. (Reynolds, 2000; Reynolds & Brown, 1984). A teacher in private school may say to a friend, 'I think private schools should pay teachers better.' The friend may reply, 'Yes, but you are biased'.

Our minds use knowledge and experience to interpret people, things and situations even in the presence of limited information. In interviews, social events or at work, we tend to form some sort of impression even if we have only a little information to interpret. These attributes can be influenced by culture, gender or race and affect subjective judgment from previous experience through our relationships (Ross, 2014). Moreover, bias can be presented differently between sectors, and the highly specific definition of such terms is unfamiliar to almost everyone. For example, 'Both in research reports and in public discourse, the scientific and popular meanings of bias are often conflated, as if even the writer or speaker had a tenuous grip on the distinction' (Reynolds and Suzuki, 2012). For this thesis, bias is defined as 'any recommendation or output from an AI system that can be considered socially unacceptable due to an ill-trained algorithm, unrepresentative data or (consciously or not) a biased programmer.

How can people use their own knowledge and experience to interpret the output of AI/ML systems?

In this thesis, we investigated four main questions, which we addressed in several experiments. For the first thesis question, we conducted the experiment in Chapter 4, where we asked participants to create a system themselves based on the knowledge they had about application designs and the potential biased output their designs sketches might hold. The participants in this experiment were all computer science majors. We asked them to use their knowledge and experience to create the system. To simplify the process, we only required a visual (conceptual) sketch of the system design. The experiment sought to identify what kinds of inputs participants thought a job application would have regarding prior experience, and whether those inputs would lead to biased outcomes.

The experiment was conducted in two phases. The first examined whether participants were going to use their own biases and preferences to create the design rather than rely on specific rules to build their designs. In the second phase, we informed them of how bias can be a problem in model designing and asked designers if they wanted to make any changes to their original work. Our result suggested that people interpret bias based on their personal values and beliefs. Most participants had the implicit tendency to respond based on expectations created by a previous experience or association. In both phases, we saw ethical bias, which was misunderstood as various systematic errors.

Bias is an unavoidable problem in AI as technology always comes from and is designed by people, which means that the machines are no more objective than we are. Understanding bias is the first step in prevention. Biases can be automated, mathematical, algorithmic, and social. Accepting these biases within ourselves then working to avoid them is one way to ensure fair implementation in any AI/ML system. The unavoidable truth is that we reflect implicit values through our work, design and research. The most significant aspect of bias in ML/AI (industry or academia) is ensuring the project (system design or research) does not have a disproportionate effect on some group of users relative to others. Hiring is not a single decision, but a process involving many small decisions that finish with selecting an applicant.

When we asked participants to consider bias in their design, their sketches were slightly better. Therefore, establishing a set of rules to be followed in each step of designing an AI system is

required. As a suggestion to prevent bias, Tripepi et al. (2010) stated that 'the correct selection of the study design, the careful choice of procedures of data collection and handling and the correct definition of exposure and outcome represent important prevention strategies for minimising systematic errors'. We suggest providing various rules that can help developers to create a debias system, which we present in this chapter.

How do programmers who create AI/ML systems make sense of the values and beliefs encoded in their algorithms?

Algorithms and automated decision-making tools are being used in our daily lives in diverse sectors including health, retail, education, recruitment and many others. Therefore, implementing a debiased and fair system is essential. It is even more important for programmers to understand this concept, so we address our second question as a way to explore how programmers comprehend the values and beliefs encoded when creating an AI/ML system. We conducted a second study where we asked several undergraduate students to create a biased machine, explaining how biased outcomes are produced. Results showed that the project raises awareness of how the developers of AI/ML might avoid taking a narrow perspective on 'bias' as a statistical rather than a social or ethical problem.

At the end of the project, most of the participants' recommendations were based on technical implementation and awareness. This is not because they were unaware of these wider concerns but because the requirements relating to the management of data and the implementation of algorithms might narrow their focus onto technical challenges. Consequently, biased outcomes can be produced unconsciously because developers are simply not attending to these broader concerns.

In addition, a responsibility exists to think about bias, but it is unclear where in a work system that responsibility lies. We might think it is the responsibility of the programmer to consider bias when they build the program, but this study suggests that it is quite difficult for the programmer to consider social bias because they are too busy thinking about algorithms and data. The work system does not identify someone responsible to think about other types of bias. However, we cannot always blame the programmer alone. We suggest that someone else should be responsible for the resulting frame interpretation within the work system. A method of linking all the frames should be developed so the bias in the resulting frame can be checked independently of the programmers. The programmers should then be made to change what they were doing to minimise that bias. Therefore, creating an accurate and effective model is vital but so is ensuring that all races, ethnicities and socioeconomic levels

are adequately represented in the data model (O'Neil 2016). Methods to debias machine learning algorithms are under development (Gianfrancesco et al., 2018) as are improvements in techniques to enhance fairness and reduce indirect prejudices that result from algorithm predictions (Kamishima et al., 2012).

How do recommendations made by AI systems affect people's decisions?

We assume bias can occur in any automation scenario where programmers unconsciously or otherwise have their own biases and preferences. Thus, we wanted to study this effect on people who interacted with a system and used its products. We examined whether persons and decision-makers who depend on these systems would copy the same mistakes. We asked our third question regarding whether people trusted the systems. The effect was tested by examining people's decisions when using hiring tools, particularly when the system offered incorrect recommendations or provided unreliable information about the candidates. We concluded that the system's evaluations and recommendations did affect the process of decision-making for those who participated in the study.

In Chapter 6, our experiment showed that AI recommendations had an effect on both the correct answers and the confidence of the participant. Even though people answered 90% correctly when the system provided a correct recommendation, over 30% of subjects made a wrong recommendation when the system provided an incorrect suggestion. Moreover, confidence was at its lowest when the system gave an incorrect recommendation, which is evidence that people considered the system's recommendation when making their decision. Even though our results showed that our participants were unbiased when we explicitly tested gender and racial data, the system could implicitly affect people's decision-making and manipulate their choices or even lead to discriminating against people using more subtle variable.

We mentioned in the conclusion to Chapter 6 that hiring processes are usually unique to the specific organisation, society or recruiter searching for candidates. However, algorithmic screening currently impacts individuals on a larger scale that was impossible to see before. Currently, platforms such as HireVue and LinkedIn Recruiter provide centralised screening databases for many fields and professions. This scale means that the impact of a biased AI system can affect people globally. We examined the effect of AI recommendations on the recruitment field, which is a domain that has understood its biases for a considerable time before the existence of AI.

AI in hiring remains a controversial subject. Many people prefer human interaction and refuse to be assessed by a computer. Job seekers generally do not trust or accept such software, see (Schellmann 2022) for examples of people experience with AI in recruitment. In contrast, some fields have seen AI systems thrive and flourish, such as e-commerce, retail and streaming services and production. Consequently, we wondered whether society or specifically the users of the AI systems have interplay with automation in certain fields more than others. The question is whether users' trust and their assessment of acceptability of the technology depends on the naturalness of the service provided by AI.

What factors are used to determine the acceptance of AI?

At the end of this thesis, we wanted to investigate whether AI's influence on people would vary depending on the type of recommendation. In our third study, the AI system affected how our participant selected candidates for the fictional job position we proposed. We sought to clarify whether interactions would vary depending on the risk of the recommendation. For example, we queried responses to the system recommending items such as a movie, diet program for weight loss or a more critical recommendation associated with a person's life such suggesting a medical procedure. Our last study explored these different types of recommendations and presented various factors used to determine the acceptability of AI systems.

We assessed the following factors to test acceptability of the AI system based on whether (1) the algorithms used to make the recommendations were explained; (2) personalised services were provided, especially for systems that provide sensitive services; (3) the system was consistently maintained and updated; (4) a clear ethical policy was provided; (5) users of the system had knowledge of what should be done when faced with different recommendations; (6) an attractive UI was present; (7) explainablity was present for the system to explain its decisions and for users to understand where the recommendations came from; (8) the level of security the system provided was very high. Our data showed that the more sensitive the recommendation was, the higher the level of security was expected from the system.

The experiment in Chapter 7 demonstrates that people interact and differently interpret various recommenders' systems with varying levels of risk. When the system provided less risky recommendations (e.g., movies), factors such as UI attractiveness and improving performance were considered more. However, recommendations from the AI system related to people's health that were believed to have a critical impact received significant concern about factors such as security, transparency, explainability, understanding the algorithms and personalised services.

The results of the four research questions in this thesis showed that interaction with AI is two-way street. In Chapter 2, several studies have examined the process of designing AI systems. (Parasuraman, 2000) investigated to what extent the process should be automated. (Davenport et al., 2020) focused on specific type of users – costumers, for instance – asking how AI may influence marketing strategies and customer behaviours. McDuff et al. (2018) suggested that many of the biases in AI exist because of data homogeneity in the field. In this research we tried to reach further and explore a different side of AI software systems. Programmers and developers of AI/ML systems influence the final output of products even when making no effort to do so. Results from these outputs influence the decision-makers who use the system to aid them in different sectors. Sometimes it is difficult for the user to accept these outputs or recommendations. The ubiquity of computers and the growth of the 'Big Data' movement (Davenport & Harris, 2007) have encouraged the rise in algorithms (Berkeley et al., 2014). However, algorithm judgement can mimic the same mistakes as people. This issue is enormously problematic, particularly when most corporations not only use AI systems as an aid tool but depend on them to make decisions. A great movement is underway to automate every aspect in our lives. We hope for some standardisation as well as guidelines about the use of AI and where and why it is used in specific scenarios.

New legal acts and laws are working to protect AI victims. An article in the Guardian addressing the EU's AI act (Naughton, 8 Oct 2022) says these laws are meant to protect people from harmful software. The article mentions an example of 'algorithms that boost misinformation and target children with harmful content; facial recognition systems that are often discriminatory'. And how crucial is it for people who work and produce these systems to be held accountable for their systems' output and have the ability to explain the results and how it was built.

This point is similar to what we are addressing in this thesis. AI algorithms work in mysterious ways. With these systems being used in critical areas such as recruitment, education and healthcare, it is essential to be aware of the consequences they may cause. Mostly, it is important for the programmers (or the developers of the AI system) to understand the consequences and work with their organizations, governments and colleagues to prevent any ethical infringement.

Contributions of the thesis

Many studies have addressed bias in human decision making (Helms, 2010; Arnott, 2006; Hastie & Dawes, 2009; Levinson & Young, 2009; Mitchell et al., 2005; Mukherjee, 2015). However, few of the studies published to date have considered how developers of AI understand what biased output means for AI systems. This thesis provides evidence that unconscious bias exists in AI development and implementation. That is, programmers/ developers misinterpret ethical bias as technical bias or systematic error and respond to this by adjusting and modifying the data or training the algorithm. Evidence for this point was provided by the results of Chapter 4 (Exploring unconscious bias in user interface designers: A case study of a job application system) and Chapter 5 (Building a Bias Machine). In this thesis, it is argued that there is need for developers to better appreciate the potential consequences, in terms of biased decisions, that can arise from the output of the AI systems they develop. While this point was made forcefully in O'Neil (2016), the results in this thesis contribute to an understanding of how and why developers might fail to consider such consequences. It is proposed that this makes a contribution to contemporary debate on the development, implementation and use of AI systems.

Most of the studies presented in Chapter 2 focus on users' responses to one type of AI system, which usually ends in acceptance or rejection of the recommendation or decision that the AI system produces. In this thesis, people's acceptance of recommendations of various types of AI systems in considered. Prior research has demonstrated that acceptance depends on human perception of the AI system and their trust in its performance (Madhavan & Wiegmann, 2007; Shin, 2021; Kim et al., 2021). In this thesis, it is shown that perception of AI system and trust also depends on the level of risk the service provided by the AI systems and the extent to which the recommendation of these systems affects people's lives. In Chapter 7, we define factors that determine users' acceptance to AI recommendation and show that these factors are not one size fits all solution; the solutions need to change depending on the context AI recommendation. This makes a contribution to the debate on explainable AI and supports the argument that it is necessary to ensure that explanation (and the manner in which a recommendation is presented) is tailored to specific types of users and contexts of use.

It is also shown that, according to Chapter 6 results, when AI systems produce incorrect recommendation, around 30% of participant do not notice (either due to lack of attention to the recommendation or misunderstanding). This finding is important because it suggests that users of AI systems are less likely than might be expected to engage in critical or sceptical review of the recommendations of AI systems. This could be because users are unduly

trusting of AI systems (Del Giudice et al., 2021; Choung et al.,2022) but the studies in chapter 6 suggest that users might not be paying attention to all of the information provided to them. This suggests that there is a need to better understand how the output of AI systems can be presented to support human decision making and enable users to review the factors that contribute to the consequences of the recommendation. In other words, while contemporary explainable AI often focuses on encouraging users to understand *how* the AI system produced its recommendation (in terms of the data or algorithm used) it is argued that a more important factor arises from understanding the potential consequences of this recommendation. As such, the study in chapter 6 provides a user perspective on the issues raised in chapters 4 and 5, which considered developer perspectives.

We believe much work is needed to make 'fair' or explainable / interpretable ML models. The technology must be examined, tested and understood if it is to be used in a broader environment. Technology should not be adopted assuming it already works – we must participate in the creation of the technology as well for AI to be successful (Weber, 2019). Bias is an unavoidable problem in many domains, e.g., epidemiological research (Tripepi et al., 2010). Some studies (Wachter et al., 2020) maintain that fairness should not be automated because AI lacks the common sense of understanding contextual equality and cannot consider local political, social and environmental factors. As such, there is the need to ensure that the output (and the development) of AI systems provides sufficient information for humans to remain the arbiters of 'fair' output.

Fairness, bias and discrimination in algorithmic systems are globally recognised as topics of critical importance. To date, a majority of research papers have addressed bias and fairness in automation. Nevertheless, a clear gap exists between statistical measures of fairness and the context-sensitive, often intuitive and ambiguous discrimination metrics and evidential requirements used in the development of AI/ML systems. Thus, we believe that any AI automation system should be able to explain itself, and programmers of these systems must follow clear steps to create AI-based systems. We suggest using some rules to help people who interact with AI have a better model that fairly represents less-biased systems.

Five rules to create an unbiased model:

- A fair system must first provide justification for the output by having the ability to explain its results, where the result came from and how it was processed. Whether the algorithm is open for verification is crucial because few companies can provide an explanation of their model's results. This point is supported by (Morar et al., 2019;

Tintarev & Masthoff, 2012; Skitka et al., 2000) work as they emphasise the importance of 'Transparency' which was explained as the extent to which the computational process behind the recommendation is visible and apparent to the human. It has been shown that increasing the transparency of recommender systems and making explanations available to the user improves decision performance. Moreover, our results in Chapter 7 (Exploring the user reaction and acceptance to different recommendation systems), where we presented several factors that determine the user acceptance to AI systems, supports this point in two factors which are (1) explaining the algorithms used to make the recommendation. (2) for the system to explain its decision and for the users to understand where this recommendation came from is essential factor.

- Second, a diverse dataset must be used. If the dataset was strongly imbalanced (gender, race, religion or age), this might affect the training of the algorithms by replicating and using these data variables even if they were not provided in the application. There were several research work we presented in Chapter 2 emphasis the critical role the data play which affect the output of AI system. Many learned models exhibit bias as training datasets are limited in size and diversity or reflect inherent human biases (Tommasi et al., 2017; Torralba & Efros, 2011; Caliskan et al., 2017). McDuff et al.,2018, also suggests that many of the biases in AI exist because of data homogeneity in the field. Moreover, in the experiment of Chapter 5 (Building a Bias Machine), our participants believed that dataset is responsible for most of the biases AI produced. P3 had a strong opinion about this by saying 'Definitely, it all depends on the data. If the dataset is not representative, then the output is not representative either. I think people produce these outputs because of a biased dataset'.
- Third, the model must be sufficiently trained before the final implementation. This point is inferred from Chapter 5 (Building a Bias Machine) experiment. As participants were convinced that mistakes can easily be made while creating AI models and these mistakes can be minimized by doing more training and supervising the system. Furthermore, both Chapter 5 and Chapter 4 experiments results confirmed that bias in AI outcomes can be produced unconsciously. Therefor, sufficient training is essential to reduce bias in the process of implementing AI systems.
- Fourth, the most appropriate algorithm must be used for the right type of model. In Chapter 5 experiment, some participants argued that while using AI /ML models, you can control what variables the algorithm should use to select applicants. algorithms are often trained to read specific formats of forms. Thus, using the same algorithm for

different dataset might not be properly extracting the correct patterns or the right information. Moreover, some algorithms require data to be prepared in particular ways or to follow particular statistical distributions, and it is possible that this could result in algorithms being applied inappropriately. More significantly, many ML algorithms can be modified through the selection of hyperparameters (which can be adjusted to improve the algorithm's performance, in terms of recall and precision or Area Under the Curve). However, such modification is left to the intuitions and experience of the developer or arises from 'trial runs' with the data until an acceptable performance has been achieved. This can mean that the algorithm becomes modified with the goal of improving performance, but that there is little scope for considering the algorithm in terms of its output and the consequences of this.

- Fifth, the social and legal aspects of the system must be understood. Every country, culture and area have their own laws. Therefore, it is essential not to violate any law or discriminate against any group of the population. This rule is stated because most of the participants in this thesis were international, and their understanding of fair, unbiased, non-discriminatory systems differed according to their laws and culture. Most of them unconsciously ignored ethical concepts while creating their designs.

We note that many of these recommendations can be seen in ongoing developments in legislation, e.g., the European Union's AI Act. This thesis contributes to the debate surrounding these developments through the empirical investigation, from a user-centred perspective, of the factors that contribute to human interaction with, and understanding of, AI systems.

Limitations of the research

- We did not explore all aspects of the automated hiring process. In our studies, we examined the application process phase (selecting candidates) where we found that most people will choose and select personal information that should not be considered when selecting an applicant for a job. We wondered whether other phases of the recruitment such as interviewing candidates or selecting the employee would be different.
- The study in Chapter 5 (Building a Bias Machine) had only three participants, who were undergraduate students. Nevertheless, we believe that many implementations of 'routine' AI/ML will be performed by junior employees in companies, that is, recent graduates. Their knowledge of methods could be similar to those of our student participants.
- Chapter 7 experiment (Exploring the user reaction and acceptance to different recommendation systems) explores the user ability to develop and understand computer designing and ML which did not serve the main hypothesis of the experiment.
- The COVID-19 pandemic slowed the data collection and social contact meetings. The effects of the pandemic, which covered around two years (my second and third), limited my participation in several conferences and publication opportunities.

Future work

Our studies leave some open questions to investigate bias in various AI/ML applications. For example, considering algorithmic bias in-depth, interviewing software has recently been developed with complete automation where the applicant is interviewed without human intervention. This method could lead to biased outcomes because the software asks the applicants several questions, which they must answer through a video call. A final judgment is then made based on the information the applicants provide. This technique is questionable because we do not know how the software works, who built it, how it was trained and whether it is reliable. Issues regarding this type of application are a good area to examine.

This research presented the recruitment process in field experiments (Chapters 3 and 5), where we tested bias and the AI system's effect on people's decisions. We believe there are several other fields where an examination of how bias is presented is interesting, as well as exploring ways to minimise it.

AI implementations are widely used in platforms that impact social life. An extension to the social impact of AI that we addressed in this work can be done by highlighting AI biases and social acceptance through social media. Recently a vast number of applications and technologies are introduced.

An interesting area to look at is the Large language Models (LLM) which have made powerful advancements in AI implementation. The new AI CHATGPT, for instance, generate human-like text, answers questions, and completes other language-related tasks with high accuracy. The AI according to Ian Sample, a science editor from The Guardian (2023):

> is fed hundreds of billions of words in the form of books, conversations and web articles, from which it builds a model, based on statistical probability, of the words and sentences that tend to follow whatever text came before. It is a bit like predictive text on a mobile phone, but scaled up massively, allowing it to produce entire responses instead of single words.

The technology is new and has so many open questions when it comes to bias and fairness. Many companies like google also developing these AI tools and experimenting with users the quality of the output. Moreover, further research can be done in terms of trusting the technology as it lacks the ability to understand what is true or false.

A case study can be created by combining the five rules to create unbiased model along with the EU AI act to explore their efficiency and usefulness. This can provide a guide of how to create a successful AI model that ensures fairness which also can be used worldwide.

References

Aamodt, A. and Nygård, M., 1995. Different Roles and Mutual Dependencies of Data, Information, and Knowledge-An AI Perspective on their Integration. Data Knowl. Eng., 16(3), pp.191-222.

Abbaspour, S. and Dabirian, S., 2019. Evaluation of labor hiring policies in construction projects performance using system dynamics. International Journal of Productivity and Performance Management.

Ackoff, R.L., 1989. From data to wisdom. Journal of applied systems analysis, 16(1), pp.3-9.

Adensamer, A., Gsenger, R. and Klausner, L.D., 2021. 'Computer says no': Algorithmic decision support and organisational responsibility. Journal of Responsible Technology, 7, p.100014.

Ahmed, F., Anannya, M., Rahman, T. and Khan, R.T., 2015, May. Automated CV processing along with psychometric analysis in job recruiting process. In 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT) (pp. 1-5). IEEE.

AI, H., 2019. High-level expert group on artificial intelligence. Ethics guidelines for trustworthy AI, p.6.

Allport, G.W. (1961), Pattern and Growth in Personality, Holt, Rinehart & Winston, New York, NY.

American Psychological Association, 2018. https://dictionary.apa.org/first-impression

Anderson, J. R. (1983) The Architecture of Cognition. Cambridge, MA: Harvard University Press.

Anjomshoae, S., Najjar, A., Calvaresi, D., & Främling, K., 2019. Explainable agents and robots: Results from a systematic. https://www.researchgate.net/profile/Davide-Calvaresi/publication/331503549_Explainable_Agents_and_Robots_Results_from_a_Systematic_Literature_Review/links/5c7db0fb92851c695054c72b/Explainable-Agents-and-Robots-Results-from-a-Systematic-Literature-Review.pdf

Arnott, D., 2006. Cognitive biases and decision support systems development: a design science approach. Information Systems Journal, 16(1), pp.55-78.

Aronson, J.K., Bankhead, C. and Nunan, D., 2018. Confounding. Catalogue of Bias. https://catalogofbias. org/biases/confounding.

Ayton, P. and Pascoe, E., 1995. Bias in human judgement under uncertainty?. The Knowledge Engineering Review, 10(1), pp.21-41.

Baber, C., McCormick, E. and Apperly, I., (2020), A framework for Explainable AI, Contemporary Ergonomics 2020

Baber, C., McCormick, E. and Apperly, I., (2021), A human-centered process model for Explainable AI, Naturalistic Decision-making, 2021

Bahner, J.E., Hüper, A.D. and Manzey, D., 2008. Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. International Journal of Human-Computer Studies, 66(9), pp.688-699.

Bainbridge, L., 1983. Ironies of automation. Analysis, design and evaluation of man–machine systems, pp.129-135.

Baker, J., Jones, D. and Burkman, J., 2009. Using visual representations of data to enhance sensemaking in data exploration tasks. Journal of the Association for Information Systems, 10(7), p.2.

Barocas, S. and Selbst, A.D., 2016. Big data's disparate impact. Calif. L. Rev., 104, p.671.

Barrick, M.R., Mount, M.K. and Judge, T.A. (2001), 'Personality and performance at the beginning of the new millennium: what do we know and where do we go next?', International Journal of Selection and Assessment, Vol. 9 Nos 1/2, pp. 9-30.

Baweja, J.A., Fallon, C.K. and Jefferson, B.A., 2023. Opportunities for human factors in machine learning. Frontiers in Artificial Intelligence, 6.

Bertrand, M. and Mullainathan, S., 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. American economic review, 94(4), pp.991-1013.

Bench-Capon, T.J. and Dunne, P.E., 2007. Argumentation in artificial intelligence. Artificial intelligence, 171(10-15), pp.619-641.

Bex, F., Modgil, S., Prakken, H. and Reed, C., 2013. On logical specifications of the argument interchange format. Journal of Logic and Computation, 23(5), pp.951-989.

Bock, D.E., Wolter, J.S. and Ferrell, O.C., 2020. Artificial intelligence: disrupting what we know about services. Journal of Services Marketing, 34(3), pp.317-334.

Bodenhausen, G.V., 1988. Stereotypic biases in social decision-making and memory: Testing process models of stereotype use. Journal of personality and social psychology, 55(5), p.726.

Bogen, M. and Rieke, A., 2018. Help wanted: An examination of hiring algorithms, equity, and bias. Upturn, December, 7.

Borgo, R., Cashmore, M. and Magazzeni, D., 2018. Towards providing explanations for AI planner decisions. arXiv preprint arXiv:1810.06338.

Bradley, S.H., DeVito, N.J., Lloyd, K.E., Richards, G.C., Rombey, T., Wayant, C. and Gill, P.J., 2020. Reducing bias and improving transparency in medical research: a critical overview of the problems, progress and suggested next steps. Journal of the Royal Society of Medicine, 113(11), pp.433-443.

Brooks, A.P., 1997. TAAS unfair to minorities, lawsuit claims. Austin American-Statesman, p.A1.

Brynjolfsson, E. and McAfee, A., 2012. Winning the race with ever-smarter machines. MIT Sloan Management Review, 53(2), p.53.

Buckley, P., Minette, K., Joy, D. and Michaels, J., 2004. The use of an automated employment recruiting and screening system for temporary professional employees: A case study. Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management, 43(2-3), pp.233-241.

Caliskan, A., Bryson, J.J. and Narayanan, A., 2017. Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), pp.183-186.

Canoy, D., 2015. A dictionary of epidemiology—The evolution towards the 6th edition. BBA clinical, 4, p.42.

Cao, G., Duan, Y., Edwards, J.S. and Dwivedi, Y.K., 2021. Understanding managers' attitudes and behavioral intentions towards using artificial intelligence for organizational decision-making. Technovation, 106, p.102312.

Castelo, N., 2019. Blurring the line between human and machine: marketing artificial intelligence. Columbia University

Cherry, K., 2021. What Is the Representativeness Heuristic? https://www.verywellmind.com/representativeness-heuristic-2795805

Chesnevar, C., Modgil, S., Rahwan, I., Reed, C., Simari, G., South, M., Vreeswijk, G. and Willmott, S., 2006. Towards an argument interchange format. The knowledge engineering review, 21(4), pp.293-316.

Chi, O.H., Gursoy, D. and Chi, C.G., 2022. Tourists' attitudes toward the use of artificially intelligent (AI) devices in tourism service delivery: moderating role of service value seeking. Journal of Travel Research, 61(1), pp.170-185.

Chichester Jr, M.A. and Giffen, J.R., 2019. Recruiting in the robot age: examining potential EEO implications in optimizing recruiting through the use of artificial intelligence. Comput. Internet Lawyer, 36(10), pp.1-3.

Choung, H., David, P. and Ross, A., 2022. Trust in AI and Its Role in the Acceptance of AI Technologies. International Journal of Human–Computer Interaction, pp.1-13

Cleary, T.A. and Hilton, T.L., 1968. An investigation of item bias. Educational and Psychological Measurement, 28(1), pp.61-75.

Cramer, H., Garcia-Gathright, J., Springer, A. and Reddy, S., 2018. Assessing and addressing algorithmic bias in practice. Interactions, 25(6), pp.58-63.

Cronan, T.P., Mullins, J.K. and Douglas, D.E., 2018. Further understanding factors that explain freshman business students' academic integrity intention and behavior: Plagiarism and sharing homework. Journal of Business Ethics, 147(1), pp.197-220.

Dalton, S., 2018. Minimizing and addressing implicit bias in the workplace: Be proactive, part one. C&RL News, 79(9), p.478.

Dardenne, B. and Leyens, J.P., 1995. Confirmation bias as a social skill. Personality and Social Psychology Bulletin, 21(11), pp.1229-1239.

Dastin, J., 2018. Amazon scraps secret AI recruiting tool that showed bias against women. San Fransico, CA: Reuters. Retrieved on October, 9, p.2018.

Davenport, T., Guha, A., Grewal, D. and Bressgott, T., 2020. How artificial intelligence will change the future of marketing. Journal of the Academy of Marketing Science, 48(1), pp.24-42.

Dawes, R.M., 2001. Rational choice in an uncertain world: The psychology of judgement and decision-making. Sage Publications.

Del Giudice, M., Scuotto, V., Orlando, B. and Mustilli, M., 2021. Toward the human–Centered approach. A revised model of individual acceptance of AI. Human Resource Management Review, p.100856.

De Bruyn, A., Viswanathan, V., Beh, Y.S., Brock, J.K.U. and Von Wangenheim, F., 2020. Artificial intelligence and marketing: Pitfalls and opportunities. Journal of Interactive Marketing, 51(1), pp.91-105

Dickson, B. (2018) 'Learning How AI Makes Decisions - PCMag UK', PC, November. Available at: https://uk.pcmag.com/news/118591/learning-how-ai-makes-decisions (Accessed: 1 April 2019).

Dietvorst, B.J., Simmons, J.P. and Massey, C., 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), p.114.

Dietz, J. and Hamilton, L.K., 2008. Subtle biases and covert prejudice in the workplace. Ivey Publishing 9B08C005.

Drucker, K., 2016. Avoiding Discrimination and Filtering of Qualified Candidates by ATS Software.

Duffy, M., 1995. Sensemaking in classroom conversations. Openness in research: The tension between self and other, pp.119-132.

Dunnette, M.D. and Borman, W.C., 1979. Personnel selection and classification systems. Annual review of psychology.

Elsbach, K.D. and Stigliani, I., 2019. New information technology and implicit bias. Academy of Management Perspectives, 33(2), pp.185-206.

Franke, U., 2022. First-and Second-Level Bias in Automated Decision-making. Philosophy & Technology, 35(2), pp.1-20.

Frieder, R.E., van Iddekinge, C.H. and Raymark, P.H. (2016) How quickly do interviewers reach decisions? An examination of interviewers' decision-making time across applicants, Journal of Occupational and Organizational Psychology, 89, 223-248.

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. JAMA internal medicine, 178(11), 1544-1547.

Gigerenzer, G., 2008. Bounded and rational. In Philosophie: Grundlagen und Anwendungen/Philosophy: Foundations and Applications (pp. 233-257). Brill mentis.

Goddard, K., Roudsari, A. and Wyatt, J.C., 2014. Automation bias: empirical results assessing influencing factors. International journal of medical informatics, 83(5), pp.368-375.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. Journal of Personality and Social Psychology, 74, 1464–1480.

Gonzalez-Sosa, E., Fierrez, J., Vera-Rodriguez, R. and Alonso-Fernandez, F., 2018. Facial soft biometrics for recognition in the wild: Recent works, annotation, and COTS evaluation. IEEE Transactions on Information Forensics and Security, 13(8), pp.2001-2014.

Gursoy, D., Chi, O.H., Lu, L. and Nunkoo, R., 2019. Consumers acceptance of artificially intelligent (AI) device use in service delivery. International Journal of Information Management, 49, pp.157-169.

Guszcza, J., Lewis, H. and Evans-Greenwood, P., 2017. Cognitive collaboration: Why humans and computers think better together. Deloitte Review, 20, pp.8-29.

Haggarty-Weir, C., 2013. Cognitive Biases and 'Doing Your Own Research', Part 1 – Decision-making, Belief and Behavioural Biases. https://mostlyscience.com/2013/02/doing-your-own-research-cognitive-biases-and-you-part-1-decision-making-belief-and-behavioural-biases/

Haggarty-Weir, C., 2013. Cognitive Biases and 'Doing Your Own Research'. Part 2 – The Social. https://mostlyscience.com/2013/03/doing-your-own-research-cognitive-biases-and-you-part-2-social-biases/

Haselton, M.G., Nettle, D. and Murray, D.R., 2015. The evolution of cognitive bias. The handbook of evolutionary psychology, pp.1-20.

Hastie, R. and Dawes, R.M., 2009. Rational choice in an uncertain world: The psychology of judgment and decision-making. Sage Publications.

Helms, J.E., 2010. Cultural bias in psychological testing. The Corsini encyclopedia of psychology, pp.1-3.

Helms, R.W., Reece, L.H., Helms, R.W. and Helms, M.W., 2011. Defining, evaluating, and removing bias induced by linear imputation in longitudinal clinical trials with MNAR missing data. Journal of biopharmaceutical statistics, 21(2), pp.226-251.

Heuer, R.J., 1999. Psychology of intelligence analysis. Lulu. com.

Hilbert, M., 2012. Toward a synthesis of cognitive biases: how noisy information processing can bias human decision-making. Psychological bulletin, 138(2), p.211.

Hoffman, R. et al. (2018) Explaining Explanation, Part 4: A Deep Dive on Deep Nets, IEEE Intelligent Systems, 33(3), pp. 87–95.

Hoffman, R. R. (Ed.). (2007). Expertise out of context: Proceedings of the sixth international conference on naturalistic decision-making. Psychology Press.

High-Level Expert Group on ArtificialIntelligence (2019) 'High-Level Expert Group on Artificial Intelligence Set Up By the European Commission Ethics Guidelines for Trustworthy Ai', European Commission. Available at: https://ec.europa.eu/digital.

Hoffman, R. et al. (2018) 'Explaining Explanation, Part 4: A Deep Dive on Deep Nets', IEEE Intelligent Systems, 33(3), pp. 87–95. doi: 10.1109/MIS.2018.033001421.Klein, G.A., Calderwood, R. and Macgregor, D., 1989. Critical decision method for eliciting knowledge. IEEE Transactions on systems, man, and cybernetics, 19(3), pp.462-472.

Huang, M.H. and Rust, R.T., 2018. Artificial intelligence in service. Journal of Service Research, 21(2), pp.155-172.

Humphrey, N. (1976). The social function of intellect. In P. P. G. Bateson & R. A. Hinde (Eds.), Growing points in ethology (pp. 303–317). London: Faber & Faber.

Jackson, A.T., 2016. Examining factors influencing use of a decision aid in personnel selection.

Jarrahi, M.H., 2018. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision-making. Business horizons, 61(4), pp.577-586.

Jian, J.Y., Bisantz, A.M. and Drury, C.G., 2000. Foundations for an empirically determined scale of trust in automated systems. International journal of cognitive ergonomics, 4(1), pp.53-71.

Joachims, T., Granka, L., Pan, B., Hembrooke, H. and Gay, G., 2017, August. Accurately interpreting clickthrough data as implicit feedback. In Acm Sigir Forum (Vol. 51, No. 1, pp. 4-11). Acm.

Kaartemo, V. and Helkkula, A., 2018. A systematic review of artificial intelligence and robots in value co-creation: current status and future research avenues. Journal of Creating Value, 4(2), pp.211-228.

Kahneman, D., 2011. Thinking, fast and slow. Macmillan.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012, September). Fairness-aware classifier with prejudice remover regularizer. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 35-50). Springer, Berlin, Heidelberg.

Kim, T.W. and Duhachek, A., 2020. Artificial intelligence and persuasion: A construal-level account. Psychological science, 31(4), pp.363-380.

Kim, J., Giroux, M. and Lee, J.C., 2021. When do you trust AI? The effect of number presentation detail on consumer trust and acceptance of AI recommendations. Psychology & Marketing, 38(7), pp.1140-1155.

Kirschner, PA et al. (eds.), 2003, Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making. Berlin: Springer.

Kizilcec, R.F., 2016, May. How much information? Effects of transparency on trust in an algorithmic interface. In Proceedings of the 2016 CHI conference on human factors in computing systems (pp. 2390-2395).

Klein, G.A., Calderwood, R. and Macgregor, D., 1989. Critical decision method for eliciting knowledge. IEEE Transactions on systems, man, and cybernetics, 19(3), pp.462-472.

Klein, G., Phillips, J.K., Rall, E.L. and Peluso, D.A., 2007. A data-frame theory of sensemaking. In Expertise out of context: Proceedings of the sixth international conference on naturalistic Decision-making (pp. 113-155). New York, NY, USA: Lawrence Erlbaum.

Klein, G., Moon, B. and Hoffman, R.R., (2006 a). Making sense of sensemaking 1: Alternative perspectives. IEEE intelligent systems, (4), pp.70-73.

Klein, G., Moon, B. and Hoffman, R.R., (2006 b). Making sense of sensemaking 2: A macrocognitive model. IEEE Intelligent systems, 21(5), pp.88-92.

Kliegr, T., Bahník, Š. and Fürnkranz, J., 2021. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. Artificial Intelligence, 295, p.103458.

Knotek, S., 2003. Bias in problem solving and the social process of student study teams: A qualitative investigation. The Journal of Special Education, 37(1), pp.2-14.

Kordzadeh, N. and Ghasemaghaei, M., 2022. Algorithmic bias: review, synthesis, and future research directions. European Journal of Information Systems, 31(3), pp.388-409.

Krishnakumar, A., 2019. Assessing the Fairness of AI Recruitment systems.

Kuo, Y.F. and Wu, C.M., 2012. Satisfaction and post-purchase intentions with service recovery of online shopping websites: Perspectives on perceived justice and emotions. International Journal of Information Management, 32(2), pp.127-138.

Langer, M., König, C.J., Sanchez, D.R.P. and Samadi, S., 2020. Highly automated interviews: Applicant reactions and the organizational context. Journal of Managerial Psychology.

Lawrence, J., Snaith, M., Konat, B., Budzynska, K. and Reed, C., 2017. Debating technology for dialogical argument: Sensemaking, engagement, and analytics. ACM Transactions on Internet Technology (TOIT), 17(3), p.24.

Lee, J.D. and Seppelt, B.D., 2009. Human factors in automation design. Springer handbook of automation, pp.417-436.

Levinson, J.D., 2007. Forgotten racial equality: Implicit bias, decisionmaking, and misremembering. Duke LJ, 57, p.345.

Levinson, J.D. and Young, D., 2009. Different shades of bias: Skin tone, implicit racial bias, and judgments of ambiguous evidence. W. Va. L. Rev., 112, p.307.

Lichtenthaler, U., 2019. Extremes of acceptance: employee attitudes toward artificial intelligence. Journal of Business Strategy.

Lodato, M.A., Highhouse, S. and Brooks, M.E., 2011. Predicting professional preferences for intuition-based hiring. Journal of Managerial Psychology.

Longoni, C., Bonezzi, A. and Morewedge, C.K., 2020. Resistance to medical artificial intelligence is an attribute in a compensatory decision process: response to Pezzo and Beckstead (2020). Judgment and Decision-making, 15(3), pp.446-448.

Logg, J.M., Minson, J.A. and Moore, D.A., 2019. Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, pp.90-103.

Lu, L., Cai, R. and Gursoy, D., 2019. Developing and validating a service robot integration willingness scale. International Journal of Hospitality Management, 80, pp.36-51.

Lu, V.N., Wirtz, J., Kunz, W.H., Paluch, S., Gruber, T., Martins, A. and Patterson, P.G., 2020. Service robots, customers and service employees: what can we learn from the academic literature and where are the gaps?. Journal of Service Theory and Practice, 30(3), pp.361-391.

Lundgren, H., Kroon, B. and Poell, R.F., 2017. Personality testing and workplace training: Exploring stakeholders, products and purpose in Western Europe. European Journal of Training and Development.

Madhavan, P. and Wiegmann, D.A., 2007. Similarities and differences between human–human and human–automation trust: an integrative review. Theoretical Issues in Ergonomics Science, 8(4), pp.277-301.

Marsh, S., 2019. Government immigration database ‘deeply sinister’, say campaigners. Home Office project will share people’s status ‘in real time with authorised users’.

Masalonis, A. and Parasuraman, R., 2003. Fuzzy signal detection theory: Analysis of human and machine performance in air traffic control, and analytic considerations. Ergonomics, 46(11), pp.1045-1074.

Maso, I., Atkinson, P., Delamont, S. and Verhoeven, J., 1995. Openness in research: The tension between self and other. Van Gorcum & Comp.

Mathieson, K., 1991. Predicting user intentions: comparing the technology acceptance model with the theory of planned behavior. Information systems research, 2(3), pp.173-191.

McDuff, D., Cheng, R. and Kapoor, A., 2018. Identifying bias in AI using simulation. arXiv preprint arXiv:1810.00471.

Minsky, M. (1986). The society of mind. New York: Simon & Schuster.

Minsky, M., 1975. A framework for representing knowledge', in The Psychology of Computer Vision, ed. P. Winston, New York: McGraw-Hill.

Militello, L., Lipshitz, R. and Schraagen, J.M., 2017. Making Sense of Human Behavior: Explaining How Police Officers Assess Danger During Traffic Stops. In Naturalistic Decision-making and Macrocognition (pp. 147-166). CRC Press.

Misuraca, G., van Noordt, C. and Boukli, A., 2020, September. The use of AI in public services: results from a preliminary mapping across the EU. In Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance (pp. 90-99).

Mitchell, T.L., Haw, R.M., Pfeifer, J.E. and Meissner, C.A., 2005. Racial bias in mock juror decision-making: A meta-analytic review of defendant treatment. Law and human behavior, 29(6), pp.621-637.

Modgil, S. and Prakken, H., 2013. A general account of argumentation with preferences. Artificial Intelligence, 195, pp.361-397.

Morar, N. and Baber, C., 2017, September. Joint human-automation decision-making in road traffic management. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 61, No. 1, pp. 385-389). SAGE Publications.

Morar, N., Baber, C., McCabe, F., Starke, S.D., Skarbovsky, I., Artikis, A. and Correai, I., 2019. Drilling into dashboards: Responding to computer recommendation in fraud analysis. IEEE Transactions on Human-Machine Systems, 49(6), pp.633-641.

Mosier, K.L., 2002. 5. Automation and cognition: maintaining coherence in the electronic cockpit. Advances in human performance and cognitive engineering research.

Mukherjee, R., 2015. Gender bias. International Journal of Humanities & Social Science Studies, 1(6), p.76.

Nagpal, S., Singh, M., Singh, R. and Vatsa, M. (2019) Deep learning for face recognition: Pride or prejudiced?. arXiv preprint arXiv:1904.01219.

Naughton, J. 8 Oct (2022) Tech firms say laws to protect us from bad AI will limit 'innovation'. Well, good. the Guardian.https://www.theguardian.com/commentisfree/2022/oct/08/tech-firms-artificial-intelligence-ai-liability-directive-act-eu-ccia?CMP=Share_AndroidApp_Other

Nickerson, R.S., 1998. Confirmation bias: A ubiquitous phenomenon in many guises. Review of general psychology, 2(2), pp.175-220.

Niemelä, M., Arvola, A. and Aaltonen, I., 2017, March. Monitoring the acceptance of a social service robot in a shopping mall: first results. In Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-robot Interaction (pp. 225-226).

Oguego, C. L., Augusto, J. C., Muñoz, A., & Springett, M. (2018). Using argumentation to manage users' preferences. Future Generation Computer Systems, 81, 235-243.

Olhede, S. and Wolfe, P. (2017) 'When algorithms go wrong, who is liable?', Significance, 14(6), pp. 8–9. doi: 10.1111/j.1740-9713.2017.01085.x.

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy, New York: Crown Publishing Group.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013) Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. Journal of Personality and Social Psychology, 105, 171–92.

Parasuraman, R. and Riley, V., 1997. Humans and automation: Use, misuse, disuse, abuse. Human factors, 39(2), pp.230-253.

Parasuraman, R., 2000. Designing automation for human use: empirical studies and quantitative models. Ergonomics, 43(7), pp.931-951.

Parasuraman, R., Sheridan, T.B. and Wickens, C.D., 2000. A model for types and levels of human interaction with automation. IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans, 30(3), pp.286-297.

Parasuraman, R. and Manzey, D.H., 2010. Complacency and bias in human use of automation: An attentional integration. Human factors, 52(3), pp.381-410.

Pardes, A. (2019). Need some fashion advice? Just ask the algorithm. https://www.wired.com/story/stitch-fix-shop-your-looks/

Pedersen, P., 1987. Ten frequent assumptions of cultural bias in counseling. Elementary school counseling in a changing world, pp.3-11.

Pena, A., Serna, I., Morales, A. and Fierrez, J., 2020. Bias in multimodal AI: Testbed for fair automatic recruitment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 28-29).

Peng, A., Nushi, B., Kıcıman, E., Inkpen, K., Suri, S. and Kamar, E., 2019, October. What you see is what you get? the impact of representation criteria on human bias in hiring. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (Vol. 7, pp. 125-134).

Pezzo, M.V. and Beckstead, J.W., 2020. Patients prefer artificial intelligence to a human provider, provided the AI is better than the human: A commentary on Longoni, Bonezzi and Morewedge (2019). Judgment and decision-making, 15(3), p.443.

PR Newswire, 2018. Montage Announces New Solution to Reduce Unconscious Bias in Hiring Process. PR Newswire Association LLC Sep 12, 2018

Quillian, L., Pager, D., Hexel, O. and Midtbøen, A.H., 2017. Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. Proceedings of the National Academy of Sciences, 114(41), pp.10870-10875.

Rai, A., 2020. Explainable AI: From black box to glass box. Journal of the Academy of Marketing Science, 48(1), pp.137-141.

Ranjan, R., Sankaranarayanan, S., Bansal, A., Bodla, N., Chen, J.C., Patel, V.M., Castillo, C.D. and Chellappa, R., 2018. Deep learning for understanding faces: Machines may be just as good, or better, than humans. IEEE Signal Processing Magazine, 35(1), pp.66-83.

Rasmussen, J., 1987. Information processing and human-machine interaction. An approach to cognitive engineering.

Reed, C., Walton, D. and Macagno, F., 2007. Argument diagramming in logic, law and artificial intelligence. The Knowledge Engineering Review, 22(1), pp.87-109.

Rempel, J.K., Holmes, J.G. and Zanna, M.P., 1985. Trust in close relationships. Journal of personality and social psychology, 49(1), p.95.

Reynolds, C.R. and Brown, R.T., 1984. Bias in mental testing. In Perspectives on bias in mental testing (pp. 1-39). Springer, Boston, MA.

Reynolds, C.R., 2000. Methods for detecting and evaluating cultural bias in neuropsychological tests. In Handbook of cross-cultural neuropsychology (pp. 249-285). Springer, Boston, MA.

Reynolds, C.R. and Suzuki, L.A., 2012. Bias in psychological assessment: An empirical review and recommendations. Handbook of Psychology, Second Edition, 10.

Ricci, F., Rokach, L. and Shapira, B., 2011. Introduction to recommender systems handbook. In Recommender systems handbook (pp. 1-35). Springer, Boston, MA.

Ross, H.J., 2014. Everyday bias. Rowman & Littlefield Publishers.

Ross, H.J., 2020. Everyday bias: Identifying and navigating unconscious judgments in our daily lives.

Russell, S. and Norvig, P., 1995. A modern, agent-oriented approach to introductory artificial intelligence. Acm Sigart Bulletin, 6(2), pp.24-26.

Samad, D., 2012. A decision support system that provides an automated short listing of potential employees in recruitment and hiring of any organization.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S. and Müller, K.R., 2016. Evaluating the visualization of what a deep neural network has learned. IEEE transactions on neural networks and learning systems, 28(11), pp.2660-2673.

Sample, I, 2023, ChatGPT: what can the extraordinary artificial intelligence chatbot do?, the Guardian. https://www.theguardian.com/technology/2023/jan/13/chatgpt-explainer-what-can-artificial-intelligence-chatbot-do-ai.

Schellmann, Hilke., 2022. Finding it hard to get a new job? Robot recruiters might be to blame. The Guardian. https://www.theguardian.com/us-news/2022/may/11/artitifical-intelligence-job-applications-screen-robot-recruiters?CMP=Share_AndroidApp_Other

Schmid, C., 2018. in-Progress: Decision Support Systems for Recruiting College Athletes.

Seegebarth, B., Backhaus, C. and Woisetschläger, D.M., 2019. The role of emotions in shaping purchase intentions for innovations using emerging technologies: A scenario-based investigation in the context of nanotechnology. Psychology & Marketing, 36(9), pp.844-862

Serna, I., Morales, A., Fierrez, J., Cebrian, M., Obradovich, N. and Rahwan, I., 2019. Algorithmic discrimination: Formulation and exploration in deep learning-based face biometrics. arXiv preprint arXiv:1912.01842.

Sheridan, T.B., Sheridan, T.B., Maschinenbauingenieur, K., Sheridan, T.B. and Sheridan, T.B., 2002. Humans and automation: System design and research issues.

Shin, D. and Park, Y.J., 2019. Role of fairness, accountability, and transparency in algorithmic affordance. Computers in Human Behavior, 98, pp.277-284.

Shin, D., Zhong, B. and Biocca, F.A., 2020. Beyond user experience: What constitutes algorithmic experiences?. International Journal of Information Management, 52, p.102061.

Shin, D., 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human-Computer Studies, 146, p.102551.

Singh, A., Rose, C., Visweswariah, K., Chenthamarakshan, V. and Kambhatla, N., 2010, October. PROSPECT: a system for screening candidates for recruitment. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 659-668).

Skitka, L.J., Mosier, K. and Burdick, M.D., 2000. Accountability and automation bias. International Journal of Human-Computer Studies, 52(4), pp.701-717.

Song, S.Y., 2017. Modelling the consumer acceptance of retail service robots.

Stanovich, K.E., 2003. The fundamental computational biases of human cognition: Heuristics that (sometimes) impair decision-making and problem solving.

Starke, S.D. and Baber, C., 2020. The effect of known decision support reliability on outcome quality and visual information foraging in joint decision-making. Applied Ergonomics, 86, p.103102.

Steinbock, B., 1978. Speciesism and the Idea of Equality. Philosophy, 53(204), pp.247-256.

Stock, R.M. and Merkle, M., 2017, March. A service Robot Acceptance Model: User acceptance of humanoid robots during service encounters. In 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops) (pp. 339-344). IEEE.

Sun, T.Q. and Medaglia, R., 2019. Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare.

Talley, S.M., 2020. Public Acceptance of AI Technology in Self-Flying Aircraft. Journal of Aviation/Aerospace Education & Research, 29(1), pp.49-64.

Taylor, D.M. and Doria, J.R., 1981. Self-serving and group-serving bias in attribution. The Journal of Social Psychology, 113(2), pp.201-211.

Thomas, J. B., Clark, S. M., & Gioia, D. A. (1993). Strategic sensemaking and organizational performance: Linkages among scanning, interpretation, action, and outcomes. Academy of Management journal, 36(2), 239-270.

Thordsen, M.L., 1991, September. A comparison of two tools for cognitive task analysis: Concept mapping and the critical decision method. In Proceedings of the Human Factors Society Annual Meeting (Vol. 35, No. 5, pp. 283-285). Sage CA: Los Angeles, CA: SAGE Publications.

Tintarev, N. and Masthoff, J., 2012. Evaluating the effectiveness of explanations for recommender systems. User Modeling and User-Adapted Interaction, 22(4), pp.399-439.

Tito, J., 2017. Destination unknown: Exploring the impact of artificial intelligence on government. Centre for Public Impact, pp.7-8.

Tommasi, T., Patricia, N., Caputo, B. and Tuytelaars, T., 2017. A deeper look at dataset bias. In Domain adaptation in computer vision applications (pp. 37-55). Springer, Cham.

Torralba, A. and Efros, A.A., 2011, June. Unbiased look at dataset bias. In CVPR 2011 (pp. 1521-1528). IEEE.

Torrington, D. and Hall, S., 1998. Human resource management and the personnel function. New perspectives on human resource management, pp.56-67.

Toulmin, S.E., 1958. The use of argument. Cambridge University Press.

Toulmin, S.E., 2003. The uses of argument. Cambridge university press.

Tripepi, G., Jager, K.J., Dekker, F.W. and Zoccali, C., 2010. Selection bias and information bias in clinical research. Nephron Clinical Practice, 115(2), pp.c94-c99.

Tversky, A. and Kahneman, D., 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. science, 185(4157), pp.1124-1131.

Van Noorden, R., 2020. The ethical questions that haunt facial-recognition research. Nature, 587(7834), pp.354-359.

Verheij, B. (2009) 'The Toulmin Argument Model in Artificial Intelligence Or : how semi-formal , defeasible argumentation schemes creep into logic', pp. 219–238. doi: 10.1007/978-0-387-98197-0.

Walsh, B. and Volini, E., 2017. Rewriting the rules for the digital age: 2017 Deloitte global human capital trends.

Wang, D. et al. (2019) 'Designing Theory-Driven User-Centric Explainable AI'. doi: 10.1145/3290605.3300831.

Weber, C., 2019. Engineering Bias in AI. IEEE pulse, 10(1), pp.15-17.

Wiener, E.L. and Curry, R.E., 1980. Flight-deck automation: Promises and problems. Ergonomics, 23(10), pp.995-1011.

Wickens, C.D., Gempler, K. and Morphew, M.E., 2000. Workload and reliability of predictor displays in aircraft traffic avoidance. Transportation Human Factors, 2(2), pp.99-126.

Woodson, T. (2018) 'Weapons of math destruction', Journal of Responsible Innovation. Routledge, 5(3), pp. 361–363. doi: 10.1080/23299460.2018.1495027.

Young, A.D. and Monroe, A.E., 2019. Autonomous morals: Inferences of mind predict acceptance of AI behavior in sac

Young, A.T., Amara, D., Bhattacharya, A. and Wei, M.L., 2021. Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. The Lancet Digital Health, 3(9), pp.e599-e611

Zimmerman, L.A., 2008. Making sense of human behavior: Explaining how police officers assess danger during traffic stops. Naturalistic decision-making and macrocognition, pp.121-140.

Appendix for the thesis

https://docs.google.com/document/d/1NyStk_7p1TFv9faIemKIlhaZnJbpssku/edit?usp=sharing&ouid=104371181972297877112&rtpof=true&sd=true