Master Thesis

Adjusting Trust in Artificial Intelligence within the European Context of Consultancy Services

Daniel Aaltonen

Student No. 133882

Supervised by
Torkil Clemmensen

COPENHAGEN BUSINESS SCHOOL
Copenhagen, Denmark

A thesis submitted in fulfillment of the requirements for the degree of MSc in Business Administration and E-Business

17th May 2021

Number of Characters: 91,955
Number of Pages: 51

Abstract

"Perfect is the enemy of good"

-Voltaire

Trust between human beings is paramount in everyday interaction. From all parties trusting in the value of money, to us trusting a contract. However, as artificial intelligence becomes more prominent in our lives, the question of how to increase trust and how to maintain trust arise. The legislators have a say in the equation. The industry professionals have a say in the equation. The everyday users have a say in the equation. The academia however has been struggling in giving out concrete examples of increasing or adjusting trust to an appropriate level. Moreover, the academia has lacked in answering the questions of what actually affect trust levels we as humans portray towards artificial intelligence. This study attempts to uncover factors which lead towards Distrust in artificial intelligence algorithms, via the means of qualitative analysis and further, by talking to industry professionals with different backgrounds. The findings suggest that the lack of ethics and the presence of bias play a role as what this paper refers to as *effectors* in eventual poor resolution between the trustworthiness or the AI and the users' trust.

Keywords: Trust in Artificial Intelligence, Effect of Bias and Lack of Ethics, Legislation, Context-dependency, Calibrated trust, Overtrust, Distrust.

Acknowledgements

The thesis would have never taken the direction it eventually took without a special person at Deloitte Netherlands, who helped me throughout the writing process. This person supported me by giving interesting insights to the field of AI, from a perspective only a few get to ever experience. The person in question knows who this paragraph is describing.

Moreover, I would especially thank my supervisor, Torkil Clemmensen from Copenhagen Business School for guiding me towards the topic of Trust in AI, from having a keen interest in the topic, to answering any mundane questions that I had whilst writing this paper. Additionally, I want to thank all my lovely colleagues at Deloitte across the world who took part in my study as interview participants, or additional and external industry professionals.

Without the nudges from my immediate family back in my teenage years towards pursuing a degree in Business, instead of Architecture as I had initially dreamed of, I would not be in a position where I am today. Most likely, I would be in a position of not having a degree as I cannot draw for the life of me.

Lastly, a great thank you goes to my significant other here in Copenhagen, Denmark, who unconditionally supported and believed in me through thick and thin, and supplied me with endless amounts of morning espressos. Without the additional boost of caffeine running through my veins and the poor quality of jokes due to caffeine-highs, this paper would have never happened.

Thank you!

Table of Contents

List of Figures

List of Tables

1

Introduction

1.1 Background

The well-documented sentiment of human bias has been observable throughout history. In more recent years, implicit association tests and field experiments have been conducted to prove biases exist, and how these biases affect outcomes of different interactions. Furthermore, the studies have demonstrated biases which humans have not been aware of prior conducting the studies. (Manyika, Silberg, & Presten, 2019).

An event from recent history, 1988 to be precise, illustrated the negative effects artificial intelligence – hereon also referred to as AI – may have if incorporated poorly into existing processes. This thesis uses a commonly understood definition of artificial intelligence, defined by IBM: "Artificial intelligence leverages computers and machines to mimic the problem-solving and decision-making capabilities of the human mind" (2020). One could say that the events which were to unfold were due to the immaturity of the technology. Back in 1988 discrimination was proven to exist in the admission process of a UK-based medical school. Prior to proving such, speculations of discrimination had been expressed for multiple years by the applicants. The computer algorithm used to assist admission officers in the admission process discriminated against women and those from ethnic minorities. Where bias enters the scene is when it was realized that the algorithm had been programmed to emulate human admissions officers' admission decisions. Noteworthy is that the algorithm had been fine-tuned up until 1979 when the tool reached a correlation of 90-95% with the decision made by the human officers. (Lowry & Macpherson, 1988).

It can be concluded that the events which unfolded back in 1988 decreased the level of trust in the use of artificial intelligence within the admission process, even when the decisions followed

similar patterns human admission officers would make. In today's world, multiple frameworks exist to ensure the fairness of AI – however, the definitions of fairness vary between the artificial intelligence softwares, still resulting in possible biased outcomes (Manyika, Silberg, & Presten, 2019). The European Commission's High-Level Expert Group on Artificial Intelligence – hereon also referred to as AI HLEG – has since its formation back in 2018 been putting efforts in proposing possible guidelines for ethical AI, which in turn would minimize unfair bias present in AI. The group has worked closely with academia, professionals and policymakers to ensure that all aspects of artificial intelligence algorithms, regulation and industry needs are met and complied with. (European Commission, 2021a).

All of this raises the question on how bias may affect the trustworthiness of AI. A study conducted by Lee & See (2014) argues that if the level of trust in the technology is unproportionally different from the actual capabilities of the technology, the lack of use will ensue, also referred to as disuse. What's more, as certain trust determinants vary from country to country, the context of the trust transaction matters greatly. Trust determinants are underlying aspects, such as wealth and education, which affect the level of trust.

This study consists of interviews conducted with Deloitte employees who work with AI on a day to day basis. Currently Deloitte offers tools and solutions for their clients in order to comply with the local laws and regulations, but more importantly to attract potential new users of their clients' tools either directly or indirectly. An example of such tool is called *Glassbox*, to which certain interviewees may refer to, and to which is referred to in the results. Glassbox is a string of code which analyses the fairness level of said artificial intelligence algorithm (Waaijer & Chronéer, n.d.).

1.2 Research Question & Objectives

The aim of this research is to shed light on the rather unclear waters on methods to adjust trust towards AI algorithms and the different effectors' impact on the trustworthiness of AI, keeping in mind that not all AI solutions have not been designed to function the same way, nor are their fundamental purposes of existence the same. By effector, this paper refers to any means, aspects, actions or the lack thereof in the context of the study. All in all, the research question culminates to the following main research question, and a sub-research question:

RQ: ”How can a user’s trust towards an artificial intelligence algorithm be adjusted in the context of European consultancy services?”

Sub-RQ: “What effect does bias in artificial intelligence algorithms have in facilitating trust in towards the algorithms?”

Another objective of this study is to provide some insight onto how trust is impacted by underlying effectors and events within the artificial intelligence algorithm, and to identify how large said impacts may be. By leveraging pre-existing studies and collected data, the outcomes of this study may be extended to cover all types of artificial intelligence implementations within Europe and possibly Western societies.

Inevitably, whilst looking into what the effects of user trust to the trustworthiness of AI are, suggestions of ways of adjusting said trust will pop-up.

The relevance of the study is supported by both the academia and the professional landscape. A contact person from Deloitte, with whom this thesis has been discussed with on multiple occasions, argues that neither the clients nor the policy makers fully grasp the importance of the topic. He continues by stating that often times a “carrot and a stick” approach is necessary when dealing with the clients, as often the downsides of not being ethically or legally compliant is not brought up nor visible in the context of AI algorithms.

1.3 Scope

Artificial Intelligence is embedded into our lives in several ways – thus it is important to delimit the study and its outcomes. Trust in artificial intelligence could be discussed from a corporate standpoint, from a user standpoint and possibly from a regulatory standpoint. However, this study is limited to analyzing the European context from the organizational standpoint of the “fixer”. By fixer, we are referring to Deloitte as a partner in helping their clients, often times large corporations, to solve their issues with artificial intelligence compliancy. By purely being compliant, often times leads to rather ethical AI, as compliancy requires the operating within the anti-discrimination laws and regulations. And by having an ethical AI, bias in it should be minimized, increasing or decreasing trust to an appropriate level. (Siau & Wang, 2020).

With the delimitations of the study, the way the study leverages the sampling methods is affected. As we're analyzing the trust spectrum users have towards AI algorithms, and how to adjust it accordingly via minimizing bias and other possible means, the selection of professionals was limited to only include employees of Deloitte.

1.4 Thesis Outline

The thesis follows a structure which should ease the reader's understanding on the studied topic. Chapter 1 introduced the topic on a high-level, teasing the reader and providing motivations to why the study is conducted and the importance of it. The chapter includes the developed research question(s) and scope delimitation, from which the remaining thesis has been built on. In chapter 2 relevant studies and frameworks have been introduced in the way of a literature review, to shed light on what the current playing field looks like, including the world's current understanding on the topic and supporting topics. The successor of said chapter is the methodology chapter, numbered as the third chapter. In the methodology chapter, the research philosophy and overall approach to data collection is discussed based on Saunders' (2015) research onion. After clarifying the research approach and data collection methods, the findings and brief analysis of the data is presented in chapter 4. From there, chapter 5 consists of the discussion section, where the results are presented in relation to what the academia knows already, and what might be expected from the industry going forward. The study concludes in chapter 6 where the summary of the results are presented and how the results may affect the existing playing field of artificial intelligence, namely answering the research question. Further, chapter 7 brings out the limitations of this study, and suggests a direction for further studies.

2

Theoretical Framework

To analyze the trust in AI, it is of utmost importance to first understand what the academia regards of trust between humans, and how trust is projected in the relationship between human and computers. Furthermore, are there common biases in AI algorithms identified by the academia, and if yes, what is said about the biases? The European Union's officials at least think so, thus the possible to-be legal frameworks created shall be presented in the last sub-chapter to see what the official policy might be going forward.

2.1 Context Dependent Trust Between People

Trust is a natural characteristic within humans and can be observed across the globe (Henderson & Churi, 2019). The widely accepted definition of trust is "a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of others" (Rousseau, Sitkin, Burt, & Camerer, 1998; Supported by Dietz, Gillespie, & Chao, 2010 and Lee & See, 2014) and can be categorized in multiple ways, some of such categorizations being government trust, business trust and most notably personal trust. (Henderson & Churi, 2019; See also Sousa, Lamas, & Dias, 2014 for a high-level categorization of trust). However, how trust is built, perceived and upheld varies based on the context present during a specific interaction. Context is often referred to, when discussing culture – it can be as high-level of notion as a geographical and/or demographical culture, but also as specific as an individual organization's culture. (Dietz, Gillespie, & Chao, 2010). Language is also a variable in the equation of trust (Henderson & Churi, 2019).

Fundamental pillars of interpersonal trust in a business context consist of six elements: Reliability and Dependability, Transparency, Competency, Authenticity, Fairness, Openness and

Vulnerability. The aforementioned elements are intertwined, and respectively affect the notion of trust an individual cast over their peer. (Jaffe, 2018; See also summarized version by Zenger & Folkman, 2019 and Sousa, Lamas & Dias, 2014).

However, trust decreases between people as social distance increases, no matter the cultural context (Buchan & Croson, 2004). The decrease in trust caused by social distance is negated partly by allowing personal communication between the parties before an action which requires trust is enacted on (Buchan, Johnson, & Croson, 2005). Contrary to universal trust determinants i.e., trust determinants independent of context, macro-factors on a geographic level such as wealth, education and strong formal institutions are non-universal and thus affect the level of trust an individual portrays towards others differently from geography to geography. (Ferrin & Gillespie, 2010). Studies have shown that the principles driving interpersonal trust provide a valid framework to analyze human-computer trust (Madhavan & Wiegmann, 2007).

2.2 Human-Computer Trust

As the complexity and calculation power of computer agents keep growing, human-computer trust has been studied from multiple angles, some of which being the trust portrayed during cooperative work settings, others identifying the trust characteristics and components. (Sousa, Lamas, & Dias, 2014; Kulms, 2018). The shift in Human-Computer Interaction from computer agents being seen merely as computational tools towards acknowledging their importance and effect on humans' work took place between 1960-1980s. (Kulms, 2018; See also Muir, 1987). It has since been identified that users who interact with computer agents project societal norms onto the agents, including gender biases. (Nass, Steuer, & Tauber, 1994).

Trustworthiness in automation can be referred to as the capability of said technology. Inappropriate levels of trust in automation, may lead to misuse or disuse of automation. (Lee & See, 2004). Automation is a broad term, which also encompasses AI (Gaynor, 2020). Facilitating the correct amount of trust with users is highly important in avoiding misuse or disuse of said technology. (Wicks, Berman, & Jones, 1999). Lee & See (2004) present a graph which visualizes the optimal relationship between the trustworthiness – also referred to as Automation Capability – of the technology and users' trust. The identified dimensions are calibration, resolution and specificity, all

of which have been incorporated into their framework based on the other academic studies. Going forward, the term *trustworthiness* will refer the automation capability, only when discussing the below framework. Otherwise, the term *trustworthiness* and any deviations of it will reflect the definition given by the Merriam-Webster thesaurus, "worthy of one's trust" (Merriam-Webster, n.d).

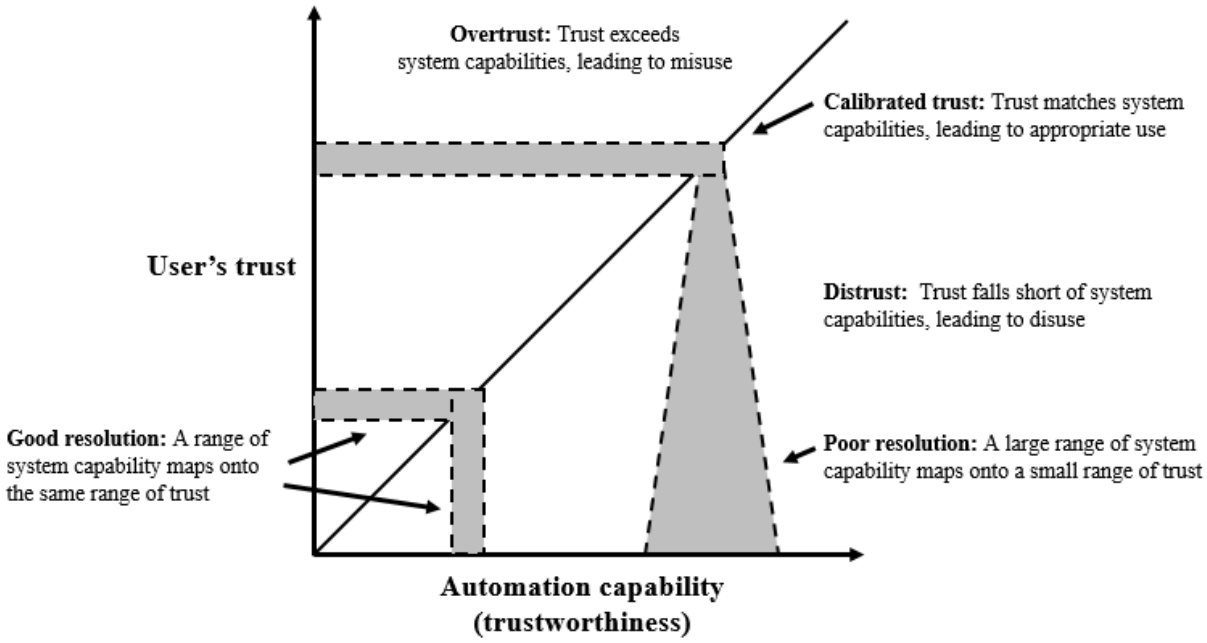

Figure 1 - Mapping the appropriate level of trust in automation (Lee & See, 2004).

The balance of user's trust in the AI algorithm and the AI's capabilities is referred to as calibration (Muir, 1987; See also Lee & See, 2004). Resolution refers to how much the user's casted trust deviates from the capabilities of the technology – high resolution meaning that there is little to no deviation. The definition of specificity may vary from paper to paper, but for the purposes of analyzing the trust in AI, the definition of temporal specificity shall be used. Temporal specificity refers to the user's trust fluctuation levels when there are changes in the trustworthiness of the technology. Low temporal specificity signifies that the user's trust levels are only affected by long-term changes to the trustworthiness of the AI. High temporal specificity understandably expresses the user's trust levels being affected by the slightest changes to the AI's trustworthiness. Appropriate level of calibration, high resolution and high temporal specificity may decrease and prevent the misuse and disuse of AI. However, it is noteworthy that trust is impacted by individual, organizational and cultural factors. (Lee & See, 2004). Specifically, the calibration of trust is influenced by environmental factors (Madhavan & Wiegmann, 2007).

What comes to decision supporting systems (DSS), the source of diagnostic information, reliability of data, and the credibility or perceived expertise, are factors which influence the level of trust shadowing AI (Madhavan & Wiegmann, 2007). It is argued that users blame technology for the mistakes the users themselves have done, and fail to recognize their role in the negative outcome of the technology at hand (Sampson, 1986). Moreover, it has been observed that users often *Overtrust* the results given by the computer agent (Madhavan & Wiegmann, 2007). This is turn may lead to the crumbling of the DSS's credibility. The reasoning behind overtrusting the computer agent stems from the perception of the computer agent being more objective and rational contra to a human counterpart. (Dijkstra, Liebrand, & Timminga, 1998). Moreover, what comes to the credibility factor of trust in AI, credibility plays a large role in the equation. What has been identified is that the perceived credibility of the data source plays a larger role than the actual credibility or expertise. This leads to a situation where computer agents are trusted less than their human counterparts when the human agents are framed as having greater expertise compared to the computer agent. Further, high initial expectations emitted by the user onto the AI accompanied by low reliability of the AI, has been shown to have a more detrimental effect of the level of trust on AI than in similar cases where the agent would be a human. (Madhavan & Wiegmann, 2007).

2.3 Artificial Intelligence

2.3.1 Ethics in Artificial Intelligence

The ethics of human beings have been discussed and pondered on since civilized and developed societies first became to existence. One of the earliest moral codes which still is available to us, is the Code of Hammurabi from around 1750 BCE. (Singer, 2021). And when Alan Turing first introduced the idea of "a machine that can learn from experience […] and alter its own instructions" it can be argued this was when articulating the ethics for AI also became relevant (Copeland, 2021). Studies conducted near the year 2020 have suggested that leveraging human cognition models would result in proper guidelines for ethical AI (Thomsen, 2019), whereas others suggest applying marketing ethics to AI algorithms would output an AI that treats its users and targets ethically. It is suggested that there should be norms and rules set in place for AI which cannot not be violated, unless the consequences of not breaking said norm or rule results in an overall worse situation for a larger

audience. (Ferrell & Ferrell, 2021). However, how is a "worse" situation really determined? Ferrell & Ferrell (2021) acknowledge the complexity of the ethical questions with AI, but nevertheless argue that if the consequence of breaking a rule adheres to a greater good principle, then such violating action should take place.

Siau & Wang (2020) present a framework for ethical AI, in which they've used sources from eight differing institutions and organizations, some of which being the United Nations Educational, Scientific and Cultural Organization (UNESCO) and the European Commission (EC). According to them, the factors identified, as shown in Figure 2, should all be considered when developing AI's ethics. It is argued that even though a framework for ethics in AI exists and has been recreated multiple times by various parties, a consistency between them should be reached, and the prior mentioned high-level rules and norms by Ferrell & Ferrell should be created as an industry standard both in business context, but also on a legal level. (Siau & Wang, 2020).

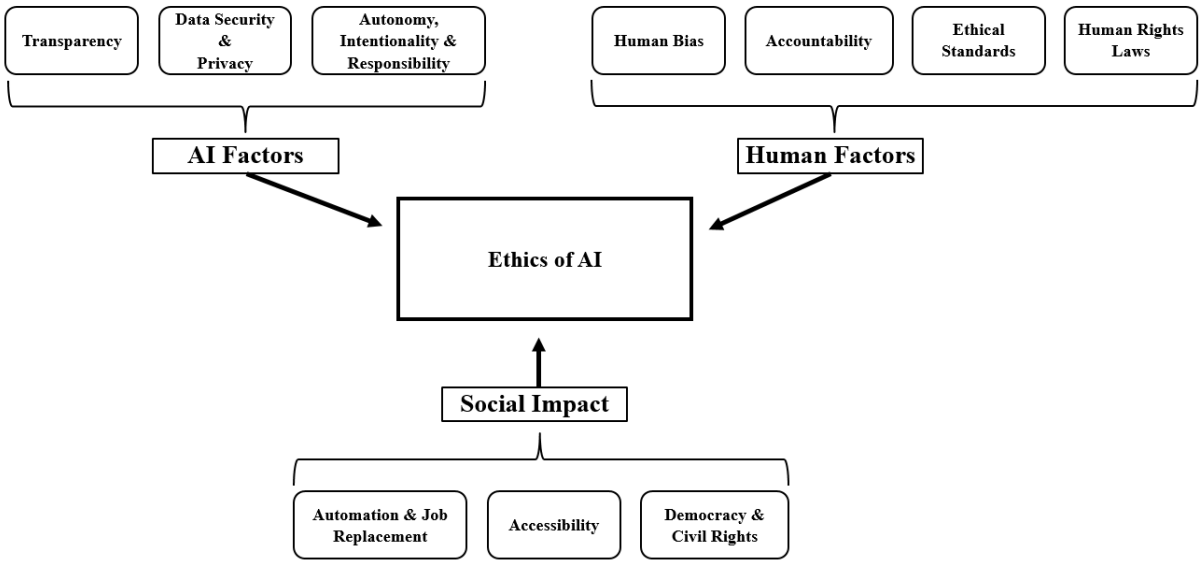

Figure 2 - Framework of building ethical AI (Siau & Wang, 2020).

The role of self-governance in AI is as crucial as legal and governmental regulation. By self-governing AI, a mean of communication and discussion is more likely to be opened, as self-governance often relies on ensuring a proper degree of ethical guidelines to exist within the party who develops AI algorithms and AI administrating party. Communication and information disclosures may support the development of societies to a level where Distrust towards AI is nearly eradicated. The communication is recommended to include forums for discussion as well as making the ethical

guidelines used by companies, industries and regulatory decisionmakers public. (Siau & Wang, 2020). Noteworthy is that creating ethics for an AI algorithms is more challenging than teaching ethics to a human being. One key reason for this statement is the artificial intelligence's inability to "easily and naturally develop empathic feelings for humans" and other agents, whereas humans may retain a certain level doubt towards AI and other agents. (Thomsen, 2019).

Similar to trust between humans, ethics may differ between geographical locations. This understandably rises the issue of having an ethical AI, which adheres to each geography's code of ethics. The fact that the development of AI is taking place in developed countries where ethics may differ from the lesser developed countries, but may also differ between developed countries examples of such being Finland and China, poses another question of a universal ethics guideline for AI. Moreover, an additional issue for generating a universal ethics guideline or framework for AI is the different applications of AI algorithms – some applications may substantially change peoples' lives (Hao, 2019; see also McIlwain, 2020) whereas others may simply generate art or commercial activity with minimal or no effect on people. (Carrillo, 2020).

2.3.2 Bias in Artificial Intelligence

Multiple studies have been conducted in recent years on human bias being present in the hiring process – even in processes where most of the initial processing is done by AI. AI bias is often referred as an "anomaly in the output of machine learning algorithms". Bias can also be considered as something that is present in the algorithm, prior to it giving an output. (Kantarci, 2021). A recent study from the Swiss labor market suggests applicants belonging to ethnic groups other than the majority Swiss population are "penalized" by lesser contact rates of up to 19% by employers. Moreover, the same study suggests than men experience a penalty of 7% when applying for female dominated professions. The same phenomenon applies for a vice versa situation as well, suggests the study. (Hangartner, Kopp, & Siegenthaler, 2021). A report from the United States outlines a similar phenomenon. The Equal Employment Opportunity Commissions (EEOC) found that African-Americans, Hispanics and women were lacking representation in the technology industry. However, the disparity can be partly attributed to under-representation of said groups within the necessary education degrees, and the other factors employers consider, such as extracurricular activities including a profile of a popular coding website, Github, or taking part in hackathons which essentially

are a gathering of people interested in technology and hacking. (Mone, 2016). Other events of bias showing itself in AI algorithms, are the cases of labeling African-Americans as Gorillas by the Google Photos software, and lightening of darker pigmented skin by another photo application (Kasperkevic, 2015; See also Morse, 2017). Bias in AI can be categorized on a high-level as being either cognitive bias or bias due to the lack of complete and inclusive data. Cognitive bias takes place during the development process of the algorithm, where the creators of the technology are unconsciously including biases into their computational models. (Kantarci, 2021).

As the training data used for algorithms are being constantly improved to be more inclusive, the effect on trust and fairness of the AI is positively impacted by minimizing the consequences of incomplete training data. When an AI's algorithm takes gender, ethnicity or other sensitive personal information into account whilst computing and executing the tasks it set out to do, structural or explicit discrimination, or even new forms of biases may be replicated or created. An AI's algorithm is most often capable of adapting to new data it receives, thus a decision made by the AI may reinforce societal biases. (Lee, 2018). Ferrell & Ferrell (2021) argue that if an AI does not have a programmed ethical component, it could produce biases unintentionally. Furthermore, a governmental report recommends AI algorithms to account most of the dimensions identified by Siau &Wang (2020) in Figure 2 during development, or post-deployment if necessary. And if said inclusions are impossible to be included, such technologies should be abandoned (Executive Office of the President, 2016). To date, computer and data scientists behind algorithmic AIs are considering the right balance between hard data and accuracy of their models, especially when the underlying data is identified to be incomplete and non-inclusive of all groups. (Lee, 2018).

In terms of racial bias in algorithms, researchers are not conclusive on where and when bias is born in the modelling process of the AI (Lee, 2018). Certain studies argue the disparity to become present already in the training data (Corbett-Davies, Pierson, Feller, Goel, & Huq, 2017). However, others suggest that bias becomes a factor only after the user has interacted with said AI (Bucher, 2012). Lee & Lee recommend transparency together with giving the user control over their digital footprint as a method of lowering the bias (Lee, 2018). McKinsey & Company introduce multiple potential ways of minimizing bias in AI algorithms, shown in Figure 3. (Silberg & Manyika, 2019). The proposed ways of minimizing bias in algorithmic AIs in practice support the academia's and European commission Expert Group's (2019a) conclusions on facilitating correct amount of trust in AI and minimizing bias in AI.

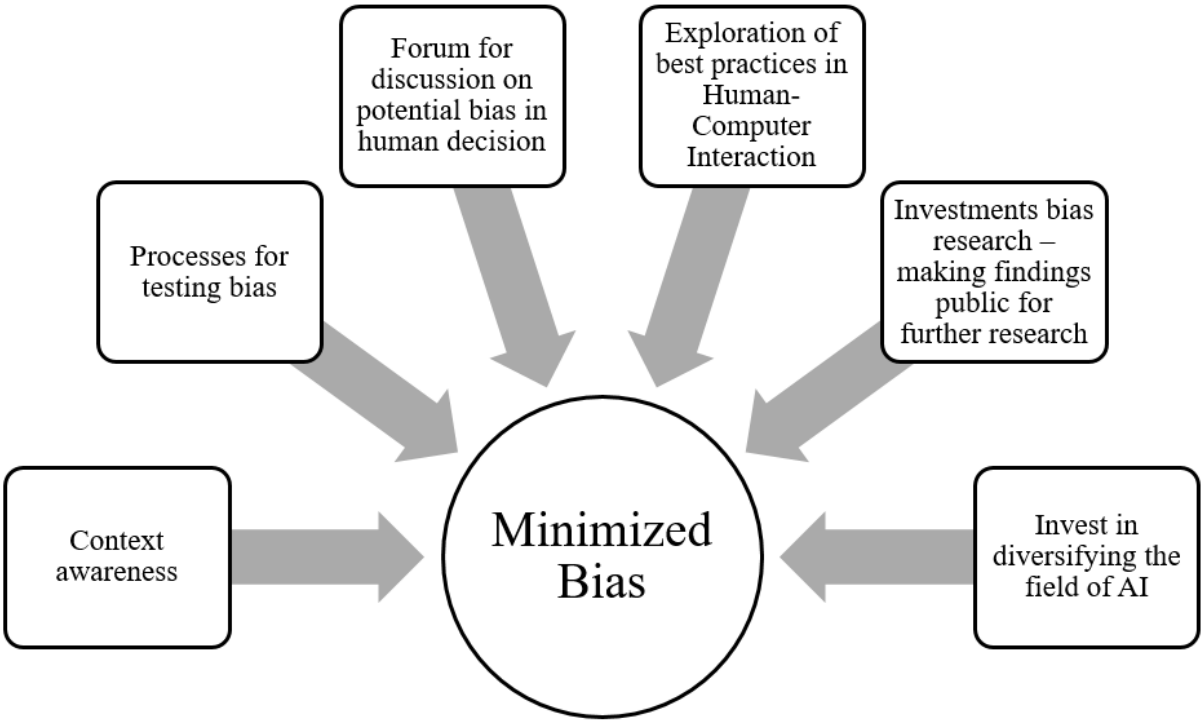

Figure 3 - Potential actions to minimize bias in AI (Re-visualized, Silberg & Manyika 2019).

2.3.3 Possible future EU-wide legislation

The European Commission's High-Level Expert Group on AI has been created in an effort to outline the necessary actions that need to be taken with artificial intelligence as a whole. They state that the implementation of AI is still at a level where it is possible to impose certain frameworks for the creation of ethical and fair AI algorithms. What is presented by the AI HLEG in their various reports is not to be taken as legally binding. (High-Level Expert Group on Artificial Intelligence (AI HLEG), 2019b).

The AI HLEG recommends artificial intelligence to be legally compliant, ethical and trustworthy in order to be effectively considered as trustworthy. Purely adhering to the ethics standard set in their report, does not constitute as a trustworthy AI, as other prior mentioned factors of their framework are not fulfilled. (High-Level Expert Group on Artificial Intelligence (AI HLEG), 2019a). However, Siau & Wang (2020) argue this by stating that by purely being compliant, often times the ethical standards are met, thus creating ethical AI. Artificial intelligence technology does not operate in a lawless vacuum, but rather in a space where national level and EU legislation apply, depending

on the context. Some of said legislation to mention on an EU level is the General Data Protection Regulation – also known as GDPR – and directives against discrimination (2004/113/EC; 2000/43/EC; 2000/78/EC; 2006/54/EC). For algorithmic AI to be trustworthy, it needs to comply with moral and ethical standards present in a local culture. By solely being legally compliant, the AI might not be on par with the ethical and moral expectations from the users, essentially decreasing its trustworthiness. The developers are required to ensure that the AI upholds the set moral and ethical standards after deployment phase and does not train itself to disregard the standards based on data. Moreover, the users must be able to trust the AI to not make unintentional mistakes to the highest degree possible. The factors impacting the unintentional actions taken by the AI may be context dependent, and situations where one action or decision might be considered acceptable and unacceptable in another context. Both the technical context and social context needs to be considered pre-deployment phase of AI. (High-Level Expert Group on Artificial Intelligence (AI HLEG), 2019b).

The AI HLEG present a process of 7 fundamental steps to achieve a trustworthy AI. The 7 steps, or requirements, are as follows: Human Agency and Oversight; Technical Robustness and Safety; Privacy and Data Governance; Transparency; Diversity, Non-Discrimination and Fairness; Societal and Environmental Well-being; and Accountability. (High-Level Expert Group on Artificial Intelligence (AI HLEG), 2019b; See also High-Level Expert Group on Artificial Intelligence (AI HLEG), 2020). The self-assessment list provided by the AI HLEG is an opportune event to check whether the AI at hand sufficiently adheres to the recommendations set forth by the group. Actors affiliated with the AI, i.e. the industry leveraging algorithmic AIs, predict the AI HLEG and European Commission's recommendations to become regulations and directives in the near future (Edwards & Nevola, 2020). Questions set forth in the self-assessment list are such as "*Is the AI system overseen by a Human-in-the-loop or Human-on-the-loop".* Most relevant question for the thesis at hand is as follows: *"Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design?"*. (High-Level Expert Group on Artificial Intelligence (AI HLEG), 2020). The High-Level Expert Group tries to provide examples on how these questions set in AI HLEG (2020) could be solved. They suggest that by following the current laws on discrimination and seclusion, many of the issues possibly raised by the question are solved. Moreover, by having a large time spread in the training data and the data used in the decision-making algorithms, the problem of creating new biases or reinforcing existing ones would be mitigated. They continue stating that having people from diverse

backgrounds working with the AI's development could impact how well biases are identified. (High-Level Expert Group on Artificial Intelligence (AI HLEG), 2020). The guidelines and papers put forth by the AI HLEG serve as a basis for the European Commission's efforts to safeguard the values and rights of users (European Commission, 2021b).

3

Methodology

3.1 Research Philosophy

From the multiple available and differing perspectives on research philosophies, this paper's research shall be conducted with an interpretivist perspective. Interpretivism is also known as anti-positivism or negativism, as it argues against positivism's fundamentals of there being a single reality. (Krauss, 2005). The stance of interpretivism of this paper lies in the notion of the non-existence of a universal trust in AI, but rather a complex set of trusts all dependent on the domain of the user.

The system of beliefs one holds whilst conducting research is called a research philosophy. Unconscious decisions and assumptions are constantly made throughout the process of research, which inevitably affect the development of knowledge. (Saunders, Lewis, Thornhill, & Bristow, 2019). Interpretivism interprets each piece of evidence as being socially constructed and it, the evidence, having its own social roots, due to which the evidence may not be universal throughout the globe, contra to positivistic research where evidence is often taken as de facto. Interpretivism is often used with qualitative research. (Elliot, Fairweather, Olsen, & Pampaka, 2016). Hence the usage of interpretivism seems the most feasible philosophy to be used to conduct this research.

3.2 Research Approach

A so-called research onion is often used in the planning of the underpinning constructs of a research. The widely used and accepted research onion introduced by Saunders et al. lays out the details that have an effect on the researchers' attitudes and approaches, and eventually outcomes. In the second layer of their "onion", the approach to theory development is articulated. Deduction, Induction and

Abduction are the approaches specified. In deduction, the researchers start out by having a hypothesis which is tested throughout their research, in order to reach a logical conclusion. A differing approach, induction, starts out by having a broad generalization from a rather narrow set of observations. From these observations and generalization, a hypothesis is formulated, from which a conclusion is drawn. (Saunders, Lewis, Thornhill, & Bristow, 2019). As this paper does not revolve around creating theory as an outcome existing theory, nor does it truly use broad generalizations to reach a generalized conclusion, the third approach, abduction, is used throughout this research. Abductive research pursues to find the likeliest solution to an issue based on observations. (Mantere & Ketokivi, 2013). The said approach does not disregard theory, but questions whether existing theory may explain everything that is observed. Moreover, the benefit of an abductive approach is its flexibility to fluctuate between an inductive and a deductive approach. (Saunders, Lewis, Thornhill, & Bristow, 2019). All in all, as an abductive approach can be summarized to be a combination of both deductive and inductive approaches, it is the most fitting option for the paper.

3.3 Methodological Choices

3.3.1 Research Purpose

"The purpose of interpretivist research is to create new, richer understandings and interpretations of social worlds and contexts." (Saunders, Lewis, Thornhill, & Bristow, 2019). In this paper, we seek to understand the underlying drivers behind trust in artificial intelligence, and how to mitigate the existing biases within them in an attempt to increase trust. Current academic research seems to have a gap in the understanding of this specific topic – however, the business industries involved in AI development and monitoring are largely invested in further generating trust in AI for a wider adoption throughout businesses, and thereby have presented possible frameworks to achieve their goals. The commonly categorized research purposes are as follows: *Exploratory, Descriptive and Explanatory* (DiscoverPhDs, 2020). By there being a limited amount of understanding around the topic specifically, this paper follows an exploratory research purpose, with the intent of shedding light and providing a better understanding on the nature of the problem. By having an exploratory research purpose, this paper does not aim towards generating a conclusive outcome, but rather to create a hypothesis. (Sue & Ritter, 2012).

3.3.2 Research Method

Hand-in-hand with the research purpose is the option for a quantitative or qualitative research method. The choice for the research method is often tied to the research purpose, approach and philosophy. The former research method means research via methods of numeric analysis, where survey and questionnaire data are often analyzed in quantitative units (Yoshikawa, Weisner, Kalil, & Way, 2008). Staying true to our prior choices on our philosophy, research approach and purpose, this paper leverages and conducts a qualitative research. Data from exploratory studies tend to be qualitative (Sue & Ritter, 2012). Saunders et al. mention that a research conducted from an interpretivist perspective typically follows an inductive approach, has small data-samples, digs deep into in-depth investigations, using qualitative methods of analysis. (Saunders, Lewis, Thornhill, & Bristow, 2019).

3.4 Research Strategy & Time Horizon

To achieve a scientifically valid research, we use Bryman's six-step process designed for a qualitative research. The process involves eight steps but can be summarized into six. The first step is to present the general research question. The presentation and identification of a research question then leads towards selecting relevant data subjects. In this paper, professionals within industries leveraging AI have been identified as relevant data subjects. The collection of relevant data is then started via means of semi-structured interviews, as to provide enough space for discussions and thoughts, not thought of by the interviewer. After the collection of data, its interpretation shall begin, by first transcribing all interviews after which the actual interpretation begins from the chosen philosophical perspective. The fifth step comprises of the two additional steps mentioned earlier. In this part of the process, the conceptual and theoretical work is conducted, which thereby results in a loop back to the interpretation of data, fourth step, by first tightening the specifications of the research question, then collecting further data. After all the above, the findings and conclusions may be written, resulting in a qualitatively valid research. (Bryman, 2012).

Mainly due to time limitations set on the research, the time horizon is cross-sectional. Saunders et al. present two options for the time horizon of a research – cross-sectional or longitudinal. Longitudinal research refers to a study which is conducted over a large span of time, which will not be the case due to time constraints and nature of the study. A cross-sectional study instead analyses

the current state of affairs in the researched topic at a specific point in time. (Saunders, Lewis, Thornhill, & Bristow, 2019).

3.5 Data Collection

Main form of primary data collection has taken place via means of semi-structured interviews. A select group of professionals within the consulting firm Deloitte have been chosen as data subjects. Said data subjects have varying demographic, educational and professional backgrounds, but mainly work with AI at the time of the study. The data subjects are currently employed by differing member firms across the globe. As a single data-collection technique and a corresponding analysis method is used, this research follows a mono method qualitative study. Semi-structured and in-depth interviews which are conducted for the research of this paper, are categorized as being unstructured, and the terms are used rather interchangeably. In semi-structured interviews, an overall theme and questions related to said theme are established, in an attempt of opening the discussion forum for thoughts and insights that the possibly prior unknown to the researcher. The benefit of a semi-structured interview is its flexibility, as it enables the option of omitting certain questions from the interview due to organizational or another context. Compared to a structured interview, where a rigid schedule is followed, semi-structured interviews compliment the research type. Between the options of conducting a non-directive interview, also known as informant interview, or a focused interview, this paper's data collection aligns with the definition of focused interviews. (Saunders, 2015).

Informant interviews consist of enabling the interviewee to freely talk about their experiences and events. Contrary to this, a focused interview in turn is an approach where the interviewer has greater control of the direction of the interview, whilst still giving the interviewee a chance to express their opinions. Furthermore, a decision of conducting one-to-one or one-to-many interviews must be derived. (Saunders, 2015). Due to the nature of the study, this research leverages one-to-one interviews as its main way of data collection.

3.5.1 Sampling

As collecting data and input from everyone affected by AI could be deemed impossible, a census shall not be done. Census refers to a data collecting method determination, where data is collected and analyzed from every possible case or group member. Instead, sampling is used to determine a

suitable set of data subjects, from whom data shall be collected. Sampling enables the reduction of data from all stakeholders and parties within each stakeholder-section, to data just from sub-groups. (Saunders, 2015). The sampled data in our case will be generalized as deemed necessary in order to answer our research question. Some researchers suggest that sampling might have a higher accuracy in terms of accuracy of results. The main argument behind this is the time allocation – instead using time collecting the data from a substantial number of participants, higher quality data collection methods may be used. (Brown, 2006; See also Barnett, 2009).

Sampling is often categorized into two techniques; *probability and non-probability sampling* (Business Research Methodology (BRM), n.d.; Saunders, 2015). In probability sampling every member of the population has an equal chance in participating in the data collection. This technique consists of random data subject selection, whereas non-probability sampling does not in the same manner. In non-probability sampling, the data subjects are chosen in a non-random matter. The sub-techniques of non-probability sampling are the: *Quota; Purposive; Volunteer and; Haphazard techniques.* The sub-techniques may have underlying sub-techniques as well. (Business Research Methodolgy (BRM), n.d.). This research leverages the use of purposive and volunteer sampling.

By leveraging purposive sampling, the researcher withholds the right to select the participants, and selection is made per judgement of the researcher. The benefits of using purposive sampling is its time efficiency and the adequate representation of the population. However, it is argued that a purposive sampling method is unscientific as it incorporates large amounts of personal bias in the selection of data subjects. To mitigate this, volunteer sampling's sub-technique snowball sampling is used in tandem with purposive sampling. Snowball sampling essentially enables the researcher to recruit hidden parts of the population into the sample. In said technique, the sample group members nominate additional members to participate in the research. A possible downfall of snowball sampling is the over-representation of a particular network. The criticism toward the snowball technique is echoed throughout all the sampling methods. If a wrongful determination of the overall population is made, a certain group may be over-represented. (Business Research Methodolgy (BRM), n.d.).

The selection of data subjects chosen for the sample group consists of professionals within the field of AI. All of them are employed by a Deloitte Touche Tohmatsu member firm – hereon Deloitte. As earlier stated, purposive and snowball sampling techniques are used. Purposive sampling is used in the initial steps of choosing the sample group – we have identified three (3) data subjects via the purposive sampling technique. Moreover, the three data subjects have further nominated

additional data subjects to be a part of the research. The criteria for being a participant in the study are as follows: minimum 2,5 years experience in the field or; research in a relevant field. Table 1 and 2 below list the data subjects, categorized per sampling method.

Name	Deloitte Member Firm & Department	Position	Approach	Years of Experience
Benjamin Chronéer	Deloitte Netherlands – Risk Advisory	Senior Consultant	Technical	2,5
Valeria Gallo	Deloitte United Kingdom – Risk Advisory	Senior Manager	Legislative	4
Andrew Joint	Deloitte United Kingdom – Tax	Partner	Legislative	20

Table 1 - Data subjects by purposive sampling.

Name	Deloitte Member Firm & Department	Position	Approach	Years of Experience
Alexander Galt	Deloitte Netherlands – Risk Advisory	Consultant	Ethical	3,5
Sebastiaan Berendsen	Deloitte Netherlands – Risk Services	Consultant	Technical	2,5
Ivana Bartoletti	Deloitte United Kingdom – Risk Advisory	Associate Director	Ethical / Legislative	12

Table 2 - Data subjects by snowball sampling.

3.6 Qualitative Analysis Method

Using an inductive approach in qualitative data analysis raises some questions on the accuracy of the results and conclusions drawn from the collected data. However, there is debate around using a

deductive approach with qualitative data analysis, as it is suggested that the theoretical framework built might restrict the identification of insights from the collected data. (Saunders, 2015). Hence, the choice of an abductive approach is supported in an effort to mitigate the issues raised by both approaches. However, as our study is of an exploratory nature, the elements of an inductive approach qualitative analysis is used.

The data collected should be analyzed as it is collected, and a conceptual framework should be built around the themes arisen, to truly identify the underlying topics within the research in question. As emphasized by Saunders, it is of utmost importance to have audio recordings of the interviews, which shall be transcribed further. The importance of audio recordings is supported by the fact that the subtle changes in the tone used in speech may affect the actual message wanted to portray. Furthermore, transcriptions assist the formulation of a conceptual framework used later in a thematic analysis. (Saunders, 2015).

A thematic analysis is often used to analyze qualitative data. A thematic analysis enables flexibility in terms of analyzing the data collected. Moreover, it is considered as a systematic and logical way of analyzing data, as it offers robust descriptions, explanations and theories of phenomena identified. It is often used to identify key themes and patterns, and to further draw and verify conclusions. (Saunders, 2015).

The process of analyzing qualitative data via thematic analysis, follows a six step procedure portrayed in Figure 4.

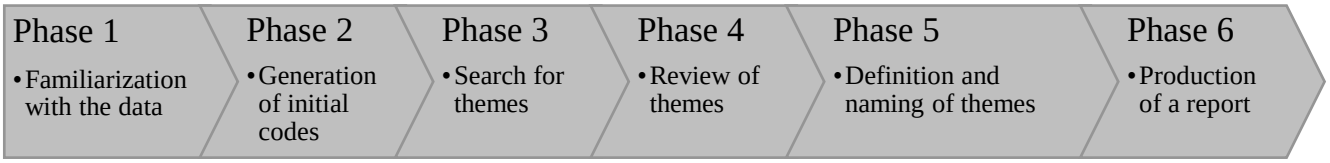

Figure 4 - Process of thematic analysis (Nowell, Norris, White, & Moules, 2017).

In order to familiarize with the data, it was chosen to transcribe the interviews manually. This way, mistakes made by an automated transcription tool are completely mitigated, increasing the accuracy of the data, noting that speech recognition software that are commercially available have an average error rate of approximately 12% on phone conversations. (Worthy, 2019). One could note

that the decision to transcribe the interviews manually due to lack of trust in the artificial intelligence is ironic, considering the research question. However, as automatic transcription tools may be inaccurate in understanding accents, the decision of manual transcription is supported. The transcripts consist of all of the elements of the discussions had, including pauses, possible mistakes and errors. By documenting each audio aspect of the interview, the subtle differences in tone are captured increasing the trustworthiness and accuracy in the data. (Nowell, Norris, White, & Moules, 2017). After familiarizing with the data and transcribing them, preliminary codes are created to analyze the data more clearly. The preliminary codes are essentially labels for the data collected, and are based on the dimensions and details identified in the theoretical frameworks. Furthermore, a number of specific words are used in the preliminary codes, which might or might not be identified in the theoretical framework. The use of the software NVivo is leveraged in identifying and reviewing the themes present in the data sets. NVivo enables the organization of complex code sets and vast amounts of data, into a visually understandable format. (Saunders, 2015).

3.7 Assessment of Quality

Reliability and Validity are metrics used to evaluate the quality of the research. Both indicate how well a research succeeds in measuring the study at hand. Reliability refers to the consistency of the measure, whereas validity refers to the accuracy of said measure. (Middleton, 2020).

As key concepts of credibility of the study, by having high reliability in a study, the study should be able to be replicated on a later stage, and should result in the same findings and conclusions. (Saunders, 2015). However, an affecting dimension to the study is the everchanging operating field of Trust and Ethics within AI. Findings of this study may thus differ from an identical study conducted in the future, as new frameworks and operational incentives may be identified. What's more, the validity of this study should correspond with the limited understanding of the academia on the topic.

4

Findings

This chapter is approached from a perspective of presenting and analyzing the insights provided by the interviewees. The process of the analysis has been detailed in Chapter 3, section 6. A thematic analysis is built based on the interviews, from which insights and correlations between the insights of interviewees are identified. Said correlations are referred to as themes.

The interviews were conducted with the intention of confirming and/or disproving arguments presented in chosen academic papers, detailed in Chapter 2, *Theoretical Framework,* from which the research question and the sub-question has been derived from. Moreover, as some of the interviews displayed unstructured characteristics in terms of interview questions, the data collected from certain data subjects may not be supported by other interviewees solely due to the fact that the discussions were unique for said interviews, not due to disagreement.

For reasons of simplicity and readability, the interviewees are referred to with their initials – shown in Table 3 below.

Name	*Initials used*
Benjamin Chronéer	BC
Valeria Gallo	VG
Andrew Joint	AJ
Alexander Galt	AG
Sebastiaan Berendsen	SB
Ivana Bartoletti	IB

Table 3 - Interviewees' names and their initials used.

This chapter leverages the following structure: first section, 4.1 analyzes the insights the data subjects had on trust in AI. After this, in section 4.2, questions around bias in AI are touched upon. From here, the interviewees have been asked to evaluate the need for regulation in the field of AI in section 4.3. The sections include direct quotations from the interviewees.

4.1 Factors Affecting Trust in AI

The interviewees have identified multiple aspects which may affect the trust levels users portray towards AI. Namely, a number of the interviewees suggest the lack of education, or knowledge on the usage and/or use cases of AI to be a key obstacle in adjusting trust towards an appropriate level, the level of calibrated trust.

> **BC:** "I think that public's, just, lack of understanding of what is actually going on, causes, […] a lack of trust, because if you don't know how something behaves, you don't know how something works. […] You have so little information to go on."

> **SB:** "Yeah, I think it's really important [educating the user] because […] from a technical point of view, but also from a user's point of view. […] But if that use will be different because the end user does not fully understand how it's needed to use, or how he or she should interpret the outcomes of the algorithm […] then it can still have adverse outcomes."

> **VG:** "As often it's the case, it's a lack of understanding, both the benefits and the limitation. And this is both true for the individuals and within organizations. There's a lot of terminology, or jargon, or even technical terms that are just bandied about. And then I think people don't really understand it, or even if they think they understand, they don't."

One of the interviewees, AJ, argued the answer to the question of what affects trustworthiness of AI from a legal standpoint. He stated that due to the ambiguity still surrounding accountability and responsibility in negative AI output scenarios, trust would not increase before the aforementioned are cleared up by legislation. However, AJ's statement was accompanied by optimism, as it was argued

that legislation has been able to deal with technologies which were emerging in past, examples of such being the internet, cloud and mobile devices. What's more, he continued by strongly opposing the idea of lawyers from external bodies, such as lobbyist, being the ones defining the playground for AI, but rather that the guardrails are set in place by legislators, who have the interest of the individuals in mind, not the corporations. VG supported AJ's position on the lack of accountability to be a driver in undermining the trust people portray onto AI algorithms, and explained her stance through an event which took place with British upper secondary students' final exams.

To contextualize the below statement, the British upper secondary final exams, A-levels, were graded by artificial intelligence algorithms in 2020, due to the COVID-19 pandemic. The results of the algorithms implementation into the grading process were devastating causing mass-hysteria, outrage and disbelief amongst the students, and their parents. The grading of the exams did not seem to correlate to any grading precedents set during prior years. (Burgess, 2020)

> **VG:** "I think that exemplified what the problem is with trust – because there was absolutely no accountability for the humans involved. They said it was the algorithm that made the mistakes. […] But as it was an automated decision, the human intervention was actually meaningless. They just accepted what this algorithm had sprouted out, even if it wasn't anything complex."

She continued by arguing that now, after the event, people associate AI algorithms with something that can affect one's life negatively to a tremendous extent, especially when there is no accountability involved. IB's arguments parallel VG's statements, yet IB argued the notion of increasing trust in AI to be two-fold: the consumer side, and the governance side. From a consumer or user perspective, the contestability of the outcome is a high value driver, a dimension affecting trust in a fundamental manner. Contestability, also referred to as transparency and accountability as IB puts it, is the main driver behind trust in algorithms. From a governance perspective however, a best practices framework was suggested by IB.

> **IB:** "… From a governance perspective, to me it's about what the European Union is proposing to do, which is simply to do some sort of […] conformity assessments or standards that need to be followed."

Without prompting towards any references to the used literature of this paper, SB mentioned users' increased trust levels towards AI compared to their human counterparts, in an observed phenomenon called *Automation Bias*, discussed by the likes of Madhavan & Wiegmann (2007).

> **SB:** "And also there's such a thing like automation bias. So when people read something from a machine, they tend to believe it, so also some awareness around that will really help mitigating the risks so that they are aware that the algorithm can also be mistaken."

However, some interviewees argued that only educating the public on AI would not suffice in building trust towards it.

> **AG:** "I think general awareness is important and this kind of digital data literacy of end users. But […] they're only going to be able to scrutinize the output that's relevant to them, in the context that's specific to them. […] One user being a bit more clued up about how it affects them isn't going to tackle this broader picture [of lack of trust]."

> **AJ:** "It's too complex. And you have to start from the lowest common denominator when you're talking about society generally. And explaining [AI], I don't think gets you far enough, when you've got to deal with a consumer who knows absolutely nothing and can't understand anything. Because they're the people that the rules should really be there to protect the most."

> **IB:** "In order to be able to understand, you have to be able to comprehend […]. And the problem is that the user understanding of AI system is important. But what is far more important is that the system is built properly. […] It's part of the solution [educating the user]. Unless there is a user who was very switched on and has studied computer science, no user will ever understand an algorithm."

During the interview held with BC, a follow-up question was asked from him, outside of the guiding interview questions – the effect of time, in building trust towards AI. BC admitted that they

had not thought about time as an affecting factor but continued by stating that it would seem like a viable dimension in adjusting trust in algorithmic AIs onto a proper level.

> **BC:** "So what I do think, I didn't think of before, but when you mentioned it now, time […] it might just be time. […] As soon as something becomes easy to do, we kind of stop thinking it's scary and mystic. […] For instance, just a couple of years ago, having a computer recognize the difference between a picture of a dog and a picture of a cat would have seemed crazy, super advanced, we just cannot do it. […] We kind of stopped thinking of it as some sort of magic machine."

Noteworthy is that IB argued against the above statement. She referred to the effects GDPR has had on our society, where a notable level of numbness has arisen towards the privacy notices implemented by the organizations.

> **IB:** "Take the example of GDPR. People have gotten used to an extractivist made model built on data extractivism and datafication of society. But what has happened is that trust has gone down, because the privacy notices become totally unreadable."

SB's response is in line with BC's statements, where he stated that once AI has had the time to mature, and once time passes on, the general public will have an increased level of trust towards AI. However, this increase in trust will only happen if the amount of scandals surrounding AI are minimized to the highest extent possible.

AG on the other hand argued that even though time may help in building trust towards AI algorithms, the responsibility of generating trust does not fall on the individual users. He argued, that as more companies start leveraging AI in a more public manner, it is in the public's interest for the public to challenge the ethical issues arising from the algorithms. What's more, he argued that it is not the onus of the user's to uphold the trust towards the algorithms, but rather the developing party's duty, as abiding by the possible frameworks and requirements set in place by regulators, trust should increase naturally.

> **AG:** "So in general, I think the rest of society needs to keep pace with that and to challenge that [ethical issues]. But I also don't think we should put the onus on the individual to try to have to understand what's going on. We don't do that with any other area of law, or ethics. When you go to the doctor, you don't have to understand all of the test that they're doing to trust that the diagnosis is correct."

An interesting point was brought up by AG, where it was argued that by having the public educated enough on AI, and its decision-making processes, unwanted counterproductive effects on the trustworthiness of the algorithm may arise. By the users being enlightened enough, the question of "why does this decision need to be made by an AI?" may arise. SB's statements support AG's arguments. Noteworthy is that the General Data Protection Regulation already has a section on this exact topic, where a user may request their claim and/or request to be handled by a human, instead of an algorithm (European Parliament and the Council, 2016). This is something that AG also mentioned in his comments. Furthermore, the protection of the so called "protected groups" of identifying features in individuals plays a role in having trustworthy AI. Protected groups include attributes of an individual which an he or she cannot be expected to change, or it would be an unreasonable attribute to base a decision on. Examples of such are age, ethnicity and gender. Both AG and SB argued that the lack of consideration of said protected groups in the algorithms, discriminating or unfair biases will be present in the outcomes of the algorithm.

A notion of misalignment of what trust means within AI is said to be a key reason why it is hard to identify and quantify the trust factors of AI. SB argued that due to the said misalignment, thus far trust in AI has not meant much. Furthermore, as artificial intelligence algorithms are often built on complex models, where the clarity of data is not necessarily up par with what might be required from said algorithm, users may have a decreased level of trust towards the technology the more aware the public becomes of the process underlying the algorithm.

> **SB:** "So I think that scandals always affect trust. So the more scandals there are, the more […] people will resent technology".

Whilst looking at trust in AI from a deeper technical level, SB stated that to increase trust in AI, one should apply machine learning validation models to review how well the algorithm adheres

to the anti-discriminatory regulations and other national-level rules set in place. One way of approaching the analysis is to use an algorithm to see what kind of relationships between variables the analyzed algorithm creates to reach a decision. The second used method is to use calculate the fairness of an algorithm. This method produces a numeric value, from which the level of fairness can be analyzed. If the fairness value is not on an expected level, the trust in the algorithm may be compromised compared to an algorithm with a higher value, as the algorithm can produce unwanted results in real-world use, thus decreasing the trust in the tool, and eventually possibly leading to a scandal, argued SB.

Where Western cultures share a rather similar conceptualization of trust, Eastern cultures may deviate from the Western cultures' understanding of such. AG stated that AI built outside of the Western cultures, namely EU, the developers may have viewed trust and bias in a manner which do not fully align with Western values. Moreover, the values, ethics and morals are also suggested to deviate between geographical locations per literature.

> **AG:** "I think if you look at […] Western versus Eastern philosophy […], you see a lot more about individualism from the west, and you want to protect individuals rights, whereas the community is something in the east. It's a community above self, and mirroring those things is kind of difficult when […] an AI is being set to make decisions on an individual person, and you're in the EU, you need to protect individual rights, whereas it might be bit more group based."

> **VG:** "If I go to Japan, there will be some norms and traditions, that as a visitor, I will need to take into consideration in my behavior. […] Even just in the UK and Italy there will be differences. […] So as potentially annoying and costly as it is, we will need to account for those because different societies will have different standards, different views of what is ethical. […] If you ask whether or not an automatic car should hit an old woman or a child, they [Japanese] will say they should hit the child because elderly people are held in a very high regard."

All of interviewees agreed that geographic trust determinants, such as morals, ethics and values need to be taken into consideration whilst developing, training, or reviewing the outputs of the AI.

4.2 Bias in AI

When asked whether there is bias present in AI, the interviewees mainly responded in cohesion, agreeing there to be bias. However, where there were possible discrepancies in their responses, they can mainly be explained by the interviewees differing approaches to thinking of bias. Which is why, the first step was to ask the participants to define bias in AI.

> **BC:** "[…] You can of course use human bias as example and see […] does the machine have bias in the sense of how a human have bias. And the answer is really no, right? Usually when you start building an algorithm, you have a blank slate. You just feed the data and it learns, however good it can. But the data you provided is usually if not always biased."

> **AJ:** "I don't see bias in as any different to any other type of bias available in other types of data analytics, analog systems etc. Garbage in, garbage out. […] There has to be bias because the AI doesn't operate in a vacuum. It operates in an analytical set, it operates in a clever analytics of in-historic data."

Based on the gathered data, it can be said that AI is not inherently biased as it is essentially just numbers from which the algorithm has learnt its behavior models. However, as the data input is more likely than not biased, technically the AI is thought to be biased.

> **SB:** "There is, of course [bias], but that's because of the differences in society. […] So it's more on the whole society. Because I think it's also unconscious bias within the developer or in the data, that is just really subtle. So that's quite difficult to really say it originated there, but it originated in society and it leaks through into an AI or into a model."

> **VG:** "I think that AI systems can be built so that they lead to bias outputs. I don't believe there is bias inherently in AI. I also believe that it's extremely naïve to think that we can build anything that is not biased in some way. Now, it can be biased, in a positive way, in a bad way. But there will always be some bias."

When asked about the possibilities of minimizing or even completely eliminating unwanted and unfair bias from AI, nearly all of the interviewees felt uncomfortable in giving out definitive responses – some even suggest that it is impossible to not have bias in AI. This is due to the fact that we as humans might not even consider certain factors in the decision making process to be a biased factor.

> **BC:** "I guess one way of looking at it, is that if it's better than the human's judgement, it's already quite good. […] We can do our best to really minimize it and try to find all the different ways it can be bias free so to say. […] But I don't think the perfection should be the enemy of good. […] Even if we minimize all of these things that we reasonably can minimize and remove as much bias as we can, that might just be good enough, in a sense, because what is the alternative? That the human is doing these decisions? […] Perfection, I think is a very nice goal. But I don't think that if it's not perfect, we cannot use it at all."

> **IB:** "In order to achieve no bias in AI, you will need a political social statement, made by the company saying "I don't care about what the data is like, I'm going to make a decision, which is based on my values"."

> **AJ:** "I wonder how achievable that is from a technical basis to be able to remove it [bias]. […] Even if you could, is it necessary or worth the effort? Or are we just looking for an improvement from the bias and inherent problems we have and the trustworthiness that we have at the moment? Because as long as it's a step forward, and better – that's surely progress."

SB argued that there will always be bias present in AI – partially because the ultimate goal of the algorithm is to be efficient and optimal, but also due to the surrounding society. Referring to prior

mentioned protected groups, it is mentioned that using vast data libraries in the algorithm may lead to a situation where some of the data leveraged may act as a proxy for biases to creep into the algorithm, even when the more simple and obvious protected groups are taken into consideration.

The context in which the underlying data of the algorithm is collected does seem to play a role in minimizing possible biases. Moreover, looping back to SB's statement on unrelated data sometimes acting as a proxy for bias, AG continued by stating that often due to poor data quality, complex statistical methods and the use of proxy data, the algorithm's outcomes may not model what it was originally intended to model, which may lead to a position of having bias, or simply a poorly designed algorithm. VG supported AG's sentiment and continued on context awareness.

> **VG:** "AI developed in Europe could very well be not suitable for an Indian market, or the Indian population. […] And how do you ensure that the way that you design it [the algorithm], the way you train it, the way you test it ensures that the target population is treated as intended? […] AI needs to be representative."

VG continued her argument of context awareness, by stating that having representation may facilitate a false sense of security. She used Italy as an example; even if one would take the dark-skinned population of Italy into consideration in building an AI for the Italian market, it would be representative of the Italian dark-skinned population, which would in turn create unwanted discrimination as dark-skinned people are a minority and are already heavily discriminated in Italy,. And all of this erodes trust in AI, according to her.

As it is argued by a number of interviewees, bias will often times, if not always be present in AI. However, when asked whether the lack of bias present in AI would affect the trustworthiness of the algorithm, the same tone continues, by several interviewees stating that the trustworthiness would not be affected.

> **VG:** "And there are some biases that we actually want to maintain. […] There are a lot of situations in life where seen objectively, and taking away our human emotions, they could be dealt with a lot more efficiency. But as a society, we don't think that's moral or ethical [to remove emotion]. And therefore, we express our bias that way. So I don't think that

removing completely the human element from an AI system would actually improve our society and make it more trustworthy."

AJ hesitantly stated that if the developers are ever able to completely eliminate bias from AI, then the trustworthiness of the algorithm would be increased. However, an alignment of expectations should still take place, where the expectations, or users' portrayed trust would need to be adjusted to ever correlate with the trustworthiness of the algorithm. Said alignment should happen prior to the elimination of bias, according to him.

4.3 Legislation

The need for regulation and legislation in the field of AI is recognized by all of the interviewees. Some argued the need for regulation to stem from the likes of AI implementation in China with their social scoring system, whereas others suggest the functionality and integrity of the AI's outputs to be compromised without legislation. The relevance of the context in which the underlying data has been gathered from is brought up my number of the interviewees. The EU is argued to be in an important position of guiding the industry.

> **AG:** "I think […] EU regulation is there to try and guidance as well. So if there's an Indian made AI system that was made within the context of India, and is had test data from India, is it then appropriate to use within the EU when the cultural and political circumstances and ethical circumstances are different?"

> **SB:** "[…] You need the regulation, but it also works as a guideline for people who develop models. […] When developing such guideline[s], people have different opinions, but the regulation is just… it is like it is. So that also helps in discussing how to do certain things."

The carrot-and-stick approach has seemed to bear fruit within the fields, where AI is used. Prior to regulations even being in the near horizon of being proposed, organizations, namely banks, were not keen in validating their AI models to see whether they have a minimized amount of bias. However, once the EU became active in regulating the behavior around the deployment of AI,

organizations woke up to the possibility of getting hit by large fines for not complying to the current legislation and upcoming legislation.

> **BC:** "It's quite funny, because when I worked a lot on Glassbox […] a lot of the companies and banks in particular […] just weren't, like ready for this solution, right? Because they were "we have some model, we don't even have that many AI models. We don't have any information from the regulator […], why should we really care". But once you see the European Union, […] starting to build the cloud above them, starting to like the regulations is coming, then you can see, like an attitude shift towards "Okay, this is quite important"."

Moreover, a minor indication of the impact the EU has had on the market, is the continuous hiring of new employees into a Deloitte team working on AI, based from the Netherlands. In addition, the United States has expressed their interest in following similar steps to what the EU is currently doing with their approach to regulating the processes around AI, according to IB. Further, the need for regulation is said to be somewhat long awaited, as there is no real incentive to action biased AI algorithms, until a scandal or a high-profile incident takes place, argued SB.

> **AG:** "On multiple levels, I think there's some of the things that have gone unnoticed about liability around AI that there needs to be really […] certainty around. So, who's liable for when it goes wrong? Is it the developer? Or is it the company in general?"

VG and AJ agreed with the above statement, by saying that the until the liability aspect is solved, trust will stagnate.

A question left unanswered by the academia, industry and the regulatory bodies according to AG is whether legislation should mandate strict rules in place on AI, and if yes, how many rules. An alternative suggested by AG is to replace some of the strict rules by good practice check and balances, third party audits and a development of a best practices guidelines. However, he continues by stating the development of best practices will only happen over time once the industry has recognized what "good, ethical, appropriate and trustworthy" AIs look like – something in which a regulatory body may help. AJ continued on the sentiment of worry AG brought up on the reach on and depth of the possible new regulations.

> **AJ:** "I think yes [we need regulation]. Because there is a trust deficiency. And it is going to inherently encumber the development of the technology until we deal with that. And we have seen that with other technologies over the years – cloud's a really good example. Until people have trust in terms of data storage etc. and the laws around it, it stops it developing at a pace it should. […] I'm always worried about the level and the depth of regulation of technologies, because by the time they come in they're out of date.

VG stated that she has not seen anything noteworthy resulting from self-regulation, and thus believes that artificial intelligence needs regulation. Furthermore, she supports AJ's comments on the slowness of legislation.

> **VG:** "I'm quite sure that self-regulation is not going to cut the mustard on this one. It's going to be really difficult. This is going to be super pervasive across all the industries and aspects of life, that you need to have some boundaries. […] One of the key challenges of policymakers and regulators is that they do something, it takes a long time to do it. And oftentimes, it's obsolete by the time even before it's agreed to come to force.

The concept of relative versus absolute performance is brought up in the context of Anti-Money Laundering. It is argued by VG that it has been demonstrated to the legislators that having an AI work on fraud detection and anti-money laundering investigations is more beneficial for most, if not all parties with good intent, as the accuracy of an algorithm in this context is often times much greater compared to a human's when analyzed through a relative performance lens, but possibly even when looked at with an absolute performance perspective.

> **VG:** "An AI for anti-money laundering is not perfect. And this can definitely lead to a lot of harm, because certain individuals may be denied access to payments or financial services on the basis of the fact that they're considered an anti-money laundering risk. And that could be unfair, if the decision is not correct. However, an AI for AML is as a whole considerably better than humans. There is relative performance rather than absolute performance.

5

Discussion

In this chapter the findings are presented in the light of the chosen theoretical frameworks. The comparison of statements given by industry professionals towards the frameworks presented by the likes of Silberg & Manyika (2019), Siau & Wang (2020) and Lee & See (2004) in their respective topics is done via either confirming or disproving the theories.

As per the findings received and gathered from the data subjects, one of the main issues of causing misalignment of trust in AI algorithms seems to be the lack of full understanding of the AI's capabilities. It is stated by the interviewees and supported by the literature (Sampson, 1986), that users tend to blame the AI algorithms once the output is not favorable, even when the reason behind an unfavorable output would be the cause of the user themselves. Moreover, increasing accountability is deemed to be factor of great importance in terms of increasing trust towards AI algorithms.

When analyzing trust in AI through the lens of Lee & See's (2004) framework of appropriately calibrated trust level, it is apparent that the industry professionals argue there to be an overall poor calibration of trust within the general public, and whereas from an organizational use standpoint trust is not evaluated with the same critical lens as is been done by the public.

Poor calibration seems to be caused by poor resolution, where a large amount of expectation is placed on the AI's capability, more specifically on a level on which the AI is incapable of operating at, or is not design to function. This is what Lee & See (2004) refer to as Overtrust, which will result in misuse. Misuse, in the thesis' context means the wrongful or inappropriate usage of artificial intelligence. Our results portray a relationship between Overtrust and what Lee & See (2004) have called Distrust. The data indicates, that the initial phase of lacking trust has in fact began from a stage where there has been Overtrust in the AI system in question. More often than not, once there is misuse due to Overtrust, the trust diminishes as a result, leading to Distrust – a state where the trust of the

user falls short of the system capabilities. An underlying factor in the equation is the temporal specificity. As the data suggests, the general public has both a high and low temporal specificity level, depending on the situation. As an example, a user may lose their trust in the algorithm based on a single news article or event – indicating high temporal specificity. However, the users' trust levels do not seem to rebound from a low state as easily as it was lost, when there is positive news published on the algorithms – indicating low temporal specificity. Moreover, the data suggests that trust in algorithmic AI also includes elements shared by interpersonal trust, namely transparency. As stated by the literature, and supported by the data gathered, organizational and cultural factors do affect all the aforementioned, hence the overall level of trust and how one may increase or adjust it to a proper level.

As the industry professionals suggest towards the inability of having a single driver of trust behind when it comes to analyzing AI and the trust the users cast on it, the conclusion is that nothing is quite black and white. The lack of exact answers and results suggest that increasing trust in AI, or adjusting it, is just as a complicated notion as increasing trust between human beings, as one must take the environmental, cultural and organizational contexts into consideration. A noteworthy pointer is the fact trust is not rebuilt in a human-algorithm interaction as easily as in a human-human interaction, as per literature. However, the data leads towards us believing that there is a step between transitioning from Overtrust to Distrust – and that affecting factor might be both the lack of ethics in AI and the unwanted biases present in AI, as described in Figure 5.

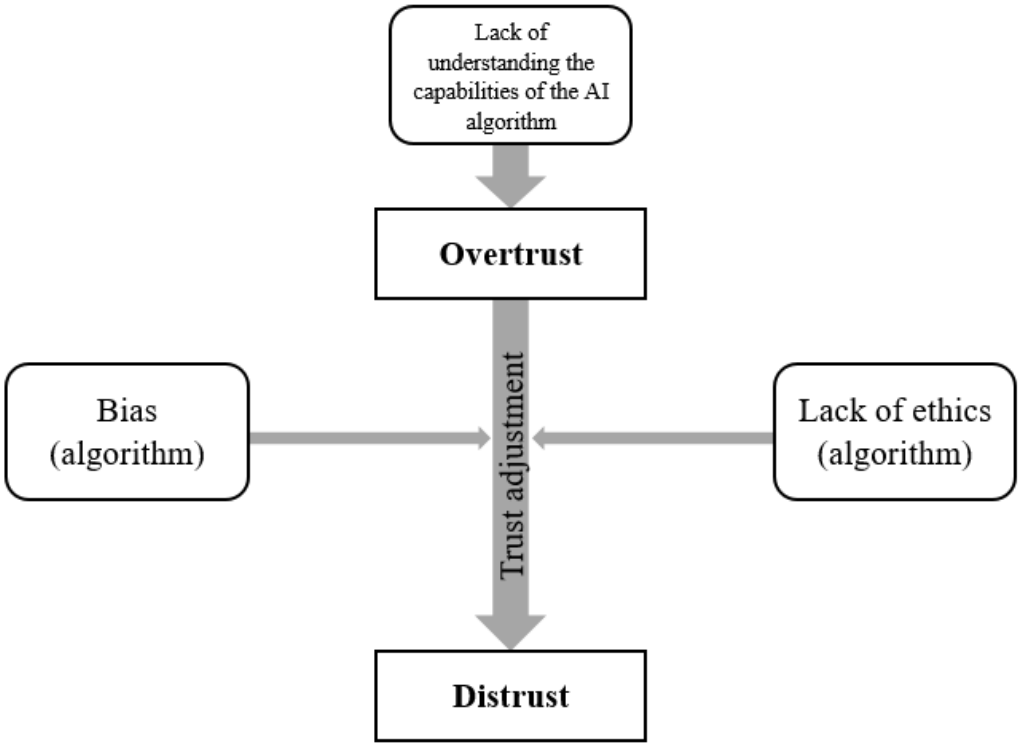

Figure 5 - Drivers affecting trust during the transition from Overtrust to Distrust.

The above figure describes the drivers which seem to affect the level of trust a user has towards an AI algorithm. The initial step in having Overtrust is initiated by the lack of knowledge of the AI's algorithm, however the data is not conclusive on whether or not educating the user on the capabilities of AI would in turn affect the level of trust one portrays in the events leading towards Distrust. Thus, the lack of ethics, including lack of accountability, morals and values is deemed to be a larger effector in the grand scheme of trust. Moreover, the bias that may creep into the algorithm due to proxy data, incomplete datasets, poor modelling, lack of review and sorts, will eventually affect the trust adjustment to a state of Distrust. The model represented in Figure 5 demonstrates the process of the users' trust shifting by using the term *trust adjustment*. Further, as the findings suggest, a scandal or an unintentional event with negative outcomes often times escalates and expedites the trust adjustment towards Distrust. The data also suggests that having a calibrated amount of trust decreases mis- and disuse of the AI algorithm, statements which are supported by Wicks, Berman & Jones (1999) and Lee & See (2004).

A possible outcome of Distrust is that of a never-ending loop, where as a result of Distrust, the users' avert the use of similar algorithms, furthering the lack of understanding, thus creating

Overtrust in the long-run, as suggested by the data. Lee & See (2004) argue the same, however, without explaining how to address the issues in a concrete manner. Figure 6 demonstrates the loop of Overtrust and Distrust, considering the erratic nature of the general publics' temporal specificity, in one event it being high and in other being low.

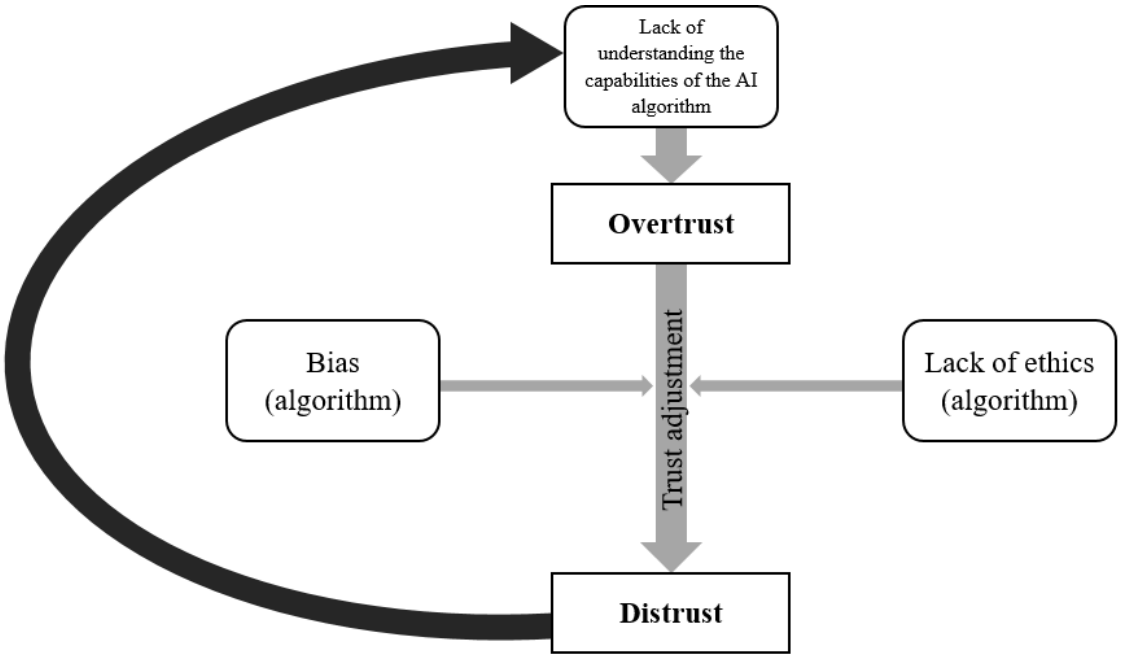

Figure 6 - The continuation of Distrust in the long term.

As stated earlier, the presence of unwanted bias and the lack of ethics do seem to affect the trust adjustment. Arguably, they are not the sole perpetrators behind a diminished level of trust. However, the effectors not identified by this study will quite possibly not exist yet considering the scope of the study. Nevertheless, the industry professionals all state legislation to be a possible solution for the above sequence in Figure 6, with certain limitations, all presented in the findings. From there, it is imaginable what the process towards having a calibrated trust in AI would look like, as presented in Figure 7.

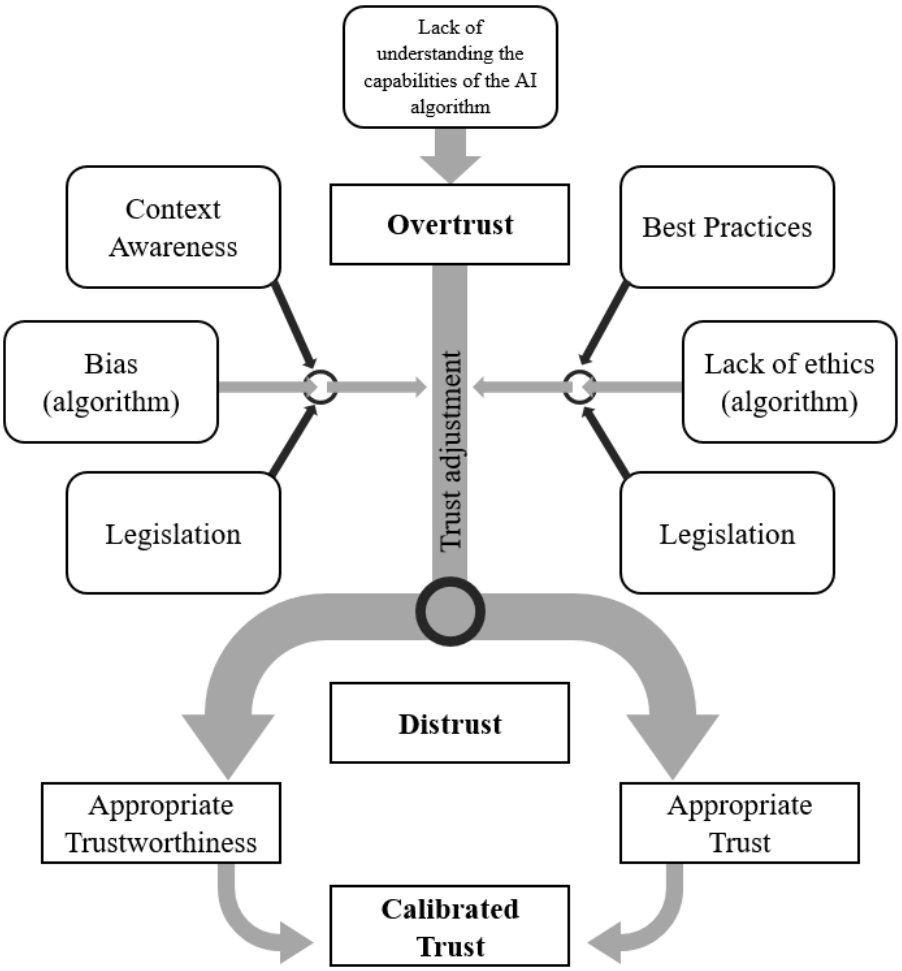

Figure 7 - The process towards calibrated trust, from a state of Overtrust.

In line with Silberg & Manyika (2019), the findings suggest towards context awareness, reviews, open forums for discussions around potential human biases to be a some, but not the sole, effectors within a solution towards having an appropriate level of trust – presented as *Context Awareness* in Figure 7. Moreover, a great amount of hope and emphasis is laid on the guardrails set up by possible future legislation around the processes of AI behavior and creation. What comes to the lack of ethics, and its effects on the event of trust adjustment, the results portray the lack of accountability, ethical standards, and possible unintentional disregard of human rights laws to be an effector in the equation, all shown as *Lack of ethics (algorithm)* in Figure 7, all of which support Siau & Wang (2020) in their argument of the mentioned being a part of a larger picture of incorporating ethics in AI. Thus, it is argued and suggested based on the results, that by having guardrails in place, similarly as with the case of bias, the lack of ethics would be dealt with in an appropriate manner and

legal level, possibly mandating all parties to comply. Moreover, by having a best practices framework, derived from possible legislation, as self-regulation is not supported by the industry professionals, the chain of effect the lack of ethics has on the trust adjustment would be disrupted.

As the trust adjustment takes place with the effectors considered, instead of leading the users towards a state of Distrust, the process is diverted towards appropriate levels of trust and trustworthiness, the prior with the users, latter with the algorithm. All of this in turn, will lead into having what Lee & See (2004) state to be calibrated trust. As is the case with the trust adjustment from Overtrust to Distrust, the events taking place leading to a calibrated trust do not happen in a vacuum. What this means is that, once the "optimal" level of trust is reached, trust adjustment will keep taking place in the background – something which the developers, the reviewers, and possibly the legislators need to consider.

6

Conclusion

Staying true to the nature of an exploratory research, the paper did not result in any absolute answers to the research questions, but rather a hypothesis. The purpose of this paper was to reach a conclusion and find concrete, but still hypothetical actions to be taken on how to adjust the trust the public has towards AI algorithms. Based on the findings, and derived models this chapter starts out by answering the research question(s).

> RQ: ”How can a user’s trust towards an artificial intelligence algorithm be adjusted in the context of European consultancy services?”
>
> Sub-RQ: “What effect does bias in artificial intelligence algorithms have in facilitating trust in towards the algorithms?”

Deloitte being the one of the largest consultancy service providers globally, the leverage the firm has in assisting their clients in building more accurate, optimal and efficient artificial intelligence algorithms is greater compared to one being built without the assistance of such a firm. However, not all aspects, or effectors, identified by this paper are under the direct influence of Deloitte. Thus, the research question will partially be answered from Deloitte’s perspective, after which a national-level perspective is taken.

Based on relevant prior studies conducted around the research question, many of the identified effectors by this paper confirm the theories presented in the literature review. However, certain dimensions in minimizing bias are given greater importance, than others shown in the literature, based on the lack of mentioning of such by the interviewees.

The paper has identified the users’ lack of understanding towards the capabilities of the used AI algorithms to be a key effector in facilitating Overtrust. Yet, the findings are not conclusive

whether increased knowledge on the algorithms' underlying functions which affect the final output of the algorithm, would diverge the drive of users' trust levels from a state of Distrust where disuse would ensue. In liaison with the existing literature, the findings suggested that the presence of unwanted bias and lack of ethics are one of the key effectors in the users' trust being adjusted towards Distrust. Moreover, once a state of Distrust and disuse is reached, rebounding towards Overtrust is difficult, but will take place eventually, by the users not fully understanding the capabilities of the algorithm, creating misuse and Overtrust.

Effectors such as context awareness and a best practices framework are aspects in which Deloitte may facilitate the increasing of trust towards artificial intelligence algorithms. As it stands, the company provides services to its clients with fairness reviews and compliance guidance in relation to the research question. The fairness reviews are conducted with in-house built tools, one of which being the tool called Glassbox. An identified effector within the lack of ethics, based on the findings, is the lack of accountability, also referred to as contestability. The results suggest that by having the mentioned effector accounted for in the AI algorithms would help in reaching calibrated trust, as trust adjustment keeps taking place throughout the use-cycle of the AI algorithm. However, the tools available to the company are limited in terms of increasing trust, as the findings suggest the effect of legislation to be in a large position as well. Noteworthy is that by legislation, the paper refers to actions and guardrails to be set in place to protect the users from unfair bias and the adverse effects of having lack of ethics involved in the AI algorithm. Hence, once possible EU-wide legislation comes to force, the implications of such will reach the aforementioned effectors.

7

Limitations and Future Research

7.1 Limitations

Unfortunately, as the findings and data differed from one another, based on the interviewees background and approach, the results provided by the paper may differ if the same study is conducted with a different sample group. The aim of having interviewees with differing approaches to the research question was to facilitate contradicting thoughts and comments, and to provide more reliability and validity. This indeed was the case, watering down the impact of concrete action suggestions. Moreover, the original intention was to conduct data collection with 12 participants, all of whom were industry professionals with an extensive experience in the field. For reasons such as redaction, existing client commitments, time constraints, and technical issues, only 6 out of the initially planned 12 interviews were successful.

During the final stages of the writing process of this paper, the European Commission came out with a proposal which include most, if not all, of the presented dimensions in the legislation sections. While it would have been a factor increasing the relevance of this study, due to time constraints and prior commitments to the structure of the paper, not excluding already conducted interviews, it was decided to leave out the proposal from the scope of this paper.

As the paper leveraged the industry professionals from within the European context, the relevance of this paper might be limited to purely the Western cultures and the algorithms built for said markets. Still, as the results portrayed by the study are rather general, and intrinsic to human cognition, the paper could be extended to cover the Eastern cultures with certain limitations, namely considering the effect of legislation to the equation of increasing trust towards the algorithm.

7.2 Future Research and Implications to Academia

The contribution this paper brings to the academia may be essential in further understanding what the drivers behind trust in AI are. An interesting approach for future research would be to conduct an ethnographic study where the identified effectors would be observed in action. The direction of the study could be to analyze the AS-IS and TO-BE states of the users' trust over a lengthy time period to evaluate whether or not the identified effectors do indeed play a role in adjusting trust to an appropriate level, or whether they are simply too high-level notions given by the data subjects of this study.

This paper aimed to confirm Lee & See (2004), Silberg & Manyika (2019) and Siau & Wang (2020) proposed frameworks or models either as a whole or aspects of each, with the results being promising in terms of confirming them. As stated, an ethnographical study would most likely present findings which this study was unable to result. Due to AI algorithms increasingly becoming a part of our everyday lives, analyzing them in different geographical contexts could portray interesting and differing results, than what this paper presented, as the interviewees were all from the Europe.

Bibliography

Barnett, V. (2009). *Sample Survey Principles and Methods.* Wiley.

Brown, R. B. (2006). *Doing Your Dissertation in Business and Management: The Reality of Research and Writing.* Sage Publications.

Bryman, A. (2012). *Social Research Methods.* New York: Oxford University Press Inc.

Buchan, N. R., Johnson, E. J., & Croson, R. T. (2005). Let's get personal: An international examination of the influence of communication, culture and social distance on other regarding preferences. *Journal of Economic Behavior & Organization*, 373-398.

Buchan, N., & Croson, R. (2004). The boundaries of trust: own and others' actions in the US and China. *Journal of Economic Behavior & Organization*, 485-504.

Bucher, T. (2012). Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. *New Media & Society, 14*(7), 1164-1180.

Burgess, M. (2020, August 20). *The lessons we all must learn from the A-levels algorithm debacle: Wired*. Retrieved May 14, 2021, from Wired Web site: https://www.wired.co.uk/article/gcse-results-alevels-algorithm-explained

Business Research Methodolgy (BRM). (n.d.). *Sampling: BRM*. Retrieved April 5, 2021, from BRM | Business Research Methodology Web site: https://research-methodology.net/sampling-in-primary-data-collection/

Carrillo, M. R. (2020). Artificial Intelligence: From ethics to law. *Telecommunications Policy, 44*(6). doi:10.1016/j.telpol.2020.101937.

Copeland, B. J. (2021, August 11). *Artificial Intelligence. Encyclopedia Britannica*. Retrieved March 20, 2021, from Britannica Encyclopedia: https://www.britannica.com/technology/artificial-intelligence

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. *Proceedings of KDD '17.* Halifax, NS.

Dietz, G., Gillespie, N., & Chao, G. T. (2010). Unravelling the complexities of trust and culture. In M. N. Saunders, D. Skinner, G. Dietz, N. Gillespie, & R. J. Lewicki, *Organizational Trust - A Cultural Perspective* (pp. 1-127). New York: Cambridge University Press.

Dijkstra, J. J., Liebrand, W. B., & Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour & Information Technology, 17*(3), 155-163.

DiscoverPhDs. (2020, September 10). *What is Research? - Purpose of Research: DiscoverPhDs*. Retrieved March 28, 2021, from DiscoverPhDs Community Blog Website: https://www.discoverphds.com/blog/what-is-research-purpose-of-research

Edwards, J., & Nevola, C. C. (2020, August). *The EU's Approach to AI - Recent Regulatory Developments: Bird & Bird*. Retrieved March 26, 2021, from Bird & Bird Website: https://www.twobirds.com/en/news/articles/2020/global/the-eus-approach-to-ai-recent-regulatory-developments

Elliot, M., Fairweather, I., Olsen, W., & Pampaka, M. (2016). *A Dictionary of Social Research Methods.* Oxford University Press.

European Commission. (2000). *Directive 2000/43/EC against discrimination on grounds of race and ethnic origin.* Brussels: European Commission.

European Commission. (2000). *Directive 2000/78/EC against discrimination at work on grounds of religion or belief, disability, age or sexual orientation.* Brussels: European Commission.

European Commission. (2004). *Directive 2004/113/EC equal treatment for men and women in the access to and supply of goods and services.* Brussels: European Commission.

European Commission. (2006). *Directive 2006/54/EC equal treatment for men and women in matters of employment and occupation.* Brussels: European Commission.

European Commission. (2021a, March 10). *High-level expert group on artificial intelligence: Policy: Shaping Europe's digital future: European Commission*. Retrieved April 14, 2021, from European Commission Web site: https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

European Commission. (2021b, March 9). *A European approach to Artificial Intelligence: Policy: European Commission*. Retrieved March 27, 2021, from European Commission Website: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

European Commission. (2021c). *Proposal for a Regulation of the European Parliament and of the Council - Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending certain Union Legislative Acts {SEC(2021) 167 final} - {SWD(2021) 84 final}...* Brussles: European Commission.

European Parliament and the Council. (2016). Regulation (EU) 2016/679 of the European Parliament [...] on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. *Official Journal of the European Union*.

Executive Office of the President. (2016). *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights.* Washington D.C.

Ferrell, O. C., & Ferrell, L. (2021). Applying the Hunt Vitell ethics model to artificial intelligence ethics. *Journal of Global Scholars of Marketing Science, 31*(2), 178-188. doi:10.1080/21639159.2020.1785918

Ferrin, D. L., & Gillespie, N. (2010). Trust differences across national-societal cultures: much to do, or much ado about nothing? In M. N. Saunders, D. Skinner, G. Dietz, N. Gillespie, & R. J. Lewicki, *Organizational Trust - A Cultural Perspective* (pp. 42 - 86). New York: Cambridge University Press.

Gaynor, M. (2020, January 29). Automation and AI sound similar, but may have vastly different impacts on the future of work. Retrieved March 18, 2021, from https://www.brookings.edu/blog/the-avenue/2020/01/29/automation-and-artificial-intelligence-sound-similar-but-may-have-vastly-different-impacts-on-the-future-of-work/

Hangartner, D., Kopp, D., & Siegenthaler, M. (2021). Monitoring hiring discrimination through online recruitment platforms. *The International Journal of Science, 589*(7843), 572-576.

Hao, K. (2019, January 21). AI is sending people to jail - and getting it wrong. *MIT Technology Review*. Retrieved March 20, 2021, from https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/

Henderson, M. T., & Churi, S. (2019). *The Trust Revolution - How the Digitalization of Trust Will Revolutionize Business and Government.* Cambridge: Cambridge University Press.

High-Level Expert Group on Artificial Intelligence (AI HLEG). (2019a). *Policy and Investment Recommendation for Trustworthy AI.* Brussels: European Commission.

High-Level Expert Group on Artificial Intelligence (AI HLEG). (2019b). *Luotettavaa tekoälyä koskevat eettiset ohjeet.* Brussels: European Commission.

High-Level Expert Group on Artificial Intelligence (AI HLEG). (2020). *The Assessment List For Trustworthy Artificial Intelligence (ALTAI).* Brussels: European Commission.

IBM. (2020, June 3). *Artificial Intelligence: IBM*. Retrieved May 14, 2021, from IBM Web site: https://www.ibm.com/cloud/learn/what-is-artificial-intelligence

Jaffe, D. (2018). The Essential Importance Of Trust: How to Build It Or Restore It. *Forbes*. Retrieved March 16, 2021, from https://www.forbes.com/sites/dennisjaffe/2018/12/05/the-essential-importance-of-trust-how-to-build-it-or-restore-it/?sh=6b35dae964fe

Kantarci, A. (2021, February 13). *Bias in AI - What it is, Types & Examples, How & Tools to fix it: AI Multiple*. Retrieved March 21, 2021, from AI Multiple Website: https://research.aimultiple.com/ai-bias/#what-is-ai-bias

Kasperkevic, J. (2015, July 1). Google says sorry for racist auto-tag in photo app. *The Guardian*. Retrieved March 21, 2021, from https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app

Krauss, S. (2005, December). Research Paradigms and Meaning Making: A Primer. *The Qualitative Report, 10*(4), 758-770.

Kulms, P. (2018). *Trust in Interdependent and Task-Oriented Human-Computer Cooperation.* Bielefeld: Bielefeld University.

Lee, J. D., & See, K. A. (2004). *Trust in Automation: Designing for Appropriate Reliance.* Iowa City: Human Factors and Ergonomics Society.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*(10), 1243-1270. doi:10.1080/00140139208967392

Lee, N. T. (2018). Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society, 16*(3), 252-260. doi:10.1108/JICES-06-2018-0056

Lowry, S., & Macpherson, G. (1988). A blot on the profession. *British Medical Journal, 296*(6623), 296.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human and human-automation trust: an integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277-301. doi: 10.1080/14639220500337708

Mantere, S., & Ketokivi, M. (2013). Reasoning in Organization Science. *Academy of Management Review, 38*(1), 70-89.

Manyika, J., Silberg, J., & Presten, B. (2019, October 29). What Do We Do About the Biases in AI? *Harvard Business Review*. Retrieved April 13, 2021, from https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

McIlwain, C. (2020, October 20). AI has exacerbated racial bias in housing. Could it help eliminate it instead? *MIT Technology Review*. Retrieved March 20, 2021, from https://www.technologyreview.com/2020/10/20/1009452/ai-has-exacerbated-racial-bias-in-housing-could-it-help-eliminate-it-instead/

Merriam-Webster. (n.d). *Trustworthy: Thesaurus: Merriam-Webster*. Retrieved May 14, 2021, from Merriam-Webster Web Site: https://www.merriam-webster.com/thesaurus/trustworthy

Middleton, F. (2020, June 26). *Reliability vs Validity - what's the difference?: Scribbr*. Retrieved April 7, 2021, from Scribbr Web site: https://www.scribbr.com/methodology/reliability-vs-validity/

Mone, G. (2016). Bias in Technology. *Communications of the ACM, 60*(1), 19-20. doi:10.1145/3014388

Morse, J. (2017, April 25). *App cretor apologizes for 'racist' filter that lightens users' skin tone: Tech: Mashable*. Retrieved March 21, 2021, from Mashable: https://mashable.com/2017/04/24/faceapp-racism-selfie/?europe=true#zeUItoQB5iqI

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 527-539.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are Social Actors. *CHI '94: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 72-78). Stanford: Stanford University.

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic Analysis: Striving to Meet the Trustworthiness Criteria. *International Journal of Qualitative Methods, 16,* 1-13.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: cross-discipline view of trust. *Academy of Management Review*.

Sampson, J. P. (1986). Computer Technology and Counseling Psychology: Regression Toward the Machine? *The Counseling Psychologist, 14*(4), 567-583. doi:10.1177/0011000086144006

Saunders, M. N. (2015). *Research Methods for Business Students.* Pearson Education UK.

Saunders, M. N., Lewis, P., Thornhill, A., & Bristow, A. (2019). Chapter 4: Understanding research philosophy and approaches to theory development. In M. N. Saunders, P. Lewis, A. Thornhill, & A. Bristow, *Research Methods for Business Students* (pp. 128-171).

Siau, K., & Wang, W. (2020, April-June). Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. *Journal of Database Management, 31*(2).

Silberg, J., & Manyika, J. (2019, June 6). Tackling bias in artificial intelligence (and in humans). Retrieved March 21, 2021, from https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans#

Singer, P. (2021, February 2). *Ethics. Encyclopedia Britannica*. Retrieved March 20, 2021, from Britannica Encyclopedia: https://www.britannica.com/topic/ethics-philosophy

Sousa, S., Lamas, D., & Dias, P. (2014). A Model for Human-Computer Trust - Contributions Towards Leveraging User Engagement. *Learning and Collaboration Technologies - Designing and Developing Novel Learning Experiences, I*, pp. 128-137. Heraklion.

Sue, V. M., & Ritter, L. A. (2012). *Conducting Online Surveys.* SAGE Publications Inc.

Thomsen, K. (2019). Ethics for Artificial Intelligence, Ethics for All. *Paladyn, Journal of Behavioral Robotics, 10*(1), 359-363. doi:10.1515/pjbr-2019-0029

Wicks, A. C., Berman, S. L., & Jones, T. M. (1999). The Structure of Optimal Trust: Moral and Strategic Implications. *Academy of Management Review, 24*(1), 99-116.

Worthy, B. (2019, October 23). *Automated or Manual Transcription Service - Which Is Better?: GMR Transcription*. Retrieved April 7, 2021, from GMR Transcription Web site: https://www.gmrtranscription.com/blog/automated-or-manual-transcription-service-which-is-better

Waaijer, R., & Chronéer, B. (n.d.). *Measuring fairness with Glassbox - A toolkit to create transparent and responsible AI*. Deloitte Netherlands. Retrieved May 16, 2021, from https://www2.deloitte.com/nl/nl/pages/risk/articles/measuring-fairness-with-glassbox.html

Yoshikawa, H., Weisner, T. S., Kalil, A., & Way, N. (2008). Mixing Qualitative and Quantitative Research in Developmental Science: Uses and Methodological Choices. *Developmental Psychology, 44*(2), 344-354.

Zenger, J., & Folkman, J. (2019). The 3 Elements of Trust. *Harvard Business Review*. Retrieved March 16, 2021, from https://hbr.org/2019/02/the-3-elements-of-trust

Appendices

Guiding interview questions

1. Introduction of themselves
2. How long have you been working at Deloitte and/or with AI?
3. When I mention “trust in AI” what comes to mind?
4. What do you think are the main aspects affecting trust in AI?
5. I’ll read out the following statement, and I’d like you to rate on a scale of 1 to 5 (5 being the highest level of agreement) on how much you agree with the statement:
   “By educating the user on the capabilities of the AI, trust in the tool is increased.”
6. How would you define “bias in AI?”
7. Do you feel there is bias present in AI? In what ways?
8. Is it achievable to have no bias in AI?
9. How would you suggest bias should be eliminated from AI?
10. Would eliminating bias in AI increase the trustworthiness of the AI?
11. Do you feel that there is a need for regulation and/or legal intervention for AI?
12. How do you feel Deloitte’s offering achieves this goal?
13. How would Deloitte be able to “fix” / increase trust in AI?

Interview Transcripts

Link to Transcriptions