LUND UNIVERSITY
School of Economics and Management

Part of You is in Your Bot

The Use of Debiasing Strategies to Prevent Cognitive Biases from Impacting Conversational Agents

by

Helene Hoffmann

Rica-Salome Strotmann

June 2020

Master's Programme in International Strategic Management

Supervisor: Pauline Mattsson

1. Abstract

Purpose: With the increasing importance of unbiased conversational agents, this research investigates how debiasing strategies can be used to prevent cognitive biases from impacting conversation agents in the development process. A research gap was identified in the literature of prevention strategies for the management of cognitive biases. Therefore, existing literature on debiasing is explored and primary data is generated through interviews with experts in the field of conversational agents to gain insights for the management of cognitive biases on prevention strategies.

Methodology: A qualitative study is conducted following a systematic approach to provide objective findings. Further, literature is screened and coherently compiled in common underlying concepts.

Findings: Debiasing is a broad research area in which authors adopt different management approaches, either by basing strategies on specific cognitive biases or on specific causes of cognitive biases. The current literature differentiates between debiasing and prevention differently than organisations and does not focus on the application field of conversational agents. Thus, literature presents different strategies for cognitive bias management than practice. In practice, the concept of debiasing and prevention is only loosely and not uniformly defined and differentiates between debiasing and prevention on the basis of the development process. Cognitive bias management is not directed at an individual but is applied at an organisational level, wherefore cognitive bias management strategies can be both debiasing and prevention. Consequently, the literature gap is not reflected in practice.

Contribution: The research found that the theoretical gap does not pose a practical issue. Additionally, this research contributes to the current literature by providing findings from a new field of application in which strategies are presented that are potentially new. These provide proposals for future research. The literature gap of prevention, however, remains theoretical. Practical implications are further identified where education about cognitive biases should be offered to all hierarchical levels, or individuals in superordinate positions should guide individuals in operational positions on the management of cognitive bias.

Keywords: cognitive bias, rationality, debiasing, prevention, conversational agent, artificial intelligence, biased AI

2. Acknowledgements

We would like to take this opportunity to thank all those who have supported us during the master's studies as well as the process of this research.

First and foremost, we would like to thank our supervisor Pauline Mattsson, who supported us with her expertise and provided us with feedback at all times. Thanks to her guidance, constructive criticism, and empathetic understanding, this master's thesis advanced significantly.

We also want to express our gratitude to all interview participants without whom this thesis could not have been created. Thank you for showing interest in this research and for contributing time and expertise. We would also like to thank the professors of the psychological faculties who gave us insight into an entirely new research stream and provided us with their expertise to gain deeper insight into the vast field of psychology.

A special thanks is owed to our families and friends, who consistently and patiently supported us. Thank you for always listening and reassuring us on our way. Finally, we would like to express gratefulness to our fellow students who accompanied us on this trip and always showed sympathy and understanding.

Helene and Rica,

01.06.2020

Table of Contents

List of Tables

List of Figures

1 Introduction

Artificial intelligence (AI) is used for decision-making by businesses and governments (Ntoutsi, Fafalios, Gadiraju, Iosifidis, Nejdl, Vidal, Ruggieri, Turini, Papadopoulos, Krasanakis, Kompatsiaris, Kinder-Kurlanda, Wagner, Karimi, Fernandez, Alani, Berendt, Kruegel, Heinze, Broelemann, Kasneci, Tiropanis & Staab, 2019) and plays an increasingly important role in society because it increases productivity and economic growth (European Commission, 2020). As AIs are increasingly utilised (Pannu, 2015), the occurrence of biased AIs increases as well. AIs can become biased by decisions made by individuals during the development process, such as in the selection of data to be inserted into the algorithm (Miller & Brown, 2018). These decisions can be influenced by individuals' own cognitive biases (CBs), which are systematic deviations from rationality (Haselton, Nettle & Andrews, 2005). Organisations already recognise the potential risk of biased AIs and acknowledge that AI brings new or increasing challenges in areas such as ethics, technology, and compliance (United States Securities and Exchange Commission, 2018a). However, only unbiased AIs will last in the long run, as some market players are intensively working on mitigation methods (IBM Research, 2020). This introduces a market for bias-free AIs driven by the industry, while laws, regulations, and policies remain unmodified as they do not keep up with the speed of technological changes (Osoba & Welser, 2017). The lack of regulatory involvement leaves organisations without common compulsory standards regarding design, development, deployment, implementation, and usage (Madiega, 2019). This lack of guidance increases the risk of introducing biased AIs to the market (Osoba & Welser, 2017).

CBs can have a positive impact on individuals by providing quick orientation in uncertain situations, enabling rapid short-term solutions, providing sufficiently robust categories that allow assessments, and motivating individuals to solve problems that might otherwise be abandoned prematurely (Tobena, Marks & Dar, 1999). However, CBs can also have a negative effect as they can lead to erroneous reasoning and poor decision-making (Dawson & Arkes, 1987). Therefore, managing CBs becomes relevant for management, as decision-making is one of its focal activities (Harrison, 1996).

CB management becomes particularly relevant when working with technologies like AI, as they can pick up CBs from individuals (Ntoutsi et al. 2019), which are then reproduced or even amplified in AIs (Karimi, Génois, Wagner, Singer & Strohmaier, 2018). The controlled transfer of biases into AI, also called system biases, is essential to bring the system closer to human behaviour (Sutton, 1992). However, an uncontrolled transfer of CBs followed by the amplification of CBs by the AI can lead to new types of biases and previously positive biases can result in negative effects (Ntoutsi et al. 2019). This can lead to a misbehaving AI whose outcome is erroneous, inequitable, and has harmful side-effects (Osoba & Welser, 2017). History has already shown that the risk of CBs can become reality, in which they lead to social prejudice and mass injustice, such as discrimination based on gender, ethnicity, or religion (Sen & Ganguly, 2005). AI, through its ability to pick up and amplify CBs, has the potential to exacerbate these effects.

Biased AI can have a long-term impact on organisations as it not only reduces the public acceptance of the technology but can also lead to brand or reputational harm (United States Securities and Exchange Commission, 2018b). This can lead to a financial loss, as customers switch to competitors, and advertisers or other business partners withdraw cooperations (United States Securities and Exchange Commission, 2018a). Failures often occur, such as Microsoft's Xiao Bing chatbot which gave racially insensitive responses (Wei, Yu & Fong, 2018). Other examples are AIs speech recognition from Apple, Microsoft, IBM, Amazon, and Google that showed race and gender biases (Tatman, 2017). It was found that these AIs have twice as many errors in transcribing African American voices compared to white American voices (Koenecke, Nam, Lake, Nudell, Quartey, Mengesha, Toups, Rickford, Jurafsky & Goel, 2020).

Conversational agents (CAs) reveal CBs in an AI as they are the interfaces that enable direct communication between the system and end-user (Osoba & Welser, 2017). CAs are designed to conduct nearly natural conversations with the end-user (Jurafsky & Martin, 2008). Here, biases may be displayed through answers of the CA. It is therefore of great importance for an organisation to manage CBs to avoid negative effects of biased CAs. Consequently, organisations need to understand how CBs occur, how they can be managed, and possibly be prevented.

1.1 Problem Statement

CBs pose a risk to CAs as they can distort rational decision-making (Arnott, 2006). In the development of a CA multiple decisions are taken of which any poses a risk that CBs may enter into the CA. CBs can be managed either before they distort the decision-making process through prevention measures, or after they have led to distortions, in the form of debiasing strategies (Arkes, 1991; Fischhoff, 1982). Based on the definition of debiasing, such strategies aim to correct decisions that have already been identified to be biased (Fischhoff, 1982) and only intervene after an initial decision was erroneous. Debiasing is thereby more than just the mere improvement of a decision-making process, as it provides new thinking strategies (Soll, Milkman & Payne, 2015). In debiasing, the initial erroneous decision is an indication to implement strategies to manage further errors, which implies a reactive approach. Prevention similarly addresses CBs, partly with similar strategies, however, strategies are implemented proactively, without the need for an existing error. Instead, the implementation of the strategies is ensured before a decision takes place. Therefore, this management approach aims to prevent an erroneous decision to be made.

Studies have extensively identified various debiasing strategies, dealing with the management of CBs (Croskerry, Singhal & Mamede, 2013; Larrick, 2004; Arkes, 1991), and with the concept of CBs in general (Tversky & Kahneman, 1974; Fischhoff, 1982; Schwenk, 1984). However, significantly fewer studies have investigated the prevention of CBs. Also, only a few of these are generalisable, such as the concept of forewarning, in which individuals are warned that CBs may occur, which may alter their perception (Bloom & Tesser, 1971). Similarly, checklists were developed to identify and prevent possible CBs before they influence decisions (Campbell-Yeo, Ranger, Johnston & Fergusson, 2009; Pannucci & Wilkins, 2010; US Government, 2009). However, other studies on preventing CBs are either very domain-specific, such as preventing

selection and allocation biases in randomised trials (Giraudeau & Ravaud, 2009), or focus only on statistical solutions, such as how to prevent CBs by ensuring that individuals with extreme values are not entirely excluded from a sample population (Gustavson, Røysamb & Borren, 2019). Therefore, managers remain without clear guidelines on how to manage CBs to avert their CAs from being distorted. Hence, it is justifiable to state the prevention of CBs has not been sufficiently addressed and its success remains unidentified.

1.2 Purpose Statement

As, on the one hand, well-founded studies on the approaches to debiasing already exist and, on the other hand, very few studies on the prevention of CBs are available, the research draws from the well-founded field of literature for the yet rather unexplored field. Consequently, the research gap is seen as an incentive to investigate whether it is possible to use debiasing strategies to prevent CBs from entering CAs. Due to the research gap, factors for the use of debiasing strategies to prevent CBs derive from empirical data. Overall, the thesis aims to investigate which strategies literature provides for debiasing and how these strategies can be used to prevent CBs from impacting CAs, by hindering them from influencing the decision-making.

Since the research area of prevention is very limited, it is further explored by examining primary data. In addition to literature, interviews with organisational experts in the field of CAs and CBs are carried out to explore further insights into the phenomenon of CB management. To limit the research scope to being feasible, the study concentrates on one application area, namely CB management for CAs.

Contributing research to this gap leads to an enrichment of the options for managing CBs impacting CAs. The research is particularly valuable for organisations that are still in the early stages of development or are constantly evolving their CA. This enables these organisations to actively choose CB prevention as a CB management approach, to which this research contributes by identifying how debiasing strategies might be useful as a method of prevention.

1.3 Research Question

Hence, the main research question, this thesis aims to answer, is:

“How can organisations prevent cognitive biases from impacting conversational agents with the help of debiasing strategies?”.

Answering this research question requires an understanding of commonly transferred CBs from an individual to CAs, causes for CBs to occur in the development process and strategies that organisations use to prevent as well as debias individuals’ CBs. Lastly, the factors determining the preventative use of debiasing strategies is of high importance.

1.4 Delimitations

This research in the area of CB management is limited by several factors as described in the following.

The study focuses only on the application area of CAs. This allows the analysis to be adapted specifically to CBs that occur in this application and therefore, can lead to suitable and relevant results. To obtain well-founded and in-depth findings, this research only addresses the influence of individuals' CBs on CAs. To emphasise the managerial relevance of CB management, a deeper understanding of the human influence on a CA is examined. It is acknowledged that also other elements can cause biases in a CA, such as data that serve as a basis for the AI (Lloyd, 2018). Additionally, external influences, such as the in-group behaviour of the development team, the organisational culture, or their internal politics, can also incorporate CBs into the CAs (Walter, Kellermanns & Lechner, 2012). However, this study only considers individuals and their CBs.

In the course of this, only qualitative management strategies are considered to keep the focus on methods that solely involve human activities. For the thesis, theoretical and practical information is examined, whereby the practical component is grounded on the information that was obtained in interviews. These do not represent an industry average, as they only provide limited insights into the CA and the CB sectors. Further, this research is limited in its time frame, thereby, it only provides a snapshot of the dynamic technology sector and prevents the achievement of theoretical saturation of primary data.

1.5 Research Outline

This study consists of six chapters. Chapter 1 introduced the topic and presented the research questions on which Chapter 2 builds, by reviewing the literature in two sections: CBs, their causes and origin, and debiasing, entailing the process description and cross-section of strategies. Following, Chapter 3 outlines the methodology and reflects the research design decisions made. This is followed by a presentation of findings in Chapter 4 and a discussion of these in Chapter 5. Finally, a conclusion is outlined, illustrating the research implications and limitations.

2 Literature Review

The review aims to provide profound insights into the literature on which the research is based, allowing an understanding of the nature and scope of the research topic. In the following, a short overview of the technological background of CAs and the behavioural decision research is given, under which this research falls. This is succeeded by a definition of CBs, their origin, and their causes. Afterwards, the process of debiasing is described, followed by a review of debiasing strategies for the management of CBs, in their respective subdivision of CB- and cause-specific strategies. Finally, a critical reflection on the literature is stated. Overall, the literature review allows a deeper understanding of the concept of CB, its possible influence on CAs, and its management, which provides a basis for the following research analysis.

2.1 Technological Background of Conversational Agents

Understanding the concept of a CA, its technological foundations and development are fundamental, as latter entails numerous decisions which can be distorted by CBs. Furthermore, acquiring an insight into the technological concept improves the understanding of the empirical data.

CAs are based on the technology of AI. There is no universal definition of AI, most definitions, however, deal with the concept in which AI simulates human intelligence, which allows it to perform a task on this level (Ahmet, 2018). CAs are assigned to the category of intelligent agents, which process information and interact in a changing environment that influences their decisions (Vila, 2005). They are defined as intelligent since they can act autonomously (Wooldridge & Jennings, 1995), meaning without direct human intervention, to achieve a given objective (Ademu & Imafidon, 2012). CAs are dialog interfaces that understand human language and respond in natural language in the form of text, speech, or any other form of interaction (Conversational Agent, 2019; Rubin, Chen & Thorimbert, 2010), enabling direct conversations between users and AI (Jurafsky & Martin, 2008; Rubin, Chen & Thorimbert, 2010). Examples of CAs are chatbots or virtual assistants (Rubin, Chen & Thorimbert, 2010).

The development process of a software, like CAs, can be exemplified by a Software Development Life Cycle (SDLCs) of which literature provides a variety of (Ruparelia, 2010). The waterfall model is presented as an example of such because it is the fundament for other SDLCs that followed (Ruparelia, 2010). It was first developed by Benington (1956) and further enhanced by Royce (1970), who's model consists of eight stages. They are subdivided into system and software requirements, design, and analysis, design, development, testing, and operation. Each stage is

connected through an iterative feedback loop. This represents one possible development process. The exact execution of the CA development depends on the organisation and project.

2.2 Behavioural Decision Research

Behavioural decision research combines literature from economics and psychology (Moore & Flynn, 2008) and traditionally focuses on judgemental heuristics and CBs (Maule & Hodgkinson, 2002). It presents three major models of decision behaviour: normative, descriptive, and prescriptive models (Bell, Raiffa & Tversky, 1988). The authors define normative models as those that address how decisions ought to be made, which focuses on rationality. Meanwhile, descriptive models are characterised by their focus on how decisions are made. The prescriptive model adapts aspects from both perspectives by explaining how decisions can be made best given existing limitations. It adapts the normative logics and combines it with descriptive findings.

This research acknowledges the concepts of descriptive decision behaviour and is aimed at investigating which strategies can be adapted to achieve a prescriptive decision behaviour while accepting that decision-making, as described in the normative stream, is not realistic and unattainable. In many cases, the decision-making process is irrational and decision-makers have incomplete and imperfect information (Simon, 1977).

Overall, the descriptive research stream explains why CBs can distort a decision. To understand descriptive decision behaviour, Simon (1955) stressed the importance to understand the perceptual, cognitive, and learning factors that make humans deviate from the homo economicus rationale, which is based on self-interest and utility maximisation Simon (1955). Simon has dominated the descriptive research stream by introducing the concepts of bounded rationality and satisficing. Bounded rationality recognises that although individuals try to make rational decisions, they often lack information, relevant problem criteria, and the mental capacity to do so (Simon, 1957). Further, Simon (1956) describes satisfying which is a human decision behaviour in which a choice is made for a satisfactory option with respect to the information available and goals prevailing rather than the option maximising utility.

2.3 Cognitive Biases

“Cognitive biases are cognitions or mental behaviours that prejudice decision quality in a significant number of decisions for a significant number of people; they are inherent in human reasoning” (Arnott, 2006, p. 59) and as one main concept of this research, an in-depth understanding is fundamental. CBs are viewed as a deviation in judgement from complete rationality (Caverni, Fabre & Gonzalez, 1990; Haselton, Nettle & Andrews, 2005; Tversky & Kahneman, 1974; Kahneman, 2011; Kahneman, Slovic & Tversky, 1982). CBs describe the human tendency to make systematic errors based on cognitive factors instead of evidence (Tversky &

Kahneman, 1974). Different scholars and research institutes have identified a varying amount of CBs, for example, Arnott (2006) presents a literature review of 37 CBs. The RECOBIA project (REduction of COgnitive BIAses in Intelligence Analysis), presents "the worldwide first methodology to classify cognitive biases" (CORDIS, 2015) and identifies 288 CBs. This information, however, remains undisclosed from the public. Overall, there is no uniform categorisation or an all-encompassing list of CBs. CBs that are relevant for this research are presented and further explained in section 2.4.2.

CBs pose a risk to distort rationality (Arnott, 2006). In a CA development process, a multitude of decisions must be taken, which heavily rely on the rationality of decision-makers. The importance of rationality is stressed by the possible amplification of CBs by the technology (Karimi et al. 2018). Decision-making is one of the most important activities carried out by managers at all levels, as it describes the behaviour of managers best and distinguishes them from other professions (Harrison, 1996). Managers can be supported in making rational decisions as the debiasing strategies presented in section 2.4.2 elaborate on.

2.3.1 Origin of Cognitive Biases

The dual system explains how CBs originate in human cognition (Croskerry, 2013). The model categorises human cognition into two distinct systems (Morewedge & Kahneman, 2010). System 1, the fast, unconscious, and intuitive system, and system 2, the slow, conscious, and effortful system (Kahneman, 2003). While system 1 includes instinctive behaviours, which are naturally programmed and crucial for everyday decisions, system 2 is unique to humans and applied for more complex decisions when a reflective and rational approach is required (Evans, 2003; Kahneman, 2003). In the latter, abstract hypothetical thinking originates from, which cannot be achieved by system 1 (Evans, 2003). This is particularly important for decisions that cannot be based on experience. Kahneman and Frederick (2002) studied both systems in terms of their decision-making abilities. They observed that system 1 is based on heuristics and associations. The responses generated in system 1 are then forwarded to system 2, which acts as a control system. However, if system 2 is not trained, it can only approve the results from system 1. Given that system 2 has sufficient capacity and motivation, it can lead to a contradictory result compared to the first intuitive result of system 1, which then is overwritten by a more normative one. This concept is also adopted by several debiasing strategies, explained in section 2.4.2, which pit the systems against each other.

2.3.2 Causes of Cognitive Biases

The causes presented in this section provide insights into reasons for CB occurrence. Understanding these improve the awareness of the complexity of CBs and consequently, identify factors that debiasing strategies must acknowledge.

Haselton, Nettle, and Andrews (2005) categorise the reason for CB occurrence in three areas: 1) heuristics, 2) error management biases, and 3) artifacts.

Since time and the ability to gather as well as process information are limited (Miller, 1956; Lichtenstein & Slovic, 1971; Arrow, 1986; Nordstrom, Williams & LeBreton, 1996), decision-makers developed modes of reasoning (Maule & Hodgkinson, 2002), by using simplifying decision strategies or cognitive shortcuts named heuristics (Newell & Simon, 1972; Tversky & Kahneman, 1974; Hogarth, 1987). Haselton, Nettle, and Andrews (2005) identify these to likely be the most common explanation for CBs, a view shared by Caputo (2013). Heuristics arise when a decision has to be made quickly, the probabilities of the decision are unknown, there are multiple goals, or insufficiently defined problems (Gigerenzer, 2008). However, heuristics can also produce correct or partially correct judgements and it may be inevitable that people adopt some of them (Bazerman & Moore, 2009). Nevertheless, it is important to emphasise that heuristics generally aim at satisfying rather than optimising solutions (Gigerenzer, 2008).

Error management biases refer to the CBs that arise if biased solutions result in lower error costs than unbiased solutions (Haselton, Nettle & Andrews, 2005). The authors base the concept on the assumption that judgement is erroneous at times and results in two error types; false-positive and false-negative. False-positive means that the presence of a condition is reported when it is not given and conversely, false-negative communicates an error when in reality none is present. The consequences of these errors differ, hence, they are associated with different costs. Error management theory suggests that optimal decisions minimise the net effect of error instead of the actual error rate.

Artifacts describe the apparent CBs caused by applying normative standards to humans or placing humans in unnatural settings which their mind is not made for (Haselton, Nettle & Andrews, 2005).

2.4 Debiasing

The following section explains the concept of debiasing and the process an individual has to undergo to achieve debiasing. An understanding of this is fundamental to comprehend how to achieve a successful correction of CBs. Afterwards, debiasing strategies are presented. Here the literature distinguishes between two approaches, CB- and cause-specific strategies, which both introduce detailed measures to manage CBs.

Literature does not provide a uniform definition of debiasing, but rather a variety of different interpretations. Debiasing describes a strategy or a set of strategies to reduce or, in the best case, eliminate the decision-makers' CBs (Arnott, 2006) thereby, enabling more rational decisions. Debiasing strategies can be applied to improve decisions, consequently, ensure a high-quality output (Milkman, Chugh & Bazerman, 2009). Arkes (1991) conceives debiasing strategies as countermeasures to errors of decision-making implying that an error has occurred before debiasing strategies are applicable, which is agreed on by Fischhoff (1982). Thus, debiasing assumes that an error already exists that requires correction (Soll, Milkman & Payne, 2015). Therefore, debiasing is understood as a strategy to limit or eliminate CBs' impact on decisions, that can be applied only after a CB has already distorted the decision-making.

CBs may not only result from decision-makers' inaccurate judgement but may also be caused by a systematically biased decision-making process (Soll, Milkman & Payne, 2015), which requires a distinction between debiasing and the mere improvement of the decision-making process. Whereas improving the process is solely about finding new information on which a decision can be built and thus be improved, debiasing is additionally about reformulating existing information for new thinking strategies (Soll, Milkman & Payne, 2015). This gives debiasing the ability to turn decision-makers into better thinkers (Croskerry, Singhal & Mamede, 2013).

2.4.1 Debiasing Process

The application of debiasing strategies requires several actions to be taken by individuals (Wilson & Brekke, 1994). To contextualise the debiasing strategies presented in the next chapter, the debiasing process is explained first. Wilson and Brekke (1994), introduced the term mental correction when referring to the process of correcting mental contamination, which is "an unwanted judgement, emotion, or behaviour because of mental processing that is unconscious or uncontrollable" (Wilson & Brekke, 1994, p. 117), that has been caused by a CB. In contrast to the authors' definition and acknowledgement that contamination can happen unconsciously, the illustration presented in the paper shows that contamination only begins with the awareness of the individual. Thus, the illustration must be adjusted to be aligned with the authors' understanding that the awareness of a CB is not necessary for one to exist, which is also consistent with findings of other scholars (Bazerman & Moore, 2009; Greenwald & Krieger, 2006; Kang, Bennett, Carbado, Casey & Levinson, 2012).

The authors present four required steps for the correction process after a CB has been triggered (see figure 1).

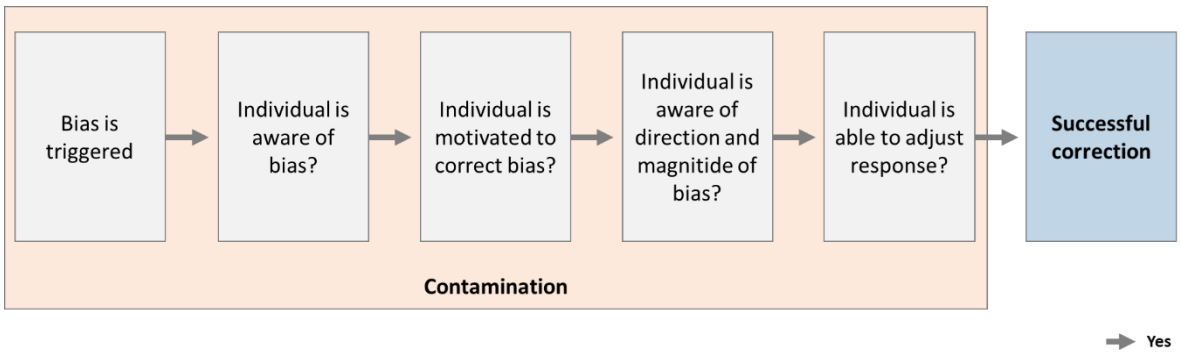


Figure 1: Process of Mental Contamination and Mental Correction (own illustration based on Wilson & Brekke, 1994)

The correction starts with the awareness of unwanted contamination. Therefore, individuals must become aware of the CB either through a direct introspective approach, for instance by self-reflection, or by assumption. An individual may assume that a CB applies without realising it. This may be the case, for example, if individuals are theoretically aware of a CB and solely suspect that CBs are inherent without actively noticing them. As a second step, the authors determine that

individuals must be willing to correct the error. Bazermann and Moore (2009) stress that individuals often find it difficult to accept their errors. Therefore, they consider the second step to be particularly important. Thirdly, individuals must be conscious of the direction and magnitude of the CB (Wilson & Brekke, 1994). Only through awareness of how the CB manifests itself, a strategy can be found to correct this pattern of behaviour. As the last step, individuals must have control over their behaviour to change it. For a strategy to allow a change, individuals must be able to adjust their behaviour permanently.

Wilson and Brekke (1994) generally apply the process to a single individual, however, they also argue that certain steps can be supported by a third person. In this context, the authors mention several studies with a variety of contradictory effects. However, it is possible to identify a link between these studies as to how far a third person can have a positive effect on the debiasing process of another individual. Wilson and Brekke (1994) explain that by forewarning an individual of CBs, a third party can facilitate several steps that an individual must take. Importantly, attention should not only be drawn to the existence of a CB, but above all, to the direction and magnitude of the CB, as otherwise, forewarning can have a negative impact (Wilson & Brekke, 1994). To ensure the effectiveness of the process, the authors note that the third person should accompany all steps of the process. The third person, however, cannot have a positive influence on the individual's motivation, as this must come from the individual directly.

2.4.2 Debiasing Strategies

The following section introduces CB management by presenting debiasing strategies from the literature. These strategies have largely been discovered through experiments, which are conducted in controlled environments and often do not correspond with real-life conditions. Research on debiasing strategies stems from the psychological research field and is predominantly tested in medical and juridical environments. This may be explained by the strong impact decisions in medicine and legislation have. Strategies in these areas partly relate to a more realistic environment. As the environment plays an important role in rationality (Gigerenzer, Todd & the ABC Research Group, 1999), the results of the practical application of the strategies might differ from the results presented in the literature. Despite their specification, these strategies can be brought into a management context. Strategies from these areas are only selected if proven to be transferable, which is considered as such if they are not restricted to their environment, for example in the case of checklists (Croskerry, Singhal & Mamede, 2013). Hence, these strategies can be extracted from their environment and applied to the management field.

Generally, it is argued in the literature that CBs can be managed from two different starting points. On one hand, several authors argue that it is necessary to initially identify a particular CB, which correction is then approached with a strategy (Arkes, Christensen, Lai & Blumer, 1987; Marks & Miller, 1987). Some authors, on the other hand, claim that the cause of a CB must first be identified before a strategy can be applied (Berger, 2005; Fischhoff, 1982; Keren, 1990). In both cases, it is pointed out that a strategy is only effective for the predetermined CB or cause.

Another approach mentioned in literature as a strategy to debiasing is critical thinking (Correia, 2018; Beaulac & Kenyon, 2014; Maynes, 2015), in which no pre-analysis of a CB nor a cause is needed. Here, individuals mitigate CBs on their own, by gradually integrating the analytical reasoning of system 2 into the intuitive reasoning of system 1 (Correia, 2018). However, its

effectiveness is widely questioned (Correia, 2018). It is argued that as a standalone strategy this approach is not realistic (Maynes, 2015), as it would require individuals to have the ability to undergo all steps of the debiasing process, as explained by Wilson and Brekke (1994), on their own. By the nature of humans (Simon, 1955), this appears to be unrealistic. As stated by Maynes (2015), in the worst case, critical thinking can lead to the occurrence of additional CBs. The effectiveness and applicability of critical thinking are widely questioned and is not compatible with the acknowledgement of the descriptive decision-making behaviour that this research adopts. Therefore, this strategy is not taken into further consideration.

This research therefore only considers literature in which strategies have been developed that are applicable to a particular CB or a particular cause of CBs. Since these strategies assume a different starting point from which strategies were created, the two approaches and their respective strategies are presented separately. This procedure is consistent with the approach of the authors of the strategies, who clearly state that the strategies only work with regards to the intended starting point since a strategy for one CB might not be effective for another (Krueger & Clement, 1994). The same applies to cause-specific strategies (Arkes, 1991; Keren, 1990; Larrick, 2004). Thus, the strategies, although they partly appear to be similar, cannot be merged without taking away their effectiveness. Overall, 110 strategies are investigated (see appendix A), on the basis of which a cross-section of the literature has been compiled for each approach, providing an overview of the wide variety of strategies.

Cognitive Bias-Specific Debiasing Strategies

Based on the approach of the authors, who argue that a CB must first be identified for the application of the strategies, the same approach is taken in this section. Therefore, relevant CBs are presented first, followed by debiasing strategies that can be applied to these CBs.

Literature addressing specific CBs relevant to CAs is not available, wherefore the scope is expanded to the complete field of AI. Still, only a few CBs that commonly impact AIs, have been identified in the literature, namely blind spot bias (Osoba & Welser, 2017; Pronin, Lin & Ross, 2002), confirmation bias (Osoba & Welser, 2017), selection bias (Challen, Denny, Pitt, Gompels, Edwards & Tsaneva-Atanasova, 2019; Ayoub & Payne, 2016), and anchoring bias (Challen et al. 2019). To enrich this overview, further CBs from the excessive pool presented in the literature are adopted, which have been assessed to either be relevant in organisational contexts or especially harmful in the context of CAs. In addition, CBs are adopted that affect individuals' intellectual ability to recognise their own susceptibility to CBs and their motivation to manage them, as this might cause damage as CBs might remain undetected. CBs deemed as additionally relevant are: sunk cost fallacy, loss aversion, group attribution error, bandwagon effect, overconfidence, self-serving, false consensus, framing, status quo, and habit bias. A definition of all relevant CBs is presented in table 1.

Table 1: Cognitive Biases Relevant for Conversational Agents

CB	**Definition**
Blind spot bias	Individuals' tendency to assume others to be more susceptible to CBs than themselves (Pronin, Lin & Ross, 2002). This extends to their ability to identify CBs in others but not themselves (Osoba & Welser, 2017; Pronin, Lin & Ross, 2002).
Confirmation bias	The tendency to seek confirmatory and neglect confuting evidence to their existing beliefs (Russo & Schoemaker, 2018).
Selection bias	The distorted selection or exclusion of data samples that do not represent the population and proper randomisation cannot be achieved (Bareinboim & Pearl, 2012).
Anchoring bias	Individuals become disproportionately influenced by an initially presented information which is seen as anker (Tversky & Kahneman, 1974).
Sunk cost fallacy	The tendency to commit resources to a cause because of an initial investment which continues even when that investment has failed to lead to the desired outcomes (Arkes & Blumer, 1985).
Loss aversion	The tendency to weigh potential losses greater than potential gain which results in the avoidance of loss (Kahneman & Tversky, 1979; Kahneman & Tversky, 1984).
Group attribution error	The tendency to attribute a group member's characteristics and preferences to the overall group (Allison & Messick, 1985).
Bandwagon effect	An increasing demand of a community caused by the mere fact that others are also consuming it (Leibenstein, 1950).
Overconfidence bias	Individuals' subjective confidence in their judgement which is higher than objectively accurate (Brenner, Koehler, Liberman & Tversky, 1996; Keren, 1991).
Self-serving bias	The tendency to attribute success to internal attributes and failure to external factors stemming from the need for esteem describes the self-serving bias (Zuckerman, 1979). Individuals tend to indulge CBs because of motivational and cognitive reasons (Shepperd, Malone & Sweeny, 2008).
False consensus bias	Individuals' overestimation of how self-related knowledge extends to others. It leads to a perceived consensus, which does not exist (Krueger & Clement, 1994).
Framing bias	Individuals' risk conception depends on the problem presentation (Kahneman & Tversky, 1979; Kahneman & Tversky, 1984). A negative frame reduces risk aversion, whereby a positive frame increases it (Kahneman & Tversky, 1984). The frame refers to the mental picture people create to simplify the

	world and is strongly built on a reference point which impacts the decision outcome (Wright & Goodwin, 2002).
Status quo bias	The phenomenon that individuals show a preference for maintaining the current situation (Samuelson & Zeckhauser, 1988).
Habit bias	A selection purely based on the fact that it has been like this before (Hogarth, 1987; Slovic, 1975).

Out of the 14 relevant CBs, half of them have proven to have greater potential to negatively influence CAs than the others and are therefore proceeded with. These are the CBs already identified by literature to be relevant in the AI field: blind spot, confirmation, selection, and anchoring bias. Additionally, the overconfidence, self-serving, and false consensus bias. The overconfidence bias is selected as it is regarded as the most significant CB (Kahneman, 2011) with the most robust findings in the decision-making literature (Lichtenstein, Fischhoff & Phillips, 1982). Its importance is further stressed by its impact on individuals' ability to comprehend their own intellectual limitations which is potentially harmful to a CA if the individual denies its vulnerability to CBs that might affect the CA. The self-serving bias is selected as it impacts individuals' ability to recognise and take responsibility for their own CBs which has a high influence on the decision-making ability. False consensus bias, on the other hand, is especially relevant as it might impact assumptions made on who the user is and how they will use the CA.

The following section presents the underlying driver behind these CBs and their respective debiasing strategies. The strategies' applicability is only ensured to be effective for its respective CB, which consequently must be identified first.

Blind Spot Bias

The self-reinforcing blind spot bias results from naive realism and excessive reliance on introspective evidence (Pronin, 2007).

Warning of the CB only enables individuals to detect and correct obvious but not subtle distortions (Stapel, Martin & Schwarz, 1998). On the one hand, training to acquire an in-depth understanding of the CB, its environment, and consequences have a positive effect, which however is not sustainable, wherefore continuous education is required (Bessarabova, Piercy, King, Vincent, Dunbar, Burgoon, Miller, Jensen, Elkins, Wilson, Wilson & Lee, 2016). On the other hand, the strategy of "considering the opposite" can be adopted to encourage individuals to rethink their decision, which draws attention to their mental processes and leads to more rational behaviour (Lord, Lepper & Preston, 1984). Implementing these strategies requires that the effectiveness is assured by continuous practices.

Confirmation Bias

Confirmation bias is explained differently by varying authors. One suggested driver is the individuals' limited ability to focus on more than one thought at a time wherefore, a difficulty to pursue alternative hypothesis testing is experienced (Nickerson, 1998). Another view of the underlying driver is the effect of desire on belief (Baron, 2000; Oswald & Grosjean, 2004). Individuals prefer positive over negative thoughts, wherefore a tendency to confirm the desired

conclusion, which is linked with positive emotions, is developed (Nickerson, 1998; Oswald & Grosjean, 2004).

The problem and consequences of confirmation bias have been extensively discussed in the literature, while the literature on their effective debiasing is rather limited (Cook & Smallman, 2008). In several experiments, the confirmation bias was mitigated by visualising information to make it easier to analyse information (Cook & Smallman, 2008; Heuer, 1999; Hillemann, Nussbaumer & Albert, 2015). Hillemann, Nussbaumer, and Albert (2015) argue that the CB can additionally be approached through feedback, by making individuals aware of possible signs of the CB. This is consistent with the approach of Wilson and Brekke (1994), whereby a third person supports an individual in the first step of the debiasing process. The third party can support the CB's detection and thus draw the attention of the CB to the individual, who then executes the mitigation independently. The authors also suggest that defending one's own hypothesis and considering other or conflicting hypotheses may mitigate the CB, as this encourages the individual to reflect. This approach addresses the individual's system 2, whereby initial decisions could be controlled and corrected if necessary.

Selection Bias

Selection bias occurs when data is partially missing and individuals must make associations to select data (Stolzenberg & Relles, 1997). The selection affects the representativeness (Munafò, Tilling, Taylor, Evans & Smith 2018).

The majority of debiasing strategies mitigate the effects of the CB statistically, as they prove effective once data has been distorted (Cortes, Mohri, Riley & Rostamizadeh, 2008; Dudík, Schapire & Phillips, 2006; Heckman, 1979; Little & Rubin, 1986; Zadrozny, 2004). Strategies addressing qualitative measures are rather limited. One possibility to eliminate the CB is the randomisation of selection, as it avoids erroneous assumptions made in deliberate selection (Berger, 2005). This strategy, however, can lead to accidentally occurring biases in the CA, for example, through misrepresentation of gender or ethnicities. Another approach is to only exclude data if there is a clear reason for it, which defines the limits of selection and thereby, reduces the CB (Berger, 2005). This approach to avoid CBs being included in the data collection requires high effort due to the mass of data investigated and establishing arguments for the exclusion of data. As selection bias only occurs if a smaller sample size than the whole data collection is selected (Howe, Cole, Chmiel & Muñoz, 2011), using the whole collection eliminates the task of selection and thus, the CB. Even though the usage of all data would eliminate the selection bias, it would also allow other biases to enter which are inherent in the data (Llyod, 2018). Therefore, selection can also be seen as a way to increase data quality. Blackwell and Hodges (1957) argue that a separation of the selector from the task can limit personal involvement and resulting pre-assumption about the selection, which can reduce this CB. The dissociation of the data collector and the user is seen as a useful strategy. However, it requires that the data use case is thoroughly understood and communicated correctly. This adds another level on which errors may occur.

Anchoring Bias

Anchoring bias is based on individuals' notion that the anchor is more relevant than the boundaries of plausible answers (Strack & Mussweiler, 1997). The effects of the anchoring bias increase with higher ambiguity, lower familiarity, relevance, or personal involvement in the decision-making, and with higher trustworthiness of the information source. (van Exel, Brouwer, Berg & Koopmanschap, 2006).

This CB can be addressed by various debiasing strategies. Whereas the ineffectiveness of raising awareness of the CB (Chapman & Johnson, 2002; Wilson, Houston, Etling & Brekke, 1996) and incentivising correct estimates (Tversky & Kahneman,1974; Wilson et al. 1996) is argued by some, others (Epley & Gilovich, 2005) found limited evidence that it can have a reduced effect. At the same time, Epley and Gilovich (2005) acknowledge that this effect is only given when the anchor is self-generated. Another approach was presented as being more reliable, whereby individuals consider the opposite (Mussweiler & Strack, 2000; Mussweiler, Strack & Pfeiffer, 2000). Additionally, providing individuals with further information and increasing their motivation to higher accuracy is suggested as a two-fold strategy (Simmons, LeBoeuf & Nelson, 2010). This strategy thus addresses Wilson and Brekke's (1994) third step in the debiasing process.

Overconfidence Bias

Overconfidence bias stems from the overestimation of individuals' own intellectual abilities (Dawes, 1980), and increase correspondingly with an increasing degree of control over and involvement in outcomes (Keren, 1991). It can further be a response to social expectations (Keren, 1991).

Arkes, Christensen, Lai, and Blumer (1987) demonstrated that the approaches, providing feedback, and questioning the decision, reduce overconfidence bias. They found that overconfidence decreases when interim feedback between the tasks is provided. This presents a simple strategy if it is feasible that feedback can be given before each decision is taken. However, this is unlikely in a business environment with always existing time constraints. In addition, Arkes, Christensen, Lai, and Blumer (1987) observed that social pressure, created, for example, through peer interaction, and asking an individual to explain or justify its answers reduces overconfidence. This is also suggested by Bhandari and Hassanein (2012), who propose that asking about past decisions reveals the thought process of individuals, thus, their behavioural patterns can be identified and improved. This approach assumes that peers have the ability to identify the behavioural pattern and that individuals are willing and able to change behaviour. These prerequisites, which link with the first stages of the debiasing process (Wilson & Brekke, 1994), must be overcome for the strategies to be successful. Differently, Blanton, Pelham, DeHart, and Carvallo (2001) directly address that overconfidence often results from an overreaction to uncertainty and suggest to create a safe environment and the feeling of certainty. Apart from this, awareness training in CBs and debiasing techniques can also mitigate the CB (Welsh, Begg & Bratvold, 2007).

Self-serving Bias

Self-serving bias has motivational and cognitive drivers (Shepperd, Malone & Sweeny, 2008). The authors explain that individuals are motivated by self-enhancement to ultimately enhance their self-worth and self-presentation allowing them to depict a desired image to others.

The self-serving bias is a very resistant CB (Babcock, Wang & Loewenstein, 1996). Still, it can be addressed with a variety of debiasing strategies. Farnsworth (2003) identified strategies varying in their degree of personal and structural intervention from which structural solutions are more effective, but also linked to higher costs. The first strategy presented is reeducation by informing individuals of the CB and its likely occurrence. This, however, can have an effect of convergence rather than increasing the decision accuracy (Farnsworth, 2003). This strategy serves as a good foundation but must be enriched with another strategy, as it only addresses the first step to successful CB correction, according to Wilson and Brekke's (1994) process. Secondly, the author suggests penalising biased individuals by moving them to a different position in the decision-

making process (Farnsworth, 2003). This is an extreme punishment, regarding that the first step of debiasing, the identification of the CB (Wilson & Brekke, 1994), is already difficult for individuals to execute by themselves (Osoba & Welser, 2017; Pronin, Lin & Ross, 2002). Additionally, allocating positions to individuals according to their CBs seems very costly in terms of resources. It is not realistic to adjust positions based on CBs as this would require constant iterations. Thirdly, biased individuals could be separated from the decision (Farnsworth, 2003). This can be done by excusing an individual from their position, requiring them to listen to other perspectives, or by adding another individual to the decision-making which balances the effect of self-serving bias on the decision and reduces the cost of decision (Farnsworth, 2003). This strategy acknowledges the difficulty to eliminate CBs, but at the same time requires CB detection and traceability to a specific individual. Nonetheless, by matching individuals a potentially safer environment is created. Individuals share responsibilities, which may limit the individual's need for self-enhancement (Shepperd, Malone & Sweeny, 2008). However, in the case that a decision of one individual does not allow the participation of another individual, this strategy is obsolete. In addition, listening to other perspectives might be beneficial as this allows individuals to recognise factors that might have been missed otherwise. Generally, this strategy succeeds in addressing the underlying drivers and not only the effects of the CB.

False Consensus Bias

False consensus bias has several drivers. Kruger and Clement (1994) argue that individuals may not have complete information and may not be aware of it. Morrison and Matthes (2011), however, explain that the cause is attributed to the individuals' need for belonging and personal importance of the related task. Wetzel and Walton (1985), on the other hand, found that three factors are particularly influential. Individuals' ego-defensiveness, their lack of cognitive abilities, and incorrect belief system that choices are based on situational factors when they in reality are based on internal variables. More specifically, Marks and Miller (1987) state that the CB is intensified in situations with like-minded people and when a decision is very significant, unusual, or stressful.

The false consensus bias is very resistant, and many common debiasing approaches such as training, forewarning, feedback, or availability of information do not result in significant improvements (Krueger & Clement, 1994). Complete elimination of the CB can only be achieved if solely causal attributions are considered in a decision-making process, as those can ensure factuality (Marks & Miller, 1987). This, however, represents a very normative approach, as the strategy neglects that CBs are inherent in humans (Arnott, 2006). It, therefore, does not acknowledge that CBs continue to influence decision-making even if individuals are compelled to only consider causal attributes. Another strategy proposed is generating alternatives to a task position, as concentration on only one position provokes false consensus and thereby mitigates the bias (Marks & Miller, 1987). This presents a more realistic strategy, but the derivation of alternatives can still be biased, which shows the difficulty of fully eliminating CBs.

Cause-Specific Debiasing Strategies

In the following, strategies are presented that the authors have related to a specific cause leading to CBs. According to the literature, identifying the underlying causes is necessary to enable the selection of applicable debiasing strategies (Arkes, 1991; Keren, 1990; Larrick, 2004), whereby the effectiveness of each strategy is dependent on the assigned cause. The causes presented by the authors differ strongly from each other. Some relate their strategies to the causes of faulty judges, faulty tasks, or a mismatch of both (Fischhoff, 1982). Keren (1990) distinguished between procedural and structure causes, while others differentiate between strategy-, association-, or

psychophysically based causes (Arkes, 1991). Almost no direct overlap in causes is found. However, parallels could be drawn in the analysis of the underlying cause. A total of 91 strategies from 12 scientists were analysed, which were subsequently grouped according to their underlying causes (see appendix A). In total, strategies could be assigned to six overarching causes: 1) lack of guidance, 2) faulty task, 3) inadequate decision, 4) lack of feedback, 5) lack of education and 6) inadequate overview.

Lack of Guidance

In literature, it was identified that a lack of guidance can be counteracted by the implementation of accountability, incentives, or guidance itself.

By holding individuals accountable for their decisions, their social obligation to appear consistent to others is initiated (Larrick, 2004). It raises the individuals' stakes, lowers self-confidence in their judgement, and encourages thoughtful decisions (Arkes, 1991). In addition, it removes the feeling of anonymity (Croskerry, Singhal & Mamede, 2013) and leads to more effort being invested in the decision-making (Larrik, 2004). Overall, the motivation rises (Kaufmann, Carter & Buhrmann 2012) and the performance improves (Croskerry, Singhal & Mamede, 2013). Problems may occur when accountability results in the argumentation only being adapted to the target group to confirm their preferences, or when it results in the use of only easily justified aspects (Larrik, 2004). Here, accountability is used as a means of pressure to oblige individuals to think rationally. But, this strategy can also have negative effects if used incorrectly or if it causes stress so that a person has to additionally rely on other CBs to cope with this stress.

Incentives offer an alternative to accountability (Arkes, 1991), as they exert less pressure and aim at individuals' voluntariness. By raising the stakes, individuals' system 2 is activated as their entire cognitive capital is needed to solve the talk sufficiently (Larrick, 2004), which is also seen by Kaufmann, Carter, and Buhrmann (2012), who state that this can improve decision performance. Nevertheless, incentives are not a universal strategy as they only have a positive effect on performance if individuals already possess the necessary cognitive capital and only its intrinsic motivation for the task is insufficient to achieve an optimal result (Larrick, 2004).

Several strategies address the lack of guidance by suggesting guidance through cognitive forcing. Since individuals find it difficult to detect their own CBs, forcing requires a double-check of actions by which CBs can be eliminated (Thammasitboon & Cutrer, 2013). These strategies require that intuitive decisions of system 1 are verified through forced actions performed by system 2 (Croskerry, 2003). These enforced actions can be checklists that prescribe a standard set of tasks or rules to be followed (Croskerry, Singhal & Mamede, 2013). For the effectiveness of this strategy, individuals do not need to be aware of CBs, as described by Wilson and Brekke (1994). Effectiveness is primarily dependent on individuals' willingness to use the strategies at hand, which can be reinforced by oversight, accountability, or incentives. Part of guidance is forewarning strategies to increase awareness for a specific CB or for specific situations in which CBs likely occur (Croskerry, 2003), which in turn increases the motivation to counteract. The forewarned person can either avoid the situation or adapt their behaviour to avoid CB occurrence (Keren, 1990).

However, forewarning is widely debated, as the behaviour to manage a CB from occurring could potentially lead to the occurrence of other CBs (Wilson & Brekke, 1994).

Faulty Tasks

The cause of faulty tasks can be mitigated by supporting individuals through the alteration of a task. Soll, Milkman, and Payne (2015) state that extensive decisions are associated with greater uncertainty, wherefore decomposing the task can reduce CBs. Higher clarity is achieved by narrowing the scope of each task component (Kaufmann, Michel & Carter, 2009). Fischhoff (1982) suggests that faulty tasks might be addressed by clarifying the tasks carefully to increase familiarity with it, avoiding confusion, or by asking fewer questions, averting individuals from falling into a pattern of answers. These strategies may limit the occurrence of some CBs caused by incomprehension or overextension, but they do not address, nor eliminate, all CBs that may distort individuals' decision-making.

Inadequate Decision

Another cause of CBs is an inadequate decision, which can be addressed by introducing a new method for selecting the optimal decision or changing the decision derivation. One possibility, proposed by Keren (1990) is giving individuals instructions, as this does not require their understanding of the underlying problem of the task. For this purpose, individuals can avoid assumptions (Keren, 1990) or a blind, random choice can be made (Arkes, 1991). However, this seems rather unrealistic as decision-makers are employed to execute well-thought-out decisions. Alternatively, Thammasitboon and Cutrer (2013) argue that the accuracy of a decision is based on expert knowledge and therefore, decisions should only be taken by respective experts in this field. Also, primary decisions could be verified by challenging it with colleagues to minimise the susceptibility of CBs in a decision (Kaufmann, Carter & Buhrmann, 2012). Such approaches may reduce the occurrence of CBs in individuals' decisions but can still be distorted by CBs of third parties. However, the probability of CB reduction is higher due to the diversity of people involved, as it allows for more perspectives and may reduce extreme cases of false mental reasoning. Further, Soll, Milkman, and Payne (2015) claim that when alternatives to a solution are generated, the greater choice reduces CBs in decision-making. Choosing from several options could limit CBs, although no guarantee is given that the decision-making process becomes bias-free along the way.

Besides a revised selection method, the derivation of the decision can also be addressed. On one side, Croskerry, Singhal, and Mamede (2013) argue that by structuring the collection of information, presenting information in a way that differs from a text, or decelerating the decision process, CBs can be countered. On the other hand, Arkes (1991) suggests, in case of an error in assessment, elements of the task are either re-linked or separated, or terms free of associations are used or a definition is added to allow a better understanding of the task. These structured approaches provide a more comprehensive understanding of the decision and reduce the probability of CB occurrence, although they do not guarantee a better understanding of the task. Also, altered terms can still be incorrectly associated, whereby only clear explanations and definitions may help.

Lack of Feedback

Another cause of CBs is insufficient feedback. Conversely, the provision of feedback can be seen as a strategy to debias this cause. Arkes (1991) suggests that feedback on the adequacy of a decision can be given daily to reduce CBs. Thammasitboon and Cutrer (2013) argue that feedback should be given systematically, according to the performance of an individual, for a sustained learning

effect and increased accuracy in the future. Both strategies are highly relevant, as it is difficult to identify CBs within oneself, implying that supervision is necessary. While daily feedback offers a quicker correction effect, systematic feedback provides greater reflection and learning, making both strategies especially valuable in combination. However, daily feedback does not seem realistic.

Feedback is especially important for the reflection on previous decisions. This can be supported by mandatory re-evaluation, for example, in form of double-checking a decision by asking the same question twice in succession with a time lag in between (Soll, Milkman & Payne, 2015), or by newly assessing the first decision (Arkes, 1991). Second opinions (Thammasitboon & Cutrer, 2013) or control groups can also be consulted for reflection and feedback (Croskerry, Singhal & Mamede, 2013). These strategies can be regarded as supportive strategies, as they do not guarantee the elimination of CBs, but, as explained above, create a learning effect. This can enable the achievement of the first step of the debiasing process (Wilson & Brekke, 1994).

Lack of Education

A lack of education as a cause of CBs and the strategy to promote education for debiasing is widely discussed in the literature. In general, this strategy refers to acquiring knowledge about the environment of a decision (Thammasitboon & Cutrer, 2013), and training in CBs, or strategies to counteract CBs (Larrick, 2004), which can involve learning about CBs that are likely to affect the specific decision-making process (Croskerry, Singhal & Mamede, 2013) and their influences (Kaufmann, Michel & Carter, 2009). These approaches, beyond providing education and raising awareness, the first step in the debiasing process (Wilson & Brekke, 1994), guide individuals to manage CBs through countermeasures.

Inadequate Overview

CBs might occur if individuals have an inadequate overview of all information available. Debiasing strategies in this category discuss that overexposure or lack of information is likely to be the reason for this and therefore, balanced information exposure is essential (Croskerry, Singhal & Mamede, 2013). Exposure control explains that information, for example previously made decisions, is not disclosed to the decision-makers to avoid any influences that could affect the rationality (Croskerry, Singhal & Mamede, 2013). Other authors also support this strategy in cases in which gaps in knowledge are identified, as they claim that additional information can be included in the decision-making process to provide a more complete understanding (Thammasitboon & Cutrer, 2013), which can reduce uncertainty (Keren, 1990). This is an essential strategy, as information is fundamental to a correct decision, however, preparatory work is required to determine the right balance of exposure.

Also, many scholars agree that a change of perspective can provide a more comprehensive overview of all perspectives (Kaufmann, Carter & Buhrmann, 2012), which reduces the likeliness of CBs. Herefore, Fischhoff (1982) suggests firstly to force individuals to openly communicate all information they hold allowing transparent processing. Next, new information from other or new perspectives can be added to the decision-making process and thereby, expand the frame. This extends the reference point and reduces CBs in the process (Arkes, 1991). This may emerge, for example, when experiences from past cases are included in the decision (Chen & Lee, 2003) or when the analysis is carried out from a future perspective (Soll, Milkman & Payne, 2015). Another approach to this involves changing the thinking style while accessing the same decision several times, for example by switching between system 1 and system 2 (Soll, Milkman & Payne, 2015).

Another strategy, "consider the opposite", attempts to debias by stimulating other impulses than those that would normally occur (Arkes, 1991; Fischhoff, 1982), similar to the strategy of approaching the decision from a worst-case scenario (Croskerry, Singhal & Mamede, 2013). By taking up the opposing proposition and evaluating why that side might be superior (Soll, Milkman & Payne, 2015), personality traits and CBs might be sidestepped (Croskerry, Singhal & Mamede, 2013). A change of perspective is a relevant strategy, as it challenges individuals' way of thinking, expands their horizons while a more diverse viewpoint reduces CB occurrence. However, if individuals are biased through a blind spot, the same CBs might still affect the alternatives considered.

In general, it can be noted that the strategies provide useful references to improve individuals' mental reasoning, in some cases with the intervention of a third person. However, some of the strategies require preparatory work, which requires substantial resources that may not always be available. Therefore, strategies must be selected according to the available resources of organisations. As previously noted, some strategies appear unrealistic, while others are not sufficient in isolation to completely eliminate CBs. Therefore, further assessment is required to determine the most effective combination of strategies.

2.5 Reflection on Literature Review

The literature distinguishes debiasing strategies between CB- and cause-specific approaches. When conducting the literature review it was recognised that several strategies in these approaches overlap, which makes their classification in such categories debatable, therefore, this structure is reflected on. A distinction may be useful in cases in which a specific CB or a specific cause has been identified, whereupon, a specific strategy should be applied. A strategy for a specific CB is only effective for that CB, and could cause harm, as the counteraction of one CB could intensify another CB that was additionally inherent in an individual (Arkes, Christensen, Lai & Blumer, 1987; Marks & Miller, 1987). However, individuals commonly carry multiple CBs that require counteraction (Gigerenzer, Todd & the ABC Research Group, 1999), or multiple factors can be identified that cause CBs. Therefore, applying a strategy only addressing one CB or one cause seems unrealistic. Moreover, the identification of a CBs and causes is costly, as knowledge must be acquired and profound analyses must be applied, wherefore a more general approach might be more realistic in practice.

Instead of the categorisation, the debiasing process presented by Wilson and Brekke (1994), for example, could serve as an underlying concept of debiasing strategies. Debiasing strategies should enable the steps in the process to support individuals in achieving a successful correction of CBs. Strategies can enable the achievement of certain steps in the process, whereby different strategies may target different steps. Some of the debiasing strategies that are presented in the following, can serve as an example of this assumption. The strategy of offering incentives, for example, may increase individuals' motivation to undergo the process and correct the CBs. Another example is that education about CBs may positively influence the awareness of the CB or of the direction and magnitude, which can facilitate these steps of the process.

Primarily, debiasing strategies address single steps, but can also go beyond that, as they may make some steps redundant. For example, the debiasing strategy of using checklists could invalidate the motivation step by forcing individuals to apply the strategy and thereby, correct the CB. Furthermore, this strategy could eliminate the need to be aware of the CB and its direction and magnitude. This demonstrates that the debiasing process and the strategies can be related to each other and could serve as an alternative to the categorisation that literature provides. However, the CB- and cause-specific knowledge could enrich the process as this knowledge remains essential due to the attached effectiveness of strategies.

2.6 Chapter Summary

The literature review generated an overview of the relevant literature as a basis for the upcoming analysis and provides an understanding of how CBs can influence decision-making and, thereby ultimately, a CA. CBs are behaviours or reasonings of individuals that deviate from complete rationality, originating from the dual process of individuals, and are caused by heuristics, error management biases, and artifacts. These CBs can influence a CA by individuals that make irrational decisions in the development process so that their CBs are transferred to the CA. To successfully achieve a correction of a CB, individuals must meet the criteria of being aware of the CB, motivated to correct it, aware of its direction and magnitude, and able to adapt their responses. Hereby, third persons can support biased individuals along the process.

Debiasing strategies can be applied after distortions have been identified to limit or, in the best case, eliminate CBs, thus making the decision-making process more rational. In literature, these debiasing strategies are approached differently, in the form of CB- and cause-specific strategies. For the CB-specific approach, strategies were presented for the most relevant CBs that can be transferred to a CA, namely strategies for 1) blind spot, 2) confirmation, 3) selection, 4) anchoring, 5) overconfidence, 6) self-serving, and 7) false-consensus bias. Cause-specific strategies address 1) lack of guidance, 2) faulty tasks, 3) inadequate decisions, 4) lack of feedback, 5) lack of education and 6) inadequate overview. Overall, it must be noted that many strategies have been identified to not serve as a stand-alone solution that guarantees the complete elimination of the CBs.

3 Methodology

The purpose of this chapter is to justify the research design selected to answer the research question. This section reviews the methodological choices and addresses the research limitations in the form of presenting implications of validity, reliability, and the role of researcher.

The methodology is based on the research onion (Saunders, Lewis & Thornhill, 2016), displayed in figure 2. It visualises the decisions a researcher must take to determine an appropriate research design by working from the outer layer to the centre. The onion consists of six layers which are analysed in the following paragraphs.

It must be considered that the research onion originates from a normative view providing clear distinctions between the different options which cannot always be reflected fully in practice.

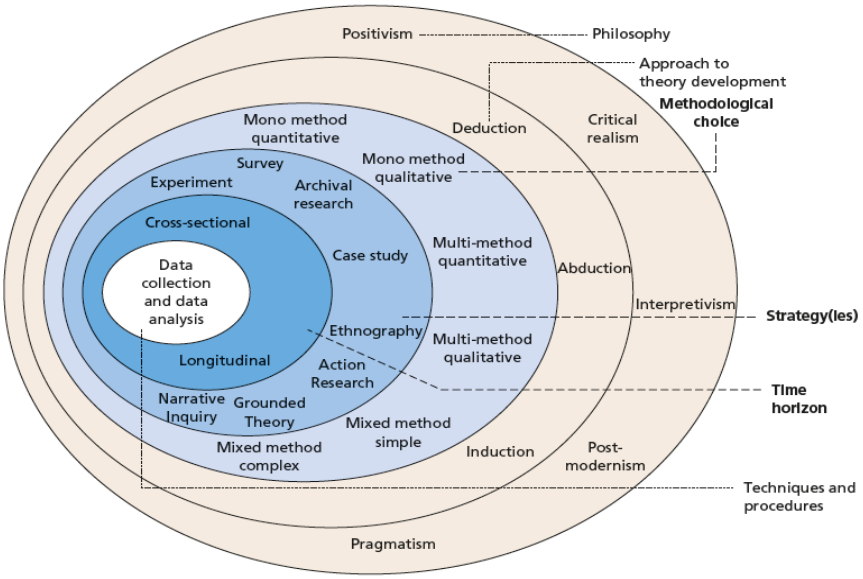


Figure 2: Research Onion (Saunders et al. 2016)

3.1 Philosophy

The research philosophy determines the "system of beliefs and assumptions about the development of knowledge" (Saunders, Lewis & Thornhill, 2016, p.124). The researchers' worldview, who in this case are the authors of the thesis, defines assumptions about the importance of information and influences the choice within the inner layers of the onion (Saunders, Lewis & Thornhill, 2016). Literature highlights four worldviews that can be adopted: positivism, realism, interpretivism, and pragmatism (Saunders, Lewis & Thornhill, 2016).

The researchers considered the research question as essence, focused on problem-solving, and aimed to contribute to future practices while adapting an interpretivist and positivist position to answer the research question. Further, the research was practice-oriented and adopted a relativistic view as it acknowledges the possible absence of absolute truth as individuals can only present one view, that can never reflect the whole picture. Nevertheless, the importance of obtaining an in-depth understanding of the participants and their truth was recognised.

Accordingly, pragmatism was chosen as an underlying philosophy based on the researchers' nature, the above-mentioned research question, and the aim of this research. According to Kelemen and Rumens (2008), in pragmatism applied concepts are only relevant if action is supported. The pragmatist worldview "arises out of actions, situations, and consequences rather than antecedent conditions" (Creswell & Creswell, 2018, p.10). Consequently, pragmatism is linked with freedom of choice which allows integration of different methods, assumptions, techniques, and approaches to suit the research problem's nature (Saunders, Lewis & Thornhill, 2016).

3.2 Approach

The approach to theory development is classified into deduction, abduction, and induction (Saunders, Lewis & Thornhill, 2016).

This research aimed at understanding how organisations can use debiasing strategies to prevent CBs from impacting their CAs. Therefore, data were collected to explore this phenomenon after which it was analysed for themes and patterns. Primary and secondary data were gathered in a continuous process.

Consequently, the research mainly adapted an abductive approach (Saunders, Lewis & Thornhill, 2016). However, it was not intended to develop a new theory, as traditionally associated with the abductive approach (Saunders, Lewis & Thornhill, 2016), due to the insufficient time to test hypotheses resulting from the data analysis.

3.3 Methodological Choice

The methodological choice is categorised into mono method, quantitative and qualitative, multi-method, quantitative and qualitative, and finally, mixed method simple or complex (Saunders, Lewis & Thornhill, 2016).

Answering the research question required insight into different variables of CB management on a deep level to make sense of the meaning. This research attempted to integrate various viewpoints by focusing on one method, interviews, which was used to generate primary data from different expertise fields and backgrounds. Addressing the assumed issue that people are only able to picture one truth, the research generated data from various input sources attempting to draw a complete picture. Information was explored and gathered in participant settings by conducting interviews. Afterwards, an interpretative approach to data was selected, whereby data collection techniques and analytical procedures were used to understand the data's meaning and to generate a theoretical contribution. This enables a rich understanding of the assessment of the transferability of theory from one context, debiasing, to another, prevention, which is linked to the abductive approach.

Hence, a mono method qualitative approach was applied as it generates the most credible, well-founded, and relevant data for this research in the given time horizon (Saunders, Lewis & Thornhill, 2016). This methodological choice was aligned with the pragmatic, problem-centered philosophy as it ensured that a comprehensive understanding of the research problem was provided, through accessing the individual's meaning (Saunders, Lewis & Thornhill, 2016). As the qualitative approach recognises the importance of individual meaning and depicting the complexity of the phenomenon (Creswell & Creswell, 2018), it also highlights the importance of creating trust and encouraging participation to access the meaning and create understanding (Saunders, Lewis & Thornhill, 2016). This aspect is addressed further in section 3.6.1.

3.4 Strategy

The research strategy refers to the plan to conduct research and to achieve the research objective. It is categorised between experiment, survey, archival research, case study, ethnography, action research, grounded theory, and narrative inquiry and is mainly deduced from the research philosophy and approach (Saunders, Lewis & Thornhill, 2016). The five latter being representative of qualitative research strategies (Saunders, Lewis & Thornhill, 2016).

A concurrent collection and analysis of data are commenced as of the point that the research idea was clarified. Hence, interview conduction and analysis took place simultaneously. This allowed the researcher to explore the two rich topics, CBs in the context of CAs, in the interviews, and to build on insights of conducted interviews to explore the phenomenon further in coming interviews. This allowed deepening the understanding further from interview to interview. Codes emerged from the data analysis which enabled information categorisation, facilitating a constant comparison, which increases consistency in coding and data analysis (Saunders, Lewis &

Thornhill, 2016). The concepts emerging from the data collection guided further processes, such as the continuous interview partners selection presented in section 3.6.1. Memo writing was further applied as an aid to maintain an overview of the researchers' thought process throughout the coding process and to conceptualise data.

Consequently, this research applied the grounded theory strategy in which the generated theory reveals an already existing reality that is made visible through a systematic collection and analysis of qualitative data (Saunders, Lewis & Thornhill, 2016). Grounded theory is part of a wider methodological approach that allows investigating the meaning that is constructed by social actors (Charmaz 2006; Glaser & Strauss, 1967; Suddaby, 2006). Therefore, strategy and techniques and procedures merge in this approach, which is addressed in section 3.6.1.

3.5 Time Horizon

The time horizon defines the period in which a study takes place, which can either be cross-sectional or longitudinal (Saunders, Lewis & Thornhill, 2016).

The research project was limited to a time frame of ten weeks hence limiting the length of the research period. Furthermore, the research investigated the current strategies to manage CBs impacting CAs at the interviewed organisations at a particular time, thereby creating a snapshot of the current management strategies.

Hence, this research represents a cross-sectional study (Saunders, Lewis & Thornhill, 2016).

3.6 Techniques and Procedures

The research techniques and procedures refer to the practicalities of data collection and analysis (Saunders, Lewis & Thornhill, 2016). The selection of which methods are best applied is based on the above-made research specifications. The following section is separated into primary and secondary data, which both serve as a foundation for this research.

3.6.1 Primary Data

The primary research was based on a qualitative mono-method. Hence, data was gathered in participant settings which, in theory, can be done by experiments, observation, and communication (Ghauri & Grønhaug, 2002). This research generated data through communication in an interview setting as it provided the opportunity to choose experts in a specific field of interest and explore this field beyond the limits of experiments and observations as it allows generating an in-depth understanding (Saunders, Lewis & Thornhill, 2016). Interviews were designed as problem-oriented

conversations, which enabled the researchers to retrieve reliable and valid findings (Saunders, Lewis & Thornhill, 2016). To understand the individuals' meaning, a subjective interview approach was applied, putting the focus on understanding the interviewee's point of view. A subjective approach implies interviews to be socially constructed by the researchers and respondents because of the co-production of the interview. Interviews are differentiated by their degree of formality and structure, one common typology categorises them as follows: unstructured interviews, semi-structured interviews, and structured interviews (Saunders, Lewis & Thornhill, 2016).

In this research, semi-structured interviews were conducted, in accordance with the choice of an exploratory, grounded theory approach. Semi-structured interviews allow a deeper understanding of the relationship between the variables (Saunders, Lewis & Thornhill, 2016). Additionally, this interview method provided insight into a detailed data set necessary for developing satisfactory conclusions (Saunders, Lewis & Thornhill, 2016). A further advantage of in-depth interviews was that the interviewee can reveal information that was not initially considered. Semi-structured interviews served to understand the research question and each interviewee's viewpoint.

Interview Preparation and Execution

As an interview preparation, research and interview topics were developed (see appendix B), based on the previously generated understanding of the research area and its environment.

To increase the validity and reliability of interviews, participants were provided with the interview topics at least 48 hours in advance of the scheduled interview time, to allow interviewees to prepare. These topics were further developed into open questions, which maintained a link to the research question. To avoid unauthentic answers, the questions were not made available before the interview. The topics and questions may have differed depending on the purpose of each interview. Open-ended questions allowed the respondent to further explain the answer, placing the response in context (Saunders, Lewis & Thornhill, 2016).

The researchers considered the interview setting, such as appropriate location and dress code, as an influencing factor on the perceived competence and credibility (Saunders, Lewis & Thornhill, 2016), and therefore, ensured that a professional ambience was created. The importance to create a pleasant environment for the interviewee to feel comfortable sharing information was acknowledged (Saunders, Lewis & Thornhill, 2016), wherefore the structured communication before the interview and the introduction phase of each interview were used to build rapport and create trust. Before commencing the interview, it was ensured that all questions were clarified.

Consistency of the research was ensured by a consistent framework of the interview procedure. The language in which the interviews were conducted was agreed with the interviewee in advance. All interviews were conducted in English. They were conducted one-on-one, except for a joint interview of P6 with its superior P7, and via remote communication channels. Respondents' availability may have influenced the time frame, but generally, interviews were conducted for 45 minutes, with slight variations possible.

To prevent ambiguity, respondents received an information sheet and a consent form. With consent given, the interviews were recorded on tape, with notes being taken at all times. The questions were

asked neutrally and kept concise. Judgemental comments from the researchers were avoided during the interviews. The researchers adopted an open attitude and ensured a pleasant and attentive atmosphere in which the interviewee can freely share its knowledge. If uncertainties arose, clarifications were requested and the indented understanding was tested.

Interview Participant Selection

Interviews were conducted with people of different cultural, academic, and professional backgrounds and organisations to create diversity in primary data sources. The research applied theoretical sampling, by selecting each interviewee based on its connection to the analytical phenomenon emerging from the coded data in the previous cases, which falls under purposive sampling, as part of the grounded theory methodology (Saunders, Lewis & Thornhill, 2016).

Purposive sampling is common when the interviewee selection is intended to identify participants who offer particularly informative insights (Neuman, 2005). A strong emphasis was placed on the participant's ability to support answering the research question. Hence, judgement is applied to select interview participants. The selection criteria driving this process were the interviewee's practical experience in the field of CAs or AI, and their expertise, which is reflected in various ways such as their job positions or publication of books. Further, the first contact was designed to establish a clear understanding of the research to allow individuals to reflect on their ability to contribute to the research, which presents another selection filter. Finally, the selection was intended to present a cross-section of people to understand diverse points of view and thereby, generate a deeper understanding of the phenomenon. As the grounded theory strategy was adapted for this research, the specific form of purposive sampling, theoretical sampling, was applied. Therefore, the selection was also driven by codes and themes emerging from the data analysis from previous interviews, which formed the basis to assess participants' suitability to advance the understanding of emerging themes. In other words, the focus was placed on "pursuing theoretical lines of enquiry" (Saunders, Lewis & Thornhill, 2016, p. 194) and participants were selected based on their ability to advance this theoretical line. Due to the limited time frame for data collection, an outline of who to sample was generated initially but also adapted continuously, while the themes for each interview were defined along the process.

Theoretical sampling allowed to simultaneously and concurrently generate, evaluate, and compare the data collected and codes identified. In theory, theoretical sampling is performed until saturation is achieved, meaning that no new ideas are revealed in the data collection and that categories and their relationship are well-established and comprehended (Strauss & Corbin, 1998). Applying theoretical saturation to the context of grounded theory, it can be translated to the generation of a contextually based theoretical explanation (Saunders, Lewis & Thornhill, 2016). However, this research was limited in its time frame, which prevented the achievement of theoretical saturation.

127 people were contacted out of which 11 agreed to be interviewed (see table 2).

Table 2: Interview Participants

ID	Current Job Position(s)	Employer	Interview Date
P1	- Artificial Intelligence Specialist - Author - Speaker	Self-employed	28.04.2020
P2	- Chair at Institute for Accountability in the Digital Age - Program Manager Dutch Government's Innovation Community - Boardroom advisor: IT Circle - Independent Consultant on Business Community Building	Government, self-employed	29.04.2020
P3	- Conversational AI Designer	Enterprise software company	29.04.2020
P4	- Chief Technology Officer	Information Technology & Services Consultancy	04.05.2020
P5	- Product Operations Specialist for Speech Recognition AI	Social Media & Technology Company	06.05.2020
P6	- European Lead Consultant Conversational AI	Business & Technology Consultancy	06.05.2020
P7	- Conversational AI Consultant	Business & Technology Consultancy	06.05.2020
P8	- Speech Recognition Engineer	Automotive Manufacturer	07.05.2020
P9	- Operations Research, Data Science & AI Manager	Information Technology & Services Consultancy	08.05.2020
P10	- Head of Analytics & AI - Female AI Ambassador	Insurance Company	10.05.2020
P11	- AI & FinTech Leader	Consultancy	12.05.2020

Further information about the interview participant's background is presented in appendix C.

Interview Analysis

In the interview analysis, qualitative data were explored, analysed, synthesised, and transformed to access the socially constructed meanings of the interview conversation (Saunders, Lewis & Thornhill, 2016). The meaning developed through the interpretation of its context. Consequently, a systematic analysis was followed to ensure high research quality and manage the ambiguity of words, which the thematic analysis provides. This approach formed the overall structure for the analysis, while aspects of the grounded theory methodology were incorporated. The four stages of the thematic analysis entail: 1) familiarisation with data, 2) coding of data, 3) categorisation of data in themes and understanding of relationships, and 4) refinement of themes (Saunders, Lewis & Thornhill, 2016). To identify each interviewee's meaning without being guided by published theory, data-driven codes were applied.

The familiarisation with the data occurred by transcribing the interviews. The word-to-word transcription applied clear speaker identifiers for each statement. Filler words were not transcribed. Interview notes recording contextual information and observations made such as the participant's non-verbal communication and tone of speech, enriched the transcription. To ensure a qualitatively high transcript, the researchers implemented a data cleaning phase which eradicated transcription errors. Subsequently, the transcription was made available to the interviewee to ensure factual accuracy. Data coding is a central component of grounded theory research. There are different typologies with varying distinguishments of the coding stages. This research applies Charmaz's (2006) coding stages, which are split in initial and focused coding. First, relevant sections were highlighted in the transcripts, after which codes were assigned. In vivo codes were applied, in accordance with the grounded theory methodology, which uses the exact wording of the participants, and therefore, does not change the intended meaning. Subsequently, codes were aggregated into code groups clustering the codes according to themes. A code network was created to visualise the interrelations of codes and their groups. Stage two and three were performed iteratively, as constant analysis and comparison were performed. In the final phase, themes were refined in accordance with the insights gained in the initial phases (Saunders, Lewis & Thornhill, 2016). An overview of generated code groups and themes is presented in appendix D.

Grounded theory further centres around memo writing. Memos were written to create a record of ideas and thoughts that developed along with the project's progress. The memos, for example, captured how much the interview participants had to be guided to maintain focus on CBs impacting CAs.

3.6.2 Secondary Data

In addition to primary data, secondary data is used to meet the research objective and answer the research question. Various types of secondary data were sourced, with the main focus on academic articles and books. These were retrieved through online databases or physical libraries that were publicly available or gained access through Lund University. This research required a deep psychological focus to explain the management of CBs, therefore several research streams were included. Mainly, psychological and management research streams were considered and a link was established between them.

As a supporting act, consultation interviews with field specialists were conducted for approximately 30 minutes each, to gain an overview of the research topic, confirm the understanding of theoretical concepts, and receive recommendations on secondary literature related to the respective field of expertise (see table 3).

Table 3: Secondary Data Interviewee Participants

Current Job Position	Employer	Interview Date
Senior Lecturer at the Department of Psychology	University	16.04.2020
Theoretical Philosophy Professor at the Department of Philosophy	University	23.04.2020
Professor at the Department of Law	University	27.04.2020

Using secondary data allowed the researchers to save resources in the form of time and money (Vartanian, 2011) while enabling the construction of the research phenomenon's holistic context and enhancing data triangulation (Saunders, Lewis & Thornhill, 2016). To assess the suitability of a secondary data source for this thesis, five criteria were assessed: 1) ability to enable answering the research question, 2) associated benefits of the secondary data are greater than the costs, 3) accessibility of the data, 4) suitability of research's purpose from secondary source with this research, and 5) quality and academic degree (Saunders, Lewis & Thornhill, 2016). In case the secondary data source was unable to meet one of the criteria, it was not considered for this research.

Noteworthy, as the relevant research fields produced an extensive volume of information over the years, which, at the same time, were partially incoherently presented. Therefore, a clear structure, which would allow presenting an overview of the data in the literature review, was required. Firstly, strategies had to be structured to their respective approach, CA-specific, and cause-specific debiasing methods. In particular, the cause-specific debiasing strategies which authors have structured with different cause categorisations required structure. To enable an aggregated overview of all relevant cause-specific debiasing strategies mentioned in the literature, the underlying causes of each existing categorisations were analysed and identified. Subsequently, debiasing strategies were regrouped based on matching underlying causes. This procedure was applied to 91 strategies, developed in 12 academic papers. The list of scholars is not exhaustive, however, it was found that further literature corresponds to the identified selection of strategies.

In contrast, to the fast information availability in the field of cause-specific debiasing strategies, the area of application of this research, CAs, has not been widely discussed in the context of CB management. This posed a challenge when defining relevant CBs to present relevant debiasing strategies for these CBs. Therefore, the scope from CA-related literature had to be extended to the overarching technology AI. Nonetheless, the data availability did not improve significantly wherefore general CB literature had to be taken into consideration and CBs had to be assessed on their relevance for CAs. This was done based on their impact on individuals' ability to understand their own intellectual limitations, the ability to recognise CBs, and the potential harm it could cause in the context of CAs.

Triangulation

The research applied data triangulation to increase confidence in the research data (Patton, 2002). The data was collected by highlighting different angles including multiple sources. Additionally, a diverse interviewee group was compiled in which, if possible, at least two interviewees received the same questions to ensure findings were generated from more than one source.

3.7 Research Limitations

Research limitations are influences that affect the credibility of research findings (Saunders, Lewis & Thornhill, 2016). The following section is divided into reliability, validity, and role of the researcher.

3.7.1 Reliability

Reliability refers to the replicability and consistency of the research (Saunders, Lewis & Thornhill, 2016). To promote reliability, it was considered if measures will yield the same results on other occasions, if similar outcomes will be reached by others, and if transparency was given in how sense was made from the raw data.

A replication of the research on another occasion will unlikely yield the same results as the research is based on data that intends to reflect a phenomenon at the time data is gathered. Since the circumstances and environment of the research were dynamic, the data collected is likely to change and may therefore not be repeatable. Further, the chosen sampling strategy was based on the researchers' judgement and did not achieve population representativeness, which imposed risk to the research's replicability. Nevertheless, the applied research methodology was identified clearly, and the path to answering the research question was elaborated, which allows the research approach to be replicated and ensures the findings' dependability (Saunders, Lewis & Thornhill, 2016). Transparency was achieved by explaining methodological choices and analysis approaches. Reliability was threatened by the lack of standardisation in semi-structured interviews, participants', and researchers' errors and biases. To manage these threats an overall interview approach was developed and methodological rigor is established.

3.7.2 Validity

The validity addresses the measures' appropriateness, analysis' accuracy, and findings' generalisability (Saunders, Lewis & Thornhill, 2016). To enrich this, Ghauri and Grønhaug's (2005) division of validity is applied. The authors categorise it into four types: descriptive, interpretative, theoretical, and generalisable, which respectively define the degree of data accuracy, data interpretation accuracy, theory outcome adequacy, and finally, the extent to which the findings can be applied to other samples.

This study established validity by implementing various measures. Generally, the methodological rigor of this study promoted validity as it entails a transparent line of argumentation for each choice, ensuring a systematic approach. To specifically promote descriptive validity the research applied data triangulation and provided interviewees with the transcript to allow correction of factual errors. Specific member checks, however, were not feasible due to the full work capacity of the interview participants. Interpretative validity was achieved by ensuring the same understanding of the underlying concept, CBs, is given before the interview commenced and by asking clarifying questions on this especially relevant and complex topic. Further, interviews were transcribed to allow contextualisation of specific statements. Further, relevant quotes were shown in the findings to maintain the exact wording and meaning. Theoretical validity was promoted by describing and explaining the studied phenomenon in-depth and ensuring the fit of the chosen theory by validating these from experts. Further, the analysis was executed in a systematic, methodologically well-argued manner while the findings were presented coherently. The generalisability in this qualitative research presented itself as problematic as no population representativeness nor theoretical saturation was achieved. Therefore, the findings are only applicable to a similar sample.

3.7.3 Role of Researcher

When conducting interviews, the researchers were an influencing component of the data generated. Generally, three biases can occur: interviewer bias, interviewee bias, and participation bias (Saunders, Lewis & Thornhill, 2016).

The interviewer bias describes the impact the researchers' choice of words, tone, and non-verbal behaviour of the researchers might have on the interviewee's answers (Saunders, Lewis & Thornhill, 2016). It further addresses potential biases that may influence the interpretation of meaning and data analysis. The interviewee's perception of the researcher may have caused the interviewee to be biased (Saunders, Lewis & Thornhill, 2016). Given the exploratory interview character, the interviewee may also be sensitive to disclose specific information. Participants may have refused to answer any question in the interview, thus, limiting the sense of intrusion that could be caused by a semi-structured interview. The willingness to participate in the interviews affects the degree of participation bias (Saunders, Lewis & Thornhill, 2016). The time investment required for participating in interviews can present a barrier to some, hence influence from whom data is collected. Culture and language differences may have represented further barriers. To lower these, the researchers were culturally reflexive by critically evaluating their role as a researcher, reflecting the relation to each interviewee and considering possible cultural differences, and identifying ways to best interact with the participants (Saunders, Lewis & Thornhill, 2016). The researcher's involvement may have led to biases, but it also enabled gaining an in-depth understanding of the phenomenon and the relationship between different topics.

3.8 Chapter Summary

This chapter provided a rationale for the research design applied to answer the research question in this cross-sectional study. A pragmatist philosophy is chosen in combination with an abductive approach. Secondary data is reviewed critically and restructured for a more comprehensive understanding, which enables a contextualisation of the subject matter and outlines the information existent on CBs and debiasing strategies. Further, primary data is retrieved in semi-structured interviews with 11 participants, allowing an in-depth data collection establishing the basis to understand the transferability of debiasing strategies to prevention. The primary data collection is characterised by a concurrent collection and analysis of data. Finally, validity and reliability measures were carried out to ensure the high quality of the dissertation.

4 Findings

In this chapter, findings obtained from the interviews are presented systematically according to topics that emerge from the empirical data. The results focus on four dimensions, namely debiasing and prevention, CBs, causes of CBs, and lastly, strategies for managing CBs. It was found that interviewees gave very fragmented answers, as they did not draw connections between the addressed dimensions. To avoid distorting the meaning of the answers, the findings are initially presented in their dimensions and are only correlated in the subsequent analysis.

4.1 Differentiation Between Debiasing and Prevention

In the following, the differentiation between debiasing and prevention of the interviewees is analysed to achieve the research objective to examine how organisations can prevent CBs impacting CAs with the help of debiasing strategies.

Two key findings can be derived from empirical data. It was found that, on one hand, interviewees have different understandings of the differentiation between debiasing and prevention strategies among themselves, and, on the other hand, the interviewees' understanding of the differentiation differs from the literature.

Firstly, interviewees show a different understanding among each other when a strategy is used for prevention, and when a strategy is intended for debiasing. One interviewee described prevention as a measure before the launch of a CA and debiasing strategies as a measure that is applied while the CA is already in operation. When asked how to prevent CBs, the interviewee responded:

"[sic] by testing and evaluation for bias [the test] will pick up cognitive biases that have been injected along the development process." - *P4.*

A second interviewee assigned prevention strategies to the pre-testing phase of the CA development and debiasing to the post-testing phase. The testing phase was used also by another interviewee as a determined step that differentiates prevention and debiasing strategies, implied in the statement:

"One thing that you should do [to prevent] is doing more testing in the development period" - *P3.*

Another interviewee described prevention as a strategy before CBs enter the algorithm and debiasing once CBs have entered:

"So that we are confident to say that we did everything we can to prevent, you know, bias in algorithms in AI." - P10.

Some of them also changed their understanding throughout the interviews, while one did not know the term debiasing at all, indicating that this is a very theoretical concept:

"I have done organisation behaviour and those modules, but I never studied psychology aside, so I am not familiar with first term [debiasing]." - P11.

This shows that there is no uniform understanding among interviewees of the distinction between debiasing and prevention. Moreover, several strategies mentioned by the interviewees, such as sharing lessons learned, indicate that CB management is approached on an organisational level, wherefore a debiasing strategy for one is a prevention strategy for another individual. This is also addressed by P10's statement:

"But you learn the really meaningful lesson that can prevent all our other colleagues to make the same mistake".

In these cases, the erroneous decision-making of one individual is used to prevent others from being impacted the same way by CBs. For the individual committing the error, the strategy is a debiasing strategy, as the lessons learned will benefit the next decision. This also explains why, when asked specifically how debiasing strategies could be used to prevent CBs from entering CAs, respondents referred to the same or stated additional strategies within their understood differentiation, but never fundamentally different strategies. Even when the question was changed and it was asked about reasons for the prevention and debiasing, the interviewees still did not show any differentiation between the two concepts. Only reasons why CB management should and should not be carried out were stated, such as the cost of resources or the constant evolution of the CA were mentioned. Consequently, no factors could be identified to assess how to use debiasing strategies for the prevention of CBs.

Secondly and most relevant, literature relates its differentiation between debiasing and prevention to a single decision of an individual and whether it has been erroneous before. Most, interviewees relate their differentiation on the basis of a development process and not an erroneous decision. Interviewees determine different points in the process to determine whether it is classified as a prevention or debiasing strategy, as the examples above demonstrate. Their distinction is not related to whether, as suggested in the literature, a previous decision of an individual was distorted or not. Therefore, the differentiation between debiasing and prevention strategies from the interviewees differs from literature.

The question of how debiasing strategies can be used for prevention is consequently answered by the finding that organisations do not differentiate according to the literature, as they apply CB management on an organisational level, nor uniformly among each other. In addition, some strategies are used simultaneously for debiasing and prevention. This means that organisations do not look at strategies in the literary categories of debiasing or prevention. They solely undertake CB management. This makes the indented research if organisations can use debiasing strategies for prevention redundant, as the problem remains on a theoretical but not on a practical level. Also,

a test in practice is invalid, as organisations do not have a uniform understanding of the differentiation, or do not differentiate at all.

Since practice and literature do not differentiate on the same level between debiasing and prevention, it is further examined whether the empirical strategies can still contribute to the literature. Even if the mentioned strategies are not only directed at an individual, they may still be valuable for this purpose. Since the literature does not provide prevention strategies, the following analysis focuses only on strategies from the CB- and cause-specific literature. This examines whether the strategies mentioned in the interview can possibly be regarded as new to the literature.

4.2 Findings on Cognitive Biases

The following section presents an overview of the findings related to CBs, subdivided into identification of CBs and overall findings. For an identification of CBs, the literature must be used to clearly assign statements to an explicit CB. The identification is essential, as part of the literature argues that strategies are only applicable to certain CBs. To be able to conduct a later analysis of the strategies identified in the interviews, CBs must be identified as a first step.

4.2.1 Identification of Cognitive Biases

Following, CBs relevant to the development of CAs are identified. Herefore, the literature must be taken into consideration to allow the identification of CBs. Firstly, this enables an understanding of which CBs are particularly likely to distort decision-making in the development process of a CA. Secondly, the identification of CBs allows a following analysis of relevant debiasing strategies, as CBs partly form their basis.

Several CBs were mentioned, implied, or demonstrated in the interviews, which are analysed according to literature, allowing their identification. In total, 11 biases are determined, out of which two were explicitly mentioned, seven were explained, and two were solely demonstrated in the interviewee's way of answering.

The blind spot bias was identified as a result of an interviewee's explanation of how others might be biased, while not reflecting on its own CBs throughout the interview. This demonstrated the clear identification of CBs in others but not oneself, indicating the blind spot bias (Osoba & Welser, 2017; Pronin, Lin & Ross, 2002). Moreover, another interviewee directly addressed the difficulty to identify own CBs.

When describing CBs, several interviewees showed a tendency to talk about CBs in data and refrained from talking about personal CBs that might impact the development. This tendency to attribute externalities for failure can be defined as the self-serving bias (Zuckerman, 1979).

The following statement suggests that mere professionalism enables individuals to ensure unbiased answers, which was added to also be linked to common sense:

"I think the professionalism of a consultant in this domain is to make sure that you build a conversational agent that gives the right answer" - P6.

This demonstrates overconfidence in one's own abilities and indicates the presence of overconfidence bias as defined by Brenner et al. (1996) and Keren (1990).

Additionally, the false consensus bias, as defined by Krueger and Clement (1994), was described by two interviewees. One interviewee states that its experience led to a decision based on the assumption that everyone else's experience is similar, which was classified as a common issue by another interviewee.

Selection bias, as defined by Bareinboim and Pearl (2012), was also determined in several interviews.

"I guess the bias you can see [is that] you limit your sample size to just specific audiences that you want, and you're not looking deep enough into different [...] or diverse enough demographic[s]" - P5.

Distorted mental reasoning leads to an incorrect assessment of, for example as one interviewee addresses, a sample size that is insufficiently and does not reflect reality or target groups are completely excluded from the data set.

Next, the framing bias was detected, in line with Wright and Goodwin (2002), when two interviewees reported about collaborations with clients. These clients had pre-set preferences in mind, for example, a gender preference of the CA. In a client-consultant relationship this is a CB that consultants might be aware of, but not act against, as stated by one interviewee, to achieve customer satisfaction. Additionally, another interviewee mentioned that an industry preference for one technical solution may frame individuals' choices, falling under the same CB.

Also, the status quo and the habit bias were described by the same interviewee who mentioned that no changes were made from the previous project. This falls under the status quo bias according to the definition of Samuelson and Zeckhauser (1988), which is also indicated by the statement below:

"It is just a balancing act, okay, we know we need to innovate, we know we need to modernise, we know we need to improve, but the minute you do that, you step off the ledge and that is scary. So, you balance between the stable conservative system and doing something new." - P2.

In the case in which individuals stick to what they are familiar with and try to change the current pattern as little as possible, the habit bias was identified (Hogarth, 1987; Slovic, 1975).

When selecting different options, loss aversion may play an important role. It was mentioned that individuals tend to choose safe options, due to the tendency to weigh potential losses higher than potential gains (Kahneman & Tversky, 1979; 1984), which distorts their rationality impacting the CA's neutral behaviour.

Assuming a solution is good because others are using or buying it, as indicated in the interviews, is the classical definition of the bandwagon effect (Leibenstein, 1950). This finding is demonstrated in P11's statement:

"Some banks and some institutions [...] burned a lot of money by just doing something that the market is saying that is cool, or we want to have this, this is the hype right now.".

Further, in the technological sector, many errors are the result of overoptimism. It was noted that the initial public enthusiasm for new technological solutions leads to a hype that raises customer expectations. This phenomenon is also based on the bandwagon effect (Leibenstein, 1950) and according to an interviewee may cause a definition of business requirements that the technology is simply not capable of, resulting in erroneous behaviour of a CA, as the CA is over-challenged. One interviewee observed that factors, such as the academic status of users, are generalised, thus, incorrectly assessed. This is caused by attributing one characteristic of a group member to the overall group, as defined in the group attribution error (Allison & Messick, 1985). The CA is then trained with these wrong assumptions, resulting in errors in the actual application, due to the gap between expectations and reality.

4.2.2 Overall Findings on Cognitive Biases

In the course of all interviews, several CBs proved a high likeliness of impacting the decision-making in the development process negatively. Their importance is reflected in their ability to influence individuals' understanding of their own intellectual limitations, their ability to recognise CBs, and the potential harm it could cause in relation to CAs. This implies their importance, wherefore these CBs are henceforth used for further analysis: 1) blind spot bias, 2) self-serving bias, 3) overconfidence bias, 4) false consensus bias, and 5) selection bias. The first three CBs are intertwined and influence individuals strongly. All three affect individuals' misleading assessment of their own intellectual abilities and whether they assign CBs to themselves, thereby, accepting their behaviour as faulty and becoming aware. These CBs are engrossed in individuals and can affect decisions strongly. Hence, they may impact the entire development process while remaining difficult to trace back. Therefore, it is fundamental to manage the CBs before they distort the CA. The false consensus bias is highly relevant because, when transferred by humans into a CA, it poses the risk that the data selected for the CA may not reach representativeness. Thus, it may discriminate against a user group. This is linked to the selection bias, which is particularly relevant for this specific application and has a great influence on the data selection for a CA and consequently, the data representativeness.

Only a very limited number of CBs were mentioned by interviewees, even though they presented a diverse interview participant selection. It is noteworthy, that mentioned CBs were rarely related to the interviewees themselves. Interviewees tended to talk about clients' CBs or CBs in general, a reflective position on their CBs, however, was rarely taken. Further, CBs were mostly described but not mentioned explicitly. Moreover, most interviewees had to be redirected to the topic of human CBs multiple times, even though it was defined and reminded of in the communication before the interview and before questioning commenced. A tendency to talk about data bias and, in particular, biases the CA displays were identified.

Overall, a difference in the reflectiveness of interviewees was identified. Especially, individuals in superordinate positions showed more reflectiveness as they concede the possibility that individuals might be influenced by their character or feelings. Additionally, they were able to explicitly name CBs. Individuals in operational job positions did not show this reflection and sometimes exposed CBs with their behaviour rather than naming or describing them. It was further analysed that individuals within the operational area have a background which is more specific to one area in the development process, while at the same time having a shorter overall work experience compared to individuals in superordinate positions.

4.3 Findings on Causes of Cognitive Biases

This section gives an overview of the findings with regard to the causes of CBs, divided into the identification and overall findings. To allow for later analyses of the strategies mentioned in the interviews, causes must be identified first.

4.3.1 Identification of Causes of Cognitive Biases

Some interviewees reflected that decisions are sometimes made on the basis of assumptions, intuition, or human gut feeling. These and other comments allow for the identification of a variety of causes for CBs from the interviews. The identified causes are presented in 13 categories emerging from the empirical data.

Lack of Guidance

Missing guidance was identified as one of the causes of CBs. A lack of superordinate guidance by law or set standards to manage CBs was mentioned. CB management is, therefore, the responsibility of the organisation and hence, often directed to the individuals involved in the development process, as indicated with the statement of P5:

"So it's up to me to kind of figure out like, oh, I need a bigger audience to test this.".

According to the interviewees, guidance is resource-intensive and consequently, often refrained from. While some organisations have no guidelines at all, others publish guidelines only sometimes, but individuals might not be aware of the exact content. One interviewee described guidelines as common sense and believes that they are automatically implemented by a good team. Another problem addressed is loosely formulated guidance, which allows room for own interpretation and personal assumptions.

The lack of sufficient guidance was seen problematic by employees at a large corporate and in smaller organisations, as stated by one interviewee. In general, there is a tendency for individuals working operationally in the development process to see the problem of supervisor guidance more

prominently than individuals in superordinate positions. Individuals working operationally notice that guidance is often lacking, as indicated by P8:

"I don't get that specific kind of support.".

Therefore, it is found that intuition must be relied upon whereby CBs can occur more easily. Conversely, too much guidance can also prevent the detection of CBs, as, for example, specified test processes remain unchanged and allow that the same blind spots inherent in their design may be overlooked repeatedly.

Unawareness

Some of the interviewees showed great unawareness about CBs, their impact, and importance in the development of a CA. It was argued that the current state of the technology, requires structured answers to be inserted in the system, leading to the statement:

"I think in the current situation it [CB] is not that much of an issue"- P7.

which was perceived likewise by their superior. It was seen as a problem that will gain importance in the future as the technology of CAs evolve, but at the current state, it was stated that professionalism and common sense ensure that unbiased answers are given by the system. Still, some reflection was identified, however, the awareness just remained on the level of acknowledging CBs in other people, not themselves.

A difference can be observed in the responses of individuals according to their job positions. Individuals who work operationally, specialised on one respective step of the CA development process, showed to be less aware of CBs and causes, and required more interview guidance. Individuals in superordinate positions provided a more comprehensive view of the development process and its related CBs and causes. Due to the existing awareness of the topic, they needed less guidance in the interview and tended to stick to the right topic. Only two interviewees working in superordinate positions were an exception to this observation, as they proved to have knowledge about the subject but missed to acknowledge its general relevance. In addition, interviewed consultants, who are confronted with CBs from clients, stated that they often simply have to accept clients' CBs without identifying their causes, as they are not in a position to question them.

Lack of Resources

CBs occur or remain undetected because resources are limited. For example, it can affect the amount of data used for training and testing which is fundamental to the level of detail and ultimately, the quality of training and testing. This problem is summarised by interviewee P5:

"I hesitate to call it a bias. It's more how much resources do you have to complete these things.".

The extent to which data is available for training and testing is another crucial aspect that can impact the prevention but also the detection of CBs that might have already entered the development process. It was stated that training and test data should be as diverse as possible to prepare CAs for a real-world environment in which a variety of data is processed. This, however, is often not given to the extent required due to the amount of resources it demands.

Stress

Stress is identified by interviewees as negatively affecting individuals' ability to assess the susceptibility of CBs. This cause was identified only by individuals who are closely connected or working in the operational area of the CA development process. Stress can be increased through time pressure as mentioned by some interviewees of which two identified that it can lead to individuals taking shortcuts, as explained by P5:

"And there could be a point where no one is questioning what's happening simply because the deadlines [are] happening too fast and then we just have to get them out the door.".

Time pressure originates, for example, from a lack of communication such as in last-minute changes or product specifications on short notice. This shows the influential effect of functioning internal communication and alignment between involved departments. A factor, identified by an interviewee, impacting the alignment of relevant parties and thereby, the level of stress, are individual KPIs which can hamper the collaboration between those departments.

Lack of Education

Several interviewees see a lack of education that creates a barrier to detect and manage CBs. It was stated that sometimes it is difficult to detect CBs because they are deeply integrated into the CA, which suggests a lack of understanding of CBs, and how they impact a CA. These findings suggest that the concept of CBs is not understood fully and that education is needed.

Human Factor

The majority of interviewees acknowledge the human factor as a cause. The interviews revealed that areas particularly susceptible to CBs are the ones allowing room for personal preferences and consequently, for CBs. Especially, small-talk creation and data selection were mentioned as highly personalisable processes. This dilemma is summarised by P3's statement:

"Part of you is in your bot".

Besides that, career aspirations, personal interests, and job satisfaction are mentioned to further impact the extent to which employees are motivated to be aware of CBs and avoid their influence.

Localisation

Another cause for CBs is the selection of preferences according to local needs. Localisation requires selection of these preferences which hold the risk for selection biases to enter. It further increases the complexity as more variables must be considered. For example, in the case of a voice assistant targeted at a wider population, theoretically, training for every relevant accent is required. The awareness of this fact is given by some interviewees. However, it is not feasible to train a system on every existing accent including their permutation, as implied by P5's statement:

"That whole bias is pretty obvious, but it's hard [...] to do accents".

This is why a selection must be made. It was noticeable in the disclosure of this cause that only interviewees working in operations identified this factor as influential.

4.3.2 Overall Findings on Causes of Cognitive Biases

From an overarching perspective, it can be concluded that the causes for CBs are on the organisational and process level and are extremely interrelated. The majority of causes, such as lack of guidance, stress, lack of education, or localisation, can be traced back to a lack of resources, which is thereby identified as the main cause of CB occurrence. CB management has not yet reached an industry-standard, wherefore it is resource-intensive and often not prioritised by organisations, which in turn results in a shortage of resources assigned to this matter. Also, missing laws or standards lead to a lacking awareness and in-transparencies in the development. For this reason, it is often not possible to precisely distinguish between the categories in the causes, and their interconnection must be taken into account. Furthermore, it was found that, compared to the statements addressing CBs, a higher overlap of statements from different interviewees mentioning the same causes could be identified.

4.4 Findings on Management Strategies

The following section provides an overview of the management strategy findings, separated into the identification of strategies, interrelation of CBs and causes with strategies, and overall findings.

4.4.1 Identification of Management Strategies

Several strategies for the management of CBs were presented in the interviews, which were grouped according to the interviewees' statements to allow an overview. As described above, no distinction can be made between prevention and debiasing, hence, the following does not distinguish between the concepts. It should also be noted that in some cases strategies are interlinked and even overlap in categories, which are discussed again at the end. In total, seven strategies are identified.

<u>Extension of the Perspective</u>

One approach to manage CBs, which was mentioned by several interviewees, is to actively involve further expertise from within the organisation or from externals of the development process, as explained by one interviewee:

> *"I think it's good to bring it [issues] up with them because they can have different perspectives that I might not have thought about because they are not regularly working with speech." - P8.*

Different ways of achieving a change of perspective were suggested. For example, talking with customer service, discussing decisions with other parties, who can add an outside perspective, or

by considering worst- and best-case scenarios. The latter can be used especially when decisions must be made quickly.

Promotion of Awareness

Awareness about biased CA was identified as a method to manage CBs by several interviewees. It can be raised through communication, for example with transparent explanations of a decision process to draw awareness to possible entry points while removing emotions and emphasise factual information. Several interviewees also stated that raising awareness about the methodological approach to building unbiased CAs improves the common understanding of CBs impacting CAs. Moreover, it can be helpful to stimulate exchange within the team by disclosing personal information, such as educational background or culture. This raises the awareness and transparency of possible CBs of team members, allowing individuals to pay closer attention to each other's CBs. Additionally, it was noted that awareness is a result of interest in motivation for and focus on the job which can be diminished when stress is too high. Reducing stress and ensuring employee motivation are, therefore, identified as a strategy to increase awareness. However, P11's statement must be noted:

"Just being aware of it doesn't mean you can avoid it".

Education

Several interviewees revealed that education about CBs can serve as a measure, across all hierarchical levels. Hereby, it was emphasised that a balance between force and willingness to learn should be ensured. Problematic to this strategy is, as indicated by P11, that:

"People who need it the most are probably not the ones going on this training or reading those books".

Additionally, one interviewee understands that the wide range of academic knowledge requires extensive effort for individuals to gather all essential information. Furthermore, it was pointed out that no best practices in this area exist from which organisations can take an example. However, one interviewee noted that learned lessons gained from previous mistakes can be shared with others to prevent them from making similar mistakes.

Provision of Guidance

Guidance to ensure that individuals perform more reflectively can be provided through organisation policies, checklists, or guidelines, as mentioned by several interviewees. As examples, the policy of double-checking the design at all times or strict adherence to an AI bias playbook were mentioned. Similar to the education strategy, there is a lack of best practices and international standards, which complicates the compilation of guidance. Another way to provide guidance can be achieved through collaborative KPIs or values. These guide decision-making and influence the outcome, while limiting the entry of CBs. Guidance can also be given by a superior team or person who has oversight. A large car manufacturer, for example, formed a so-called "halo-team" that

assembles general lessons learned from across the organisation and passes these on to individuals in need. This soft approach helps to ensure an open attitude towards issues, like CBs, which in turn facilitates the collection and targeted placement of feedback.

Provision of Feedback

Feedback, which can be given by third parties outside the decision-making process, was identified by interviewees as a strategy. Feedback can be given, for example, by customers, or, as one interviewee estimates to be most common, on a personal level within the organisation. Both positive and negative feedback can be used to calibrate the decision-making process and make individuals less biased. Since it is difficult for individuals to identify CBs themselves, feedback is essential for the management of CBs, as stressed by one interviewee. Double-checking decisions can represent a further method of providing feedback. In the interviews, two different approaches were identified. Either the development process is systematically designed that after each stage a new person double-checks the decisions or by a partner analyst double-checking the respective decisions.

Accountability

The enforcement of accountability promotes awareness and motivation to reduce or eliminate CBs and was mentioned in several interviews.

"Accountability is a fundamental prerequisite for technology" - P2.

and ensures trust in a decision. One interviewee indicated that by holding individuals accountable, safety, fairness, and explainability of the technology can be ensured. The introduction of a Hippocratic oath or communicating the consequences of misbehaviour are further approaches to enforce accountability which reduces the likelihood of CBs in decisions.

Conversational Agent-Specific Strategy

Since the technology of a CA is highly complex and difficult to understand, transparency is seen as a significant part in the management of CBs. Seldomly, data entailing all variables are available wherefore data must be selected. Several interviewees stressed that the correct data selection is important as it holds the risk of CB occurrence. Depending on the determined target group, it is crucial to find data that reflects the true population of users. One interviewee stated that this can be achieved through data transparency by establishing a chain of custody including the origin and collection method of the data.

Another essential component to the management of CBs is testing the CA which detects and directs awareness to biases potentially caused by CBs. Interviewees identified a large number of errors that become apparent in the testing, as this phase allows error identification which can subsequently be managed, suggesting refinement as another strategy. Testing, however, rarely focuses on the identification of CBs in particular. Several interviewees stated that as the live environment is particularly crucial to evaluate the CA's performance, but rarely integrated into pre-production

testing, quality assurance also takes place after production. The more diverse the tests, the higher the chance of detecting and eliminating CBs.

Another approach can be to simply avoid decisions and pass them on to the CA users, whereby CBs are sidestepped. Also, one interviewee suggested that ambiguity in conversations can be restricted. With this practical approach, answer-response pairs can be created, which allows straight-forward conversations and thereby, limits ambiguity and possible integration of CBs. Moreover, the CA can be restricted to only being able to give answers to an unambiguous conversation, which therefore decreases the possibility for CBs to be transferred into the CA.

4.4.2 Interrelation Between Cognitive Biases, Causes, and Strategies

As already addressed in the introduction of section 4, interviewees did not align CBs, causes of CBs, and strategies. Therefore, it was analysed whether the mentioned strategies of the interviewees address the identified CBs and the mentioned causes of CBs. To determine their relations, CBs and subsequently, causes are interrelated with the strategies. The aim is to assess whether the strategies are effective against the identified CBs and causes.

Firstly, the identified CBs, namely 1) blind spot bias, 2) overconfidence bias, 3) self-serving bias, 4) false-consensus bias, and 5) selection bias, are interrelated with the mentioned strategies. For this purpose, the drivers of the respective CB had to be considered to assess if the mentioned strategies address the identified CB. The following presents the highlights of the findings as presented in figure 3.

Strategy Categories	Strategies	Blind Spot	Over-confidence	Self-serving	False Consensus	Selection
Extension of the Perspective	Further expertise	✕		✕	✕	
	Change of perspective	✕		✕	✕	
	Worst-/best-case scenarios					
Promotion of Awareness	Transparent decision process					
	Explanations about the CA					
	Communication CA methodology					✕
	Personal exchange within the team		✕	✕		
	Reducing stress					
	Ensuring employees' motivation					
Education	Education	✕	✕	✕		✕
	Sharing lessons learned		✕	✕		✕
Provision of Guidance	Policies, checklists or guidelines	✕			✕	✕
	Collaborative KPIs					
	Oversight	✕	✕	✕	✕	✕
Provision of Feedback	Feedback	✕	✕	✕	✕	✕
	Double-checking	✕	✕	✕	✕	✕
Accountability	Accountability				✕	✕
CA-specific Strategy	Transparent data selection				✕	✕
	Diverse testing				✕	✕
	Refinement after testing				✕	✕
	Decision avoidance					
	Restricted application					

✕ Cognitive bias and strategy correspond

Strategy makes cognitive bias redundant

Figure 3: Interrelation Between Cognitive Biases and Strategies (own illustration)

The strategies oversight, feedback, and double-checking address all identified CBs and can be applied to detect individuals' CBs with the help of the involvement of a third person. This person can identify CBs for individuals, eliminating the need for them to identify their own CBs. The CA-specific strategies of diverse testing and refinement after testing, only detect and address biases in CAs that may be caused by CBs. The strategies, therefore, do not address CBs directly. Similarly, the strategies to avoid decisions or to restrict the application are not directed at the CBs, but make their management redundant. Either because the decision is not taken by an individual, or because it is unlikely that the CA discloses CBs, due to the limited scope, which reduces the likelihood that CBs enter and ultimately be displayed. CB education can contribute to the mitigation of CBs in almost all cases as it can make individuals aware of CBs and provide strategies to manage CBs.

Almost all strategies mentioned from the interviewees address one or more CBs, except for the strategies of explanations about CAs, reducing stress, transparent decision process, ensuring employees' motivation, and collaborative KPIs. These, therefore, have no direct effect on CBs.

After linking CBs and strategies, the same analysis approach was conducted by relating causes of CBs and strategies (see figure 4). Therefore, the mentioned causes, namely 1) lack of guidance, 2) unawareness, 3) lack of resources, 4) stress, 5) lack of education, 6) human factor, and 7) localisation were interrelated with the mentioned strategies.

Strategy Categories	Strategies	Lack of Guidance	Un-awareness	Lack of Resources	Stress	Lack of Education	Human Factor	Localisa-tion
Extension of the Perspective	Further expertise		✕				✕	✕
	Change of perspective		✕				✕	
	Worst-/best-case scenarios							
Promotion of Awareness	Transparent decision process	✕						
	Explanations about the CA		✕			✕		
	Communication CA methodology	✕	✕			✕		✕
	Personal exchange within the team		✕					
	Reducing stress		✕		✕			
	Ensuring employees' motivation		✕				✕	
Education	Education	✕	✕			✕		
	Sharing lessons learned		✕			✕		✕
Provision of Guidance	Policies, checklists or guidelines	✕	▨			▨		▨
	Collaborative KPIs	✕			✕			
	Oversight	✕	✕			✕		✕
Provision of Feedback	Feedback		✕			✕		✕
	Double-checking		✕					✕
Accountability	Accountability						✕	
CA-specific Strategy	Transparent data selection							✕
	Diverse testing		✕				✕	✕
	Refinement after testing							✕
	Decision avoidance	▨	▨	▨	▨	▨	▨	▨
	Restricted application	▨	▨	▨	▨	▨	▨	▨

✕ Causes and strategy correspond

▨ Strategy makes cause redundant

Figure 4: Interrelation Between Causes of Cognitive Biases and Strategies (own illustration)

The analysis showed that each strategy, except for worst- and best-case scenarios, addresses a minimum of one mentioned cause of CBs. The strategies of decision avoidance restricted application and partly, policies, checklists, and guidelines do not directly address causes, but rather make them redundant. Similar to the interrelation analysis of CBs and strategies, these strategies bypass the causes' influences. From the perspective of the causes, the strategies address all causes except the lack of resources. Lack of resources was not addressed but was prominently mentioned and is seen as one of the main drivers of the other causes.

At first glance, it appears that the interviewees mentioned strategies in accordance with the causes they mentioned. However, not every match of strategy and cause was mentioned by the same interviewee. For example, the strategy addressing the localisation of CAs was introduced by P4 who addressed the chain of custody to ensure a transparent data selection and unintentionally provided a strategy to the cause mentioned by P5. Further, it was observed that in some cases causes mentioned, like the individuals' job satisfaction impacting the motivation to manage CBs,

were not picked up upon anymore when the conversation shifted to CB management strategies. This shows that even if strategies and causes matched, this match may not be attributable to an individual interviewee.

4.4.3 Overall Findings on Strategies

Overall, it might appear like the majority of strategies target general operations, however, CA-specific strategies were discussed more extensively and in-depth in the interviews. Testing, in particular, was highlighted by the majority of interviewees as crucial for CB management. Often the interviewees did not specify when and partly why a strategy is in place, for example in the strategies of education, change of perspective, or feedback. None of the interviewees mentioned that pre-analyses, with regards to the identification of CBs or causes of CBs, are conducted before strategies are applied. This implies that strategies are used rather broadly than selectively to address as many CBs and causes as possible. This can also be supported by the analysis of the relation between causes of CBs and strategies, as the link between them is not always made by the interviewees individually.

In general, strategies were often given based on examples rather than in abstraction. Additionally, several strategies are interlinked. Education and awareness are related as they build on each other, for instance, when individuals are taught about CBs they may be more aware. Also, the strategies feedback and guidance are interrelated, as feedback provides guidance for individuals.

Almost half of the strategies, such as double-checking, feedback, and oversight, consider additional individuals to be involved in the management process. This suggests that organisations recognise that in many cases a single individual has difficulty in identifying and managing its own CBs. The tendency to apply strategies involving third parties is supported by the interviewee confirming the difficulty of detecting own CBs.

4.5 Chapter Summary

Overall, the research has found that CBs affecting CAs are mainly those that make individuals blind to CBs or other perspectives, and CBs address a core activity in the CA development process, which are 1) blind spot bias, 2) self-serving bias, 3) overconfidence bias, 4) false consensus bias, and 5) selection bias. The main cause leading to CB occurrence is a lack of resources, which restricts individuals in their actions and is interrelated with various other causes. Further causes are lack of awareness or lack of education, which are highly correlated. Besides, a difference in the level of awareness of CBs was detected depending on different job positions. Individuals in a superordinate position showed a greater degree of awareness than those in operational positions.

Most significantly, it is found that practice does not distinguish at all, nor uniformly between debiasing and prevention. When a differentiation was made, the differentiation was based on the development process and not on the decision-making process, as in the literature. Additionally, several strategies mentioned by the interviewees were seen as both debiasing and prevention strategies.

5 Discussion

In the following, empirical findings are discussed and linked to literature. Firstly, the difference between debiasing and prevention is addressed and its practicality is questioned. Subsequently, the limited reflectiveness and awareness of interviewees regarding CBs are addressed and related to different job positions. Also, the causes of CB occurrence are examined in more detail and linked to the literature. Particular focus is placed on the question of whether the lack of guidance is caused by the companies themselves or whether it is a problem for the entire industry. Thereupon, the lack of resources, as the main reason for the occurrence of CB, is debated. The section ends with a discussion on management strategies. In addition to examining tendencies in the direction of the empirical strategies, previous analyses of CBs, cause, and strategy interrelations are compared with the literature. This enables an assessment of whether the strategies mentioned in the interviews can contribute to the literature.

5.1 Fluid Concept of Cognitive Bias Management

This research has found that organisations distinguish rather fluidly between prevention and debiasing measures. Strategies represent preventative measures for one individual and debiasing measures for another.

However, differentiating between these strategies could be more effective, as the strategies may be applied more consciously as they will be designed to target one specific purpose. It is debatable whether a strategy that addresses both, debiasing and prevention, is effective since literature meticulously requires pre-analysis of what the strategy targets, a particular CB or cause, suggesting that the target is highly relevant. Nor addressed literature the use of debiasing strategies for prevention, which makes a mutual utility debatable. By differentiating the strategies, organisations would be able to assess the effectiveness of the strategies more clearly, as the line where one strategy ends and the other begins would be drawn more strictly. It might, therefore, be easier to trace the effectiveness of each strategy. At the same time, CB management is only one of many components in the development of CAs, wherefore a strategy may be particularly likely to be applied if it allows addressing multiple uses, as this reduces costs. Therefore, it might be that organisations do not differentiate to increase the feasibility of CB management. A separation would require more strategies, as they have to be specifically designed to target either debiasing or prevention. This, however, might not be feasible given that organisations lack resources. Therefore, the decision to differentiate between debiasing and prevention strategies can be assumed to be based on a choice between feasibility and effectiveness.

5.2 Reflectiveness and Awareness

One topic arising multiple times throughout this research is individuals' reflectiveness and awareness of CBs and subsequent topics such as CB causes and debiasing strategies.

The findings and analysis have surfaced that there is a higher overlap between the respondents' statements regarding the causes of CBs than the one regarding CBs. Moreover, CBs and strategies are presented descriptively or with the use of examples. This shows a lack of ability to express both

precisely. In addition to that, CBs, the causes of CBs, and strategies were not addressed coherently by the interviewees. Their fragmented responses did not indicate that relations between these were drawn, even though the analysis presented that they interrelate. This may be either because interviewees find it difficult to establish a link, due to the depth of the topic and their limited knowledge, or because the interview questions may not have been sufficiently aligned to inquire about a relation.

The limited awareness can stem, on the one hand, from the fact that CB related knowledge is mostly found in academic, psychological literature, and might not be part of the background from individuals working in the field of CAs, which is confirmed by literature as it only identifies four AI-related CBs (Ayoub & Payne, 2016; Challen et al. 2019; Osoba & Welser, 2017; Pronin, Lin & Ross, 2002). While on the other hand, the mass of information available in the psychological literature, might create a barrier to address the topic, as it might seem overwhelming and requires education. Therefore, it might be simpler to identify which factors influence one's ability to reflect and think rationally. Moreover, causes already indirectly explain what can be done to overcome CBs, which creates a more straightforward picture. The causal link, therefore, offers two pools of knowledge from which the interviewees were able to extract insights from, the causes itself and the causes which are addressed by strategies that they are aware of.

By addressing causes, the individuals are less animated to talk about their or other’s mental contamination, which is required when talking about CBs, and are able to direct the fault to other externalities. This might make it easier to talk about, even though this is usually regarded as a behaviour impacted by the self-serving bias.

The identification of CBs requires greater reflectiveness, which is hard to achieve due to identified CBs such as blind spot, overconfidence, and self-serving. These CBs may also have had an impact on the mention of strategies, many of which are strategies that require third-party execution. Therefore, the focus is placed on an external attribute, rather than the interviewee itself, which also may have been due to the self-serving bias. This also concerns the statements regarding causes. When addressing CA-specific causes, such as the localisation, individuals closer to the development of the CA showed more awareness.

Still, interviewees revealed more AI relevant CBs than literature (Ayoub & Payne, 2016; Challen et al. 2019; Osoba & Welser, 2017; Pronin, Lin & Ross, 2002), which allowed the identification of not yet mentioned CBs in the context of CAs, namely self-serving bias, overconfidence bias, and false consensus bias. This shows the importance of literature to identify CA-relevant CBs as the interviewees were already analysed to be partially affected by CBs hindering their ability to reflect on CBs, and still presented more than literature. Moreover, it is noticeable that especially individuals in superordinate positions are more reflective of CBs as they proved to be more self-aware. Individuals working operationally, even though most of them work directly on the CA's core, show less self-awareness and thereby, less reflectiveness. Reasons for this could be that individuals in superordinate positions may be experienced professionals and have a more comprehensive overview of the entire development process and all parties involved. Thereby, it is more evident to them if they make a mistake themselves. However, particularly individuals working on the CA should reflect on their CBs, or at least be guided accordingly, as their impact

is significant. As it was analysed that individuals in operation have a more limited background, it might suggest that this restriction is reflected in their awareness of CBs.

It could be argued that individuals in superordinate positions are more reflective as they have received training for this. At the same time, the importance of operational individuals may be undermined and disregarded. This problem is shown in a study that reveals that only 24% of employees in non-management positions stated that their professional background prepared them very well for their job position, whereas the number of manager positions was at least 12% higher (Statista, 2018). At the same time, this study showed that as the hierarchy level increases, more additional training is offered. This goes hand in hand with an interviewee's statement that training is usually provided to people who do not need it most. It is important to note that studies should be interpreted cautiously, as they only give an average result for the US and do not provide precise figures for CB training. As a result, it could be concluded that it is important in an organisation to achieve CB awareness along the entire hierarchy chain.

It could also be argued that it might be sufficient for only one individual in a management oversight position to have awareness of CBs. Given that awareness is the first step towards the elimination of CBs, and that strategies often require the involvement of a third party (Wilson & Brekke, 1994), individuals in superordinate positions could take responsibility for the implementation of CB management strategies. Due to superordinates' training experience, they are able to take responsibility for raising awareness among their subordinated employees, implementing strategies to counteract CBs, and guide to subordinate positions. This could reduce the degree of awareness required by each subordinate and still ensure efficient CB management in an organisation. However, operational workers clearly identify a lack of guidance. For effective CB management, organisations should therefore either emphasise CB training for all levels in the hierarchy or ensure guidance for individuals working operationally.

5.3 Organisational Causes of Cognitive Biases

Part of the literature refers to cause-specific debiasing strategies. Therefore, it is relevant to determine causes that are not yet addressed by literature which could indicate that respective cause-specific strategies can contribute to it.

Generally, the causes identified in the interviews address a different context than causes presented in the literature, namely heuristics, artifacts, and error management (Haselton, Nettle & Andrews, 2005). Empirical causes are practical and business-related, while literature approaches causes from a psychological perspective (Haselton, Nettle & Andrews, 2005). The empirical causes represent an extension to literature but are on a more organisational level than the psychological causes, heuristics, error management.

Nevertheless, a similarity between empirical causes and causes presented from authors of cause-specific debiasing strategies is found. Here, empirical causes correspond to literature in some cases, but may also represent an extension, as seen in figure 5.

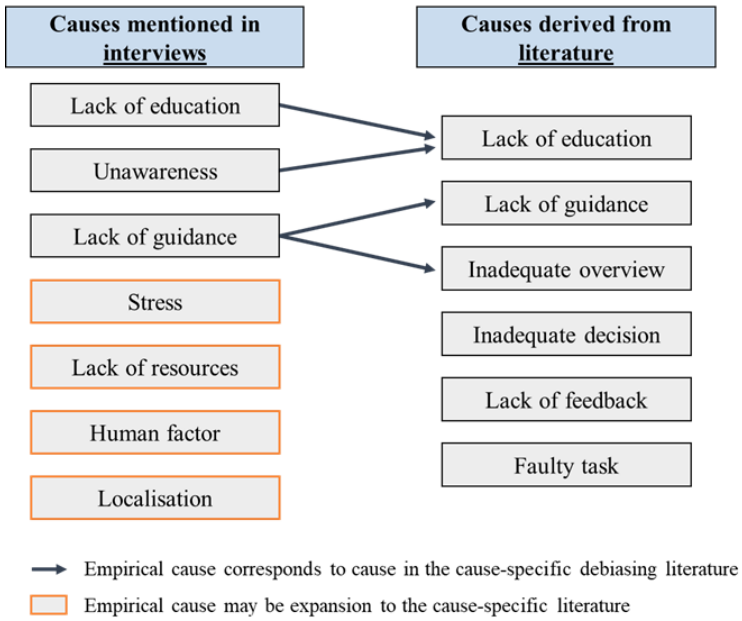


Figure 5: Cause Comparison (own illustration)

While interviewees describe causes focused on organisational- and process-level, the literature considers the task-level and the individual as the main cause of CB occurrence. This shows how practice considers a wider management context. It is noteworthy that the interviewees mention stress, a lack of resources, human factor, and localisation, which are not found in the literature and therefore, may contribute to it. These causes stem from a business perspective, which may explain why debiasing strategies in literature, which are mostly found in a psychological context, do not address these. Stress, lack of resources, and the human factor are closely linked to the explanation of why individuals are not rational decision-makers. Stress and resources both address time and monetary limitations. The human factor, on the other hand, mostly addresses individuals' motivation. Therefore, it is noteworthy that debiasing strategies do not address these aspects, as they are clearly integrated in descriptive decision-making (Simon, 1955). Localisation, on the other hand, is very specific to CAs and shows that literature provides a general approach to managing CBs, which however misses out on specific problems in certain application areas.

Besides that, a high level of agreement is analysed under the cause of a lack of guidance. At first glance, it might seem easily addressed by establishing guidelines, however, it was found that organisations rarely actively develop those. An explanation for this can be given by Osoba and Welser (2017), who observed a lack of industry standards and laws which leave organisations without a common, compulsory standard (Madiega, 2019). Consequently, superordinate individuals do not have guidance to provide to operational individuals, who see the lack of guidance as an issue resulting in CBs. This demonstrates how the speed of technology is faster than regulatory entities who cannot keep up (Osoba & Welser, 2017), which results in the chain of causation explained. Therefore, the regulatory, legal guidance can be seen as highly influential. It is important that these are established by the state or industry as an overall guide for organisations (Djelic, Kourula, Moon & Wickert, 2016). In case this guideline is not given, organisations must develop such themself. However, according to the findings, organisations do not have or want to invest the resources required that are needed to create guidelines themselves. Consequently, the

lack of guidance causes CB occurrence, as this research demonstrated. Therefore, organisations should consider carefully in advance whether it is more efficient to invest in guidance to limit the occurrence of CBs, or to debias after the detection of an error. Alternatively, there might not be an incentive for organisations to manage CBs at all, as legal requirements are not in place yet (Madiega, 2019) and as the system is intransparent (Miller, Katz & Gans, 2018) which makes it very difficult for clients to detect if CBs were transferred, if they did not develop the CA themselves.

Analysing the causes mentioned in the findings, revealed that a lack of resources plays a significant role in practice. The deficiency of resources is closely related to other causes and thereby, often prevents CB management. Surprisingly, this factor was neither addressed by the debiasing strategies in literature nor the strategies mentioned by the interviewees. Factors like resource allocation are assigned to the literature field of quality management (Yang, 2006) and certainly, influence the management of CBs. However, debiasing literature does not pick up on this. Additionally, it should be scrutinised why organisations do not address resources with their strategies. It can be argued that resources might be allocated on a superior level (Noda & Bower, 1996), wherefore the interviewees might deem their influence as too small. It might also depend on the organisation's attitude towards quality management. If CB management is not aligned with the organisation-wide objectives, then this lack of resources might be intentional and no focus is drawn to it from a superior perspective, as strategically the focus should remain on actions supporting the organisation's vision (Dess & Miller, 1996).

5.4 Management Strategies

The strategies mentioned by the interviewees are not specified in their relation to either a CB or a cause. They tend to focus on broad objectives and are strongly interlinked. On the one hand, a reason might be that a pre-analysis of CBs or causes are resource-intensive and is therefore not carried out since a lack of resources was simultaneously emphasised. On the other hand, strategies such as feedback or accountability may already be in place and may incidentally prove to be beneficial to CB management. Alternatively, it could be related to the fact that specific strategies simply might not cover problem areas sufficiently and are therefore not applied. A finding in support of the latter argument is that strategies are often interlinked, which suggests that they should be used in combination to be effective. This would correspond with Wilson and Brekke (1994), who present several steps towards debiasing, that might be viewed in relation to some supporting strategies. Nevertheless, it could again be concluded that it is difficult for organisations to determine a specific strategy that is effective in its individual application.

The mention of strategies for avoiding decisions and limited application stands out as striking findings. Both strategies are rather exceptional compared to the other strategies mentioned, as they do not address causes and CBs, but make them redundant. For aiming at redundancy, they provide a very effective approach. However, their application may impact the CAs quality and thereby, hinder the technological development of an organisation. By restricting the CA, organisations may remain behind the state of the art and thus, potentially behind the competition. By avoiding the decision, organisations may hamper technological development. Therefore, the permanent application of these strategies is rather unrealistic.

5.5 Strategy Expansion to Literature

In section 4.4.2 the interrelations of the strategies mentioned in the interview with regard to CBs and causes from the interviews are already analysed. This analysis is used to discuss to which extent strategies can contribute to either CB-specific literature or cause-specific literature.

Strategy Categories	Strategies	Blind Spot	Over-confidence	Self-serving	False Consensus	Selection
Extension of the Perspective	Further expertise	✕		✕	✕	
	Change of perspective	✕		✕	✕	
	Worst-/best-case scenarios					
Promotion of Awareness	Transparent decision process					
	Explanations about the CA					
	Communication CA methodology					✕
	Personal exchange within the team		✕	✕		
	Reducing stress					
	Ensuring employees' motivation					
Education	Education	✕	✕	✕		✕
	Sharing lessons learned		✕	✕		✕
Provision of Guidance	Policies, checklists or guidelines	✕			✕	✕
	Collaborative KPIs					
	Oversight	✕	✕	✕	✕	✕
Provision of Feedback	Feedback	✕	✕	✕	✕	✕
	Double-checking	✕	✕	✕	✕	✕
Accountability	Accountability				✕	✕
CA-specific Strategy	Transparent data selection				✕	✕
	Diverse testing				✕	✕
	Refinement after testing				✕	✕
	Decision avoidance					
	Restricted application					

✕ Empirical strategy may be expansion to the literature
✕ Empirical strategy corresponds to the literature of the respective CB
✕ Empirical strategy corresponds to the literature of another CB
Empirical strategy makes CB redundant

Figure 6: Literary Interrelations Between Cognitive Biases and Strategies (own illustration)

Considering the strategies with regard to the CBs identified in the interviews and comparing those strategies with the literature of CB-specific strategies, it can be noted that several of them do not appear in the literature and are therefore potentially new. The interrelations are very comprehensive; hence, the following only discusses the most important highlights, whereby figure 6 supports the comprehension. Appendix E presents a detailed picture including further information with reference to the literature.

The colouration in figure 6 visualises with crosses in orange that 11 strategies are potentially new to CB-specific literature, addressing each of the five CBs. Blue crosses in the figure indicate that a matching strategy that may address the respective CB already exists in the literature addressing one of the other identified CBs. Therefore, these strategies are not entirely new but may contribute to strategies for that respective CB. The 11 entirely new strategies mainly fall into the strategy categories of CA-specific, perspective expansion, and guidance. Furthermore, the strategies double-checking and oversight, which can each enable CB detection through third party involvement as stated in section 4.4.3, are not addressed by literature, hence, they are regarded as a valuable expansion to the CB-specific literature. However, detection is only possible if the third party is capable of identifying CBs in others, which could be achieved through education and training.

Similarly, cause-specific strategies from the literature are compared with strategies from the interviews in relation to the stated causes. To create a complete picture, the cause comparison of section 4.4.2 is also taken into consideration (see figure 7). Again, the discussion focuses only on the highlights, whereby the detailed analysis can be found in appendix F.

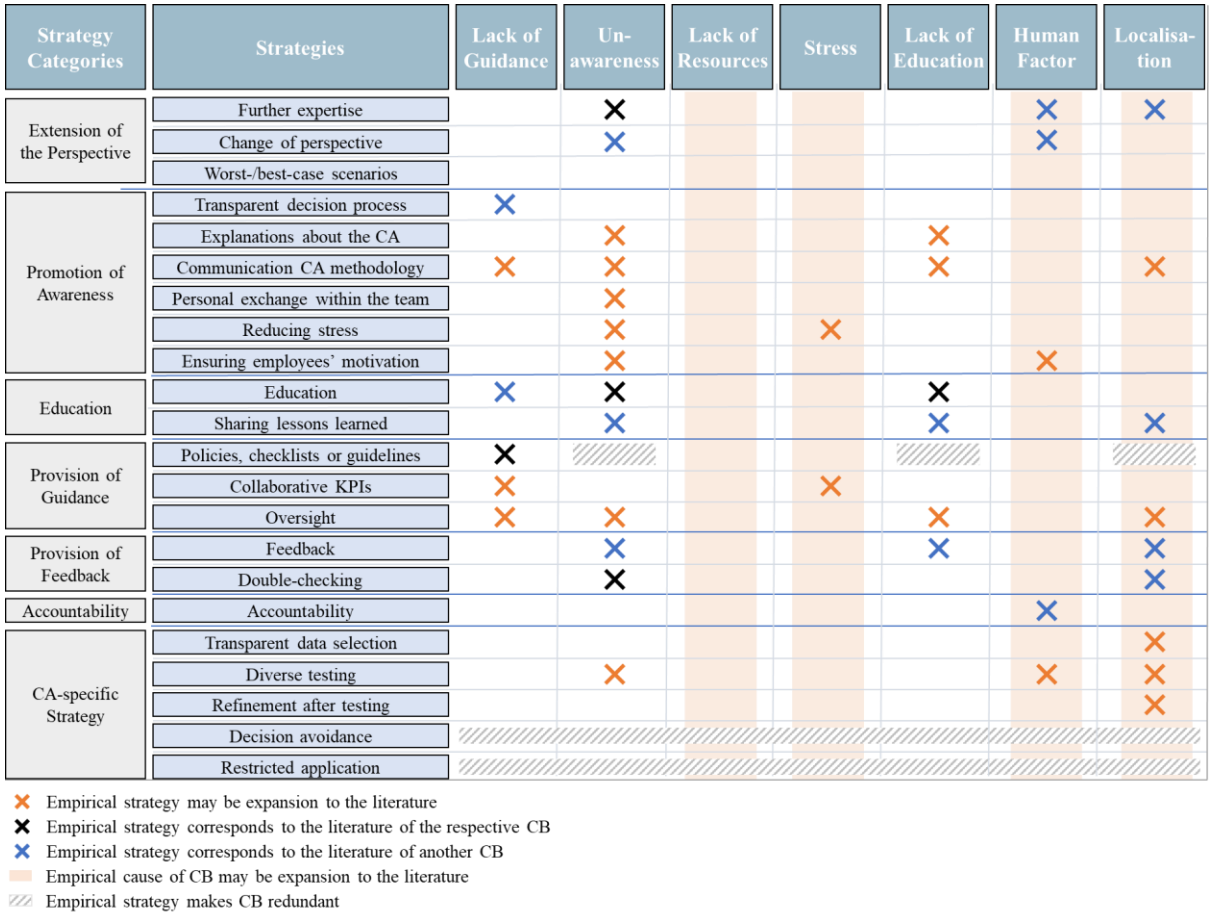


Figure 7: Literary Interrelation Between Causes of Cognitive Biases and Strategies (own illustration)

The comparison of section 5.3 implies already which causes are not mentioned in the literature. Here, it can be noted that the causes of stress, lack of resources, human factor, and localisation and their respective strategies do not appear in the literature at all or not in the same respective cause categories. In general, only very few strategies mentioned in the interviews can be found in the same cause categories as the literature, as indicated in the figure. This means, conversely, that many strategies match the literature, but are not found in the same literary cause categorisation. Thereby, they may contribute to the literature. Overall, 12 strategies are entirely new to cause-specific literature. It is particularly important to point out that CA- and organisation-related strategies are not to be found in the literature.

In both comparisons above, it becomes evident that especially CA-specific strategies are not taken into consideration from CB-specific nor cause-specific literature. Those strategies mentioned by the interviewees, such as diverse testing or explanations about the CA, are strongly related to the application. Since there is no literature discussing debiasing in the CA context, the strategies obtained in the interview are especially relevant and may contribute to the limited scope of literature.

Generally, a tendency is observed in which organisations focus on CA and organisational related strategies. These often represent a possible expansion to literature, which is supported by the fact that academic insight into the topic of CB management in CAs is extremely limited. On the one hand, it could be that literature might move too slowly compared to the pace in the tech sector, which means that strategies might not be detected and adopted quickly enough. On the other hand, it could imply that empirical strategies may not prove effective once they have been tested. Irrespective of the literature, it is logical and relevant for organisations to approach CB management at these levels, as these areas represent the decision-making environment by which rationality is influenced (Gigerenzer, 2008). Still, the effectiveness of the strategies remains uncertain until assessed.

It is noticeable that in the comparison of strategies mentioned in the interviews and strategies from CB- and cause-specific literature, the strategies CB education, double-checking, and feedback effectively address almost all CBs and causes. This implies that these can be widely used for CB management. All three strategies address the first step of the debiasing process (Wilson & Brekke, 1994) and ensure that individuals become aware of the possibility of CBs and their occurrence. Education and partial feedback are applied to train individuals in recognising CBs. Double-checking and partial feedback is carried out when individuals are not yet able to recognise CBs independently and therefore, an intervention of a third party is needed.

6 Conclusion

When answering the main research question of how organisations can manage the prevention of CBs impacting CAs with the help of debiasing strategies, it can be concluded that organisations do not differentiate between the two approaches, as literature does. Organisations do not differentiate based on erroneous decisions but consider stages in the development process as determining factors. In this respect, they consider the organisation and its workforce as the basis of CB management and do not specifically approach debiasing from the literary perspective of the individual. As a result, prevention and debiasing are approached differently in each organisation. Besides, it is identified that CB management is approached in a way that strategies can be both prevention and debiasing. Thus, organisations approach CB management more generally and address multiple uses within one strategy, therefore, the gap remains literary.

Strategies are often not consciously associated with CBs and causes of CB occurrence, thus implying a limited reflectiveness of the interviewees. The strategies used for CB management are often designed for the specific application, the CA. In this context, the most emphasised activity is the diverse testing of the application.

Following, the theoretical implications address CBs, causes of CB occurrence and strategies that represent a possible extension to the literature. These should be further tested to assess their relevance, effectiveness, and generalisability. It was found that the self-serving bias, overconfidence bias, and false consensus bias are not yet considered by literature in a CA- nor AI-related context. Strategies for CB management that specifically focus on the application area of CAs and on organisational aspects are not found in the literature either. The identified cause of the lack of resources was emphasised but not addressed with appropriate strategies by the organisations nor in the literature. Further research is needed to explore reasons for this and, if required, to develop strategies that address this cause. A lack of guidance in organisations is also clearly defined as a cause for the occurrence of CBs, which is partly a result of missing standards in the industry. To address this cause, further research is necessary to determine industry guidelines. Overall, more research is needed to explore the fluid concept debiasing and prevention in practice.

As managerial implications, it is summarised that for organisations it may be important to consider supporting CB management more strongly with a combination of strategies that conclusively address all steps of the debiasing process. This could facilitate advanced CB management by allowing a more targeted use of strategies. It is particularly important that all individuals within the organisation are considered in the management of CBs. Education should be provided for individuals at all hierarchical levels. If this is not feasible, it should be considered that individuals in superordinate positions provide guidance to individuals in operational positions.

As a limitation, it should be noted that this qualitative research is based on interviews with 11 experts. All interviewed experts come from the field of CAs or AIs as it was aimed to investigate

the CB management in that area. This may have limited the psychological perspective of the research. However, experts in the field of psychology which have insights into the development process of CAs are scarce. Generally, the primary data do not provide a sufficient level of representativeness, wherefore the findings are not generalisable and must be researched further. Additionally, the publishing dates of the secondary data must be addressed. As this research mainly draws information from a relatively mature research field that lacks newer insights, some information might be outdated. An unusual factor that must be considered is that the thesis was conducted during a worldwide pandemic (Covid-19) which prevented face-to-face interview conduction. This might have limited the reception of implied meaning of respective statements. Finally, after extensive discussion of CBs, it has to be noted that this research may also have been distorted by CBs of the individuals conducting this research. As a result, several aspects of the research, such as questions asked in the interviews, may be affected by CBs and therefore, limited the rationality with which the research was carried out.

References

Ademu, I. O., & Imafidon, C. (2012). Agent-Based Computing Application and Its Importance to Digital Forensic Domain, *14th International Conference on Artificial Intelligence (ICAI'12)*

Ahmet, C. (2018). Artificial Intelligence: How advanced machine learning will shape the future of our world. Shockwave Publishing

Allison, S. T., & Messick, D. M. (1985). The Group Attribution Error, *Journal of Experimental Social Psychology*, vol. 21, no. 6, pp. 563-579

Arkes, H. R. (1991). Costs and Benefits of Judgment Errors: Implications for debiasing, *Psychological Bulletin*, vol. 110, no. 3, pp. 486-498

Arkes, H. R., & Blumer, C. (1985). The Psychology of Sunk Cost. *Organizational Behavior and Human Decision Processes*, vol. 35 no. 1, pp. 124-140

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two Methods of Reducing Overconfidence, *Organizational Behavior and Human Decision Processes*, vol. 39, no. 1, pp. 133-144

Arnott, D. (2006). Cognitive Biases and Decision Support Systems Development: A design science approach, *Information Systems Journal*, vol. 16, no. 1, pp. 55-78

Arrow, K. J. (1986). Rationality of Self and Others in an Economic System, *The Journal of Business*, vol. 59, no. 4, pp. 385-399

Ayoub, K., & Payne, K. (2015). Strategy in the Age of Artificial Intelligence, *Journal of Strategic Studies*, vol. 39, no. 5-6, pp. 793-819

Babcock, L., Wang, X., & Loewenstein, G. (1996). Choosing the Wrong Pond: Social comparisons that reflect a self-serving bias, *Quarterly Journal of Economics*, vol. 111, no. 1, pp. 1-19

Bareinboim, E., & Pearl, J. (2012). Controlling Selection Bias in Causal Inference, *JMLR Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 22, pp. 100-108

Baron, J. (2000). Thinking and Deciding, 3rd edn, Cambridge: Cambridge University Press

Bazerman, M. H., & Moore, D. A. (2009). Judgment in Managerial Decision Making, 7th edn, Hoboken: John Wiley & Sons

Beaulac, G., & Kenyon, T. (2014). Critical Thinking Education and Debiasing (AILACT Essay Prize Winner 2013), *Informal Logic*, vol. 34, no. 4, pp. 341-363

Bell, D. E., Raiffa, H., & Tversky, A. (1988). Decision Making: Descriptive, normative, and prescriptive interactions, New York: Cambridge University Press

Benington, H.D. (1956). Production of Large Computer Programs. *IEEE Annals of the History of Computing*, vol. 5, no. 4, pp. 350-361

Berger, V. W. (2005). The Reverse Propensity Score to Detect Selection Bias and Correct for Baseline Imbalances, *Statistics in Medicine*, vol. 24, no. 18, pp. 2777-2787

Bessarabova, E., Piercy, C. W., King, S., Vincent, C., Dunbar, N. E., Burgoon, J. K., Miller, C. H., Jensen, M., Elkins, A., Wilson, D. W., Wilson, S. N., & Lee, Y.-H. (2016). Mitigating Bias Blind Spot via a Serious Video Game, *Computers in Human Behavior*, vol. 62, pp. 452-466

Bhandari, G., & Hassanein, K. (2012). An Agent-Based Debiasing Framework for Investment Decision-Support Systems, *Behaviour & Information Technology*, vol. 31, no. 5, pp. 495-507

Bhushan, N., & Rai, K. (2004). Strategic Decision Making: Applying the analytic hierarchy process, London: Springer

Blackwell, D., & Hodges, J. L. (1957). Design for the Control of Selection Bias, *The Annals of Mathematical Statistics*, vol. 28, no. 2, pp. 449-460

Blanton, H., Pelham, B. W., DeHart, T., & Carvallo, M. (2001). Overconfidence as Dissonance Reduction, *Journal of Experimental Social Psychology*, vol. 37, no. 5, pp. 373-385

Bloom, R., & Tesser, A. (1971). On Reducing Experimenter Bias: The effects of forewarning, *Canadian Journal of Behavioural Science / Revue Canadienne des Sciences du Comportement*, vol. 3, no. 2, pp. 198-208

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in Probability and Frequency Judgments: A critical examination, *Organizational Behavior and Human Decision Processes*, vol. 65, no. 3, pp. 212-219

Campbell-Yeo, M., Ranger, M., Johnston, C., & Fergusson, D. (2009). Controlling Bias in Complex Nursing Intervention Studies: A checklist, *Canadian Journal of Nursing Research*, vol. 41, no. 4, pp. 32-50

Caputo, A. (2013). A Literature Review of Cognitive Biases in Negotiation Processes, *International Journal of Conflict Management*, vol. 24, no. 4, pp. 374-398

Caverni, J.-P., Fabre, J.-M., & Gonzalez, M. (1990). Cognitive Biases, 1st edn, Elsevier

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial Intelligence, Bias and Clinical Safety, *BMJ Quality & Safety*, vol. 28, no. 3, pp. 231-237

Chapman, G. B., & Johnson, E. J. (2002). Incorporating the Irrelevant: Anchors in judgments of belief and value, in Gilovich, T., Griffin, D., & Kahneman, D. (eds), *Heuristics and Biases: The psychology of intuitive judgment.* Cambridge: Cambridge University Press. pp. 120-138

Charmaz, K. (2006). Constructing Grounded Theory: A practical guide through qualitative analysis. SAGE Publications

Chen, J. Q., & Lee, S. M. (2003). An Exploratory Cognitive DSS for Strategic Decision Making, *Decision Support Systems*, vol. 36, no. 2, pp. 147-160

Conversational Agent. (2019). in *DeepAI*, Available online: https://deepai.org/machine-learning-glossary-and-terms/conversational-agent [Accessed 28 March 2020]

Cook, M. B., & Smallman, H. S. (2008). Human Factors of the Confirmation Bias in Intelligence Analysis: Decision support from graphical evidence landscapes, *Human Factors*, vol. 50, no. 5, pp. 745-754

CORDIS (2015). Final Report Summary - RECOBIA, Available online: https://cordis.europa.eu/project/id/285010/reporting [Accessed 7 April 2020]

Correia, V. (2018). Contextual Debiasing and Critical Thinking: Reasons for optimism, *Topoi*, vol. 37, no. 1, pp. 103-111

Cortes, C., Mohri, M., Riley, M., & Rostamizadeh, A. (2008). Sample Selection Bias Correction Theory, in Freund Y., Györfi L., Turán G., & Zeugmann T. (eds), *Algorithmic Learning Theory*, Springer, pp. 38-53

Creswell, J. W., & Creswell, J. D. (2018). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches, 5th edn, Los Angeles: SAGE Publication

Croskerry, P. (2003). Cognitive Forcing Strategies in Clinical Decisionmaking, *Annals of Emergency Medicine*, vol. 41, no. 1, pp. 110-120

Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive Debiasing: part 1 and part 2, *BMJ Quality & Safety*, vol. 22, no. 2, pp. 58-72

Dasny, P. P. (2019). The Code of Intelligence: Knowledge is not wisdom and, wisdom is not intelligence. StreetLib

Dawes, R. M. (1980). Confidence in intellectual judgements vs. confidence in perceptual judgements, in Lantermann E. D. & Feger H. (eds), *Similarity and Choice*. Vienna: Hans Huber

Dawson, N. V., & Arkes, H. R. (1987). Systematic Errors in Medical Decision Making: Judgment limitations, *Journal of General Internal Medicine*, vol. 2, no. 3, pp. 183-187

Dess, G. G., & Miller, A. (1996). Strategic Management, 2nd edn, New York: McGraw-Hill Inc.

Djelic, M. L., Kourula, A., Moon, J., & Wickert, C. (2016). Government and the Governance of Business Conduct: Implications for management and organization, *Organization Studies*

Dudík, M., Schapire, R. E., & Phillips, S. J. (2006). Correcting Sample Selection Bias in Maximum Entropy Density Estimation, *Advances Neural Information Process Systems*

Epley, N., & Gilovich, T. (2005). When Effortful Thinking Influences Judgmental Anchoring: Differential effects of forewarning and incentives on self-generated and externally provided anchors, *Journal of Behavioral Decision Making*, vol. 18, pp. 199-212

European Commission. (2020). On Artificial Intelligence - A European Approach to Excellence and Trust, white paper, Available online: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf [Accessed 24 April 2020]

Evans, J. St. B. T. (2003). In Two Minds: Dual-process accounts of reasoning, *Trends in Cognitive Sciences*, vol. 7, no. 10, pp. 454-459

Farnsworth, W. (2003). The Legal Regulation of Self-Serving Bias, *U.C. Davis Law Review*, vol. 37, no. 2, p. 567-602

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (eds), *Judgment Under Uncertainty: Heuristics and biases*. New York: Cambridge University Press, pp. 422-444

Ghauri, P., & Grønhaug, K. (2002). Business Research Methods in Business Studies: A practical guide, 2nd edn, Sydney: Prentice Hall

Gigerenzer, G. (2008). Why Heuristics Work, *Perspectives on Psychological Science*, vol. 3, no. 1, pp. 20-29

Gigerenzer, G., Todd, P. M., & the ABC Research Group (1999). Simple Heuristics That Make Us Smart, New York: Oxford University Press

Giraudeau, B., & Ravaud, P. (2009). Preventing Bias in Cluster Randomised Trials, *PLoS Medicine*, vol. 6, no. 5

Glaser, B. G., & Strauss, A. L. (1967). The Discovery of Grounded Theory: Strategies for qualitative research, Chicago: Aldine Publishing Company

Greenwald, A. G., & Krieger, L. H. (2006). Implicit Bias: Scientific Foundations, *California Law Review*, vol. 94, no. 4, pp. 945-967

Gustavson, K., Røysamb, E., & Borren, I. (2019). Preventing Bias from Selective Non-Response in Population-Based Survey Studies: Findings from a Monte Carlo Simulation Study, *BMC Medical Research Methodology*, vol. 19, no. 1

Harrison, E. F. (1996). A Process Perspective on Strategic Decision Making, *Management Decision*, vol. 34, no. 1, pp. 46-53

Haselton, M. G., Nettle, D., & Andrews, P. W. (2005). The Evolution of Cognitive Bias, in Buss D. M. (eds), *The Handbook of Evolutionary Psychology*. John Wiley & Sons Inc, pp. 724-746

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error, *Econometrica*, vol. 47, no. 1, pp. 153-161

Heuer, R. J. (1999). Psychology of Intelligence Analysis, Washington, D.C.: Center for the Study of Intelligence, Central Intelligence Agency

Hillemann, E.-C., Nussbaumer, A., & Albert, D. (2015). The Role of Cognitive Biases in Criminal Intelligence Analysis and Approaches for their Mitigation. *Proceedings of the European Intelligence and Security Informatics Conference*

Hogarth, R.M. (1987). Judgment and Choice: The psychology of decision, 2nd edn, Chichester: John Wiley & Sons, Ltd.

Howe, C. J., Cole, S. R., Chmiel, J. S., & Muñoz, A. (2011). Limitation of Inverse Probability-of-Censoring Weights in Estimating Survival in the Presence of Strong Selection Bias, *American Journal of Epidemiology*, vol. 173, no. 5, pp. 569-577

IBM Research. (2020). AI and Bias - IBM Research - US, Available online: https://www.research.ibm.com/5-in-5/ai-and-bias/ [Accessed 15 April 2020]

Jurafsky, D., & Martin, J. H. (2008). Speech and Language Processing, 2nd edn, Upper Saddle River: Pearson Prentice Hall

Kahneman, D., & Frederick, S. (2002). Representativeness Revisited: Attribute Substitution in Intuitive Judgment, in Griffin, D., Kahneman, D., & Gilovich, T. (eds), *Heuristics and Biases: The Psychology of Intuitive Judgment*, Cambridge: Cambridge University Press, pp. 49-81

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An analysis of decision under risk, *Econometrica*, vol. 47, no. 2, pp. 263-291

Kahneman, D., & Tversky, A. (1984). Choices, Values, and Frames, *American Psychologist*, vol. 39, no. 4, pp. 341-350

Kahneman, D. (2003). A Perspective on Judgment and Choice: Mapping Bounded Rationality., *American Psychologist*, vol. 58, no. 9, pp. 697-720

Kahneman, D. (2011). Thinking, Fast and Slow, New York: Farrar, Straus and Giroux

Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (1982). Judgment Under Uncertainty: Heuristics and biases, Cambridge University Press

Kang, J., Bennett, M., Carbado, D., Casey, P., & Levinson, J. (2012). Implicit Bias in the Courtroom, *UCLA Law Review*, vol. 59, pp.1124-1187

Karimi, F., Génois, M., Wagner, C., Singer, P., & Strohmaier, M. (2018). Homophily Influences Ranking of Minorities in Social Networks, 1, *Scientific Reports*, vol. 8, no. 1, pp. 1-12

Kaufmann, L., Carter, C. R., & Buhrmann, C. (2012). The Impact of Individual Debiasing Efforts on Financial Decision Effectiveness in the Supplier Selection Process, *International Journal of Physical Distribution & Logistics Management*, vol. 42, no. 5, pp. 411-433

Kelemen, M. L., & Rumens, N. (2008). An Introduction to Critical Management Research. London: SAGE Publications Inc.

Keren, G. (1990). Cognitive Aids and Debiasing Methods: Can cognitive pills cure cognitive ills?, *Advances in Psychology*, vol. 68, pp. 523-552

Keren, G. (1991). Calibration and Probability Judgements: Conceptual and Methodological Issues, *Acta Psychologica*, vol. 77, no. 3, pp. 217-273

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial Disparities in Automated Speech Recognition, *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684-7689

Krueger, J., & Clement, R. W. (1994). The Truly False Consensus Effect: An Ineradicable and Egocentric Bias in Social Perception, *Journal of Personality and Social Psychology*, vol. 67, no. 4, pp. 596-610

Larrick, R. (2004). Debiasing, in D. J. Koehler & N. Harvey (eds), *Blackwell Handbook of Judgment and Decision Making*, Blackwell Publishing, pp. 316-337

Leibenstein, H. (1950). Bandwagon, Snob, and Veblen Effects in the Theory of Consumers' Demand, *The Quarterly Journal of Economics*, vol. 64, no. 2, pp. 183-207

Lichtenstein, S., & Slovic, P. (1971). Reversals of Preference between Bids and Choices in Gambling Decisions, *Journal of Experimental Psychology*, vol. 89, no. 1, pp. 46-55

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980, in Kahneman, D., Slovic, R., & Tversky, A. (eds), J*udgment under uncertainty: Heuristics and biases*, Cambridge: Cambridge University Press, pp. 306-354

Little, R. J. A., & Rubin, D. B. (1986). Statistical Analysis with Missing Data, New York: John Wiley & Sons, Inc

Lloyd, K. (2018). Bias Amplification in Artificial Intelligence Systems, ArXiv

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the Opposite: A Corrective strategy for social judgment, *Journal of Personality and Social Psychology*, vol. 47, no. 6, pp. 1231-1243

Madiega, T. (2019). EU Guidelines on Ethics in Artificial Intelligence: Context and implementation, European Parliament

Marks, G., & Miller, N. (1987). Ten Years of Research on the False-Consensus Effect: An empirical and theoretical review, *Psychological Bulletin*, vol. 102, no. 1, pp. 72-90

Maule, A. J., & Hodgkinson, G. P. (2002). Heuristics, biases and strategic decision making. *The Psychologist*, vol. 15, no. 2, pp. 68-71

Maynes, J. (2015). Critical Thinking and Cognitive Bias, *Informal Logic*, vol. 35, no. 2, pp. 183-203

Milkman, K. L. Chugh, D., & Bazerman, M. H. (2009). How Can Decision Making Be Improved?, *Perspectives on Psychological Science*, vol. 4, no. 4, pp. 379-383

Miller, D. D., & Brown, E. W. (2018). Artificial Intelligence in Medical Practice: The question to the answer?, *The American Journal of Medicine*, vol. 131, no. 2, pp. 129-133

Miller, F. A., Katz, J. H., & Gans, R. (2018). The OD Imperative to Add Inclusion to the Algorithms of Artificial Intelligence, vol. 50, no. 1, pp. 6-12

Miller, G. A. (1956). The Magical Number Seven plus or Minus Two: Some limits on our capacity for processing information, *Psychological Review*, vol. 63, no. 2, pp. 81-97

Moore, D. A., & Flynn, F. J. (2017). The Case for Behavioral Decision Research in Organizational Behavior, *Academy of Management Annals*, vol., 2 no. 1, pp. 399-431

Morewedge, C. K., & Kahneman, D. (2010). Associative Processes in Intuitive Judgment, *Trends in Cognitive Sciences*, vol. 14, no. 10, pp. 435-440

Morrison, K. R., & Matthes, J. (2011). Socially Motivated Projection: Need to belong increases perceived opinion consensus on important issues, *European Journal of Social Psychology*, vol. 41, no. 6, pp. 707-719

Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M., & Davey Smith, G. (2018). Collider Scope: When selection bias can substantially influence observed associations, *International Journal of Epidemiology*, vol. 47, no. 1, pp. 226-235

Mussweiler, T., & Strack, F. (2000). The Use of Category and Exemplar Knowledge in the Solution of Anchoring Tasks, *Journal of Personality and Social Psychology*, vol. 78, no. 6, pp. 1038-1052

Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the Inevitable Anchoring Effect: Considering the opposite compensates for selective accessibility, *Personality and Social Psychology Bulletin*, vol. 78, no. 6, pp. 1038-1052

Neuman, W. L. (2005). Social Research Methods: Qualitative and quantitative approaches, 6th edn, Boston: Pearson

Newell, A., & Simon, H. A. (1972). Human Problem Solving, 1st edn, Englewood Cliffs: Prentice-Hall

Nickerson, R. S. (1998). Confirmation Bias: A ubiquitous phenomenon in many guises, *Review of General Psychology*, vol. 2, no. 2, pp. 175-220

Nordstrom, C. R., Williams, K. B., & LeBreton, J. M. (1996). The Effect of Cognitive Load on the Processing of Employment Selection Information, *Basic and Applied Social Psychology*, vol. 18, no. 3, pp. 305-318

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Broelemann, K., Kasneci, G., Tiropanis, T., & Staab, S. (2019). Bias in Data-Driven Artificial Intelligence Systems - An introductory survey, *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3

Osoba, O., & Welser, W. (2017). An Intelligence in Our Image: The risks of bias and errors in artificial intelligence, Santa Monica: RAND Corporation

Oswald, M. E., & Grosjean S. (2004). Confirmation Bias, in Pohl R.F. (eds), *Cognitive Illusions: A handbook on fallacies and biases in thinking, Judgement and Memory*. New York: Psychology Press, pp. 79-96

Pannu, A. (2015). Artificial Intelligence and Its Application in Different Areas, *International Journal of Engineering and Innovative Technology (IJEIT)*, vol. 4, no. 10, pp. 79-84

Pannucci, C. J., & Wilkins, E. G. (2010). Identifying and Avoiding Bias in Research, *Plastic and Reconstructive Surgery*, vol. 126, no. 2, pp. 619-625

Patton, M.Q. (2002) Qualitative Research and Evaluation Methods. 3rd edn, Thousand Oaks: SAGE Publications

Pronin, E. (2007). Perception and Misperception of Bias in Human Judgment, *Trends in Cognitive Sciences*, vol. 11, no. 1, pp. 37-43

Pronin, E., Lin, D., & Ross, L. (2002). The Bias Blind Spot: Perceptions of bias in self versus others, *Personality and Social Psychology Bulletin*, vol. 28, no. 3, pp. 369-381

Royce, D. W. W. (1970). Managing the Development of Large Software Systems, in IEEE WESCON, pp. 1-9

Rubin, V. L., Chen, Y., & Thorimbert, L. M. (2010). Artificially Intelligent Conversational Agents in Libraries, *Library Hi Tech*, vol. 28, no. 4, pp. 496-522

Ruparelia, N. (2010). Software Development Lifecycle Models, *ACM SIGSOFT Software Engineering Notes*, vol. 35, no. 3, pp. 8-13

Russo, J. E., & Schoemaker, P. J. H. (2018). Overconfidence, in Augier, M. & Teece, D. J. (eds), *The Palgrave Encyclopedia of Strategic Management*, London: Palgrave Publishers Ltd, pp. 1236-1246

Samuelson, W., & Zeckhauser, R. (1988). Status Quo Bias in Decision Making, *Journal of Risk and Uncertainty*, vol. 1, no. 1, pp. 7-59

Saunders, M., Lewis, P., & Thornhill, A. (2016). Research Methods for Business Students, 7th edn, Harlow: Pearson Education Limited

Schwenk, C. R. (1984). Cognitive Simplification Processes in Strategic Decision-Making, *Strategic Management Journal*, Wiley Blackwell, vol. 5, no. 2, pp. 111-128

Sen, P., & Ganguly, D. (2005). Towards Socially Responsible AI: Cognitive bias-aware multi-objective learning, Available online: http://arxiv.org/abs/2005.06618 [Accessed 20 April 2020]

Shepperd, J., Malone, W., & Sweeny, K. (2008). Exploring Causes of the Self-Serving Bias, *Social and Personality Psychology Compass*, vol. 2, no. 2, pp. 895-908

Simmons, J. P., LeBoeuf, R. A.,, & Nelson, L. D. (2010). The Effect of Accuracy Motivation on Anchoring and Adjustment: Do people adjust from provided anchors?, *Journal of Personality and Social Psychology*. vol. 99, no. 6, pp. 917-932

Simon, H. A. (1955). A Behavioral Model of Rational Choice, *The Quarterly Journal of Economics*, vol. 69, no. 1, p. 99

Simon, H. A. (1956). Rational Choice and the Structure of the Environment, *Psychological Review*, vol. 63, no. 2, pp. 129-138

Simon, H. A. (1957). Models of Man. Wiley

Simon, H. A. (1977). The New Science of Management Decision, Englewood Cliffs: Prentice-Hall

Slovic, P. (1975). Choice between Equally Valued Alternatives, *Journal of Experimental Psychology: Human Perception and Performance*, vol. 1, no. 3, pp. 280-287

Soll, J. B., Milkman, K. L., & Payne, J. W. (2015). A User's Guide to Debiasing, in G. Keren & G. Wu (eds), *The Wiley Blackwell Handbook of Judgment and Decision Making*, Chichester: John Wiley & Sons Ltd, pp. 924-951

Stapel, D. A., Martin, L. L., & Schwarz, N. (1998). The Smell of Bias: What instigates correction processes in social judgments?, *Personality and Social Psychology Bulletin*, vol. 24, no. 8, pp. 797-806

Statista (2018). Employees in the U.S. 2017. Available online: https://www.statista.com/study/49784/employees-in-the-us/ [Accessed 20 May 2020]

Stolzenberg, R. M., & Relles, D. A. (1997). Tools for Intuition about Sample Selection Bias and Its Correction, *American Sociological Review*, vol. 62, no. 3, pp. 494-507

Strack, F., & Mussweiler, T. (1997). Explaining the Enigmatic Anchoring Effect: Mechanisms of selective accessibility, *Journal of Personality and Social Psychology*, vol. 73, no. 3, pp. 437-446

Strauss, A., & Corbin, J. (1998). Basics of Qualitative Research: Techniques and procedures for developing grounded theory, 2nd edn, Thousand Oaks: Sage Publications Inc.

Suddaby, R. (2006). From the Editors: What grounded theory is not, *Academy of Management Journal*, vol. 49, no. 4, pp. 633-642

Sutton, R. S. (1992). Adapting Bias by Gradient Descent: An incremental version of delta-bar-delta, *Proceedings of the Tenth National Conference on Artificial Intelligence*, pp. 171-176

Tatman, R. (2017). Gender and Dialect Bias in YouTube's Automatic Captions, in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, Valencia, Spain, 2017, Valencia: Association for Computational Linguistics, pp. 53-59

Thammasitboon, S., & Cutrer, W. B. (2013). Diagnostic Decision-Making and Strategies to Improve Diagnosis, *Current Problems in Pediatric and Adolescent Health Care*, vol. 43, no. 9, pp. 232-241

Tobena, A., Marks, I., & Dar, R. (1999). Advantages of Bias and Prejudice: An exploration of their neurocognitive templates, *Neuroscience & Biobehavioral Reviews*, vol. 23, no. 7, pp. 1047-1058

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and biases, *Science*, vol. 185, no. 4157, pp. 1124-1131

United States Securities and Exchange Commission. (2018a). Document, *Annual Report - Microsoft Corporation*, Available online: https://www.sec.gov/Archives/edgar/data/789019/000156459018019062/msft-10k_20180630.htm [Accessed 22 April 2020]

United States Securities and Exchange Commission. (2018b). Document, *Annual Report - Alphabet Inc.*, Available online: https://www.sec.gov/Archives/edgar/data/1652044/000165204418000007/goog10-kq42017.htm [Accessed 22 April 2020]

US Government. (2009). A Tradecraft Primer: Structured analytic techniques for improving intelligence analysis, Available online: http://doi.apa.org/get-pe-doi.cfm?doi=10.1037/e587102011-001 [Accessed 26 April 2020]

van Exel, N. J. A., Brouwer, W., Berg, B., & Koopmanschap, M. A. (2006). With a Little Help from an Anchor: Discussion and evidence of anchoring effects in contingent valuation, *Journal of Socio-Economics*, vol. 35, no.5, pp. 836-853

Vartanian, T. P. (2011). Secondary Data Analysis, Oxford University Press

Vila, L. (2005). Formal Theories of Time and Temporal Incidence, in Fisher, M., Gabbay, D., & Vila, L. (eds): *Handbook of Temporal Reasoning in Artificial Intelligence*, Amsterdam: Elsevier, vol. 1., pp. 1-24

Walter, J., Kellermanns, F. W., & Lechner, C. (2012). Decision Making Within and Between Organizations: Rationality, politics, and alliance performance, *Journal of Management*, vol. 38, no. 5, pp. 1582-1610

Wei, C., Yu, Z., & Fong, S. (2018). How to Build a Chatbot: chatbot framework and its capabilities, in Proceedings of the 2018 10th International Conference on Machine Learning and Computing (ICMLC 2018). Association for Computing Machinery, pp. 369-373

Welsh, M. B., Begg, S. H., & Bratvold, R. B. (2007). Efficacy of Bias Awareness in Debiasing Oil and Gas Judgments, in Proceedings of the 29th Annual Cognitive Science Society, McNamara, D. S. & Trafton, J.G. (eds), pp. 1647-1652

Wetzel, C. G., & Walton, M. D. (1985). Developing Biased Social Judgments: The False-Consensus Effect, *Journal of Personality and Social Psychology*, vol. 49, no. 5, pp.1352-1359

Willingham, D. T. (2007). Critical Thinking: Why is it so hard to teach?, *Arts Education Policy Review*, vol. 109, no. 4, pp. 21-32

Wilson, T. D., & Brekke, N. (1994). Mental Contamination and Mental Correction: Unwanted influences on judgments and evaluations, *Psychological Bulletin*, vol. 116, no. 1, pp. 117-142

Wilson, T. D., Houston, C. E., Etling, K. M., & Brekke, N. (1996). A New Look at Anchoring Effects: Basic anchoring and its antecedents, *Journal of Experimental Psychology: General*, vol. 125, no. 4, pp. 387-402

Wooldridge, M., & Jennings, N. R. (1995). Intelligent Agents: Theory and practice, *The Knowledge Engineering Review*, vol. 10, no. 2, pp. 115-152

Wright, G., & Goodwin, P. (2002). Eliminating a Framing Bias by Using Simple Instructions to 'Think Harder' and Respondents with Managerial Experience: Comment on 'breaking the frame', *Strategic Management Journal*, vol. 23, no. 11, pp. 1059-1067

Yang, C. (2006). The Impact of Human Resource Management Practices on the Implementation of Total Quality Management: An empirical study on high-tech firms, *The TQM Magazine,* vol. 18, no. 2, pp. 162-173

Zadrozny, B. (2004). Learning and Evaluating Classifiers under Sample Selection Bias, in *International Conference on Machine Learning ICML*, pp. 903-910

Zuckerman, M. (1979). Attribution of Success and Failure Revisited, or: The Motivational Bias Is Alive and Well in Attribution Theory, *Journal of Personality*, vol. 47, no. 2, pp. 245-287

Appendix A: Complete Overview of Strategies Investigated

Cognitive Bias	Strategies	References
Blind Spot Bias	Consider the opposite	Lord, Lepper & Preston (1984)
	Constant CB training	Bessarabova, Piercy, King, Vincent, Dunbar, Burgoon, Miller, Jensen, Elkins, Wilson, Wilson & Lee (2016)
	Warning of blind spot bias	Stapel, Martin & Schwarz (1998)
False Consensus Bias	Consider solely causal attributions	Marks & Miller (1987)
	Generate alternatives	Marks & Miller (1987)
Over-confidence Bias	Asking about past decisions	Bhandari & Hassanein (2012)
	Awareness training in biases and debiasing techniques	Welsh, Begg & Bratvold (2007)
	Feedback and questioning the decision	Arkes, Christensen, Lai & Blumer (1987)
	Safe environment and the feeling of certainty	Blanton, Pelham, DeHart & Carvallo (2001)
	Social pressure and asking an individual to explain or justify its answers	Arkes, Christensen, Lai & Blumer (1987)
Selection Bias	Consider whole data collection	Howe, Cole, Chmiel & Muñoz (2011)
	Exclude data which has a clear reason to be excluded	Berger (2005)
	Random selection	Berger (2005)
	Separate selector from task	Blackwell & Hodges (1957)
Self-Serving Bias	Penalties	Farnsworth (2003)
	Reduce occasion for CB by legal standards	Farnsworth (2003)
	Reeducate individual about CBs	Farnsworth (2003)
	Separate biased individual from decision	Farnsworth (2003)

Cause Category	Strategy	References
Faulty task	Ask fewer questions	Fischhoff (1982)
	Clarify instructions	Fischhoff (1982)
	Demonstrate alternative goal	Fischhoff (1982)
	Demonstrate impossibility of task	Fischhoff (1982)
	Demonstrate overlooked distinction	Fischhoff (1982)
	Demonstrate semantic disagreement	Fischhoff (1982)
	Task decomposing	Kaufmann, Carter & Buhrmann (2012); Fischhoff (1982)
	Decomposing (focos on decision structure)	Kaufmann, Carter & Buhrmann (2012); Kaufmann, Michel & Carter (2009)
	Break down large intervals in sections	Soll, Milkman & Payne (2015)
	Break down large intervals in smaller ones	Soll, Milkman & Payne (2015)
	Break down large intervals in time frames	Soll, Milkman & Payne (2015)
Inadequate decision	Change the linkage of related elements	Arkes (1991)
	Choosing by chance	Arkes (1991)
	Add or alter associations by giving instructions	Arkes (1991); Fischhoff (1982)
	Recalibration	Croskerry, Singhal & Mamede (2013)
	Slowing down strategies	Croskerry, Singhal & Mamede (2013)
	Sparklines	Croskerry, Singhal & Mamede (2013)
	Structured data acquisition	Croskerry, Singhal & Mamede (2013)
	Challenging (focus on decision dissent)	Kaufmann, Carter & Buhrmann (2012)
	Discouraging guessing	Keren (1990); Fischhoff (1982)
	Offering an alternative response mode	Keren (1990); Fischhoff (1982)
	Generate alternatives	Soll, Milkman & Payne (2015)
	Apply competency-based certification to making diagnostic	Thammasitboon & Cutrer (2013)

Cause Category	Strategy	References
Inadequate overview	Add new gains or losses to those currently under consideration	Arkes (1991)
	Change one's reference point	Arkes (1991)
	Judging the first-order judgement	Arkes (1991)
	Reframe losses as gains (or gains as losses)	Arkes (1991)
	Consider the opposite	Arkes (1991); Fischhoff (1982)
	Analogical thinking: Recall past experience and cases	Chen & Lee (2003)
	Envision future state of business environments	Chen & Lee (2003)
	Reflect on and examine the assumptions and belief system	Chen & Lee (2003)
	Exposure control	Croskerry, Singhal & Mamed (2013)
	Get more information	Croskerry, Singhal & Mamede (2013)
	Metacognition, decoupling, reflection, mindfulness	Croskerry, Singhal & Mamede (2013)
	Rule out worst-case scenario	Croskerry, Singhal & Mamede (2013)
	Consider the control	Croskerry, Singhal & Mamede (2013); Larrick (2004); Soll, Milkman & Payne (2015)
	Make knowledge explicit	Fischhoff (1982)
	Perspective shifting	Kaufmann, Carter & Buhrmann (2012); Kaufmann, Michel & Carter (2009)
	Gain whole picture before decision made	Keren (1990)
	Structure modifying methods	Keren (1990)
	Change thinking style	Soll, Milkman & Payne (2015)
	Dialectical bootstrapping	Soll, Milkman & Payne (2015)
	Double check by asking the same question twice	Soll, Milkman & Payne (2015)
	Prospective hindsight	Soll, Milkman & Payne (2015)
	Apply metacognition to facilitate reflective practic	Thammasitboon & Cutrer (2013)
	Apply metacognition to minimize affective biases	Thammasitboon & Cutrer (2013)
	Consider mandatory second opinion on error-prone diseases	Thammasitboon & Cutrer (2013)
	Identify and close gaps in specific knowledge and skills	Thammasitboon & Cutrer (2013)
	Increase availability of point-of-care access of current knowledge	Thammasitboon & Cutrer (2013)

Cause Category	Strategy	References
Lack of education	Professional training	Arkes (1991); Soll, Milkman & Payne (2015)
	Acquisition of knowledge	Croskerry (2003); Kaufmann, Carter & Buhrmann (2012)
	Affective debiasing	Croskerry, Singhal & Mamede (2013)
	Bias inoculation	Croskerry, Singhal & Mamede (2013)
	Cognitive tutoring systems	Croskerry, Singhal & Mamede (2013)
	Simulation training	Croskerry, Singhal & Mamede (2013)
	Specific educational interventions	Croskerry, Singhal & Mamede (2013)
	Training on theories of reasoning and medical decision making	Croskerry, Singhal & Mamede (2013)
	Educate from childhood	Fischhoff (1982)
	Plan on error	Fischhoff (1982)
	Recalibrate their responses	Fischhoff (1982)
	Rely on substantive experts	Fischhoff (1982)
	Replace individual	Fischhoff (1982)
	Create awareness	Kaufmann, Michel & Carter (2009)
	Training in representation	Larrick (2004)
	Training in biases	Larrick (2004); Fischhoff (1982)
	Training in rules	Larrick (2004); Soll, Milkman & Payne (2015)
	Accumulate illness scripts to improve accuracy of pattern recognition	Thammasitboon & Cutrer (2013)
	Actively engage in continuing medical education	Thammasitboon & Cutrer (2013)
	Consult and learn from experts	Thammasitboon & Cutrer (2013)
	Engage in deliberate practice via simulation for targeted improvement	Thammasitboon & Cutrer (2013)
	Learn about intuitive decision-making and its pitfalls	Thammasitboon & Cutrer (2013)
	Learn and apply sciences of diagnostic decision-making	Thammasitboon & Cutrer (2013)
	Participate in experiential learning	Thammasitboon & Cutrer (2013)
	Provide targeted training on common errors identified inpractice settings	Thammasitboon & Cutrer (2013)

Cause Category	Strategy	References
Lack of feedback	Rapid feedback	Arkes (1991); Fischhoff (1982)
	Feedback system	Thammasitboon & Cutrer (2013)
Lack of guidance	Accountability	Arkes (1991); Croskerry, Singhal & Mamede (2013); Kaufmann, Carter & Buhrmann (2012); Larrick (2004)
	Higher the stakes	Arkes (1991); Fischhoff (1982)
	Understand possible consequences of decisions	Chen & Lee (2003)
	Awareness of specific scenarios in which error is known to occur	Croskerry (2003)
	Conduct a secondary search or survey	Croskerry (2003)
	Checklists	Croskerry, Singhal & Mamede (2013)
	Cognitive forcing strategies	Croskerry, Singhal & Mamede (2013)
	General rules	Croskerry, Singhal & Mamede (2013)
	Standing rules	Croskerry, Singhal & Mamede (2013)
	Stopping rules	Croskerry, Singhal & Mamede (2013)
	Incentives	Kaufmann, Carter & Buhrmann (2012); Keren (1990); Larrick (2004)
	Forewarning	Keren (1990); Fischhoff (1982)
	Develop cognitive forcing strategies	Thammasitboon & Cutrer (2013)
	Use diagnostic checklists	Thammasitboon & Cutrer (2013)
	Use of clinical guidelines and clinical algorithms	Thammasitboon & Cutrer (2013)

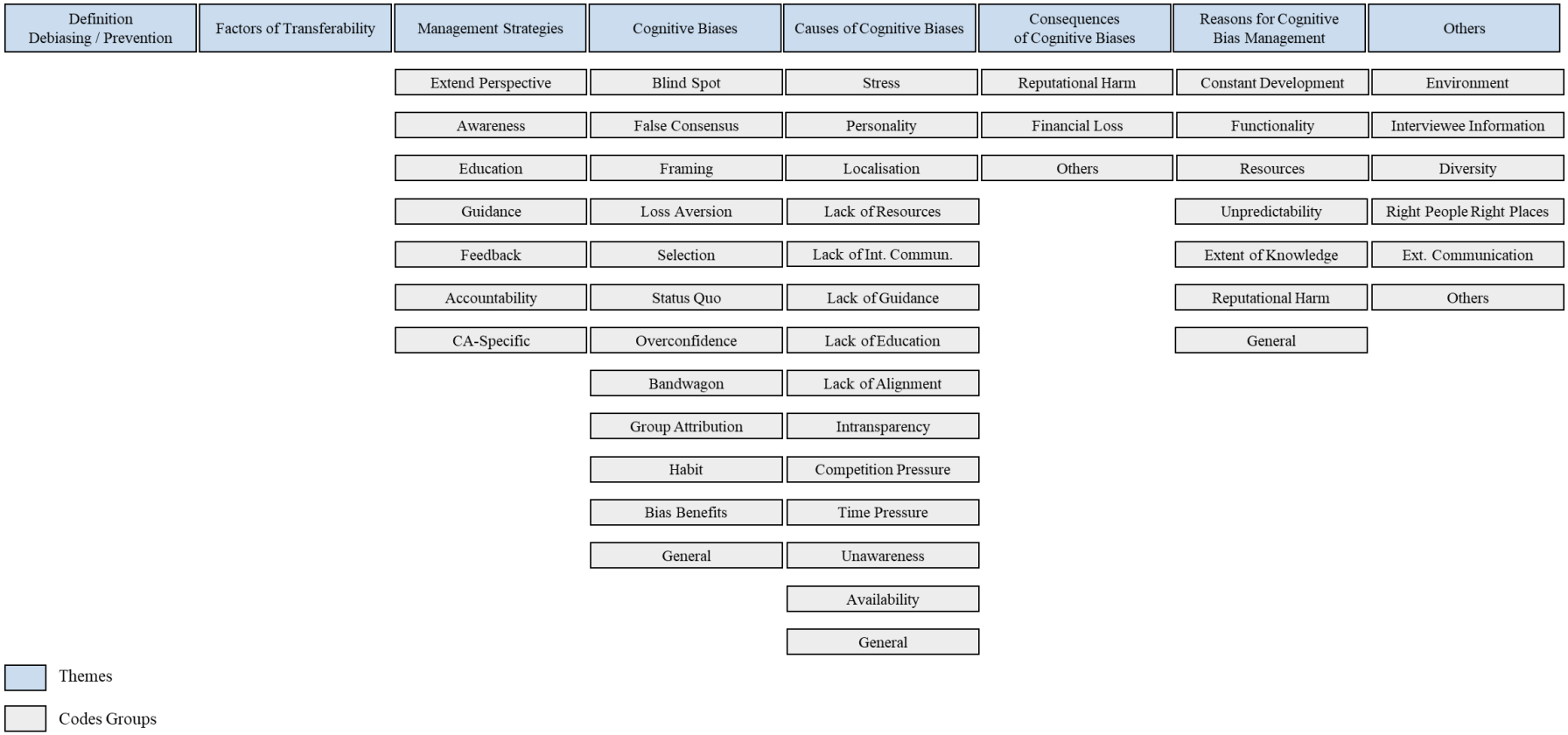
Appendix B: Interview Topics

- Development process
- Decision-making
    - Decisions in the development process
- CBs entering the development process
    - Roles
    - Activities
    - CA-development area
- Causes for CBs to enter
- CB management
    - Approaches
        - Prevention
        - Debiasing
    - Reasons for debiasing vs. prevention
    - Understanding of prevention and debiasing
    - CB management strategies

Appendix C: Participant Background Information

ID	Background	Nationality
P1	- 10+ years AI researcher (including natural language processing) - 10+ years AI Project Manager - 10+ years AI Developer	American and French
P2	- 20+ years in IT industry	Dutch
P3	- 5+ years work experience in chatbots - Design conversational solutions for different companies, business areas and languages - Research on natural language processing and conversational assistants	Spanish
P4	- 20+ years work experience in testing - Involvement with the international standards community, working on reports and standards relating to bias	British
P5	- 3+ years work experience in speech recognition / AI and assistant agent logic - Educational background in engineering and technology management	American
P6	- 5+ years work experience in technology architecture and consultancy - Educational background in engineering	Dutch
P7	- 4+ years work experience in technology consultancy - Educational background in strategic management	Dutch
P8	- 4+ years natural language processing - Educational background in audiology and speech-language pathology and neuroscience	American
P9	- 7+ years operations research experience - Doctor in philosophy and educational background management engineering	Italian
P10	- 6+ years work experience as data scientist in AI and machine learning experience in AI - Ambassador for more females in technology and machine learning	Australian
P11	- 11+ years work experience in consulting with focus on finance and technology	German

Appendix D: Code Cluster



Appendix E: Detailed Literary Interrelations Between Cognitive Biases and Strategies

Strategy Category	Strategy	Blind Spot	Overconfidence	Self-serving	False consensus	Selection
Extension of the Perspective	Further expertise	Addressing overreliance on introperspective		Only if expertise is external, so it gives proof that mistake derives from internal reasoing of individual	Effectfive if expert reveals "real" knowledge	
	Change of perspective	Addressing overreliance on introperspective		Proof that mistake derives from internal reasoning of individual	Change of self-related knowledge perspective	
	Worst- and best-case scenarios					
Promotion of Awareness	Transparent decision process					
	Explanations about the CA					
	Communication CA methodology					Provides education and raises awareness
	Personal exchange within the team		Provides safe environment as detailed background of team is disclosed (Blanton et al., 2001)	Provides safe environment as team background is disclosed (Blanton et al., 2001); Depends on team composition; either competition or safe environment		
	Reducing stress					
	Ensuring employees' motivation					
Education	Education	Education in suspectibility of individuals' CBs and strategies to detect CBs (Bessarabova et al., 2016)	Mitigates overestimation; manages social expectations as explanation that CBs are inherent in all individuals is given (Welsh, Begg & Bratvold, 2007)	Education in suspectibility of individuals' CBs and strategies to detect CBs (Farnsworth, 2003)		Education in suspectibility of individuals' CBs and strategies to detect CBs (Bessarabova et al., 2016: Farnsworth, 2003; Welsh, Begg & Bratvold, 2007)
	Sharing lessons learned		Provides safe environment, decrease social expectations that failure are not acceptable	Show that everyone makes mistakes and that they originate from internal sources		Increases awareness and provides strategies on hand how to handle data selection best

Empirical strategy may be expansion to literature

Empirical strategy corresponds to literature, as explained in the text

Empirical strategy corresponds to the literature of another cognitive bias, as explained in the text

Empirical strategy makes cognitive bias redundant

Strategy Category	Strategy	Blind Spot	Overconfidence	Self-serving	False consensus	Selection
Provision of Guidance	Policies, checklists or guidelines	Only effective if missing perspective is pointed out			Requirement of fact-based decision or need to gather all data	CB can be minimised if guidance is given how to select properly
	Collaborative KPIs					
	Oversight	Only effective if missing perspective is pointed out; additionally guidance can be given	Can detect CB	Can detect CB	Can detect CB; Only if not same-minded person and expert of truth	Can detect CB; Can be approached if guidance is given
Provision of Feedback	Feedback	Pointing out missing perspective (Arkes et al., 1987)	Can detect CB; manages social expectation (Arkes et al., 1987)	Can detect CB (Arkes et al., 1987)	Can detect CB; Only if not same-minded person and expert of truth (Arkes et al., 1987)	Pointing out CB (Arkes et al., 1987)
	Double-checking	Pointing out missing perspective (Arkes et al., 1987)	Can detect CB	Can detect CB	Can detect CB; Only if not same-minded person and expert of truth	Pointing out CB (Arkes et al., 1987)
Accountability	Accountability				Raises motivation to check if data is complete	Raises awareness (especially with combination education as system 2 is triggered)
CA-specific	Transparent data selection				Raises awareness of incomplete information (Berger, 2005)	Raises awareness of missing data; awareness higher if selection has to be argued (Berger, 2005)
	Diverse testing				Can detect	Can deteced
	Refinement after testing				Can adjusted bias of CA	Can adjusted bias of CA
	Decision avoidance	Makes CB redundant as decision is not done by individual				
	Restricted application	Prevents entrance and disclosure of CB				

Empirical strategy may be expansion to literature

Empirical strategy corresponds to literature, as explained in the text

Empirical strategy corresponds to the literature of another cognitive bias, as explained in the text

Empirical strategy makes cognitive bias redundant

Appendix F: Detailed Literary Interrelations Between Causes of Cognitive Biases and Strategies

Strategy Category	Strategy	Lack of Guidance	Unawareness of CBs	Lack of Resources	Stress	Lack of Education	Human Factor	Localisation
Extension of the Perspective	Further expertise		Third person can raise awareness of a CB (Thammasitboon & Cutrer, 2013)				Third person can validate preferences due to personality (Thammasitboon & Cutrer, 2013)	Experts increase knowledge how to localise properly (Thammasitboon & Cutrer, 2013)
	Change of perspective		Third person can raise awareness of a CB (e.g. Kaufmann, Carter & Buhrmann, 2012)				Third person can validate preferences due to personality (e.g. Kaufmann, Carter & Buhrmann, 2012)	
	Worst- and best-case scenarios							
Promotion of Awareness	Transparent decision process	Transparency leads to a clarity of responsibilities and influencing factors, which provides guidance (Fischhoff, 1982)						
	Explanations about the CA		Raises an individual's awareness			Eliminates possibility that CBs arise due to lack of knowledge		
	Communication CA methodology	Explanation of how to develop bias-free CA prescribes essential steps	Raised awareness through communication of bias-free CA			Eliminates possibility that CBs arise due to lack of knowledge		Eliminates possibility that CBs arise due to lack of knowledge
	Personal exchange within the team		Raises awareness of team members' CBs					
	Reducing stress		Reducing the stress leads to higher retentiveness		Reduces stress			
	Ensuring employees' motivation		Increases likelihood that individual is more aware of its CB				Increases likelihood that the individual is more motivated to contribute to a high quality	
Education	Education	Taught strategies to mitigate CBs can be used as guidance (e.g. Croskerry, 2003)	Education leads to increase of awareness of CBs (e.g. Croskerry, 2003)			Education can help individuals to learn about CBs, their occurrence, and strategies to mitigate them (e.g. Croskerry, 2003)		
	Sharing lessons learned		Exchanging experience leads increase of awareness of CBs			Education can be provided from the experience of others		Learning from others increase knowledge on localisation

Empirical strategy may be expansion to literature

Empirical strategy corresponds to literature, as explained in the text

Empirical strategy corresponds to the literature of another cause, as explained in the text

Empirical strategy makes cause redundant

Strategy Category	Strategy	Lack of Guidance	Unawareness of CBs	Lack of Resources	Stress	Lack of Education	Human Factor	Localisation
Provision of Guidance	Policies, checklists or guidelines	Provide a guideline of actions to be executed or factors to be included in a decision (e.g. Croskerry, Singhal & Mamede, 2013)	Makes CB redundant			Makes CB redundant		Makes CB redundant
	Collaborative KPIs	Provide understanding of key factors imporant to the company which work can be guided by			Better alignment of departments leading to reduced stress			
	Oversight	Third person can raise awareness of an CB detected in a biased individual and guide the mitigation	Third person can raise awareness of an CB detected in a biased individual			Third person can educate biased individual		Third person can raise awareness of an CB detected in a biased individual
Provision of Feedback	Feedback		Third person can raise awareness of an CB detected in a biased individual (e.g. Thammasitboon & Cutrer, 2013)			Third person can educate biased individual (e.g.Thammasitboon & Cutrer, 2013)		Third person can raise awareness of an CB detected in a biased individual (e.g. Thammasitboon & Cutrer, 2013)
	Double-checking		Third person can raise awareness of an CB detected in a biased individual (e.g. Arkes, 1991)					Third person can raise awareness of an CB detected in a biased individual (e.g. Arkes, 1991)
Accountability	Accountability						Increases the stakes and therefore the individual is more encouraged to work properly (e.g. Arkes, 1991)	
CA-specific	Transparent data selection							Transparency can mitigate CBs as higher clarity given
	Diverse testing		Can detect				Can detect	Can detect
	Refinement after testing							Improvement of localisation possible
	Decision avoidance	Makes CB redundant as decision is not done by individual						
	Restricted application	Prevents entrance and disclosure of CB						

Empirical strategy may be expansion to literature

Empirical strategy corresponds to literature, as explained in the text

Empirical strategy corresponds to the literature of another cause, as explained in the text

Empirical strategy makes cause redundant