La borsa di dottorato è stata cofinanziata con risorse del
Programma Operativo Nazionale Ricerca e Innovazione 2014-2020, risorse FSE REACT-EU
Azione IV.4 "Dottorati e contratti di ricerca su tematiche dell'innovazione"
e Azione IV.5 "Dottorati su tematiche Green"

UNIVERSITA' DEGLI STUDI "G. d'Annunzio"
Chieti-Pescara

SCUOLA SUPERIORE "G. d'Annunzio"

DOTTORATO DI RICERCA IN

Accounting, Management and Business Economics

CICLO

XXXVII

EXPLORING BIAS IN ARTIFICIAL INTELLIGENCE THROUGH A SOCIO-TECHNICAL PERSPECTIVE

Dipartimento di Economia Aziendale

Settore Scientifico Disciplinare SECS-P/10

Dottorando
Dott. Marco Smacchia

Coordinatore Prof.ssa
Prof.ssa Federica Ceci

Tutor
Prof. Stefano Za

Anni Accademici 2022/2025

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION....1

Chapter I: The Role of Artificial Intelligence in Business and Management: A Literature Review of Evolving Perspectives....4

Chapter II: Challenges of AI Implementation: Bias, Fairness, and Ethical Considerations in Theory and Practice....6

Chapter III: Empirical Approaches to AI Bias: Identification, Measurement, and Mitigation Strategies....7

Chapter IV: Manifestations of Bias in AI: A Case Study on Gender Bias in Neural Machine Translation and Large Language Models....8

CHAPTER I....12

EXPLORING THE DISCUSSION OF ARTIFICIAL INTELLIGENCE IN BUSINESS AND MANAGEMENT STUDIES....12

Introduction....12

Theoretical Background: AI literature review....15

Computational Literature Review....17

Research Method....19

Results....23

Descriptive analysis of bibliometric data....23

Topics exploration with co-word analysis....26

Impact analysis....29

Content analysis....31

Discussion....34

Conclusion....42

CHAPTER II....47

ARTIFICIAL INTELLIGENCE BIAS, FAIRNESS AND ETHICS: A COMPUTATIONAL LITERATURE REVIEW....47

Introduction....47

Theoretical Background....49

Previous Literature Reviews....50

Computational Literature Review....51

Material and methods....52

Results....55

Impact Analysis....55

Focus on the Senior Scholars List of Premier Journals of AIS....57

Content Analysis....60

Discussion....63

Conclusion 75
CHAPTER 3 79
EXPLORING AI BIAS AND MITIGATION CHALLENGES USING A SOCIO-TECHNICAL PERSPECTIVE 79
Introduction 79
Theoretical Background 82
AI regulations and guidelines 83
Methodology 85
Results and discussion 87
Future work and expected results 91
CHAPTER IV 94
DOES AI REFLECT HUMAN BEHAVIOUR? EXPLORING THE PRESENCE OF GENDER BIAS IN AI TRANSLATION TOOLS 94
Introduction 94
Theoretical Background 96
Methodology 98
Results 101
Analysis of AI translation tools 101
Analysis of the survey 105
Discussion 108
Conclusion 111
CONCLUSIVE REMARKS: A SOCIO-TECHNICAL PERSPECTIVE ON FUTURE CHALLENGES 115
REFERENCES 120

LIST OF TABLES

INTRODUCTION....1

Table 1. Summary of the development 9

CHAPTER I 12

EXPLORING THE DISCUSSION OF ARTIFICIAL INTELLIGENCE IN BUSINESS AND MANAGEMENT STUDIES 12

Table 1. Literature review comparison....16

Table 2. Most recurring journals 25

Table 3. Top 20 Journals ranked by citation count and AJG Area 31

Table 4. Topic description and possible guiding question for future research....41

CHAPTER II 47

ARTIFICIAL INTELLIGENCE BIAS, FAIRNESS AND ETHICS: A COMPUTATIONAL LITERATURE REVIEW 47

Table 1. Literature review comparison....50

Table 2. Top 20 Journals ranked by citation count and AJG Area 56

Table 3. Top 20 Journals ranked by production and AJG Area 57

Table 4. Top 10 impactful articles of the Senior Scholars list of premier journals....58

Table 5. Topics title and description 63

Table 6. Meso-topics titles and possible future research opportunities 74

CHAPTER 3 79

EXPLORING AI BIAS AND MITIGATION CHALLENGES USING A SOCIO-TECHNICAL PERSPECTIVE....79

Table 1. Information about the selected sample for the preliminary interviews. 86

Table 2. Description of the interview protocol 90

CHAPTER IV 94

DOES AI REFLECT HUMAN BEHAVIOUR? EXPLORING THE PRESENCE OF GENDER BIAS IN AI TRANSLATION TOOLS 94

CONCLUSIVE REMARKS: A SOCIO-TECHNICAL PERSPECTIVE ON FUTURE CHALLENGES 115

Table 1. Socio-technical perspective on AI bias challenges 117

REFERENCES....120

LIST OF FIGURES

INTRODUCTION....1

Figure 1. Research project overview....10

CHAPTER I....12

EXPLORING THE DISCUSSION OF ARTIFICIAL INTELLIGENCE IN BUSINESS AND MANAGEMENT STUDIES....12

Figure 1. Research protocol framework....22

Figure 2. Number of publications per year since 1983....24

Figure 3. Most productive countries....24

Figure 4. Most productive authors....26

Figure 5. Main keyword used in the dataset over time....27

Figure 6. Authors' keyword co-occurrence....28

Figure 7. Impact of the AJG fields for production and citation (total, normalized, and local normalized)....29

Figure 8. Topic Coherence Score....32

Figure 9. Number of publications per year since 1983 divided by AJG fields (see research method paragraph for AJG Area correspondence)....33

Figure 10. AJG areas inside each topic and articles distribution over time....36

Figure 11. Topic affinity....39

CHAPTER II....47

ARTIFICIAL INTELLIGENCE BIAS, FAIRNESS AND ETHICS: A COMPUTATIONAL LITERATURE REVIEW....47

Figure 1. Research protocol adapted from Smacchia and Za, (2022)....52

Figure 2. Selection of the number of topics (K)....60

Figure 3. Number of publications per year since 1983 divided by AJG fields (see research method paragraph). Full legend available at: https://bit.ly/3K1EbML....62

Figure 4. Intertopic distance representation: macro-topics (left) and meso-topics (right) 65

Figure 5. Hellinger's Distance (topic similarity)....67

Figure 6. Meso-topics distributions according to the nature and focus of the study ...69

CHAPTER 3....79

EXPLORING AI BIAS AND MITIGATION CHALLENGES USING A SOCIO-TECHNICAL PERSPECTIVE....79

Figure 1. Graphical representation of the five-step methodology adopted....87

Figure 2. AI bias categories intersection and typologies....89

CHAPTER IV....94

DOES AI REFLECT HUMAN BEHAVIOUR? EXPLORING THE PRESENCE OF GENDER BIAS IN AI TRANSLATION TOOLS....94

Figure 1. Research protocol....100

Figure 2. Jobs distribution across genders (tools taken separately)....104

Figure 3. Gender distribution across jobs taken singularly (tool aggregate data)....105

Figure 4. Gender distribution across jobs taken singularly (questionnaire data).....106

Figure 6. Gender distribution across male and female group jobs (tools vs questionnaire) 110

CONCLUSIVE REMARKS: A SOCIO-TECHNICAL PERSPECTIVE ON FUTURE CHALLENGES 115

REFERENCES.....120

INTRODUCTION

In recent years, society has become increasingly accustomed to living in an automated world, where identifying an unfamiliar object or song takes seconds, cars operate autonomously, and goods can be purchased and delivered within hours. These advancements, among numerous others, have been enabled by the rapid development of technologies rooted in Artificial Intelligence (AI) (Haenlein & Kaplan, 2019). AI is becoming increasingly prominent on the global stage, gaining growing attention from scholars, practitioners, policymakers, and society at large as interactions with AI systems continue to expand across various domains. This trend has become particularly evident with advancements in Generative AI, which have been marked by the introduction of platforms capable of generating text, images, and videos. For instance, tools like ChatGPT rapidly gained millions of users, demonstrating remarkable capabilities in solving a diverse range of tasks (Banh & Strobel, 2023; Y. Qin et al., 2024). While numerous definitions of AI exist, one that is particularly broad and has received considerable scholarly attention is provided by Haenlein & Kaplan (2019). They define AI as "*a system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation*". This definition focuses on the versatility of AI systems, which has been further amplified by recent advancements in data availability, computational power, and adaptive capabilities. Indeed, the exponential growth of data generated from diverse sources and applicable across a wide range of domains, along with enhanced computational power and the ability of AI systems to solve specific tasks, has significantly increased the potential applications of AI. Today, these capabilities enable AI to be employed across nearly every sector, where it is utilized to perform a wide array of functions and solve complex challenges (Collins et al., 2021; Shobana & Kumar, 2015; Tsai et al., 2015).

AI has emerged as one of the most influential topics of the last decade, profoundly affecting society at all levels, from governments and organisations to individual employees (D'Cruz et al., 2022; Makarius et al., 2020). It is a pivotal component in digital transformation processes (Xu et al., 2018), and its applications span many organisational functions, supporting a variety of decision-making and automation-based tasks. As AI becomes integral to future job roles, ensuring that its outputs remain free from discriminatory or unfair biases has become a relevant issue (Hunkenschroer & Luetge, 2022; Mehrabi et al., 2021). Notably, AI systems, when applied to address specific challenges, may inadvertently magnify existing issues by threatening rights,

opportunities, and equity. In some cases, they not only perpetuate or amplify pre-existing disparities but also create new inequalities (Hoffmann, 2019). As organisations and governments increasingly adopt AI technologies, the decisions made by these systems have the potential to simultaneously impact large populations, heightening concerns around ethics, fairness, and algorithmic bias (Zuiderwijk et al., 2021). There are numerous instances illustrating the consequences of AI bias. For example, Datta et al., (2015) found that Google Ads tools displayed fewer advertisements for high-paying jobs to women users compared to men. Similarly, there have been cases where AI systems disproportionately associated criminal profiles with the physical characteristics of Black men[1].

Addressing issues of bias and fairness is, therefore, significant. This has prompted a reevaluation of AI-based systems in favour of approaches that prioritize these values (Ntoutsi et al., 2020). Organizations designing and deploying AI technologies have a significant responsibility to mitigate these challenges by ensuring their systems are transparent and explainable (D'Cruz et al., 2022; Martin et al., 2019).

The growing adoption of AI technology in society has led to a corresponding increase in academic interest, fostering new research, theories, and questions (Makarius et al., 2020; Smith, 2020). Over the past four years, studies examining the societal impact of AI have expanded exponentially, with particular emphasis on ethical considerations and fairness (Mehrabi et al., 2021). Issues such as trust, transparency, and the explainability of AI systems have gained significant importance among various stakeholders. Consequently, expertise in areas such as responsible AI is becoming increasingly vital for professionals in this field (Meghan Rimol, 2021).

Despite significant advancements, ethical considerations are not consistently integrated into the AI lifecycle, which encompasses the design, development, and deployment of AI artifacts. This gap between technological innovation and ethical or responsible implementation highlights the need for further research into methodologies and tools that embed ethical principles, bias mitigation strategies, and transparent processes in AI projects. As the field continues to evolve at a rapid pace, addressing these challenges remains a relevant and urgent priority (Haenlein et al., 2022; Stahl et al., 2016; Vakkuri et al., 2019). Moreover, while biased

[1] Leonardo Nicoletti and Dina Bass (2023), Humans Are Biased. Generative AI Is Even Worse. Stable Diffusion's text-to-image model amplifies stereotypes about race and gender - here's why that matters. Retrieved October 2024 at: https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

outcomes in AI often stem from errors in algorithms or mismanagement of training data, it is also relevant to examine the broader context and understand why such technical errors occur. Although computational approaches to bias mitigation are essential, AI artifacts are created within complex socio-technical environments shaped by organizational, individual, and environmental factors (Dignum, 2023). These factors influence the values of designers, organizational processes, and user interactions, ultimately impacting the fairness and accuracy of AI systems (Akter et al., 2021; Dignum, 2018; Schwartz et al., 2022; Someh et al., 2019). Therefore, addressing AI bias effectively requires a socio-technical perspective and a greater focus on empirical research into its origins and mitigation (Kordzadeh & Ghasemaghaei, 2022; Mikalef et al., 2022).

Building on the socio-technical perspective (Bostrom & Heinen, 1977b, 1977a), the identification and mitigation of biases in AI require a holistic approach that acknowledges both technical and social dimensions. As previously discussed, while biases in AI systems may derive from technical limitations, such as insufficient training data or algorithmic design flaws, these technical issues are deeply intertwined with and shaped by the broader social biases embedded in human society. This interdependence illustrates the need for "joint optimization" (Sarker et al., 2019) to ensure that efforts to correct technical inaccuracies do not overlook the social structures and organisational dynamics that amplify them. By examining AI systems through this twofold lens, researchers and practitioners can more effectively identify sources of bias, simultaneously advancing technical improvements and fostering the adapting progression of the society and organisation (e.g., inclusive policy-making and stakeholder engagement). For instance, the adoption of a socio-technical approach in addressing AI bias could facilitate in mitigating the perpetuation of historical biases. By acknowledging the reciprocal influence between technical solutions (e.g., AI tools for human resource scouting) and social dynamics (e.g., existing biases in hiring practices and organisational culture), this perspective ensures that AI models do not reinforce outdated patterns but instead evolve in alignment with contemporary principles of fairness, inclusivity, and ethical standards. Ultimately, such a harmony between technical and social components can lead to AI implementations that not only enhance technological progress but also contribute to more equitable outcomes for society at large.

In light of what is stated, the aim of this thesis is to explore and advance the understanding of bias in AI systems, integrating theoretical and empirical analysis. More specifically, the thesis aims to answer the following questions:

RQ1: How has the academic discourse on artificial intelligence in organizational studies is evolving, and what key themes and research gaps are emerging?

RQ2: What processes do companies adopt to identify and mitigate bias in AI solutions? And how do these processes vary across the different phases of the AI lifecycle?

The research begins with an exploratory phase grounded in a comprehensive review of existing literature. This phase aims to examine the role of AI in organisational and societal contexts, focusing on how biases manifest in decision-making processes and algorithmic outputs. By employing computational literature review (CLR) techniques, this part of the thesis identifies key themes, trends, and gaps in the current academic discourse on AI bias and ethics. It highlights the dual impact of AI on technical and societal dimensions, uncovering behavioural and practical implications that extend beyond traditional computational approaches.

Building on these insights, the second part of the research project adopts an empirical approach, advancing the debate through the collection and analysis of primary data. This phase is focused on case studies, interviews, and experimental methodologies to investigate the emergence and mitigation of bias in AI systems. By focusing on specific contexts, the research explores the connections between technical and socio-cultural factors. Empirical findings are used to propose novel frameworks for identifying and mitigating bias, offering insights for practitioners, policymakers, and academics. Through this dual structure, the thesis aims to contribute both theoretically and practically to the ongoing discourse on AI bias, enhancing the understanding of how to better mitigate the emergence of bias within AI systems.

This thesis is part of a research project started in January 2022 and concluded in January 2025. It is structured around four chapters, each corresponding to a specific paper developed collaboratively by my supervisor and me over the course of the project. Each chapter represents a distinct phase of the research, reflecting the progression of the study. The following paragraphs provide context for the individual chapters, which are arranged to offer a cohesive and comprehensive overview, ensuring fluidity in the development of the argument.

Chapter I: The Role of Artificial Intelligence in Business and Management: A Literature Review of Evolving Perspectives

The first chapter corresponds to the initial research conducted within the research project, providing a comprehensive overview of the current debate on AI within organisational and managerial studies. A

preliminary analysis of the dataset was presented at the Workshop di Organizzazione Aziendale (WOA) 2022. Subsequently, the analysis was finalised, and the full study was presented at the international conference of the Association for Information Systems (ICIS) 2022[2]. This research employed the CLR method to perform an impact analysis and a topic modelling analysis of relevant literature using the Latent Dirichlet Allocation (LDA) technique. Specifically, computational methods were employed to conduct a literature review on a large corpus of data. This approach facilitated a comprehensive mapping of the field, enabling the integration of quantitative research, focused on descriptive and computational analyses, with qualitative research, involving an in-depth examination of the most relevant papers within the dataset. Fifteen major topics in the AI discourse regarding organisational studies were identified and analysed in-depth. Each was categorised as declining, stable, or emerging, offering a comprehensive view of the current debate. The paper also highlighted two primary branches of research: technical and societal aspects, the latter of which have gained increasing importance in recent years. Furthermore, a set of guiding questions was proposed to foster future research directions, particularly those focusing on emerging topics. Within this first chapter, a section is dedicated to a related study presented as a separate paper. This second paper was presented at the Italian Chapter of the Association for Information Systems (ItAIS) 2022 and published in a Springer volume as part of the conference proceedings[3]. It involves a bibliometric literature review, providing an overview of the current debate on AI in organisational studies by conducting a bibliometric analysis of the same dataset of 1102 used for the CLR. The dataset included publications from journals within the organisational and management fields listed in the Academic Journal Guide. The analysis revealed a significant increase in publications addressing societal and ethical issues related to AI. Through co-word analysis, key topics in the research domain were mapped and described, many of which are experiencing renewed interest among scholars. The two studies are complementary, integrating quantitative and qualitative methods to offer a comprehensive understanding of the topic. Both contributions underscore the emergence of a branch of literature over the past four years that

[2] Smacchia, M. and Za, S., "Artificial Intelligence in Organisation and Managerial Studies: A Computational Literature Review" (2022). In Proceedings of the 43rd International Conference on Information Systems (ICIS), Copenhagen Denmark. https://aisel.aisnet.org/icis2022/soc_impact_is/soc_impact_is/6.

[3] Smacchia, M., and Za, S. (2024). Investigating the Artificial Intelligence Debate in Organisation and Management Studies. In: Braccini A.M., Pallud J., Pennarola F. (eds) Technologies for Digital Transformation. ItAIS 2022. Lecture Notes in Information Systems and Organisation, vol 64, pp 27-47. Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-031-52120-1.

focuses on societal and ethical issues in AI. From the results of these two literature reviews, a subsequent CLR was developed to explore AI bias, fairness, and ethics in organisational and managerial studies. The first CLR, which explores the debate within the field of AI and organisational and management studies, has been revised to incorporate methodological and presentation improvements developed during the second CLR. This updated version is currently being prepared for publication as a book chapter. For this reason, certain sections, such as the research protocol, may appear similar. Both papers have been revised during the development of this document to minimise repetition and enhance readability.

Chapter II: Challenges of AI Implementation: Bias, Fairness, and Ethical Considerations in Theory and Practice

Building on the findings of the first research on AI and organisational studies, the second chapter is devoted to a second CLR focused on the discourse surrounding AI bias, ethics, and fairness. In this study, the query, computational techniques, and analytical methods were further refined. Specifically, the research protocol implemented in the previous CLR was revised to provide greater specificity regarding the steps required to identify the research topic and objectives, manage and analyse the data, and present the results. More structured and reliable tools were employed for data analysis, enabling a deeper examination of unstructured data. As a result, eighteen topics were identified and analysed in depth, leading to the discovery of five overarching categories, which were further grouped into two macro-categories, both related to the socio-technical aspects of the AI discourse. These findings revealed distinct clusters within the debate, addressing the identification and resolution of AI bias from both technical and societal perspectives. Additionally, recurring themes were identified, highlighting the influence of human and institutional values on AI outcomes, an area that has attracted increasing scholarly attention in recent years. Following feedback received during the conference presentation, an extended version of the paper was developed, improving the introduction, analysis, and discussion of results. Two new sections were added to describe the contribution of the AIS Senior Scholars list of premier journals to the debate on AI bias, fairness, and ethics. A theoretical background section was also included to better articulate the significance of the AI bias phenomenon. Moreover, the methodology and discussion sections were refined through an additional analysis iteration, which uncovered previously unnoticed patterns. The preliminary version of this CLR was presented at the European Conference of the

Association for Information Systems (ECIS) 2023[4]. Subsequently, an extended version, incorporating the feedback received, was developed and is currently under review in the Journal Communications of the Association for Information Systems (ISSN: 1529-3181).

Chapter III: Empirical Approaches to AI Bias: Identification, Measurement, and Mitigation Strategies

The third chapter marks the beginning of the second part of the research project, which focuses on empirical research. The first study conducted, which corresponds to this chapter, is a qualitative analysis that investigates the challenges of identifying and mitigating AI bias through a socio-technical lens. Specifically, the study examines how biases are identified and addressed throughout the AI lifecycle, with the goal of developing a theoretical framework that systematically incorporates stakeholder participation into bias mitigation strategies. The research employed grounded theory, utilising semi-structured interviews with AI professionals involved in various stages of the AI lifecycle. Preliminary interviews were conducted with five AI experts, with an average interview duration of 53 minutes. The findings revealed several insights. During the design phase, challenges were identified in data exploration and collection, particularly in reviewing state-of-the-art literature and selecting appropriate datasets. The development phase highlighted the significance of continuous feedback iterations and technical challenges, such as data correction and improving model precision. Notably, a previously unconsidered form of bias related to hardware, such as the impact of graphics card drivers on deep learning models, was identified. In the deployment phase, field testing emerged as essential for uncovering hidden biases, with stakeholder engagement playing a pivotal role. In addition, this study enabled the identification of various types of bias, contributing to the refinement of the framework developed by the U.S. National Institute of Standards and Technology (NIST) for identifying and mitigating bias (Schwartz et al., 2022). A short paper presenting preliminary findings and data analysis was presented at the European Conference of the Association for Information Systems (ECIS) 2024[5]. Subsequently, additional data were gathered and analysed, resulting in an advanced version of the study that focuses on more specific aspects, such as AI regulation. This version was presented to the European Group for Organizational Studies (EGOS)

[4] Smacchia, M. and Za S., "Exploring Artificial Intelligence Bias, Fairness and Ethics in Organisation and Managerial Studies" (2023). In Proceedings of the 31st European Conference on Information Systems (ECIS), Kristiansand, Norway. https://aisel.aisnet.org/ecis2023_rp/362.

[5] Smacchia, M., Za S., and Arenas, A. E., "Identifying AI Bias and Mitigation Challenges Through a Socio-Technical Perspective" (2024). In Proceedings of the 32nd European Conference on Information Systems (ECIS), Paphos, Cyprus, 2024. https://aisel.aisnet.org/ecis2024/track03_ai/track03_ai/12.

Colloquium 2024[6]. Currently, the research is based on pilot interviews that have been instrumental in refining the research protocol and interview questions. A full version of this study is currently under development.

Chapter IV: Manifestations of Bias in AI: A Case Study on Gender Bias in Neural Machine Translation and Large Language Models

The fourth chapter focuses on a specific case study examining the influence of social norms and values on widely used AI tools. The study investigates the presence of gender bias in neural machine translation systems and large language models (LLMs). The research aims to analyse how these AI tools handle gender translation in sentences where the subject's gender is not explicitly mentioned, specifically when translating from Italian (a language where subject omission is grammatically acceptable) to English (where explicit subjects are required). The study explores the extent of gender bias in common AI translation tools, the impact of query formulation on translation outputs, and comparisons between AI-generated and human translations. For the analysis, ten job titles were selected based on gender distribution statistics. Five of these jobs were predominantly held by women and five by men, ensuring all job titles were gender-neutral in Italian. Three sentences of varying complexity were created for each job, all featuring gender-neutral phrasing in Italian. These sentences were translated into English using five AI tools: Google Translate, DeepL, Microsoft Azure, GPT 3.5, and Text DaVinci-003. To compare AI outputs with human translations, a survey was conducted with 42 Italian participants who were tasked with completing English sentences by adding gender-specific pronouns and adjectives. The results identified two distinct types of bias. The first, termed "converging bias," occurred when translation tools consistently used the same gender throughout all translations after being influenced by the initial translation choice. The second was gender bias, particularly evident in translations involving traditionally male-dominated occupations. Human translations largely mirrored the AI outputs, though they exhibited slightly less bias and made greater use of gender-neutral language. The findings highlight how AI translation tools can reflect and reinforce existing societal gender biases, particularly in occupational contexts. The alignment between AI and human translations suggests that these biases are deeply rooted in social constructs rather than being purely technical in nature. An early version of this research was presented at the

[6] Smacchia, M., Za S., and Arenas A. E., "Exploring AI bias and mitigation challenges using a socio-technical perspective" (2024). In Proceedings of the 40th European Group for Organizational Studies Colloquium (EGOS 2024), Milan, Italy.

Italian Chapter of the Association for Information Systems (ItAIS) 2023 and published in a Springer volume as part of the conference proceedings[7]. The analysis is currently being updated with new LLM models and Spanish language sentences to provide a more comprehensive understanding of the issue.

A timeline of the research project, along with a brief overview of each chapter's contribution to the overall study, is shown in Table 1 and Figure 1.

Table 1. Summary of the development

Period of the research project	Chapter	Paper	Type of study	Outcome
2022	I	• Smacchia, M. and Za, S., "Artificial Intelligence in Organisation and Managerial Studies: A Computational Literature Review" (2022). In Proceedings of the 43rd International Conference on Information Systems (ICIS), Copenhagen Denmark. • Smacchia, M., and Za, S. (2024). Investigating the Artificial Intelligence Debate in Organisation and Management Studies. In: Braccini A.M., Pallud J., Pennarola F. (eds) Technologies for Digital Transformation. ItAIS 2022. Lecture Notes in Information Systems and Organisation, vol 64, pp 27-47. Springer, Cham.	Computational literature review (CLR) + Bibliometric analysis	ICIS 2022 ItAIS 2022
2022/2023	II	• Smacchia, M. and Za S., "Exploring Artificial Intelligence Bias, Fairness and Ethics in Organisation and Managerial Studies" (2023). In Proceedings of the 31st European Conference on Information Systems (ECIS), Kristiansand, Norway. https://aisel.aisnet.org/ecis2023_rp/362.	Computational literature review (CLR)	ECIS 2023 CAIS (under review)
2023/2024	III	• Smacchia, M., Za S., and Arenas, A. E., "Identifying AI Bias and Mitigation Challenges Through a Socio-Technical Perspective" (2024). In Proceedings of the 32nd European Conference on Information Systems (ECIS), Paphos, Cyprus, 2024. • Smacchia, M., Za S., and Arenas A. E., "Exploring AI bias and mitigation challenges using a socio-technical perspective" (2024). In Proceedings of the 40th European Group for Organizational Studies Colloquium (EGOS 2024), Milan, Italy.	Case based research	ECIS 2024 EGOS 2024 Full version in progress

[7] Smacchia, M., Za, S., & Arenas, A. (2024). Does AI Reflect Human Behaviour? Exploring the Presence of Gender Bias in AI Translation Tools. In Digital (Eco) Systems and Societal Challenges: New Scenarios for Organizing (pp. 355-373). Cham: Springer Nature Switzerland.

Table 1. (Continued)

2023/2024	IV	• Smacchia, M., Za, S., & Arenas, A. (2024). Does AI Reflect Human Behaviour? Exploring the Presence of Gender Bias in AI Translation Tools. In Digital (Eco) Systems and Societal Challenges: New Scenarios for Organizing (pp. 355-373). Cham: Springer Nature Switzerland.	Exploratory study	ItAIS 2023 Full version in progress

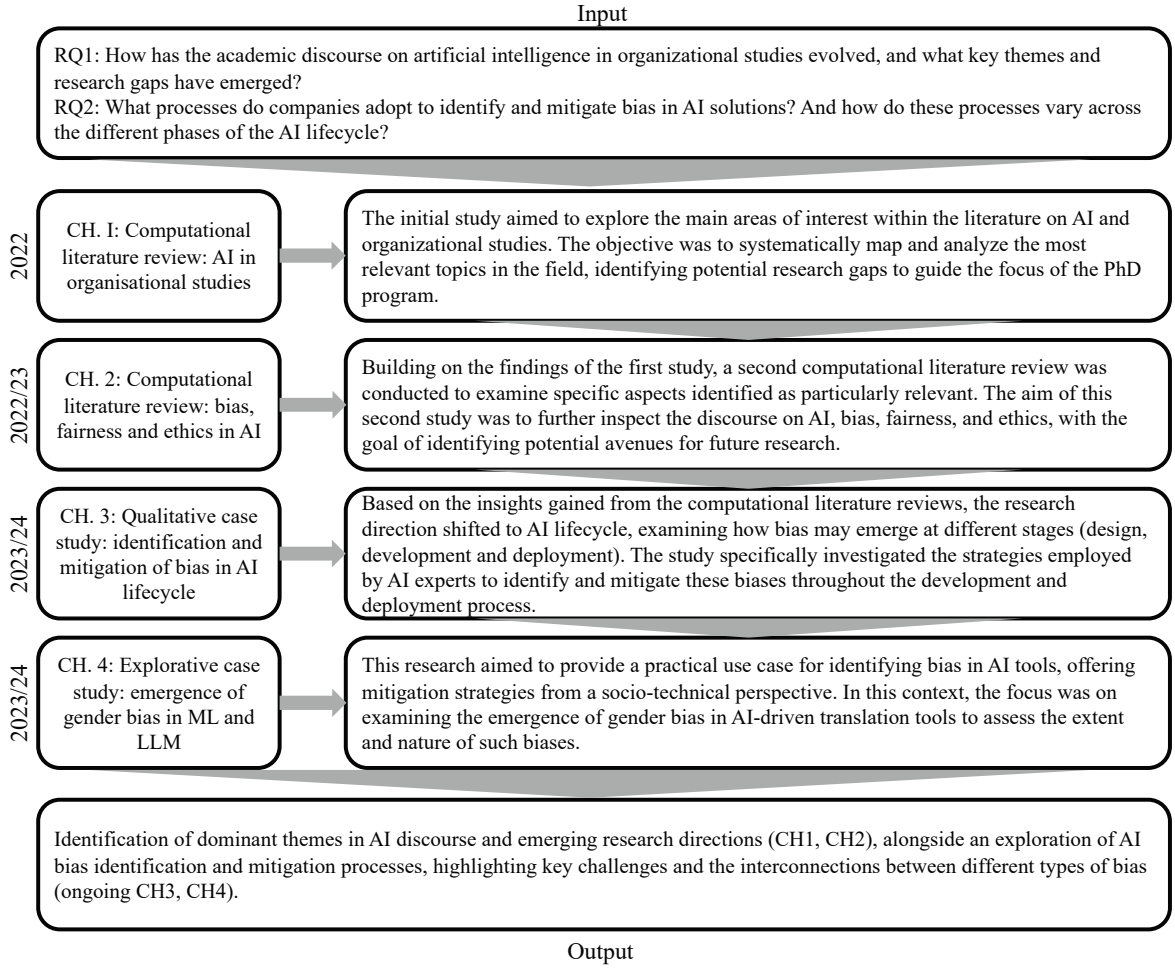

Figure 1. Research project overview

CHAPTER I

Exploring the Discussion of Artificial Intelligence in Business and Management Studies

Abstract. Nowadays, the number of organisations inside the 4.0 industry adopting new technologies based on Artificial Intelligence (AI) is continuously increasing. Scholars are not neglecting this phenomenon, investigating related topics, not only from a technological point of view, but also in the organisational and societal domains. The goal of this research is to develop a complete overview of the current discussion on AI in business and management studies. To this end, we adopted the Computational Literature Review (CLR) method to conduct an impact and topic modelling analysis of the relevant literature using the Latent Dirichlet Allocation (LDA) technique. As a result, we identified 15 topics concerning the artificial intelligence discussion in business studies, providing a detailed description of each of them and identifying which one is declining, stable or emerging. We also recognised two main branches of research regarding technical and societal aspects, where the latter has become increasingly important in recent years. Finally, focusing on all the topics, we proposed a set of guiding questions that might foster future research directions. This study provides insights to scholars and managers interested in AI and could also be used as a guide to performing CLR.

Keywords: Artificial Intelligence, Machine Learning, Computational Literature Review, Latent Dirichlet Allocation.

Introduction

Artificial Intelligence (AI) has become relevant within the global landscape, across a variety of fields. It is experiencing a wave of increasing interest and investment from both industrial and academic sectors (Stanford, 2023). AI is an umbrella term, that encompasses many approaches and techniques. It could be defined as "a system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation" (Kaplan & Haenlein, 2019a). AI can therefore be designed, developed and embedded in a multitude of technologies (Zhu et al., 2021). This technology has extensive potential for different applications, capable of addressing many complex tasks across different domains (D'Cruz et al., 2022). Researchers and developers are trying to innovate solutions that are increasingly

precise, reliable, and powerful. While current AI applications, known as artificial narrow intelligence (ANI), are predominantly confined to specialized domains, there is a blooming effort among experts to overcome these limitations. The goal is to create versatile AI artefacts capable of functioning across multiple areas and capable of performing a wide spectrum of tasks (Kaplan & Haenlein, 2019b). The transition to more advanced forms of AI poses many challenges from a technological, organisational, and societal perspective. It is therefore relevant to study the impact of these new technologies and the possible implications of their adoption (Zhu et al., 2021). Indeed, AI is among the most relevant topics of the last decade and is reshaping human interaction in various domains, from everyday tasks to working life (Makarius et al., 2020). Especially from an organisational perspective, AI is being implemented for the automation of many processes within companies (ibid). Bhaskar Ghosh & Paul R. Daugherty, (2019) state that about 80% of companies have adopted AI technology in their core business, with a 70% increase in five years. The outcomes of this broad implementation of new technologies based on AI algorithms are manifold. On the one hand, many companies are experiencing a technological revolution, by exploiting the growing computational capacity of devices together with the advantages of new technologies, they are reaching levels of efficiency that were unimaginable just a few years ago (Schwab, 2016). On the other hand, the implementation of AI technologies and their adoption often lead to some forms of resistance from employees and managers (Glikson & Woolley, 2020; M. H. Huang & Rust, 2018). In fact, AI is perceived by many as a possible replacement for their work, especially when applied to solving analytical and communication tasks. In these cases, although AI is a continuous source of innovation, it could have many effects on human employment and human behaviour (ibid.). According to Makarius et al., (2020), AI is profoundly changing the world of work. Moreover, due to the recent COVID-19 pandemic, companies have accelerated their AI adoption plans[8]. As a result, there is a need for more research investigating the effects of adopting various AI solutions in many fields of business organisation.

Since the introduction of these new technologies, studies on the use of AI have increased significantly, especially in recent years, fostering new theories and opening up new questions (Makarius et al., 2020). Although in the business and organisational sphere, technology has been considered one of the fundamental factors in shaping the forms of organisation and its functions, the number of studies investigating the impact

[8] Joe McKendrick (2021), AI Adoption Skyrocketed Over the Last 18 Months, Harvard Business Review. Retrieved May, 2022: https://hbr.org/2021/09/ai-adoption-skyrocketed-over-the-last-18-months.

of new technologies on business processes has declined over time, at least until the advent of the Industrial Revolution 4.0. In fact, starting less than a decade ago, the interest of scholars in such issues has increased, along with the interest in investigating issues related to new AI-based technologies within organisational environments (Bailey et al., 2019). For exploring the content of this increasing number of papers in the literature, traditional methods of exploration could not be very effective in comprehensively grasping the nuances of broad contexts within this rapidly evolving field. The sheer volume and complexity of AI-related research now necessitate more advanced approaches to effectively identify patterns inside large corpora of articles (Antons & Breidbach, 2018).

In light of these assumptions, we believe that it may be of interest to map and evaluate the existing literature on how the AI discussion has been developed in business and management studies, in order to identify the most prevalent themes and research gaps that may lead to future studies (Tranfield et al., 2003). More precisely, our goal is to identify and describe the relevant topics within this domain over time, tracing their evolutionary trajectory and analysing their progression. Additionally, we also aim to forecast potential future research directions, particularly focusing on those topics that are currently widespread in academic and industry discourse. Therefore, our objective is to address the following research questions:

How has the discussion of artificial intelligence evolved within business and managerial studies? And what are the most significant topics?

What are the emerging topics related to AI in business and managerial studies that could drive future research directions?

Given the large number of scientific articles related to the topic of interest, among several possible literature review approaches, we decided to perform a Computational Literature Review (CLR). The decision to adopt this specific methodological approach was driven by the need to conduct both a quantitative and qualitative analysis of the literature within a large dataset. Given the large volume of data to be examined, we opted for an inductive, top-down approach. This strategy was chosen not only to synthesise the literature on the topic but also to ensure a comprehensive outlook on the subject (J. Xie & Jain, 2024). This method, therefore, seems to be more appropriate than others, such as bibliometric analysis, which instead uses mainly a

quantitative approach to describe, evaluate and monitor published research (Lamboglia et al., 2020). Although our research employed bibliometric techniques to describe our dataset, we decided to perform a topic modelling analysis rather than alternative bibliometric methods, such as co-word analysis. This decision was strategically aligned with our primary objective: to discover the latent themes within the literature rather than mapping the explicit connections among key terms. Topic modelling, with its ability to statistically infer hidden themes from patterns of word usage, offered a more pertinent approach for delving into the topics concerning the domain (Blei, 2012). This methodology was also favoured over a systematic review due to our primary aim of conducting an exploratory analysis encompassing a large dataset. The systematic literature review approach, while meticulous and comprehensive, is more suited for focused inquiries. Given the extensive breadth of our topic and the volume of data at hand, employing a systematic review would have presented significant challenges. Furthermore, even though other literature review methods are supported by software, CLR goes further by using text mining and Machine Learning (ML) algorithms to examine the content of the article and allow the machine to automatically perform time-consuming tasks. Finally, each topic is better described and contextualised by analysing in depth the related most representative papers accordingly.

Our research provides an outlook on the evolution of the discourse surrounding AI within the domain of business and management. The findings underscore the growing relevance and novelty of societal issues linked to AI, as well as their profound impact on society at large. We have identified several research questions related to the most emergent topics in this domain. Through our study, we aim to offer insightful guidance to both scholars and practitioners who are interested in this field or who wish to conduct a similar analysis to uncover the latent themes within a particular domain. The next section provides further details on the methodology and research protocol. The presentation of the results with their implications and discussion closes the contribution.

Theoretical Background: AI literature review

Since the studies concerning AI have increased exponentially in recent years, it is not surprising that systematic literature review (Borges et al., 2021; Grover et al., 2022; Kitsios & Kamariotou, 2021; Toorajipour et al., 2021; Trunk et al., 2020; Zuiderwijk et al., 2021) and especially bibliometric analysis (Dhamija & Bag, 2020; Reis et al., 2019; Verma et al., 2021) are widely used as review methods. Interestingly, the papers cover a broad

range of topics, indicating that AI, inside the organizational setting, is being applied in different areas for different purposes.

Dhamija & Bag, (2020), conducted a bibliometric analysis of 1854 articles on AI in operations management. Through a network analysis they discovered the presence of 6 clusters on the predominant topic in the AI field. Another bibliometric analysis in conjunction with intellectual network analysis was carried out on 1580 articles from 1982 to 2020 to review the literature on marketing (Verma et al., 2021). Trunk et al., (2020) conducted a systematic literature review on 55 articles and built a conceptual model explaining how AI can be used for strategic decision-making. Moreover, the authors discover that to successfully use AI technology, human decision-makers have to become interpreters of the results of the machine and not only supervisors of predefined processes. Using the same review method, Kitsios & Kamariotou, (2021) provided a research agenda about AI and ML in organizational strategy and decision-making processes. Grover et al., (2022) combined literature review with social media analysis to study the effect of AI on operations management. Thanks to this hybrid research they were able to find different insights in both fields of analysis. Another systematic literature review was performed on 41 articles to provide a review of the state-of-the-art literature on AI in organizational strategy. Particular attention was given to Decision Support System (DSS), customer and employee engagement, automation and new products and service offering (Borges et al., 2021). Toorajipour et al., (2021) analysed 64 articles to review the use of AI in Supply Chain Management (SCM). They found interesting results regarding the most used AI techniques in SCM and also the SCM subfield that could have the highest potential if enhanced by AI. A further important topic is represented by the public sector. Two literature reviews explore how AI is impacting government and policymakers by exploring risks, implications and opportunities (Reis et al., 2019; Zuiderwijk et al., 2021).

Table 1. Literature review comparison

Article	Methodology	Focus	Findings
Dhamija and Bag (2020)	Bibliometric Literature Review (Network analysis) 1854 articles published between 2018 - 2019	Review AI and related domains in combination with operation management	Identified six clusters that focus on predominant themes for present and future research
Verma et al. (2021)	Bibliometric Literature Review (Co-citation and Co-occurrence analysis) 1580 articles published between 1982 - 2020	Review of AI inside marketing sector	Identified most relevant sources and researchers in the field, along with relevant topic and future research directions

Table 1. (Continued)

Trunk et al. (2020)	Systematic Literature Review and Content Analysis 55 articles published between 2016 and 2020	Review of adoption of AI for decision making under uncertainty	Identified a conceptual framework for AI adoption inside the organizational processes for decision making under uncertainty.
Kitsios and Kamariotou (2021)	Systematic Literature Review 81 articles published before 2020	Review of the correlation between AI and corporate strategy	Identified four sources of value creation concerning the use of AI in business strategy along with gaps for future researches
Grover et al. (2022)	Systematic Literature Review and Social Network Analytics 181 articles published between 2010 and 2019 combined with Twitter data	Review of the use of AI in operation management	Identified a framework for AI implementation inside organizations and best practices of AI tools adoption inside operation management
Borges et al. (2021)	Systematic Literature Review 41 articles published between 2009 - 2020	Review of the integration of AI to organizational strategy	Identified four sources of value creation concerning AI inside organizational processes along with gaps for future researches for each of the sources
Toorajipour et al. (2021)	Systematic Literature Review 64 articles published between 2008 - 2018	Review of AI inside Supply Chain Management	Identified AI techniques and applications that can contribute to Supply Chain Management development
Reis et al. (2019)	Bibliometric Literature Review and Content Analysis 74 articles published before 2018	Review of AI in government services: analysis of countries spending on AI	Identified areas of research concerning the use of AI in the public sector and EU - USA policies comparison
Zuiderwijk et al. (2021)	Systematic Literature Review 26 articles published between 2010 - 2020	Review of the implication of the use of AI inside Public Governance	Developed a research agenda on the implementation and use strategies of AI inside public governance and on the risks of AI use inside public sector

Table 1 summarises the methodology (bibliometric vs systematic), the focus, and the findings of recent literature reviews on AI. Specifically, a bibliometric literature review could be considered a quantitative approach, analysing many articles, while a systematic literature review adopts a mainly qualitative approach by examining a small set of articles. Indeed, due to the difficulty of qualitatively revising a large number of contributions, the systematic literature reviews mentioned previously have used a restricted pool of articles that rarely exceed 100 articles (Grover et al., 2022). With a CLR, we want to go further and try to qualitatively investigate a large sample of articles to provide a timely and insightful analysis of the current literature on AI in business organisational settings. Further details on the CLR method are provided in the following subsection.

Computational Literature Review

In recent years, there has been a proliferation of new journals and conferences leading to an exponential increase in the existing literature (Mortenson & Vidgen, 2016; Wagner et al., 2022). Furthermore, due to this growth in literature, scholars are relying heavily on literature reviews to inspect a particular research field

(Badger et al., 2000). The increasing complexity and breadth of the scientific literature, along with the growing importance of unbiased reviews, has led to the need for systematic and easily replicable literature reviews (Antons et al., 2021). Systematic Literature Review (SLR) is conducted using a systematic, rigorous and easily reproducible standard (Okoli & Schabram, 2010), this method is preferred to analyse today's extensive literature (Rowe, 2014) and to avoid any bias in article selection that is arbitrary in "Non-Systematic" reviews (Martin Kunca, 2018). However, the difficulty of conducting this type of research has increased significantly, with many researchers becoming discouraged by the time and effort required for such analyses and opting to focus on empirical studies (Mortenson & Vidgen, 2016). These problems have made it necessary to adopt new methods that allow the best practices of systematic literature review to be combined with computational methods (Antons et al., 2021) that would, on the one hand, speed up content analysis and, on the other hand, broaden the scope of reviews by identifying and extracting knowledge that would be precluded by manual analysis (Boyd-Graber et al., 2017). One method that enables researchers to analyse large volumes of documents in a rigorous and timely manner is CLR. defined as:

"*A structured process intended to augment human researchers' information processing capabilities through the use of machine learning algorithms that help analyse the content of a comprehensive text corpus in a specific knowledge domain (e.g., a research topic, academic journal, or scientific field) in a way that is scalable and real-time capable.*" (Antons et al., 2021)

CLR allows the automatic analysis of a dataset by identifying themes through topic modelling. It is sometimes referred to as non-linear principal component analysis because it finds latent components (called topics) that can explain variance in the data (Hindle et al., 2020). Topic models are algorithms capable of discovering themes within a large number of unstructured documents by analysing the connections between words contained therein (Blei, 2012). An example of a topic model is Latent Dirichlet Allocation (LDA), a probabilistic generative model for the collection and analysis of unstructured data. LDA is applied to find topics within a text on the basis of links between words and then classify texts according to the relevance of the topics found within them (Blei et al., 2003).

Since the literature review process is composed of both time-consuming mechanical tasks and creative tasks, the use of computational techniques such as AI algorithms could allow the researcher to concentrate more on the steps that require human intellect, knowledge and expertise (Wagner et al., 2022). This process

consists of an AI-human augmentation that helps the user to speed up the time required to perform part of the analysis and, in parallel, have more resources to dedicate to the salient part of the research to acquire better insights (ibid).

We perform the CLR following the guidelines provided by Antons et al., (2021), where they describe a six-step process analysis:

1) Define a conceptual goal that motivates the review.
2) Operationalize the CLR by defining the boundaries that are going to be inspected.
3) Choose a computational technique that best suits the conceptual goal.
4) Perform content analysis by preparing the data and deploying the computational technique.
5) Generate original insights by observing the results provided by the computational analysis
6) Present the findings in a clear and accessible way.

In this research, we use the guidelines as a baseline to develop the Figure 1 framework we adopted to run our CLR.

Research Method

The main purpose of this contribution, and thus our conceptual goal, is to explore and summarize the discussion on AI in business and management studies. Hence, it was quite relevant to operationalize our literature review by identifying the appropriate list of journals on which to perform the query in order to select the papers to create our dataset (Pütz & Werner, 2024). The Academic Journal Guide (AJG) (https://charteredabs.org/academic-journal-guide-2021/) of the Association of Business Schools (ABS) provides a list of journals classified according to the field of studies to which they belong and their ranking. With regard to the fields, we considered those belonging to the organizational, management and business sphere, such as: “Entrepreneurship and Small Business Management” (ENT-SBM), “General Management, Ethics, Gender and Social Responsibility” (ETHICS-CSR-MAN), “Human Resource Management and Employment Studies” (HRM&EMP), “Information Management” (INFO MAN), “Innovation” (INNOV), “Management Development and Education” (MDEV&EDU), “Organizational Studies” (ORG STUD), “Public Sector and Health Care” (PUB SEC), “Social Sciences” (SOC SCI) and “Strategy” (STRAT). For these fields, in order to get more relevant results from the research, we considered journals with AJG ratings of 4* (grade

4 journals that are recognized worldwide as exemplars of excellence), 4 (all journals with rating 4 publish the most original and best-executed research) and 3 (3-rated journals publish original and well-executed research papers and are highly regarded). The data were collected within Scopus (https://www.scopus.com/) since all the journals we decided to incorporate into the search were available in the academic database. Specifically, we performed a query in which we included all the chosen journals by entering their ISSN code and the keyword "Artificial Intelligence". We did not use time or other restrictions. The query returned 1175 articles published between 1983 and 2022; citation and abstract data were exported in CSV format. We then described our dataset using the R tool Bibliometrix (https://www.bibliometrix.org/) to extract some preliminary information (Aria & Cuccurullo, 2017).

Our first task was the choice of the computational technique to be used. We initially performed an impact analysis using the R programming language through the CLR library available on GitHub (https://github.com/rvidgen/clr) to get quantitative data on the information contained in our dataset. Through this analysis, it is possible to estimate the impact of a given paper, author and journal. To measure these dimensions, the tool uses various metrics such as citation count, the impact factor (total citation count divided by the total number of papers) and the h-index which is commonly used to assess the impact of researchers (Hirsch, 2005). The h-index refers to the following ratio: a researcher with an index of 10 has published 10 papers each of which has been cited at least 10 times. The index is not devoid of problems, e.g., potentially relevant new articles are underestimated due to fewer citations, or authors who have received many citations for an article are penalized compared to authors who have fewer citations in many articles. However, it is generally accepted to measure the impact of an author or journal.

We then focused on the content analysis of the papers by performing topic modelling using the Latent Dirichlet Allocation analysis with the Text-Analytics_LDA tool (https://github.com/tqx94/Text-Analytics_LDA). We adopted LDA because it is the most popular method used for topic modelling (Jelodar et al., 2019), and also, this method seems to have a higher level of reliability and accuracy compared to the others (Shadikur Rahman et al., 2020). LDA is an unsupervised generative probabilistic model of a corpus that allows the identification of a set of topics among multiple documents (Blei et al., 2003). Each document is considered as a set of words that can be combined to form subsets of latent topics. The model assumes that the corpus includes k topics and then distributes these topics across each document to see which one fits best. In

this way, the analysis is more efficient because it avoids cross-checking every word with every document. LDA algorithm can be summarised as follows:

$$P(\boldsymbol{W}, \boldsymbol{Z}, \theta, \varphi, \alpha, \beta) = \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{t=1}^{N} P(Z_{j,t}; \theta_j) P(W_{j,t}; \varphi_{z_{j,t}})$$

Where α represents the document-topic density (the percentage that a document is associated with a determined topic, if $\alpha < 1$ the documents tend to diverge to the single topics) β represents the topic-word density (the percentage that a word is associated with a determined topic) and θ with φ representing multinomial distributions of topics over documents and words over topics.

In order to perform the analysis, we had to clean our dataset. First, we deleted all the variables except for those related to abstracts and document IDs, then we removed punctuation and stop-words and tokenised the corpus to build the Document Term Matrix (DTM), which is a matrix that contains words and documents as dimensions. We also removed the search term "Artificial Intelligence", along with other recurring words that might yield incorrect results, such as "Elsevier", "Springer", "Research", and "Findings" (Mortenson & Vidgen, 2016). When creating the DTM, we decided to tokenise the text using one or two words to avoid ambiguity during topic analysis. In fact, some words have a different meaning when taken individually (e.g., Information and System instead of Information System). Since LDA is an unsupervised technique, the number of topics to be used was chosen a priori. To better evaluate this choice, the tool calculated a coherence score to find the best number of topics. Topic coherence measures the degree of semantic similarity between high-scoring words in the topic. We then represented the words of each topic in a word cloud and, according to the coherence score that the algorithm gave to every document, we used topic modelling results to analyse the most representative contributions for each topic. Our goal was to discover how AI is implemented and deployed within different topics and also to detect the most popular fields of research in recent years.

Figure 1 describes the research protocol we adopted to explore the literature using the CLR method. We started from Antons et al., (2021) six-step model by describing for each one the specific activities performed. In the analysis, taking into consideration Wagner et al., (2022) framework, we also provide additional information regarding the process of defining the research objectives, the literature search phase and the dataset cleaning. Moreover, we highlight the three tasks requiring iterations. More specifically, the first iterative task regards the definition of the right query for defining the dataset. Since this type of analysis is

often characterised by a dataset with thousands of articles, it becomes very difficult to perform a consistency check for all articles. Defining the query is therefore a fundamental step in order to minimise noise within the dataset that could negatively affect the outcome of the study. Subsequently, another iterative task is represented by the content analysis phase, to properly set the parameters of the tools used for identifying computationally the set of topics discussed in our dataset. In particular, the resulting topics are analysed and generated several times (modifying the parameters) to understand if they are significant or could indicate potential bias in the dataset. Finally, the third iterative task concerns the interpretation and the in-depth analysis of the most relevant papers associated with each topic, that was performed by the authors of this study.

The final part of this research protocol involves finding possible patterns in the literature to gain insights into the selected domain.

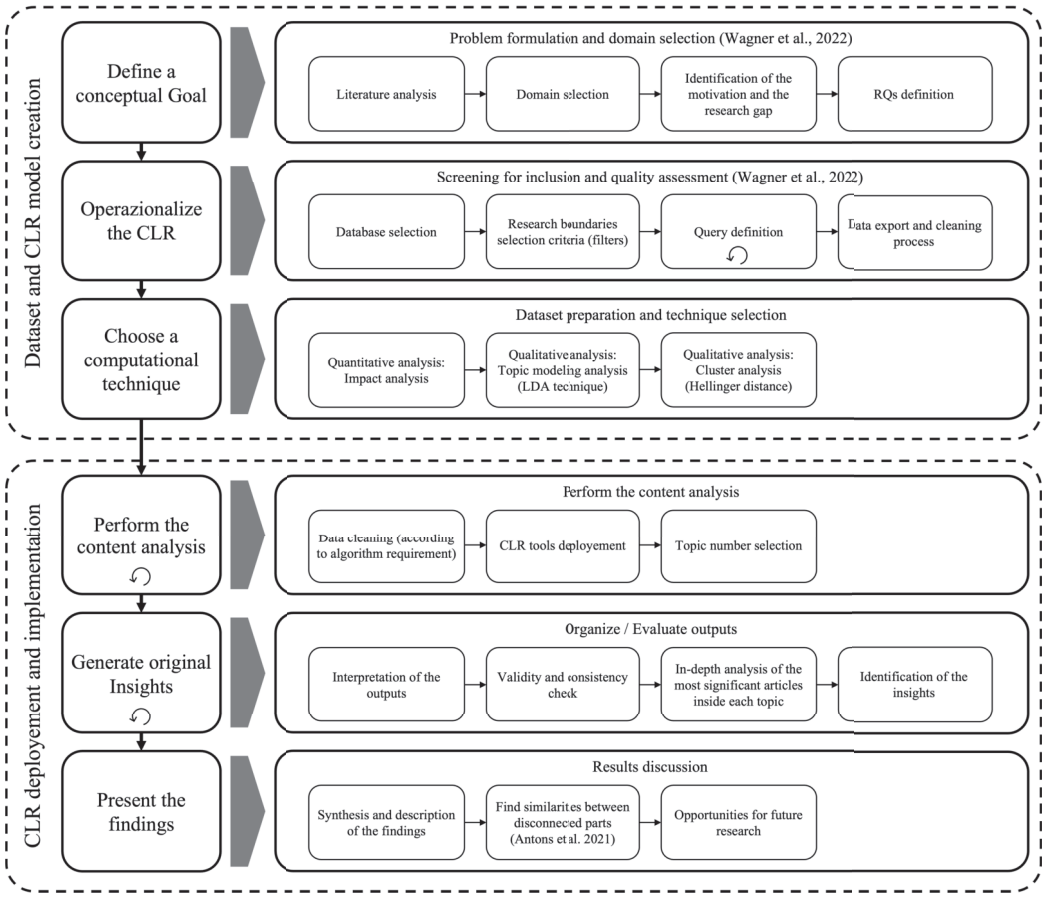

Figure 1. Research protocol framework

Results

In the subsequent paragraphs, we present the analyses performed on the dataset. First, we provide a quantitative descriptive analysis of the bibliometric data along with an impact analysis, examining the influence of the journals under consideration. Following this, we conduct a content analysis to explore the various topics related to our study.

Descriptive analysis of bibliometric data

The analysis was conducted in February 2022. The dataset comprises 1175 articles, reduced to 1102 after the data refining process, published in 86 sources with an average citation per document of 30,9. The oldest article was published in 1983. According to Figure 2, since the first article, annual publications in the journals under review have rarely exceeded 25 articles, with the only exception in 1994. In 2012 and 2013, there was an increase in publications to over 75 articles annually, but the trend settled once again between 25 and 50 articles in the following year. Finally, in 2021 there was a very significant increase in the number of published contributions to over 200, corresponding to 17.9% of the total number of articles investigated. The graph shows that this trend will likely continue in 2022, as the dataset was created in mid-February 2022, and the articles are only slightly smaller than the total in 2020 and 2019.

By looking at Figure 3, we can assess that Artificial Intelligence in organisation studies is a topic with worldwide relevance. The graph shows the most productive countries, according to the author's affiliation, and distinguishes between papers where the authors are from the same country (Single Country Publication - SCP) and papers whose authors come from different countries (Multiple Country Publication - MCP). By looking at the graph, the debate on Artificial Intelligence inside organisation studies is mainly developed in the USA. In fact, they have a greater number of publications concerning the other countries (323 papers). The United States is followed by the UK (74 papers) and China (64 papers). All the most productive countries have both single and multiple collaborations. It is interesting to note that China and Iran are the only countries where MCP are significantly higher than SCP, and China is also the second most collaborative region (38 papers) after the USA (42 papers).

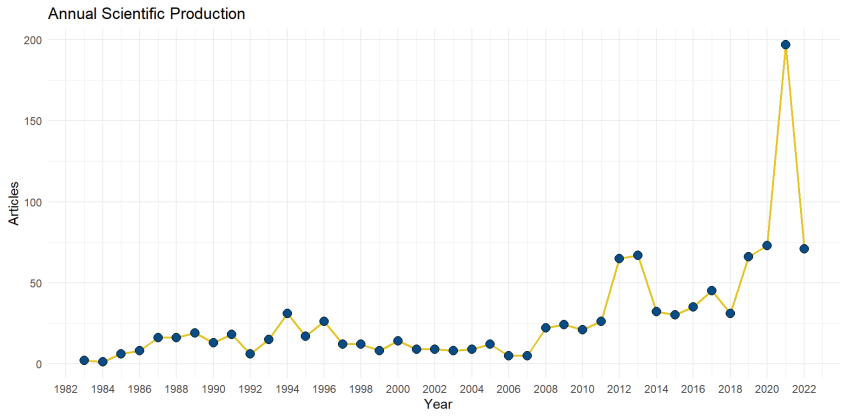

Figure 2. Number of publications per year since 1983

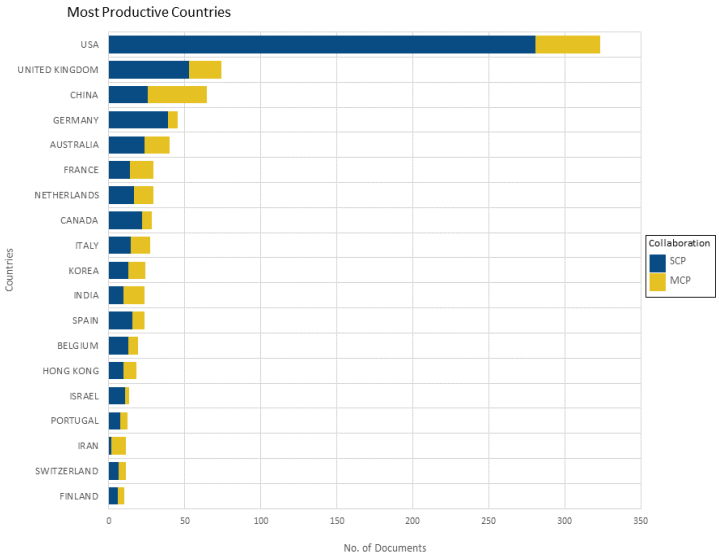

Figure 3. Most productive countries

Figure 4 represents the authors' production over time. The dot size expresses the number of published articles, while the intensity of the colour indicates the total citations per year. Over time, the two most productive authors were Chen H. and Delen D., with 8 and 7 articles. Chen H. published from 1994 to 2010, while Delen D. from 2011 to 2022. Chen H., Benbasat I. and Delen D. are also the authors that have the highest impact within the dataset, with an h-index of 8, 6 and 5, respectively. Benbasat I and Chen H. are also the authors with the highest citation count (956 and 941). Chen H. and Delen D. are also the authors that published the articles with the highest impact (659 and 384 citations). Both papers were published in the journal Decision Support System in 2004 and 2013. It is also interesting to note that, with the exception of Chen H., the articles that

have had the most influence on the literature have been published in the last decade. Finally, similar to what is represented in Figure 2, the results show how the topic interest has grown over time, especially in the previous two years.

Studies on the application of artificial intelligence-based technologies in organisational settings are published in various journals. In our search, we selected the best sources within the AJG ranking closely related to organisation studies. Table 2 highlights the journal with the highest number of publications concerning the research domain examined, the relative AJG ranking, and the AJG field.

The most prolific journals are Decision Support Systems, with 362 out of the 1102 articles and Journal of the American Society for Information Science and Technology (JASIST), with 93 papers. The first four sources account for about 55% of the total number of documents, all belong to the Information Management branch, and all have a ranking of 3. Looking at the field of study and the journal ranking, we can see that most papers (770) belong to the Information Management category. The ranking of the ten most impactful journals also includes the Journal of Business Research, Journal of Environmental Management and Risk Analysis, which belong to the categories of General Management, Ethics, Gender and Social Responsibility and Social Sciences.

Of the 770 articles belonging to information management, 60 were published in journals with a rating of 4*, 51 in journals with a rating of 4 and 659 in journals with a rating of 3. It is very interesting to note that the second most prolific category is social science, with 115 papers, of which two were published in journals rated 4*, 32 in journals rated 4 and 81 in journals rated 3. Moreover, most of the papers in this field have been published in the last four years. The results could suggest that academics increasingly inspect AI's social and ethical implications inside organisational settings. This trend may be explained by several factors, such as the change in humans' perception of AI-related technologies.

Table 2. Most recurring journals

Sources	Articles	AJG Rating	AJG Field of Study
Decision Support Systems	362	3	Info. Man.
Journal of the American Society for Information Science and Technology	93	3	Info. Man.
Information and Management	84	3	Info. Man.
Information Systems Frontiers	66	3	Info. Man.
Journal of Business Research	57	3	Gen. Man., Ethics, Gender and Soc. Resp.

Table 2. (Continued)

Information Systems Research	39	4*	Info. Man.
Journal of Environmental Management	31	3	Social Sciences
Risk Analysis	27	4	Social Sciences
Government Information Quarterly	23	3	Info. Man.
Journal of Management Information Systems	21	4	Info. Man.

They are moving from being seen as mere objects to being perceived as digital beings (Dignum, 2018). This phenomenon paves the way to new research frontiers in which AI is compared to human beings. Another factor may be the acceleration of Digital Transformation in response to the COVID-19 pandemic, which increased positive perceptions about digital work (Nagel, 2020). This last point could explain why the number of publications has also increased in the fields of General Management, Ethics, Gender and Social Responsibility, Human Resource Management, and Employment Studies.

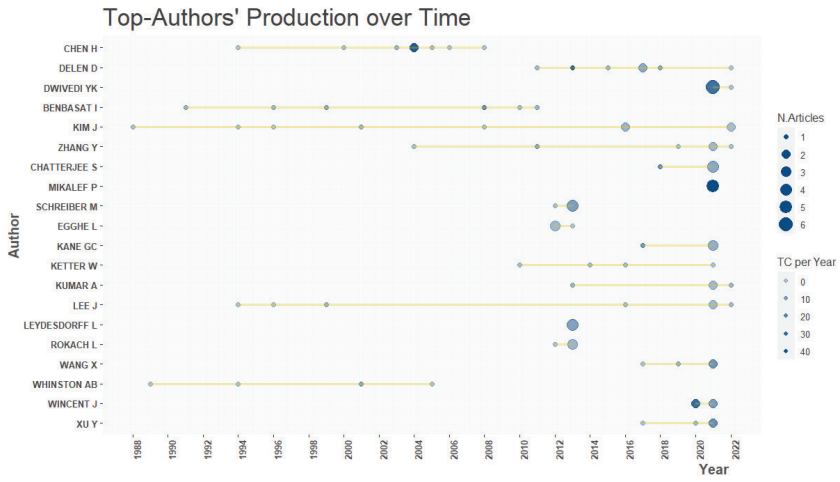

Figure 4. Most productive authors

Topics exploration with co-word analysis

In this section, we explore the most relevant topics in our dataset by analysing the most recurring keywords used by authors. The analysis could provide insights into the most related topics to our research domain and their contribution to the literature. We divided the topic analysis into three parts:

1) The trend of the most used keywords.
2) Authors' keywords co-occurrences.
3) Co-words analysis by using the thematic map to draw clusters of keywords based on the centrality and density metrics.

The analysis started with the distribution of the most used author keyword with at least ten occurrences in the dataset (see Figure 5).

After artificial intelligence, decision support systems are the most used keyword in the dataset, with 217 occurrences, followed by machine learning, with 97 occurrences. Expert systems, big data, decision support, digital transformation, big data analytics, neural networks and text mining complete the top 10 of the most used keywords with more than 20 occurrences (39, 26, 26, 23, 21, 20, 20 respectively).

By looking at the frequency of the ten most used keywords, it can be seen that most of them were used mostly in the last four years. For instance, the use of the keywords artificial intelligence, machine learning, neural network and text mining increased from 2019. The only keywords that go against the trend are decision support systems, expert systems and decision support. They were used the most in the period 2007 - 2013, 1985 - 1995 and 2010 - 2017 respectively. Finally, it is no surprise that big data, digital transformation and big data analytics are the most recent keywords. They have started to be used in the last decade.

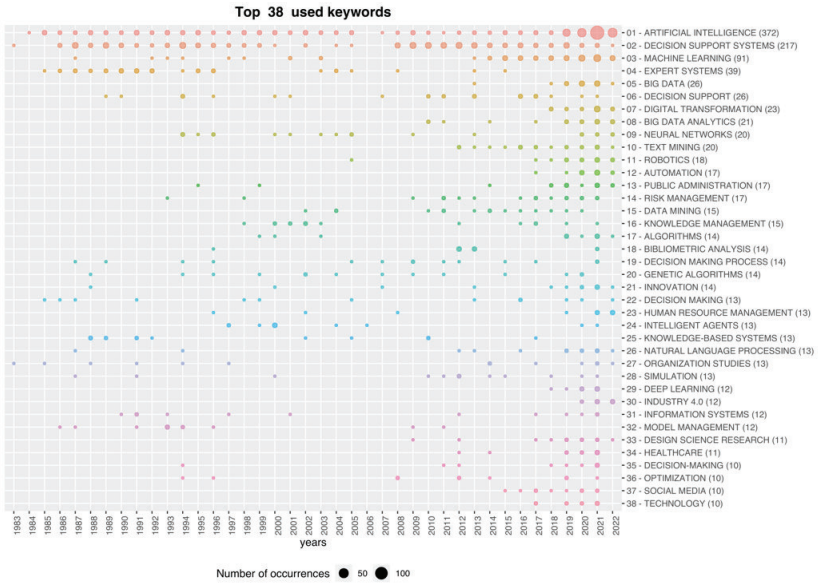

Figure 5. Main keyword used in the dataset over time

For what concern the other 28 keywords, the frequency appears regular and in line with the annual production, with the majority of occurrences in the last four years. There are some exceptions, for instance, intelligent agents, model management and knowledge management, that were more used in the mid-1990s and early 2000s. The most recent keyword among the most used is industry 4.0 first appeared in 2020.

The co-occurrence network (see Figure 6) was created from the keyword with more than ten occurrences. The nodes within the graph represent keywords that are connected to each other if they are both used in the same document (co-occurrence). The thickness of the ties reflects the number of papers in which the keyword pair appears. By looking at the graph, we can see the presence of three different clusters strictly related between them:

- The blue cluster focus on artificial intelligence, the settings where AI is adopted and the related technologies.
- The yellow cluster refers to the decision support systems, where they are used and the systems in which they are implemented.
- The red cluster contains many technologies related to machine learning and two settings where ML algorithms are broadly used, such as social media and healthcare sectors.

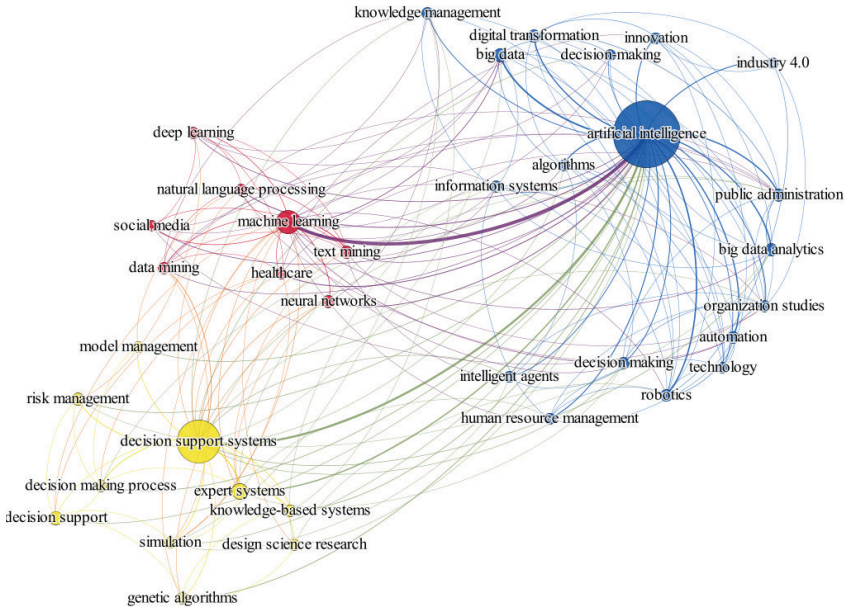

Figure 6. Authors' keyword co-occurrence

Impact analysis

This study is useful to better identify the impact, in terms of citations, of the articles, authors and journals analysed. After the data cleaning process, the analysis was carried out on 1148 articles, only 105 of them (88 out of 105 were published between 2021 and 2022) have zero citations. This can be interpreted as a sign of quality with regard to the examined papers. In the appendix, we show the 20 most cited articles. They were mainly published in the late 1990s and early 2000s. The most cited article is published in Information Systems Research and is focused on studying how the alignment between business strategy and information systems could improve business performance (Sabherwal & Chan, 2001). The exceptions are represented by some recent contributions, published in the middle of the last decade, dealing with topics that are still relevant today, such as the use of simulations within the supply chain (Tako & Robinson, 2012), the use of big data analytics for decision support systems and predictive policing (Demirkan & Delen, 2013; Gerber, 2014). One of the newest articles that belong to the top 20 of the most cited contributions is an introduction of a special issue of California Management Review by Haenlein & Kaplan, (2019) in which they give a definition of AI (the same given at the beginning of the article) and summarise some of the articles of the issue.

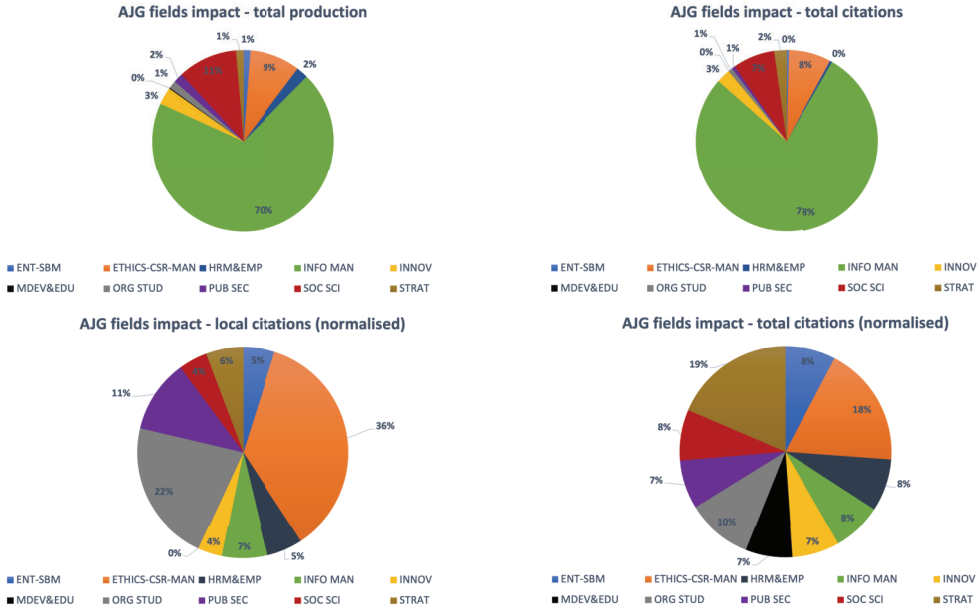

Figure 7. Impact of the AJG fields for production and citation (total, normalized, and local normalized)

On the upper side, Figure 7 graphically represents the distribution of the number of articles (left) and the number of citations (right) between the selected AJG fields. The two graphs express a similar path concerning

the field Information Management which accounts for respectively 70% of the articles in the dataset and 80% of the total citations. The fields Social Sciences and General Management, Ethics, Gender and Social Responsibility are the second most productive and impactful fields. The first one has more articles, but the same percentage of citations with respect to the second one. All AJG fields (excluding Information Management) have the percentage of 'productivity' articles higher than the percentage of 'impact' citations. Only the Innovation and Strategy fields reverse this trend, which could represent the presence of impactful articles within these research areas. The lower side of Figure 7 illustrates the impact of the AJG areas concerning the total and the local citations of the document normalised. The normalisation was done by calculating the average citations per year of each document, summing the citations of the document that belonged to a specific AJG field, and dividing them by the total number of articles in that field (mean of the average citations per year). Results show a considerable difference with the upper graph. Concerning the total citations of the documents, after the normalization the three most impactful fields are Strategy, General Management, Ethics, Gender and Social Responsibility, and Organizational Studies. This graph not only shows the support and the impact of all the AJG fields under exam in the AI debate but also underlines the importance of the fields dealing with social, organisational and ethical studies within the domain. Although the three most impactful fields in the lower right graph have only developed in recent years and have a limited number of articles, they reached a significant relevance in the AI domain, a sign that there is increased attention to issues concerning the effect and implications that AI can have on humans and organisations. Moreover, the graph underlines the impact of the field Strategy, one of the few that has a percentage of productivity lower than the percentage of citations. In Figure 7, we also analyse the AJG field's impact concerning the local citations (documents cited from the articles in the dataset) normalised. Results show the considerable relevance of the field of General Management, Ethics, Gender and Social Responsibility, followed by Organizational Studies and Public Sector. These three fields contribute in a significant way to the development of the discussion concerning AI inside business and management studies, as they are the most cited from the top journals of the AJG ranking contained in our dataset.

We have also represented the top 20 journals in terms of number of citations and impact (the ratio of a journal's total citations to its number of publications within the analysed data). Not surprisingly, most of the journals with the highest number of citations are specialized in Information Systems (IS). The majority of them

belong to the information management field of the AJG ranking. Interestingly, the journal with the highest number of citations (almost four times as many as the second) is Decision Support System (see Table 3). This highlights the importance of AI when it comes to decision-making support, especially in light of the technological development in recent years. In fact, we have entered the cognitive generation of decision support, which implements, within the platforms, all the new technologies that are positioned within the digital transformation, such as AI (Hugh J. Watson, 2017). Lastly, Information Systems Research, although it has only 39 articles, is the third most cited journal (3468 citations). This trend is in line with the relevance of this journal within the information systems field. As regards the journals with the highest impact, we noticed a greater heterogeneity of research domains. They range from the AJG field of social sciences to strategy. However, the Information Management component remains of paramount importance.

Table 3. Top 20 Journals ranked by citation count and AJG Area

Rank	Journal	Cited	AJG Area
1	Decision Support Systems	14476	Information Management
2	J. of the American Soc. for Inf. Sc. & Tech.	4434	Information Management
3	Information Systems Research	3468	Information Management
4	Information and Management	1600	Information Management
5	Journal of Management Inf. Systems	854	Information Management
6	Journal of Business Research	716	General Man., Ethics, Soc. Resp.
7	California Management Review	680	General Man., Ethics, Soc. Resp.
8	Government Information Quarterly	577	Information Management
9	Journal of Environmental Management	570	Social Science
10	Information Systems Frontiers	508	Information Management
11	Environmental Science and Technology	500	Social Science
12	Risk Analysis	488	Social Science
13	Industrial and Corporate Change	460	Social Science
14	Research Policy	456	Innovation
15	Strategic Management Journal	454	Strategy
16	Journal of the Ass. for Information Systems	421	Information Management
17	Technovation	383	Innovation
18	Long Range Planning	301	Strategy
19	Information and Organization	198	Information Management
20	Journal of Strategic Information Systems	166	Information Management

Content analysis

The analysis was performed by constructing a topic model of the articles' abstracts. After the data cleaning process, we determined the number of topics (K) for our model. In order to select the proper number of topics, we relied on the computation of the coherence score and on a manual analysis. Figure 8 graphically represents the coherence score of our topic modelling analysis. Based on the score, we selected 15 topics that were manually revised to assess whether they were consistent with the words contained in them and whether two or

more, similar topics could be merged. We then named the topics and carried out some descriptive statistics to better describe them.

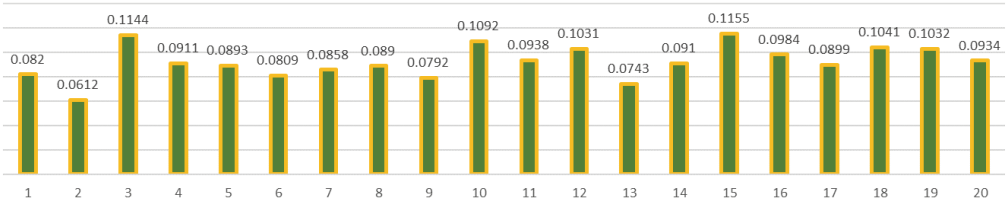

Figure 8. Topic Coherence Score

For assigning the name to each topic, we considered the 15 most frequent keywords characterizing each topic and the corresponding most relevant articles. The 15 topics are related to the different research areas in AI reviews, and they can be summarized as follows: 1) Societal Impact of Artificial Intelligence, 2) Data Intelligence and Accessibility, 3) AI Implications in Public Services, 4) DSS/Human Relationship, 5) Expert Systems Design and Development, 6) Project Evaluation, 7) IS Development Models, 8) Machine Learning Techniques, 9) Environmental Studies, 10) Business Innovation towards AI, 11) Product/Service Improvement, 12) Research Evaluation, 13) Decision Support System Development, 14) AI Influence and Acceptance, 15) AI Implications in Management.

The clusters related to the topics found are uniformly populated by a similar number of papers (m = 76.53, sd = 17.15). The only outlier is represented by topic 5 (Expert Systems Design and Development), which has 123 articles. For each topic, we represented the number of publications between 1983 and 2022 to determine the most recent trends in academic research. At the same time, we inspected the most important journals inside each topic to see which AJG research field is predominant. Not surprisingly, the topics that have had the highest number of publications in recent years such as Societal Impact of Artificial Intelligence, Environmental Studies, Business Innovation towards AI, AI Influence and Acceptance and AI Implications in Management, reflect many subjects considered of paramount importance by various companies in light of the digital revolution. Another area that has seen an article increase in recent years is the AI Implications in Public Services (topic 3), and this research field had another peak in the late 1980s coinciding with performance management reform in the public sector (West & Blackman, 2015). Finally, the topic Machine Learning Techniques has several peaks throughout the inspected period, but in the last period more articles have been published on the subject. As for the journals contained in each topic cluster (m = 18.6, sd = 9.96), it is not

surprising that for almost every topic the most important AJG research field was Information Management. Topic 10 - Business Innovation towards AI and Topic 9 – Environmental Studies were the only exceptions, the former with nearly 50% of the journals related to Human Resource Management and Employment Studies and General Management, Ethics, Gender and Social Responsibility, and the latter with 86% of the journals related to Social Sciences. Business Innovation towards AI, AI Implications in Public Services and AI Implications in Management were also the topics with the highest number of journals, i.e. 32, 35 and 39, respectively. Many topics focusing on a specific argument were represented by the relevant AJG field with a significant percentage. For example, in topic 3 – AI Implications in Public Services, 22% of the articles refer to the AJG Public Sector field, in topic 6 – Project Evaluation, topic 11 – Product/Service Improvement, topic 14 – AI Influence and Acceptance and topic 15 – AI Implications in Management, a high percentage of the AJG research domains is represented by General Management, Ethics, Gender and Social Responsibility. Overall, the most represented topics in the AJG field are Information Management, with 808 articles, Social Sciences with 118 articles and General Management, Ethics, Gender and Social Responsibility with 96 papers (see Figure 10). It is also very interesting to note that until 2018 almost all articles published on AI belonged to the AJG Information Management research area. Since 2019, however, research has also increased considerably in other areas and especially in journals belonging to the areas of General Management, Ethics, Gender and Social Responsibility and Social Sciences (see Figure 9).

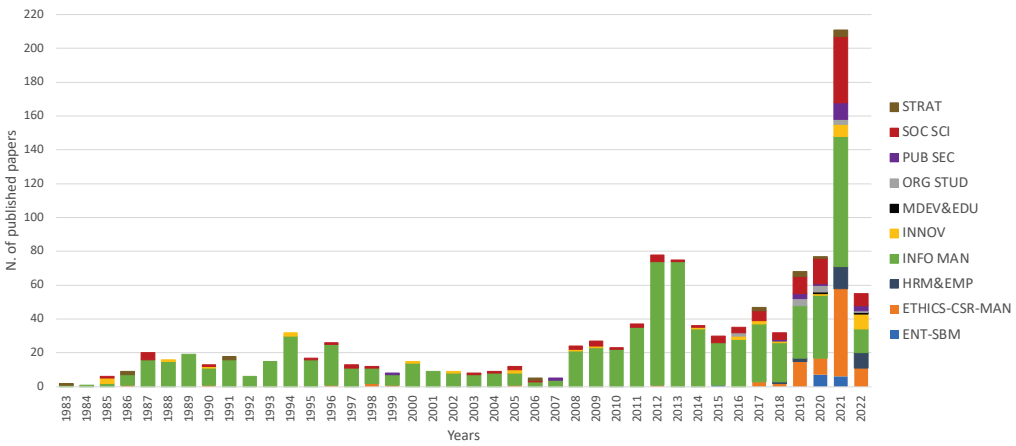

Figure 9. Number of publications per year since 1983 divided by AJG fields (see research method paragraph for AJG Area correspondence)

Discussion

After performing impact and content analysis, we decided to go one step further by specifically studying the content of each topic. We believe that a more in-depth manual study of the literature in each cluster helps to better understand the results of the topic modelling analysis and provides better insights. In our research, we created a table in which, according to the score assigned by the algorithm to each document, we entered the most significant papers for each topic, we then downloaded the entire text corpus of the articles and proceeded to a deeper analysis to better identify the topics covered for each cluster. This further study allowed us to refine the titles and extend our knowledge of each topic. We therefore give a brief description of each topic in the following lines, citing the most significant sources for each of them.

Topic 1 – Societal Impact of Artificial Intelligence: It primarily discusses the relationship between our social structure and new AI-based technologies, as well as the possible consequences of their massive adoption in our society. The consequences are not always positive; several ethical, social and financial problems may arise. The risks include the development of digital beings that do not respond to the moral laws of civil society (Mulgan, 2019), the problems of acceptance of new technologies by users accustomed to using old ones, and the role that governmental structures will play in this transition (Lundvall, 2017). The topic also discusses the use of AI to improve the reliability and user experience of various communication tools such as forums and blogs (Mousavi et al., 2020; Zach, 2012).

Topic 2 – Data Intelligence and Accessibility: Attention is paid to the use of search and clustering tools to improve user experience (UX). In particular, we mention two case studies on the development of multilingual search tools able to facilitate access to business intelligence information by non-English speakers (Chung et al., 2004; Jialun Qin, Yilu Zhou, 2006), and a case concerning the development of new phrase clustering methods for document classification (Xiaoyan Cai, 2011). Elizabeth Foss et al., (2013), on the other hand, focus on the Internet search habits of a group of young Americans to make recommendations to designers, educators and researchers to better educate young people in research.

Topic 3 – AI Implications in Public Services: The focus is on the increasing introduction of AI tools in public administration and its results. Vogl et al., (2020) state that the adoption of AI tools within United Kingdom local authorities, on the one hand, automates some processes, but on the other hand, creates a new form of relationship with administrators called algorithmic bureaucracy in which they learn from each other.

Wilson, (2022) conducted an analysis of AI implementation strategies in 16 countries, and his results denote a lack of attention towards public engagement which is important to counterbalance the ethical and social risks related to AI. Af Malmborg & Trondal, (2021) studied the adoption processes of European Union (EU) AI policies by the Nordic countries, and their findings describe a high level of reception by the countries under examination. However, countries with more organisational capacity are found to adopt EU policies on AI in a more effective manner. Finally, Gao et al., (2021) inspect the Open Government Data (OGD) phenomenon by identifying four stages of development. In this research, the author identifies a further phase in which AI analytics and Internet of Things (IoT) will enable more value to be created from open data, therefore it will be important to study the design of infrastructures for data collection and analysis.

Topic 4 – DSS/Human Relationship: The distribution of papers within this topic has several peaks over the 40 years examined, indicating that the topic has always been of interest to researchers. The articles focus on the role of DSS in facilitating people's tasks. The most recent papers investigate the confirmation bias reducing the effect of DSS along with increased satisfaction in decision-making and the effect of interactive analytic dashboards on users' cognitive abilities and tasks (H. H. Huang et al., 2012; Nadj et al., 2020).

Topic 5 – Expert Systems Design and Development: The discussion focuses on the development and integration of expert systems, where AI is the basis of these systems that simulate human intelligence to perform support tasks and strategic planning activities (Beeri & Spiegler, 1996; Goul, 1987).

Topic 6—Project Evaluation: The most significant articles on this topic focus on using AI to improve risk management and project evaluation methods. Specifically, unsupervised statistical techniques such as principal component analysis are applied to improve project control methodologies (Colin et al., 2015). Hu et al., (2013) propose a framework for intelligent risk planning software that reduces project risks by predicting project outcomes.

Topic 7 – IS Development models: Different types of information systems and their underlying models are analysed. Pendharkar, (2007) examines the performance of distributed AI systems for production scheduling, while Laitila & Virtanen, (2022) provide two new approaches to the use of the ranked nodes method used in DSS systems.

Topic 8 – Machine Learning Techniques: this topic focuses on the development and use of Machine Learning Techniques. ML is used for multiple purposes, some examples of which include the use of algorithms

to develop DSS systems based on imbalanced datasets (Piri et al., 2018), the development of techniques for bankruptcy prediction (K. C. Lee et al., 1996; Smiti & Soui, 2020) and the study of various approaches to improve the accuracy of predicting the success of movies (K. Lee et al., 2018).

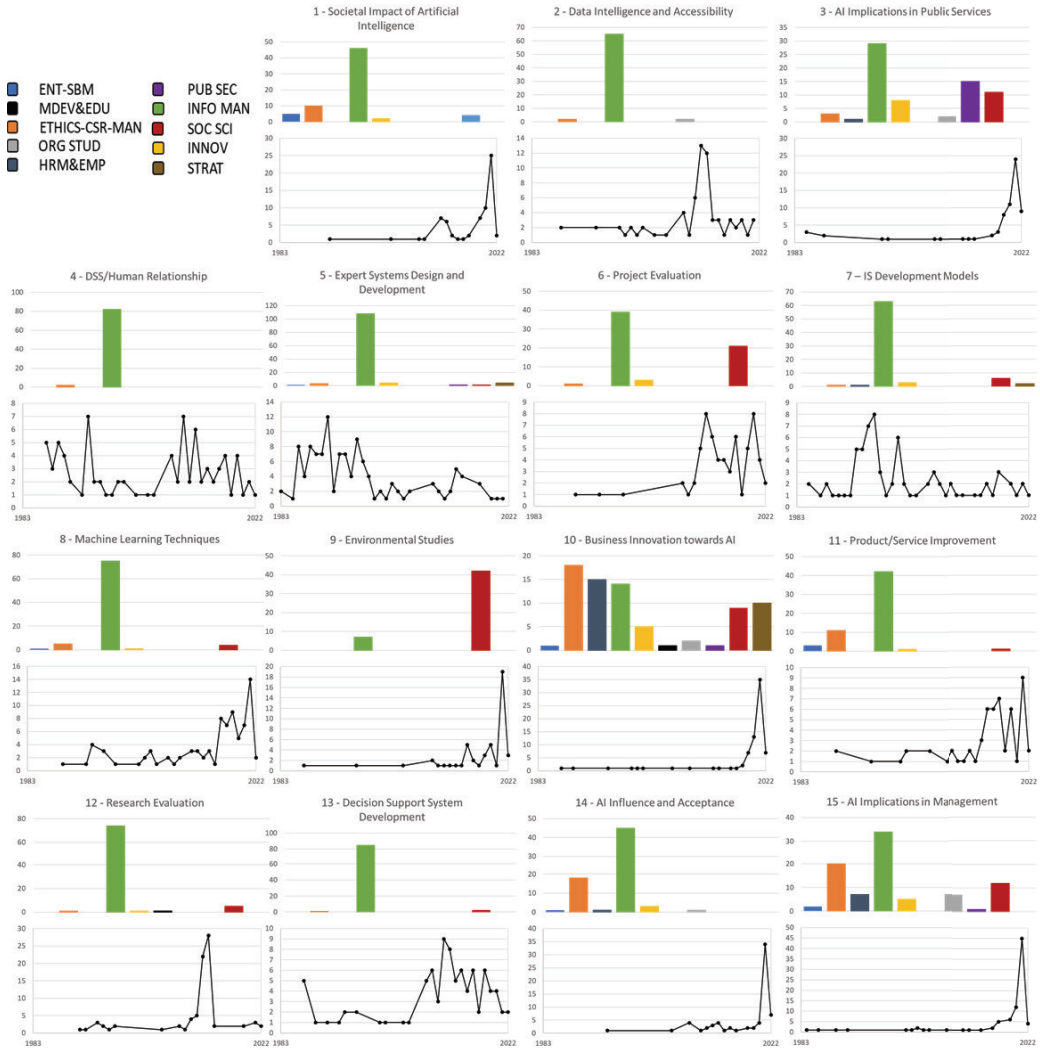

Figure 10. AJG areas inside each topic and articles distribution over time

Topic 9—Environmental Studies: This topic discusses the use of AI algorithms in environmental studies. It particularly discusses measuring the contamination of various chemical agents in water and air (Dashti et al., 2018; Elzain et al., 2021).

Topic 10 – Business Innovation towards AI: This topic looks at the impact of AI and related technologies on companies. Various areas are examined although, among the most significant articles, greater

attention is given to the implementation method of new technologies and the effects they have on workers. Sjödin et al., (2021) conducted an empirical study on six manufacturing companies engaged in Digital Servitization and were able to identify the key AI capabilities for the development of a business model. Felten et al., (2021) theorize and validate a new measure to study the effects of AI on employment and organizations. These measures can be used to identify the sectors that are most affected by AI or the effect it has on markets and organizations. Holm & Lorenz, (2021) studied the effect of AI on workers in Denmark and found a difference in impact depending on job position and AI function.

Topic 11 – Product/Service Improvement: This topic refers to factors that threaten or contribute to the improvement of a given product or service within a market. Grewal et al., (2021) give a framework to understand the advantages and disadvantages of AI within the B2B and B2C sectors. Wang et al., (2018) focus on the mobile app market to study the impact copycats have on original apps, suggesting a machine learning method to detect them. Lappas et al., (2016) study the effect that fake reviews have on online visibility within the hospitality industry.

Topic 12—Research Evaluation: This topic address particular research techniques, especially those regarding measurement method refinement. However, we were unable to find any explicit reference to AI within this topic. After a deeper analysis of the papers in this topic, we noticed that the articles were included in the dataset because the keyword "Artificial Intelligence" was listed among the indexed keywords that Scopus had assigned to each paper (indexed keywords are chosen by Scopus and are standardised to vocabularies derived from thesaurus).

Topic 13 – Decision Support System Development: These articles focus on the development of decision support systems to improve the effectiveness and efficiency of various tasks. Suzuki, (2009, 2012) develops a system that saves the cost of petrol for couriers by altering the route of trucks so that they can fill up at cheaper petrol stations, while Bandeira et al., (2009) develop a system for the efficient distribution of containers within facilities.

Topic 14 – AI Influence and Acceptance: The main discussion in this topic refers to the interaction between humans and AI-based technologies and, consequently, to the user's reaction. Cheng et al., (2021) study consumer response and trust in chatbots within e-commerce websites and identify attitudes towards them in different situations. McLean et al., (2021) focus on the attributes of AI voice assistants that most influence

consumer brand engagement, such as social presence and perceived intelligence. Upadhyay et al., (2021), on the other hand, add to the research on digital entrepreneurship by identifying factors that contribute to the entrepreneur's intention to embrace AI.

Topic 15 – AI Implications in Management: This topic discusses the impact that new technologies based on AI have within companies and the practices that management should adopt to improve the outcomes of digitization. Larson & DeChurch, (2020) identify 12 leader implications from digital technologies that affect teamwork organization. Johnson et al., (2021) discuss the possibility of task automation in management processes and its implications. Teodorescu et al., (2021) explore the debate regarding Human – ML augmentation, especially the concept of ML fairness to reduce automated decision bias. They argue that the IS research on the argument must be partially redesigned and expanded, thus opening new possibilities for IS researchers. Finally, Pereira et al., (2021) review the literature on the impact of AI in the workplace, contributing to the study of the effects of new technologies on human resource function.

Another objective of the content analysis was to assess the accuracy of topic modelling by detecting closely related topics. We constructed a dendrogram using Hellinger's Distance to determine the topic that could be merged based on the similarity of the probabilistic distributions. In light of the results derived from the in-depth examination of each topic, a very interesting result emerged, allowing us to identify two macro topics within our research. Looking at Figure 11, when cutting the dendrogram at the highest point, we note the presence of two clusters. The first one is composed of the following topics: 1) Societal Impact of Artificial Intelligence, 2) Data Intelligence and Accessibility, 3) AI Implications in Public Services, 10) Business Innovation towards AI, 11) Product/Service Improvement, 12) Research Evaluation, 14) AI Influence and Acceptance, 15) AI Implications in Management. The second one encompasses: 4) DSS/Human Relationship, 5) Expert Systems Design and Development, 6) Project Evaluation, 7) IS Development Models, 8) Machine Learning Techniques, 9) Environmental Studies, 13) Decision Support System Development. Looking at the contents of each topic, one can easily see that the first cluster mainly refers to societal studies. In fact, the various topics cover recurring themes such as the study of the impact that AI has on companies, public administration and society. Secondly, they investigate how AI-based technologies can be used to improve work at different levels of employment and to advance both new and widely used products and services. Finally, it is crucial to note that a major recurring topic, especially in topic 14, is the personification of AI and the study

of the human-AI relationship. The articles in the topics of the second cluster are mainly technical in nature. They focus on the study of algorithms and statistical techniques to improve AI artefacts. The only exception is Topic 4) DSS/Human Relationship, which is located in the second cluster due to its close similarity with Topic 13) Decision Support System Development. This allocation is due to the semantic similarity of the two topics. In particular, the recurring presence of common words biases the algorithm into placing the topic within the second cluster even though the main focus is on social studies. The two topics are, in fact, complementary and respectively investigate the effect that DSS have on their users and the development of new applications for the enhancement of operations.

Interestingly, most of the papers within the topics in the first cluster were published in recent years. This is in line with what is seen in Figure 9. Since 2019, research on AI has increasingly specialised in the social field, inspecting the impact and relationship between AI technology and humans. This phenomenon may be explained by several factors, such as the change in humans' perception of AI-related technologies. They are moving from being seen as mere objects to being perceived as digital beings (Dignum, 2018). This paves the way to new research frontiers where AI is compared to human beings. Another factor may be the acceleration of Digital Transformation in response to the Covid-19 pandemic, which increased positive perceptions about digital work (Nagel, 2020). This could be an explanation of why the publications in the Public Sector and Human Resource Management areas increased in 2021 (Figure 9).

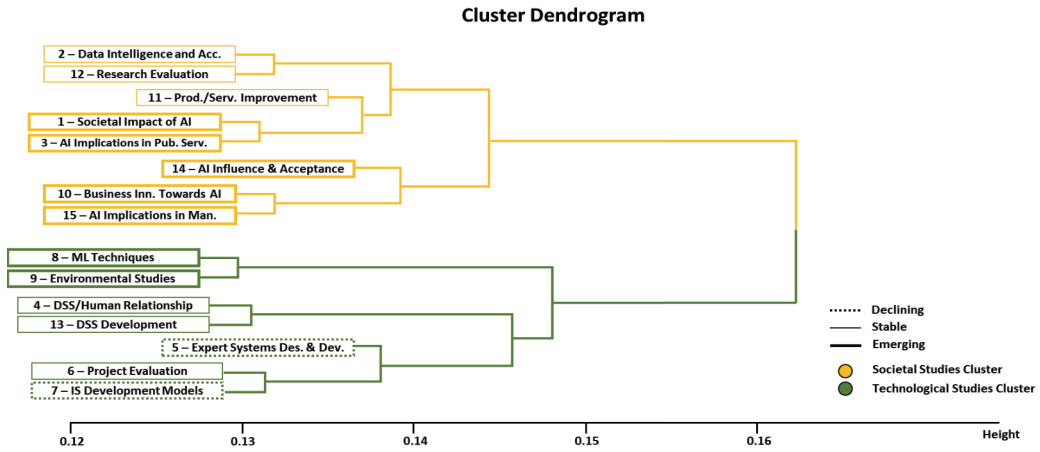

Figure 11. Topic affinity

In Figure 11 we highlighted the presence of the two macro clusters and, referring to the trend of publications over the years (Figure 10), we specified for each topic whether it was declining, stable over time or emerging.

We then focused on all the topics and for each of them, based on an in-depth analysis, we tried to identify possible research opportunities (Table 4). In line with what was mentioned previously, we can also note that the societal study cluster does not include any topics that have a declining publication trend. This seems to confirm that the societal research branch of AI is currently emerging within the business discussion.

The studies that we have taken into consideration to discover possible research opportunities reflect the results we have previously observed. We can distinguish inside them two main needs of further research: the first one concerns the continuous improvement of artificial intelligence algorithms to make them more accurate, reliable and unbiased, while the second focuses on the effects that this new technology can bring to the social sphere, with specific reference to the organizational context.

Among the possible research questions, we noticed that in many topics, belonging to both macro-clusters, the phenomenon of AI human augmentation is frequently observed. In particular, it might be relevant to focus on the relationship that would be created between a human and a machine when the latter is brought into the workplace and is assigned to work previously performed by humans (Pereira et al., 2021; Teodorescu et al., 2021; Upadhyay et al., 2021). In addition, many studies give importance to the consequences that the use of AI could have on users' cognitive abilities and knowledge, since too much reliance on this technology can lead towards a complete automation of decisions, causing users to lose the ability to make decisions (Nadj et al., 2020; Pereira et al., 2021). Within this strand of research, much attention is also paid to the problems of transparency and democracy of algorithms (C. D. Johnson et al., 2021; Teodorescu et al., 2021). These aspects, often caused by bias during the design of applications at various levels, can lead to serious problems of inequality and discrimination (Lundvall, 2017). Therefore, another possible aspect to be investigated lies in the need to make artificial intelligence algorithms more explainable and to investigate possible consequences (Alam & Mueller, 2022; C. D. Johnson et al., 2021; Vössing et al., 2022). Other studies also focus on the use of AI for the improvement of environmental and social sustainability. As these two topics have been much discussed within the literature in recent years (Nishant et al., 2020), it may be relevant to study the consequences of using AI for addressing societal and environmental issues (Elzain et al., 2021; Suzuki, 2012). Finally, some of the papers refer to the need for regulations together with precise guidelines for the development of applications based on AI algorithms (Cabitza et al., 2021; Vogl et al., 2020). Since culture differs depending on which part of the world is involved, it might be interesting to understand the efforts that

various regulators and policymakers are making in different parts of the globe to address this issue (af Malmborg & Trondal, 2021). In the following table, we seek to suggest possible future research directions for each identified topic.

Table 4. Topic description and possible guiding question for future research

Topic	Description	Future Research Questions
Topic 1 – Societal Impact of Artificial Intelligence	Impact that AI has in our society and possible future societal and individual consequences	What are the effects of new technologies on the world economic system and social inequality? What role does AI play in the development (or restriction) of human knowledge and cognitive capabilities? What are the ethical implications of using AI to moderate blogs and social networks? What are the effects on human behaviour and human cognition?
Topic 2 – Data Intelligence and Accessibility	Use of AI tools for UX improvement	Can AI give a degree of empowerment in designing digital artifacts compared to the past? What are the most affected factors? What is the impact that AI has in the development of software for content protection and education? What are the factors contributing to its successful implementation in this field?
Topic 3 – AI Implications in Public Services	Results on AI adoption in the public sector	What are the organizational and regulatory implications of adopting AI-based technologies in the public sector? Do national government strategies influence and foster AI adoption by companies?
Topic 4 – DSS/Human Relationship	Debate on user – DSS interaction	What is the impact and what are consequences that DSS based on artificial intelligence algorithms capable of autonomously developing reports have on workers' cognitive abilities and knowledge?
Topic 5 – Expert Systems Design and Development	Development and integration of expert system applications to support tasks and planning activities	During the digital era, an aspect of utmost importance concerned the correct implementation of an information system within an organization in order to have a competitive advantage. Nowadays, since many companies are adopting information systems based on AI algorithms, it is of paramount importance to compare the different case studies in the literature in order to understand: what are the factors that most contribute to the correct implementation and use of an AI application within an organization?
Topic 6 – Project Evaluation	The use of AI in project evaluation processes	What are the performances and reliability of an application based on deep learning algorithms designed for project evaluation under data overload and high-risk conditions? Which are the main issues concerning the role of project manager in dealing with AI project evaluation tools?
Topic 7 – IS Development models	AI model performance measurement and development	Today, AI is used as a support for the development of many software, what are the most relevant aspects and/or criticalities resulting from human-machine collaboration? Is the impact of AI-human collaboration limited only to the technical level or does it have an effect on behavioural or psychological spheres?
Topic 8 – Machine Learning Techniques	Development and testing of ML techniques and applications	Is it possible to define guidelines for adopting/developing ML algorithms in new and still unexplored organizational contexts? Since AI tools are becoming pervasive in organizations and society, how can we shape the concept of "AI ecosystem"? What are its main characteristics? Which aspects should we investigate and with which perspective?
Topic 9 – Environmental Studies	Development of AI applications for environmental issues	How can AI be applied to measure and prevent environmental phenomena? How can AI be used to improve simulation models of environmental phenomena, and what could its social and political impacts be?

Table 4. (Continued)

Topic 10 – Business Innovation Towards AI	AI impact inside companies and results of AI tools adoption	Since AI represents a breakthrough technology for organizations, how does AI affect IT values? How should we rethink theories and concepts related to IT values? How and to what extent do AI capabilities change depending on the industry in which companies operate? Does it always represent a disruptive technology, or could it sometimes be just a supportive or sustaining technology?
Topic 11 – Product/Service Improvement	AI used in organizational contexts to enhance products, services and practices	What are the consequences of using AI to detect and counteract opportunistic behaviour among competitors? What are the most important factors that enable organizations to use AI to engage with various stakeholders? What are the consequences related to the use of a more explainable AI for corporate communication to the outside world? Does it apply to every type of company?
Topic 13 – Decision Support System Development	AI used to improve DSS effectiveness and efficiency.	What role does AI play in improving decision-making processes to make business development more sustainable? Beyond the economic implications, which other aspects require further investigation, such as the social and the environmental ones? How could be possible to find a balance among the three sustainability pillars?
Topic 14 – AI Influence and Acceptance	Debate on AI/Human interaction process	What are the effects of the increasing use of AI applications on individuals from an organizational, societal, and psychological perspective? How is the Human/AI difference perceived in different work environments? What are the main variables affecting the perception of Human/AI differences? What could be the variables that influence (either positively or negatively) the adoption of AI applications by individuals?
Topic 15 – AI Implications in Management	Implications that AI implementation brings to management	What are the psychological and operational implications of hybrid human/AI teamwork? What aspects can make the adoption of AI more transparent, fair and democratic? What are the effects of AI human augmentation in organizational settings?

Conclusion

In this research, we performed a CLR to explore the AI discussion inside business and management studies. In parallel, thanks to the impact and content analysis, we also recognized in which journals and research areas the discussion has developed over the years. Adopting an algorithm for topic modelling, we identified 15 topics concerning the debate on AI in the corpus. We then confirmed our findings with an in-depth analysis of the most significant literature within each topic. The results show a trend in recent years where research is much more focused on the impact of AI on business, organizations and its societal and individual implications on different stakeholders. Based on topic similarity, it is also possible to distribute the 15 topics into two macro groups respectively representing the sociological and technical strands of research.

The contributions of the research are manifold. From a theoretical point of view, we provided an outlook of the discussion undertaken inside the most impactful journals of the AJG ranking, by describing descriptively and qualitatively the evolution of the underlying topic, their content and what are the main strands of research. These aspects could provide a resource, particularly for early-career academics, by highlighting

the most prominent theoretical frameworks applied in this field. This information is important also in understanding the variations of each topic and how they are evolving to address the unique challenges presented by AI in various contexts.

From a theoretical point of view, our study offers an overview of the discourse as reflected in the most influential journals listed in the AJG ranking. We have meticulously described both descriptively and qualitatively the evolution of the central themes, their content, and the main research strands within this domain. This detailed analysis serves as a resource, especially for early-career academics, by elucidating the predominant topic in the field. Such insights could be relevant not only to discerning the aspects of each topic but also for understanding their dynamic evolution in response to the unique challenges posed by AI across diverse contexts. Moreover, the CLR allows us to qualitatively analyse a large amount of data, providing several advantages compared with other approaches. First of all, a very large number of articles can be examined in terms of both content and impact. CLR also allows scholars to conduct research in a time-saving manner, as it is quite difficult to analyse thousands of documents manually in a limited amount of time. This method also enables a more efficient approach to the investigation of a specific research domain. The CLR method also allows multi-domain literature to be analysed in a rigorous and replicable manner. Finally, we hope that this article will help other researchers who intend to conduct a similar analysis by adopting the CLR method. Consequently, we believe it is important to describe each step in detail to make the research as replicable as possible. To this aim, we described in detail our proposed framework integrating the contribution provided by Antons et al., (2021) and Wagner et al., (2022), adding some phases that could improve the clarity of the research and the accuracy of the results. Additionally, following an iterative approach, we have developed a series of potential future research questions aimed at delineating the key challenges in the domain, especially focusing on socio-technical dimensions. These questions are intended to provide guidance for future studies, offering perspectives and insights for academics delving into this swiftly changing area.

From a managerial perspective, this research contributes to the identification of two distinct research strands and their publication trend. The study can offer insights to managers facing decisions on the adoption of AI-based technologies by guiding them through the most impactful research studies published in leading international journals so that they can be aware of its potentialities as well as some related criticalities. In addition, many societal studies (first macro-cluster), along with possible research opportunities, focus on the

ethical and psychological implications of AI adoption in organizations. These aspects are part of the Responsible AI phenomenon, which aims, among other things, at making the principles of AI adoption more transparent, fair and democratic. It is therefore of paramount importance that managers focus on the needs of internal and external stakeholders in their AI adoption agenda[9]. In addition, it must be considered that due to the dramatic increase in available data and the growth in computational power of technological devices, AI algorithms can perform more challenging tasks, being able to represent a phenomenon in an increasingly complex manner (Collins et al., 2021). However, as power and functionality grow, the potential problems represented by bias and discriminatory and violent sentences grow as well, representing a risk that should not be underestimated (Ntoutsi et al., 2020). A Time investigation recently reported that the company Open AI developer of Chat GPT, one of the most popular AI tools of the moment, outsourced the task of defining and labelling hate, violent and racist sentences to feed to the AI system in order to build a filter against this kind of contents. Like other companies in the field, such as Meta, Google, and Microsoft, Open AI approached Sama, a company that employs people from Uganda, Kenya, and India, with the aim to define inappropriate content and make digital tools safe for the end users[10]. As many public and private organisations are beginning to use AI applications, the decisions of such systems could influence many people, increasing the scope of potential problems related to social bias and ethical issues and the necessity to address them (Zuiderwijk et al., 2021). Nevertheless, many questions remain unanswered: “How do the customs and culture of people predisposed to labelling inappropriate content affect the outcome of AI applications?”, “Is there a proper way to train an AI-based application with values that can be shared across different populations?”, “It is possible to reach an understanding about a set of rules and policies that could regulate the design and the development of AI applications?”. Since AI reflects the values of its developers, it will be significant to address these potential issues in the future. Our hope is that this article will provide insight into the growing importance of social studies within the AI debate and raise awareness among managers and policymakers. Finally, we provide some insight about many critical aspects of AI implementation that could be addressed in future research.

[9] David Kiron, Elizabeth Renieris, and Steven Mills (2022), Why Top Management Should Focus on Responsible AI, MIT Sloan Management Review. Retrieved October 2023: https://sloanreview.mit.edu/article/why-top-management-should-focus-on-responsible-ai/.

[10] Billy Perrigo (2023), Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic, Time. Retrieved October 2023: https://time.com/6247678/openai-chatgpt-kenya-workers/.

Despite some similarities with bibliometric analysis, such as impact analysis and dataset description, CLR goes further and provides the ability to qualitatively analyse articles by giving information about their contents. However, this research has some limitations. First of all, the analysis is only based on abstracts and not on the entire content of each paper. Even though the purpose of the abstract is "to facilitate quick and accurate identification of the topic of published papers" (Peter Luhn, 1958), we argue that a more in-depth analysis of the contents of the text corpus can improve the final results by providing further insights. Another limitation concerns the breadth of the research, as we only examined a small portion of the publications on this topic. A further step could be to extend the search beyond the AJG ranking and the Scopus database and cross-reference other research databases such as Web of Science or Business Source Ultimate. Future studies might also focus on the use of different analysis techniques and tools that could lead to further insights and reduce algorithm biases. Nowadays, the proliferation of scientific literature increases the need for systematic, replicable and rigorous literature reviews, as well as the resources needed to conduct them (Badger et al., 2000). This study was the starting point of a research project, providing a first attempt to combine the rigorousness of literature reviews and the advantages of computational techniques. Finally, we used the CLR method as we needed to analyse a large number of articles in a semi-automatic way, trying to obtain, as a preliminary result, a qualitative evaluation of their content.

CHAPTER II

Artificial Intelligence Bias, Fairness and Ethics: A Computational Literature Review

Abstract. Due to the increasing adoption of AI technology in our society, this research aims to develop a complete overview of the current debate on artificial intelligence bias, fairness and ethics in organisational and business studies. To this end, we adopted the Computational Literature Review (CLR) method to conduct an impact and a topic modelling analysis of the relevant literature, using the Latent Dirichlet Allocation (LDA) technique. As a result, we identified and analysed 18 topics related to the selected domain. We further classified those topics into 5 categories creating a clear distinction between the social and the technical nature of a bias and its origins. Finally, focusing on the emerging topics, we proposed a set of guiding questions that might foster future research directions. This study provides insights to scholars and managers interested in AI bias and ethical issues and could be used also as a guide to perform CLR.

Keywords: Algorithmic Bias, AI Fairness, Computational Literature Review, AI Bias, Artificial Intelligence.

Introduction

In the organisational and business context, the implementation of AI technologies, along with their adoption, often leads to some form of resistance, frequently related to a lack of trust from employees and managers (Glikson & Woolley, 2020; M. H. Huang & Rust, 2018). Langer and König (2021) state that although trust in AI could be mainly related to its effectiveness and efficiency, it could not be confirmed when considering ethical issues. Such issues require careful attention to bias and fairness, and to this end, AI-based systems are being accordingly reconsidered in favour of new approaches (Ntoutsi et al., 2020). In this scenario, firms that design and implement these technologies must play an important part in addressing these issues in order to make AI more explainable and transparent (D'Cruz et al., 2022; Martin et al., 2019).

Due to the increasing adoption of AI technology in our society, the interest of academics has also increased, leading to new research, theories and questions (Makarius et al., 2020; Smith, 2020). Over the last four years, studies on the societal impact of AI have increased exponentially, with a specific focus on ethical implications and fairness (Smacchia & Za, 2022). Trust, transparency and explainability of AI are becoming more and more relevant to many stakeholders, and expertise in phenomena such as responsible AI will become

essential to everyone working in this field (Meghan Rimol, 2021). In addition, ethical aspects are not always applied during the development of AI applications, revealing a disparity between technological and ethical advancement. As a result, further research is needed on methods and tools to implement ethics into AI solutions design and development as the field still seems to be emerging (Haenlein et al., 2022; Stahl et al., 2016; Vakkuri et al., 2019).

In light of this, we aim to map and evaluate the organizational and business literature discussing AI bias, fairness and ethical implications (Tranfield et al., 2003). More specifically, the objective of the present document is to explore and describe the evolutionary trajectory of topics, as well as identify the journals and articles that are distinguished for their contributions and relevance. Furthermore, our objective extends to uncovering patterns within the literature that might be challenging to identify through the examination of a restricted subset of articles. For this reason, we conduct a comprehensive review of a broad selection of literature, in order to gain a holistic understanding of the phenomena, seeking to answer the following research questions:

RQ1: What are the most relevant themes that emerge from the current body of literature within the domain of AI bias, ethics and fairness? And what are the relationships between them?

RQ2: What are the topics related to AI bias, ethics and fairness that could foster future research directions?

To achieve this research objective, we decided to perform a computational literature review (CLR). This method allows us to conduct both quantitative and qualitative analysis across a large dataset by adopting an inductive (top-down) approach, to provide a comprehensive overview of the domain (J. Xie & Jain, 2024). CLR utilizes topic modelling to uncover latent themes by statistically analysing word usage patterns, thus aligning with our aim (Blei, 2012). We preferred this approach to other literature review approaches, such as bibliometric analysis, which predominantly employs quantitative methods to explore and investigate published research (Lamboglia et al., 2020). In this context, the CLR approach proved also to be more appropriate than a systematic review or a meta-synthesis, which, despite their rigour, are better suited for narrower inquiries and wouldn't be efficient for our analysis due to their intensive nature. Furthermore, CLR, with respect to other

methodologies implements text mining and machine learning algorithms, automating time-consuming tasks in the analysis and detection of significant patterns. This integration enhances the efficiency and effectiveness of extracting insights from voluminous data sets. (Antons et al., 2021). The following theoretical background section provides i) further details on AI bias, fairness and ethics, ii) a summary of previous literature reviews and iii) a description of the adopted research method. The presentation of the results, as well as their implications and discussion, closes the contribution.

Theoretical Background

The significance of concepts such as AI bias and unfairness has grown significantly in recent years due to the widespread adoption of this technology (Smacchia & Za, 2022). While AI offers numerous societal benefits, it is important to adopt an ethical approach in AI design and development and use it to mitigate potential negative outcomes due to different kinds of biases. Such an approach not only helps firms identify socially acceptable opportunities but also prevents serious mistakes (Floridi et al., 2018). AI bias, akin to the broader concept of bias, can be defined as a divergence in the outcomes generated by a particular system from the established ground truth, leading to disproportionate advantages or disadvantages for certain groups or individuals (Akter et al., 2021; Mehrabi et al., 2021). Although this definition is widely accepted in academic literature, the conceptualisation of AI bias presents challenges, particularly in certain fields. These challenges originate from different factors, such as historical, geographical, and ideological considerations, that complicate the establishment of a universally accepted ground truth (Brickell, 2006). Addressing the presence of biases in AI is a complex and challenging task. Bias can be present in various stages of the AI lifecycle and could derive from various sources, such as the training data, the algorithms used in the AI solutions, or even through the misuse of AI artefacts[11]. Furthermore, some kinds of biases are deeply embedded inside society and can be particularly difficult to identify and mitigate (Schwartz et al., 2022). These biases, regardless of their source, can lead to significant and sometimes critical issues. Instances of AI biases are extensively documented in both the press and academic literature. For example, algorithmic biases in recruitment software can lead to a preference for male candidates in traditionally men-dominated roles. Similarly, AI algorithms

[11] Michael Chui, James Manyika, and Mehdi Miremadi: What AI can and can't do (yet) for your business. Retrieved April, 2024 from: https://www.mckinsey.com/capabilities/quantumblack/our-insights/what-ai-can-and-cant-do-yet-for-your-business.

used for mortgage evaluations often result in higher interest rates for race minority applicants[12]. In parallel, biases embedded in historical data can cause serious issues in healthcare, where AI systems may deliver better services to white patients simply because they represent the majority of the data (J. Huang et al., 2022). In the judicial system as well, predictive policing algorithms can unfairly target specific races or genders based on historical crime data, leading to disproportionate enforcement against these groups (Richardson et al., 2019). While significant steps have been taken to mitigate bias and ensure algorithmic fairness, the rapid evolution of AI technology demands a proactive and continuous effort to develop and refine strategies for bias mitigation. It is important to enhance the accuracy and fairness of AI systems and ensure they deliver equitable benefits across diverse societal groups (Gichoya et al., 2023; Richardson et al., 2019).

Previous Literature Reviews

We examined the existing literature reviews on this research topic, summarised in Table 1. We performed a query on Scopus, Web of Science, and Google Scholar and selected all the literature reviews that were in line with AI ethics, bias, and fairness in organisational and business studies. Analysing previous literature reviews, we found out that they are systematic literature reviews based on a limited set of papers. Our aim is to give a broader overview of the domain, considering a larger dataset composed of more than a thousand articles, trying to investigate the topic using both a quantitative and a qualitative perspective. This approach could benefit both researchers and practitioners who are willing to have a broader overview of the most discussed topics, relevant sources, and most cited articles in the most impactful journals in business studies. Moreover, it could be useful to have a comprehensive overview of the patterns between topics and fields of study emerging from the literature that are difficult to find within a narrow dataset.

Table 1. Literature review comparison

Authors	Methodology	Focus	Findings
Meek et al., (2017)	Literature review of AI ethics contextualised into four categories using PEST tools.	Identify the central ethical issues related to AI and then isolate management recommendations by analysing those issues.	The authors provide a timeline of the most important milestones concerning Artificial Intelligence and a description of several ethical issues related to AI. In conclusion, they also make some recommendations to address AI ethical issues along with gaps for future studies.

[12] Mitra Best, AI bias is personal for me. It should be for you, too. Retrieved April, 2024: https://www.pwc.com/us/en/tech-effect/ai-analytics/artificial-intelligence-bias.html.

Table 1. (Continued)

Riazy et al., (2021)	Systematic lit. Review & case study 56 articles gathered from different databases published before 2021.	Fairness and explainability in learning analytics.	The authors identified methods to measure and mitigate discrimination in learning analytics. In the second part of the paper, they apply their findings to mitigate discrimination problems in an open dataset (students with disability).
Akter et al., (2021)	Systematic literature review and thematic analysis on 40 papers published between 2016 and 2020.	Sources of algorithmic biases in data-driven information (DDI).	The authors provide a framework to describe three major algorithmic biases in DDI phases together with guidelines to address these biases focusing on data, method and managerial capabilities.
Khan et al., (2022)	Systematic lit. review of 27 primary studies published before February 2021.	AI ethics principles and factors that could negatively impact the adoption of AI.	The authors recognised 22 ethical principles and 15 challenging factors concerning artificial intelligence.
Kordzadeh and Ghasemaghaei, (2022)	Systematic lit. review and thematic analysis of 56 papers published between 2010 and 2019 (scholarly journal and conference proceedings).	Algorithmic bias.	The authors discovered eight theoretical concepts concerning algorithmic biases and studied their relationships. They also pointed out that most studies conceptually discussed the ethical, legal and design implications of algorithmic bias, whereas only a limited number of studies empirically examined them.
van Giffen et al., (2022)	Systematic problem-centred literature review of 68 articles.	Machine learning biases in business, IS and marketing studies	The authors identified eight distinct ML biases and mapped them. They also proposed twenty-four bias mitigation methods and a conceptual model to illustrate the application of ML algorithms in marketing that helped them to analyse the biases in a case study.
Ashok et al., (2022)	Systematic lit. review and qualitative synthesis of 59 papers published in 43 journals between 2018 and 2021	Ethical use of AI in digital technologies	The authors found fourteen digital ethics implications associated with digital technologies archetypes. Then they mapped every archetype based on the presence of the ethical implications found.

Computational Literature Review

In this study, we use these guidelines as a baseline, adapting the framework proposed by Smacchia and Za (2022) to run our CLR (figure 1). For each of the six steps we provide additional information regarding the activities to carry out the analysis. Moreover, we highlight three tasks that require iterations to minimise possible bias. First, the definition of the query is a critical part of the study because the creation of the dataset to be analysed depends on it. Since this type of analysis is often conducted on a large number of articles, it becomes quite difficult to examine the consistency of each paper. It is, therefore, essential that the query is repeated several times, paying attention to the inclusion and exclusion parameters of the articles to avoid generating unnecessary noise within the dataset. Similar to the query definition part, the content analysis step also requires some testing before the parameters for analysis can be set correctly. In particular, the topics resulting from the analysis are generated and analysed several times to assess their relevance in the reference

domain. Finally, the interpretation of the results, which is conducted by the authors, is reiterated many times. The analysis is performed following a bottom-up approach, first studying individual topics and then repeating the analysis for other possible clusters of topics at a higher level.

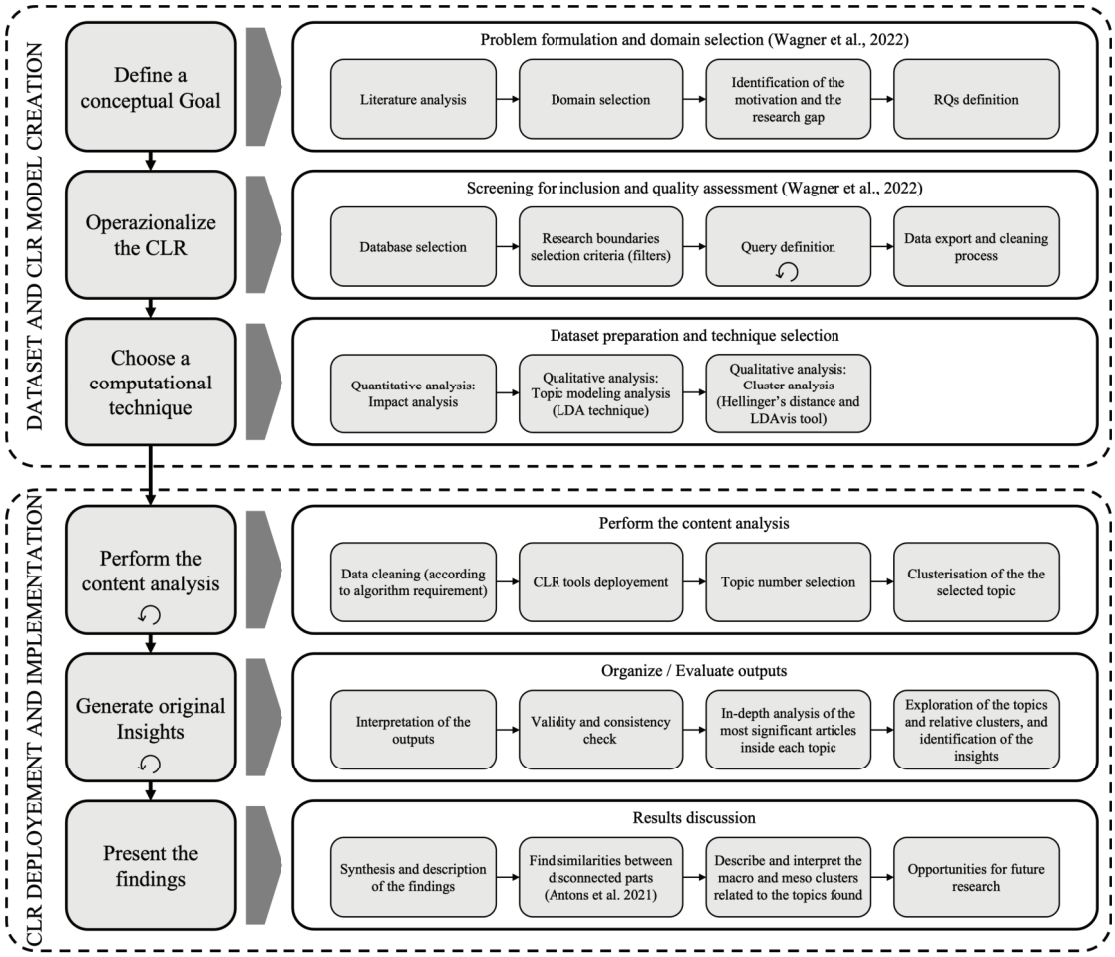

Figure 1. Research protocol adapted from Smacchia and Za, (2022)

Material and methods

The main purpose of this contribution, and thus our conceptual goal, is to explore and summarise the debate on AI bias, fairness and ethics in organisational and business studies. Hence, it was quite relevant to operationalise our literature review by identifying the appropriate list of journals to query in order to select papers to create our dataset. The Academic Journal Guide (AJG) of the Association of Business Schools (ABS) provides a list of journals classified according to the field of studies they belong to and their ranking. We included all the journals in the ranking. The data were collected within Scopus since all the journals we decided

to incorporate into the search were available in this academic database. Specifically, we performed a query in which we included all the chosen journals by entering their ISSN code and the following query:

TITLE-ABS ("deep learning" OR "artificial intelligence" OR "machine learning" OR "neural networks") OR AUTHKEY (“deep learning" OR "artificial intelligence" OR "machine learning" OR "neural networks")) AND (TITLE-ABS ("bias" OR "injustice" OR "ethics" OR "fairness" OR "trust") OR AUTHKEY ("bias" OR "injustice" OR "ethics" OR "fairness" OR "trust")) AND ISSN…

We chose not to include Scopus index keywords because they are often inaccurate and generate a lot of noise in the dataset. We did not use time or other restrictions. The query returned 1217 articles published between 1981 and 2022; citations and abstracts data were exported in CSV format. The downloaded dataset was also revised and cleaned. In particular, the occurrences lacking author information (errata articles as well) were deleted and, where possible, the papers with missing abstract were integrated. In the end, the analysis was conducted on 1198 articles. We then analysed our dataset using the R tool Bibliometrix for a preliminary description.

Our first task was to choose the computational technique to be used. We initially performed an impact analysis using the R programming language through the Bibliometrix package (Aria & Cuccurullo, 2017) to get quantitative data on the information contained in our dataset. Through this analysis, it is possible to estimate the impact of a given paper, author and journal. To measure these dimensions, the tool uses various metrics such as citation count, the impact factor (total citation count divided by the total number of papers) and the h-index which is commonly used to assess the impact of researchers (Hirsch, 2005). We then focused on the content analysis of the papers by performing topic modelling using the Latent Dirichlet Allocation analysis with the lda package. We adopted LDA because it is the most popular method used for topic modelling (Jelodar et al., 2019). Also, this method seems to have a higher level of reliability and accuracy compared to others (Shadikur Rahmane, 2020). LDA is an unsupervised generative probabilistic model of a corpus (Blei et al., 2003) that allows the identification of a set of topics among multiple documents. Each document is considered as a set of words that can be combined to form subsets of latent topics. The model assumes that the corpus includes k topics, and then distributes these topics across each document to see which one fits best. In this way,

the analysis is more efficient because it avoids cross-checking every word with every document. LDA algorithm can be summarised as follows:

$$P(\boldsymbol{W}, \boldsymbol{Z}, \theta, \varphi, \alpha, \beta) = \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{t=1}^{N} P(Z_{j,t}; \theta_j) P(W_{j,t}; \varphi_{z_{j,t}})$$

Where α represents the document-topic density (the percentage that a document is associated with a determined topic, if $\alpha < 1$ the documents tend to diverge from the single topics) β represents the topic-word density (the percentage that a word is associated with a determined topic) and θ with φ representing multinomial distributions of topics over documents and words over topics.

In order to perform the analysis, we had to clean our dataset. First, we deleted all the variables except for those related to abstracts and document IDs, then we removed punctuation and stop-words and tokenized the corpus to build the Document Term Matrix (DTM), which is a matrix that contains words and documents as dimensions. In the DTM rows correspond to the documents, and columns correspond to the terms, therefore it is essential to inspect the frequency of terms inside a document collection. We also removed all the terms in our query, along with other recurring words that might yield incorrect results, such as “Elsevier”, “Springer”, “Research” and “Findings” (Mortenson & Vidgen, 2016). When creating the DTM, we decided to tokenise the text using one or two words to avoid ambiguity during topic analysis. In fact, some words have a different meaning when taken individually (i.e. Information and System instead of Information System). Since LDA is an unsupervised technique that discovers patterns from untagged data, the number of topics to be used was chosen a priori. To assess the value of K we used the cross-validation procedure based on the perplexity that measures how well a probabilistic method can predict a sample (lower levels of perplexity can better predict a sample). To this end, we used the R package ldatuning, setting the algorithm for values from 1 to 100. To discover the value we used four metrics, CaoJuan2009 (Cao et al., 2009), Arun2010 (Arun et al., 2010), Griffiths2004 (Griffiths et al., 2004) and Deveaud2014 (Deveaud et al., 2014), where the first two has to be minimised, while the others has to be maximised. After the cross-validation procedure, we further analysed the results to assess the correctness and the significance of each topic and, hence, to validate the number “K” selected. According to the theta parameter (distribution of topics over documents) that the algorithm assigned to every document, we used topic modelling results to analyse the most representative contributions for each topic (the complete list is available at: https://bit.ly/3K1EbML). More specifically, the theta parameter

represents the probability that a document is contained in a certain topic. Hence, the documents selected were the most representative of each topic. Our goal was to discover how AI is implemented and deployed within different topics and also to detect the most popular fields of research in recent years. Figure 1 describes the research protocol we adopted to explore the literature using the CLR method. The part dealing with output interpretation and article in-depth analysis within the framework should follow an iterative cycle to have more precise and significant results. In fact, it is necessary to repeat the analysis several times to reach an output that is as relevant and reliable as possible. To find hidden patterns in the analysed literature we used the R package LDAvis (Sievert & Shirley, 2015). This interactive interface allows the visualisation of topics estimated by the LDA algorithm and it could be used to see the most associated keywords within each topic as well as a global distribution of these topics, including their similarities and differences.

Results

In this section are presented the results obtained from the impact analysis and the content analysis described in the methodology section.

Impact Analysis

The impact analysis was conducted to have information about the impact of the areas in which papers related to the selected domain are published. Moreover, thanks to the citation count and H-Index we assessed the impact of individual articles and the journals, gaining insight into the trend of publications and which field has acquired a relevant position in the debate. As regards article citations, only 230 (less than 20%) of them have zero citations (only 13 articles published before 2021) underlining the relevance of the papers in our dataset. The 20 most cited papers (the complete list is available at: https://bit.ly/3K1EbML) are distributed between 1998 and 2020 (10 articles published before 2016 and 10 articles published after). It can be noted that the oldest articles are represented mainly by technical studies with the aim of developing or enhancing AI algorithms. The majority of the newest studies adopt a social or organisational perspective investigating issues concerning the evolution of AI or the effect and/or the impact of AI adoption, suggesting a widening of the AI research stream integrating the technical issues with societal perspectives. We have also represented the top 20 journals in terms of number of citations with the H-Index, AJG Area and ranking (Table 2). Looking at the relevant AJG fields of the most cited journals we can observe a great variety of research domains, which is a

sign of a wide breadth of topics discussed concerning AI-related ethical issues, fairness and biases. Specifically, the most representative field in the ranking is Information Management followed by Operations Research and Management Science, Economics, Econometrics and Statistics and General Management, Ethics, Gender and Social Responsibility. Looking at both the AJG field and the source name, a distinction can be made between the journals that are focused on societal and organisational studies and those that focus more on technical and mathematical research. This confirms the same distinction recognised for the most representative articles.

Table 2. Top 20 Journals ranked by citation count and AJG Area

Journal	H-Index	Citations	AJG Area	AJG Rank
Annals Of Statistics	5	3326	Econ. and Statistics	4*
Expert Sys. with Applications	25	2719	Information Management	1
Computers in Hum. Behavior	18	1125	Information Management	2
Int. Journal of Inf. Man.	12	1119	Information Management	2
Ethics and Inf. Technology	16	944	Information Management	1
Journal of Cleaner Production	12	687	Tourism and Sector Studies	2
Journal of Service Man.	4	567	Tourism and Sector Studies	2
Euro. J. of Operational Res.	7	478	Op. Res. and Manag. Sci.	4
Journal Of Business Research	11	471	Man., Ethics and Soc. Resp.	3
Decision Support Systems	11	427	Information Management	3
Business Horizons	7	422	Man., Ethics and Soc. Resp	2
Reliability Eng. & Sys. Safety	4	410	Op. Res. and Manag. Sci.	3
Evolutionary Computation	5	407	Op. Res. and Manag. Sci.	3
Management Science	4	397	Op. Res. and Manag. Sci.	4*
J. of the Acad. of Market. Sci.	1	366	Marketing	4*
Econometrics Journal	2	360	Econ. and Statistics	3
Int. J. of Human Comp. Stud.	6	343	Information Management	2
IEEE Trans. on Evo. Comp.	7	342	Op. Res. and Manag. Sci.	4
Info. Processing and Manag.	9	270	Information Management	2
Psychological Review	3	261	Psychology (General)	4

In the impact analysis we also inspect the most prolific journals in order to assess their interest on the analysed topic (Table 3). Although there is an overlapping with the previous table, there are some differences that deserves attention. Concerning the number of contributions, the field Information Management is prominent with eleven journals contained within the top 20, followed by the fields Operations Research and Management Science and General Management, Ethics, Gender and Social Responsibility. With regard to the Information Management field, the majority of journals are ranked 2 and 1, while the only journals ranked 4 belong to Operations Research and Management Science. Overall, there is a wide variety of journals and fields interested in the topic, given the 417 sources that have published at least one paper concerning AI bias, ethics and fairness. The top 20 journals in Table 3 account for about 35% of the total publications with the Information Management field that is the most interested in the topic and also inspects it from both a technical and a societal

point of view. Similar to what was observed in the most cited journals, also in Table 3 other fields that are mainly involved in social, psychological and ethical studies are starting to investigate this phenomenon as well.

Table 3. Top 20 Journals ranked by production and AJG Area

Journal	Contributions	Citations	AJG Area	AJG Rank
Expert Sys. with Applications	91	2719	Information Management	1
Ethics and Inf. Technology	52	944	Information Management	1
Computers in Hum. Behavior	39	1125	Information Management	2
Info. Processing and Manag.	22	270	Information Management	2
Information Systems Frontiers	21	187	Information Management	3
Journal of Business Research	20	471	Man., Ethics and Soc. Resp.	3
Int. Journal of Inf. Man.	18	1119	Information Management	2
Journal of Business Ethics	16	219	Man., Ethics and Soc. Resp.	3
Decision Support Systems	15	427	Information Management	3
IEEE Trans. on Cybernetics	15	130	Op. Res. and Manag. Sci.	3
Journal of Cleaner Production	15	687	Tourism and Sector Studies	2
Euro. J. of Operational Res.	13	478	Op. Res. and Manag. Sci.	4
IEEE Trans. on Evo. Comp.	12	342	Op. Res. and Manag. Sci.	4
E. Commerce Res. and App.	11	163	Information Management	2
Expert Systems	11	53	Information Management	2
British J. of Edu. Technology	10	122	Information Management	2
Business Horizons	9	422	Man., Ethics and Soc. Resp.	2
Human Factors	9	23	Psychology (General)	3
Int. J. of Human Comp. Stud.	9	343	Information Management	2
Int. J. of Production Research	9	152	Ops. and Tech. Management	3

Focus on the Senior Scholars List of Premier Journals of AIS

In our analysis, we have also decided to focus on the most impactful articles of the Senior Scholars list of premier journals of the Association for Information Systems (AIS). We gathered all the documents belonging to the list and calculated the average citations per year (the total number of citations divided by years of publication). Since the journals on this list are considered the most relevant within the IS field, our aim is to give an overview of their contribution to the debate on AI ethics, bias and fairness. Table 4 contains the ten articles ordered by average citations per year. The most impactful one was published by the Journal of Management Information Systems in 2018. Moreover, half of the articles belong to the journal Decision Support Systems which is also the most prolific and cited venue along the AIS senior scholar list. Inside the table, there are also two contributions published by Information Systems Research, one belonging to MIS Quarterly and one to European Journal of Information Systems.

Table 4. Top 10 impactful articles of the Senior Scholars list of premier journals

Meso Topic	Authors	Title	Year	Journal	Avg. cit. per year
3	N. Kumar et al. (2018)	Detecting Review Manipulation on Online Platforms with Hierarchical Supervised Learning	2018	Journal of Management Information Systems	15
3	Eslami et al. (2018)	Which online reviews do consumers find most helpful? A multi-method investigation	2018	Decision Support Systems	14
3	Zhang et al. (2018)	A computational framework for understanding antecedents of guests' perceived trust towards hosts on Airbnb	2018	Decision Support Systems	12
4	B. Kim et al. 2020)	Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information	2020	Decision Support Systems	11
1	Zhang et al. (2020)	A text analytics framework for understanding the relationships among host self-description, trust perception and purchase behaviour on Airbnb	2020	Decision Support Systems	10
2	Teodorescu et al. (2021)	Failures of fairness in automation require a deeper understanding of human–ml augmentation	2021	MIS Quarterly: Management Information Systems	9
2	Geva and Zahavi (2014)	Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news	2014	Decision Support Systems	8
4	Ransbotham et al. (2016)	Special section introduction: Ubiquitous IT and digital vulnerabilities	2016	Information Systems Research	8
3	Schanke et al. (2021)	Estimating the impact of "humanizing" customer service chatbots	2021	Information Systems Research	7
4	Mikalef et al. (2022)	Thinking responsibly about responsible AI and 'the dark side' of AI	2022	European Journal of Information Systems	5

Inside the contributions taken into exam, a distinct duality can be detected. From one side, many articles discuss and investigate the dark side of AI, underlying how in recent times it is necessary to pay attention to unintended and dangerous consequences resulting from the implementation and use of an AI artefact. On the other side, AI techniques are used to address ethical and trust issues that could undermine people in various settings.

As AI's importance and use are increasing, so is the number of potential issues that could cause serious damage to individuals, organizations, and society. Aspects such as increasing decision-making autonomy, ubiquitous presence, lack of transparency, and increase of available data worsen the scenario if an AI artefact produces biased outcomes, because, for instance, it could affect an increasing people and could take into account sensitive data (Ransbotham et al., 2016). Moreover, it could be difficult to assess the presence of bias

and define mitigation strategies in many cases (e.g., there is no error in the data, but the identified issue is related to underlying societal behaviour or beliefs) (ibid). Due to the increasing attention to such issues, concepts such as responsible and explainable AI have become highly relevant in the IS field. It is considered of paramount importance to investigate from a socio-technical standpoint ways to make AI outcomes more transparent, fair, secure and accountable, trying to address possible issues (Mikalef et al., 2022). In this scenario, there are papers that focus more on how to reduce and mitigate bias from a technical point of view, proposing novel methods and algorithms that allow an increasing performance or understandability of the outcome (Geva & Zahavi, 2014; B. Kim et al., 2020). Other studies focus on inspecting the effect that AI has on people and organizations and proposing guidelines and suggestions to managers and regulators to address possible issues. Teodorescu et al. (2021) explore and explain how human-AI augmentation could contribute to enhancing fairness. To this end, they provide an explanation of criteria that are used to improve fairness in ML models. Afterwards, they build a matrix in which they recognise four different kinds of AI augmentation settings and for each of them provide the relevance for the IS field and suggestions to managers on how to address every situation. Schanke et al. (2021) focus on assessing the impact of anthropomorphism on AI chatbots. The authors conduct a field experiment with a clothing retailer company to see how people react when a chatbot is humanized. They also give recommendations on when it is appropriate to anthropomorphise a chatbot.

In Table 4 there are also some papers that use AI to mitigate ethical issues and inspect people's perceptions concerning different settings in online platforms. N. Kumar et al. (2018) The article is focused on developing a robust method for detecting online review manipulation. They use supervised learning techniques and combine univariate and multivariate distributions derived from the reviews metadata to build a model that could detect more effectively the presence of manipulation. The paper addresses an important topic for both consumers and firms, since fake reviews could represent a serious harm. Other contributions are focused on the use of AI to inspect people's attributes concerning different online platforms, such as the perceived trust of guests towards hosts on Airbnb (Zhang et al., 2018, 2020) and which characteristics of the online review have more impact on the perceived usefulness of the reviews (Eslami et al., 2018). In the discussion section we provide further explanation about these articles in light of our findings and give avenues for future research direction.

Content Analysis

After the data cleaning process, we performed a topic model analysis on the articles' abstracts. Since the LDA technique is an unsupervised learning algorithm, we had to determine the number of topics (K) a priori. Figure 2 graphically represents the distribution of topics (the y-axis indicates the level of perplexity, and the x-axis represents the number of topics).

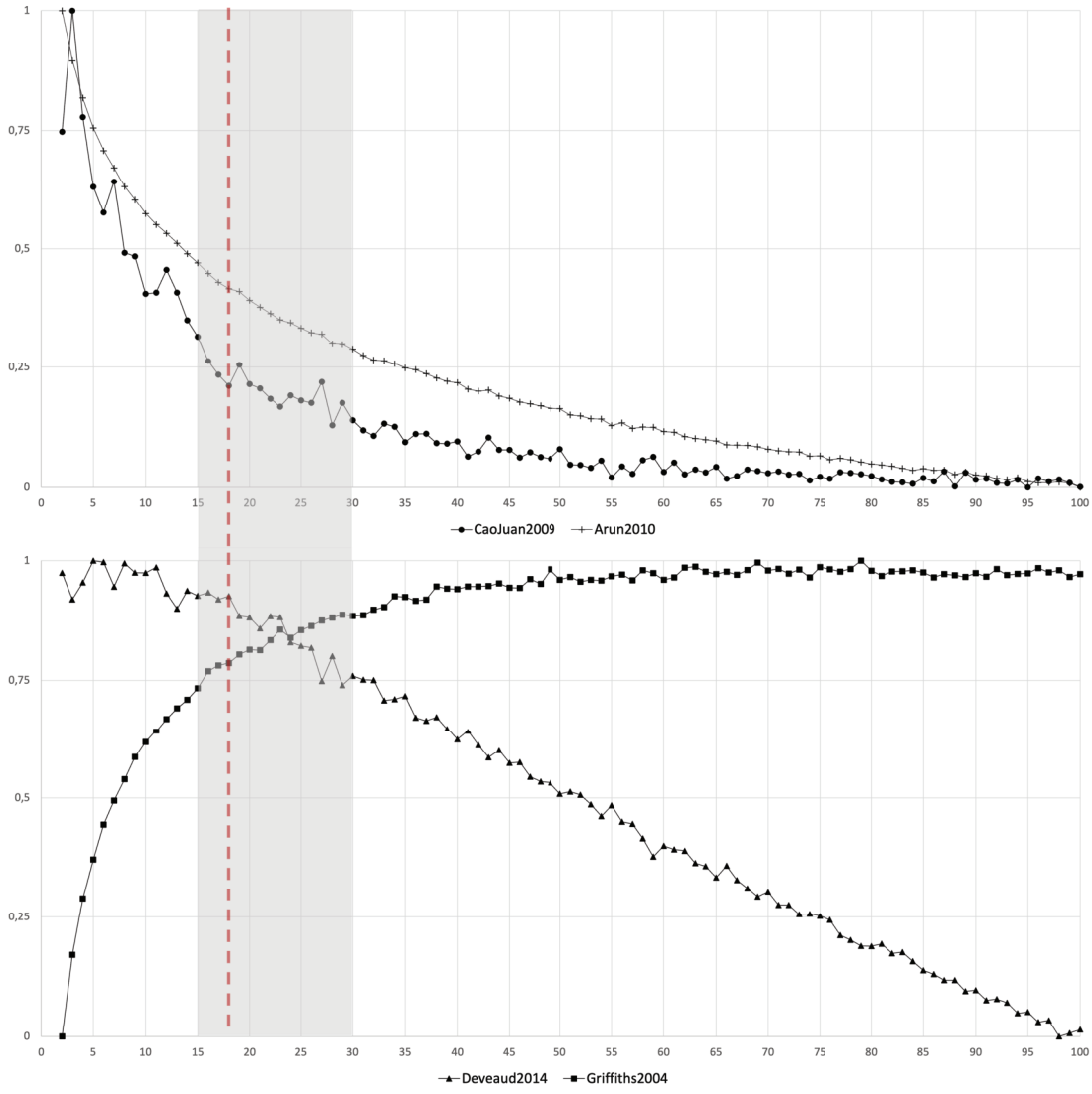

Figure 2. Selection of the number of topics (K)

To assess the correct number of topics to select, we selected the range at which the four distributions were closest. Based on the results represented in Figure 2, our range went from 15 to 30 topics. To decide how many of them to consider, we did an iterative analysis inspecting the results of the LDA distribution for each number

of topics in our range. Looking at the results, we noticed that for a higher number of topics selected, there was a strong similarity between. Looking at the results, we noticed that the higher the number of topics (K) selected, the more similar individual topics were. After several iterations, we concluded that the correct number to select was 18 topics. As a further step, we manually inspected the single topics to determine the consistency and accuracy with respect to the documents contained inside them and also to see if two or more topics could be merged together.

The 18 topics are related to different aspects of AI bias, fairness and ethics inside organisational and business settings. They can be summarised as follows: 1) AI to predict customer intentions and the implications on privacy, 2) AI to predict system performances, 3) Trends, performances and biases of artificial neural networks (ANN), 4) Content analysis on AI tools, 5) Machine learning for prediction and classification, 6) Using digital technologies to overcome barriers inside enterprises, 7) AI to control network performances, 8) Learning analytics in education, 9) Reducing bias in ML algorithms, 10) Machine anthropomorphism, 11) Natural language processing (NLP) and image recognition training to reduce error, 12) Bright and dark side of AI in the public sector, 13) AI tools for monitoring social media, 14) AI & big data in finance, 15) ML to predict quality of service in healthcare, 16) AI fairness in decision-making, 17) AI becomes human, 18) Explainable AI to build trust between human and machine (see Table 5 for further details).

The number of documents contained in each topic varies between 40 and 80 (m = 66.56, sd = 15.91). The only two outliers are represented by topics 1 - AI to predict customer intentions and the implications on privacy and 2 - AI to predict system performances, which include 115 and 86 articles, respectively. For each topic, we represented the number of publications between 1981 and 2022 to determine the most recent trends in academic research. At the same time, we inspected the most important journals inside each topic to see which AJG research field is predominant. Every topic found had a relevant number of publications in the years 2021 and 2022, with the number of articles increasing by up to ten times. For instance, topic 17 - AI Becomes Human and topic 6 - Using Digital Technologies to Overcome Barriers Inside Enterprises went from an average of 1.1 and 2.4 articles, respectively, in 2019 to 14 and 34 articles in 2022. With particular regard to topic 17 - AI becomes human, the interest of academics dealing with organisational and social studies has significantly increased in recent years due to technological advances in the field of machine learning and the growing concern over the invention of a sentient AI (Korteling et al., 2021; Tiku, 2022).

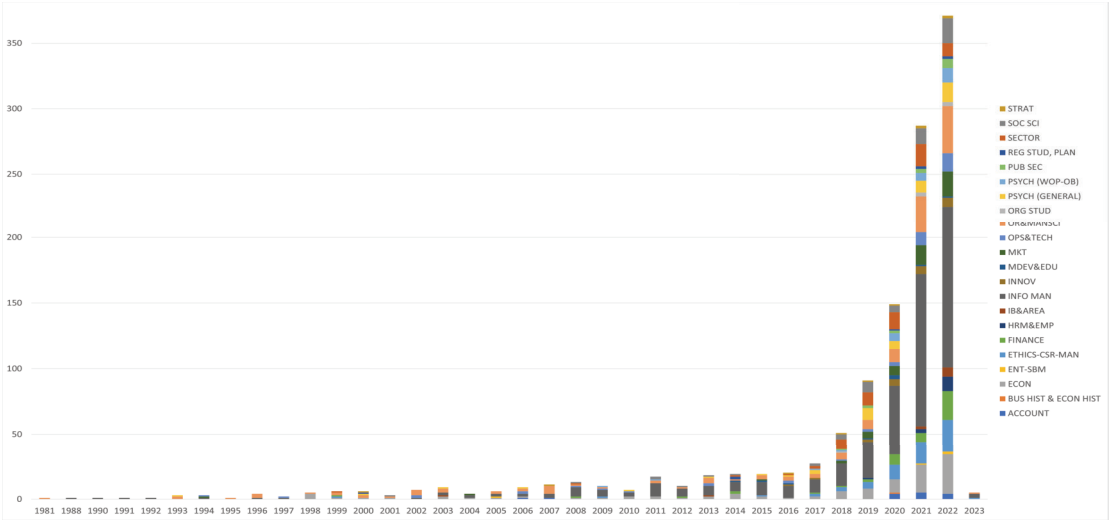

Figure 3. Number of publications per year since 1983 divided by AJG fields (see research method paragraph). Full legend available at: https://bit.ly/3K1EbML

Within each topic there are several journals (m = 42.78, sd = 8.61), and as expected the most represented AJG field inside them is Information Management, confirming that it is the AJG field in which AI issues are mostly debated. Exceptions are: topic 15 - ML to predict the quality of service in healthcare in which the most representative fields (37%) are Social Sciences and Economics, Econometrics and Statistics; topic 3 - Trends, performances and biases of artificial neural networks (ANN) with nearly 40% of the articles contained inside the research field of Operations Research and Management Science; topic 14 - AI & big data in finance; and topic 9 - Reducing bias in ML algorithms have respectively 31% and 56% of the articles belonging to the field Finance and Economics, Econometrics and Statistics. It can also be noted a convergence between the content of the topic and the most representative AJG fields. Overall, the three most important fields inside the dataset are Information Management, Operations Research and Management Science, Economics, Econometrics and Statistics and General Management, Ethics, Gender and Social Responsibility with 434, 137, 105 and 67 articles, respectively. The situation in the AJG ranking areas (Figure 3) is similar to what was observed for the topics. The number of publications in these fields has increased significantly in recent years, a sign that AI is becoming an increasingly relevant phenomenon in all business and organisational fields, as well as the ethics and the issues of the prejudice related to that technology. Some areas such as Business and Economic History, Human Resource Management and Employment Studies, Organisational Studies, Public Sector and Health Care had no publications until 5 years ago. This aspect could be an indicator of the increasing interest in

addressing ethical problems related to AI in both private and public organisations with attention to human resource management.

Discussion

After performing the impact and content analysis, we manually analysed the most relevant articles in each topic. A more in-depth analysis of the literature in each cluster helps to better understand the results of the topic modelling analysis and provides better insight. According to the theta value (distribution of topics over documents), we selected and analysed the most relevant papers inside every topic. As a result, we defined a more appropriate title and description for each topic, not only based on the list of keywords provided by the algorithm (Table 5).

Table 5. Topics title and description

Topic	Topic description
T1 - AI to predict customer intentions and implications on privacy	The topic focuses on the use of AI tools to analyse what factors are most impactful in predicting customer intentions and preferences, with a focus on privacy issues.
T2 - AI to predict system performances	The topic focuses on the use of ML and Artificial Neural Network (ANN) models to predict the performance of a specific system, especially those connected to renewable energy.
T3 - Trends, performances, and biases of ANN	The topic focuses on the study of the performances of neural network algorithms and on the description of the latest trends in the field.
T4 - Content analysis on AI tools	The theme of this topic is related to studies that use primary and secondary data with the aim of describing the performances of AI and ML tools.
T5 - Machine learning for prediction and classification	The topic focuses on the development and study of new ML models for prediction and classification.
T6 - Using digital technologies to overcome barriers inside enterprises	The theme concerns the use of digital technology to overcome institutional bias, operational issues and communication difficulties inside companies. A few studies focus on the barriers such as security, performances and standardisations emerging during a digital transformation process.
T7 - AI to control network performances	The theme is related to the use of AI tools such as ML algorithms to detect and mitigate faults within a network.
T8 - Learning analytics in education	This topic is mainly focused on the role of AI tools in the educational sector. More specifically, high emphasis is given to learning analytics tools and the role they play in supporting teachers in doing their job.
T9 - Reducing bias in ML algorithms	This topic is related to ML algorithms, trying to refine these algorithms by reducing biases with the introduction of different computational techniques.
T10 - Machine anthropomorphism	This topic is related to the concept of automation anthropomorphism and its impact on humans.
T11 - NLP and image recognition training to reduce error	The topic focuses on the development and training of NLP and image recognition algorithms to avoid bias and improve accuracy. The main aim is to use those algorithms to find reliable patterns that reflect human judgement.
T12 - Bright and dark side of AI in the public sector	The main theme is the impact of AI applications in the public sector, empirically analysing the pros and cons of AI adoption and implications.
T13 - AI tools for monitoring social media	The papers in the topic apply AI techniques to explore the behaviour and intentions of companies and users on social media.
T14 – AI & big data in finance	The topic examines the application of AI and ML tools in the financial sector.

Table 5. (Continued)

T15 - ML to predict quality of service in healthcare	This topic focuses on the use of ML algorithms to predict the quality of services in the healthcare sector and to monitor parameters related to the improvement of human welfare.
T16 - AI fairness in decision-making	The topic is related to a delicate theme, the fairness of algorithmic decision-making and its impact on minorities.
T17 - AI becomes human	Papers in this topic investigate whether AI could be seen, perceived and treated like a human being. Some relevant papers propose tests to evaluate AGI.
T18 - Explainable AI to build trust between human and machine	The human lack of trust in AI applications is the main theme of this topic. Particular attention is paid to the financial and healthcare sector in which the decisions made by an AI application could have serious implications.

Having clarified the content of the papers assigned to each topic, we tried to study their similarity to find hidden patterns inside the literature analysed. We used the R package LDAvis to explore the similarities between topics and have a complete overview of them. The intertopic distance map (Figure 1) refers to the degree of difference or similarity between topics in a given text corpus, while the marginal topic distribution represents the overall prevalence of each topic in the corpus of documents being analysed. Moreover, we built a dendrogram using Hellinger's Distance, a metric used to quantify the difference between two probability distributions. The dendrogram was useful to determine the similarity of the topics. Based on the results shown in Figure 4 and on the previous in-depth topic analysis, we were able to group the topics into two main categories of discussion: one where the bias emerges directly from the output of the algorithm (technical nature of the bias), and the other one where biases could be ascribed to socio-cultural components (socio-cultural nature of the bias).

Macro-topic 1 – Technical nature of the bias. This cluster is composed of topics 2) AI to predict system performances, 3) Trends, performances and biases of artificial neural networks (ANN), 5) Machine learning for prediction and classification, 7) AI to control network performances, 9) Reducing bias in ML algorithms, 11) Natural language processing (NLP) and image recognition training to reduce error. This set of topics focuses on technical and mathematical studies concerning AI algorithms. It concentrates on the enhancement of the mathematical and statistical techniques on which these algorithms are based. Thus, an attempt is made to reduce and address errors during the development and testing phase that may have adverse effects on the outcome of the systems in which they are used.

Macro-topic 2 – Socio-cultural nature of the bias. This cluster is composed of topics 1) AI to predict customer intentions and the implications on privacy, 4) Content analysis on AI tools, 6) Using digital technologies to overcome barriers inside enterprises, 8) Learning analytics in education, 10) Machine

anthropomorphism, 12) Bright and dark side of AI in the public sector, 13) AI tools for monitoring social media, 14) AI & big data in finance, 15) ML to predict quality of service in healthcare, 16) AI fairness in decision-making, 17) AI becomes human, 18) Explainable AI to build trust between human and machine. In contrast to the first macro cluster, studies here are mainly managerial and sociological in nature, specialising in the research on the implications of artificial intelligence algorithms when applied in different contexts. The focus of this cluster is on empirical studies in different sectors in which various AI-based applications are implemented and adopted. The majority of these contributions consider different forms of organisational settings.

Figure 4. Intertopic distance representation: macro-topics (left) and meso-topics (right)

In Figure 4 and Figure 5, another distinction emerges concerning the selected topics. Although it is less noticeable than the distinction just described (technical vs. socio-cultural nature), a difference can be seen between topics that are located above the horizontal axis (PC1) and those located below. More specifically, the lower topics focus on the application of AI to predict and empirically describe various aspects, such as the

performance of systems and processes or to analyse various factors impacting people in different contexts. The topics above the horizontal axis, on the other hand, are aimed at correcting different types of bias within AI artefacts. Some articles address the problem from a technical standpoint by developing or adjusting algorithms and manipulating training data to reduce errors in the AI outcome and increase overall performance. Others focus on describing the impact of AI biases on societies and organisations and how they are addressed at a societal level. Overall, two different strands can be identified, we named the first one "efficacy perspective" as it is dedicated to the use and improvement of AI for predictive purposes to enhance the performances and, consequently, to reduce possible issues in the outcome of AI artefacts. The second, named "critical perspective", is more focused on the exploration, description and mitigation of bias. Although the papers in the dataset belong to the topic of AI bias, fairness and ethics, the topics above the horizontal axis focus more on increasing the effectiveness of a determined process using AI. The topics below the axis are more focused on mitigating AI bias and on discussing the consequences that the dark side of AI is bringing and could bring. Naturally, the closer the topic is to the horizontal axis, the more they are borderline between the two groups. By analysing the individual topics, we were able to achieve a better overview of their content and assign them to a specific strand. This is the case of topic 9, which, despite being located below the horizontal axis, is more closely associated with the critical perspectives group. Both strands have studies that explore the phenomenon from both a primarily technical and social perspective. which will be discussed in more detail in the description of meso-topics (Figure 4).

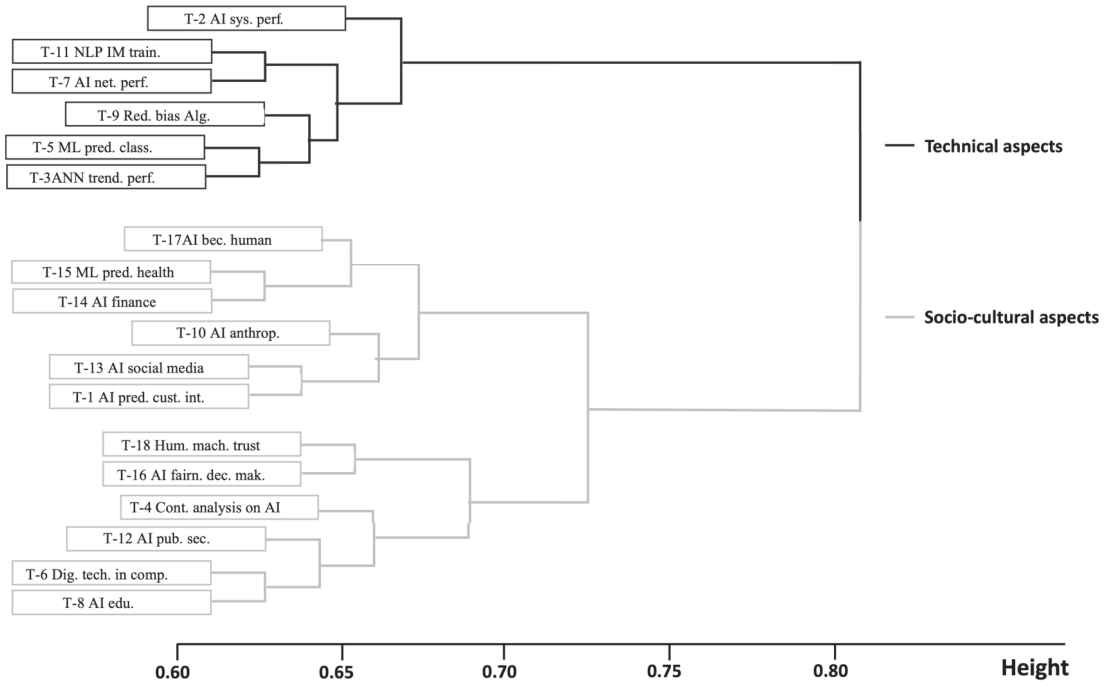

Figure 5. Hellinger's Distance (topic similarity)

Comparing the evidence produced by the cluster analysis with the results of the in-depth analysis performed manually, we were able to identify a further list of sub-categories, providing an intermediate classification of the papers, in between the two macro-topics and the 18 single topics, called meso-topics. We then identified two meso-topics belonging to Macro-topic 1 (Technical nature of the bias) and three meso-topics for Macro-topic 2 (Socio-cultural nature of the bias). Figure 6 provides a graphical representation of the distribution of the meso-topics. They are divided according to the nature of the studies (macro-topics) and the focus of the contributions to the topics. More specifically, we identify three different levels of focus in the studies (AI performance, AI affecting human behaviour, and Humanizing AI behaviour). The first one, which contains the majority of meso-topics, is related to the assessment and the further enhancement of AI performances. The research in this area investigates from both a technical and socio-cultural point of view the development of new AI tools and the results of AI implementation in various settings, describing benefits and issues and making recommendations for further improvement. The remaining areas are more related to socio-cultural studies, which are focused on inspecting how AI affect human behaviour, reactions and emotions. While one area considers AI in general, the other one is strictly focused on AI tools that resemble human aspects and behaviour. We provide a detailed description of the meso-topics in the following paragraphs. Finally, looking

at the meso-topics classification and description, we had the chance to draft a list of possible future research directions, summarised in Table 6. More details about each meso-topic are provided below.

Meso-topic 1 Algorithmic bias reduction. This cluster contains the topics 3) Trends, performances and biases of artificial neural networks (ANN), 7) AI to control network performances, 9) Reducing bias in ML algorithms, 11) Natural language processing (NLP) and image recognition training to reduce error. This set of topics belongs to the first macro-topic, it is mainly focused on AI algorithms improvement for bias reduction. More specifically, the inspected literature inside the meso-topic focuses on the development, training and implementation of various AI models to enhance objectivity, reliability and performance. The boundaries of the subjects explored are very broad and concern many AI subsystems, such as neural networks, natural language processing and machine learning. Since AI is used in complex and rapidly changing environments such as finance and robotics (G. Kumar et al., 2021; Zhou et al., 2022), it is essential to continuously improve the underlying algorithms with the application of techniques that enhance the robustness and performance of the models (Liu et al., 2022). Similar attention is also given to the process of reducing bias in inference techniques on which many machine-learning tools are based (Chernozhukov et al., 2022; Semenova & Chernozhukov, 2021). In the topic, there are also many studies focusing on the development of AI algorithms capable of monitoring and enhancing the performance and reliability of sensors to avoid faults that could cause serious damage to systems (Saeed et al., 2021; J. C. Da Silva et al., 2012). Relevance is also given to the development and training of NLP and image recognition algorithms to avoid bias and improve accuracy. The main goal is to use these algorithms to find reliable patterns that reflect human judgement. The literature about NLP tools aims to develop and train ML algorithms to analyse the text and generate reliable, trustworthy and precise results avoiding biases (Hagen, 2018; F. Wang et al., 2022). Other studies concerning image recognition algorithms are related to the development of tools to detect misinformation and improve facial recognition (Ghai et al., 2021; W. Xie et al., 2021).

Meso-topic 2 Predictive algorithms. This group includes topic 2) AI to predict system performances, and topic 5) Machine learning for prediction and classification. This meso-topic is dedicated to the study, development and enhancement of predictive algorithms. The focus is on the use of ML and ANN models to predict the performances of a certain system and to find new ways to improve accuracy. The considered research gives considerable importance to environmental sustainability by studying assets such as renewable

energy systems and sustainable processes. Studies try to predict the performance of energy systems using various parameters to reduce error and better predict performances (Abujazar et al., 2018; Khatib et al., 2011), while others focus on data forecast required to assess the feasibility of an energy system construction (Al-Shamisi et al., 2013; Ghimire et al., 2019). Attention is also given to project performance. Lind and Sulek, (2000), in fact, propose a neural network model to predict software project duration and compare it using a regression model. Results show that the neural network model outperforms the regression model in accuracy, bias forecasting and model fit. Concerning sustainable processes and practices, some research inspects the use of neural networks and logistic regression models to predict waste disposal rates and customer choices (van Wezel & Potharst, 2007; Vu et al., 2022). Another important aspect is related to the study of new methods to address the issues created by imbalanced datasets. A dataset is imbalanced when there is a high disparity of classes in the variables. These differences between the available data are problematic for AI algorithms, especially in large datasets. For this reason, many studies seek to find new methods to reduce bias when working with this kind of dataset (Bhowan et al., 2013; Y. Huang et al., 2022).

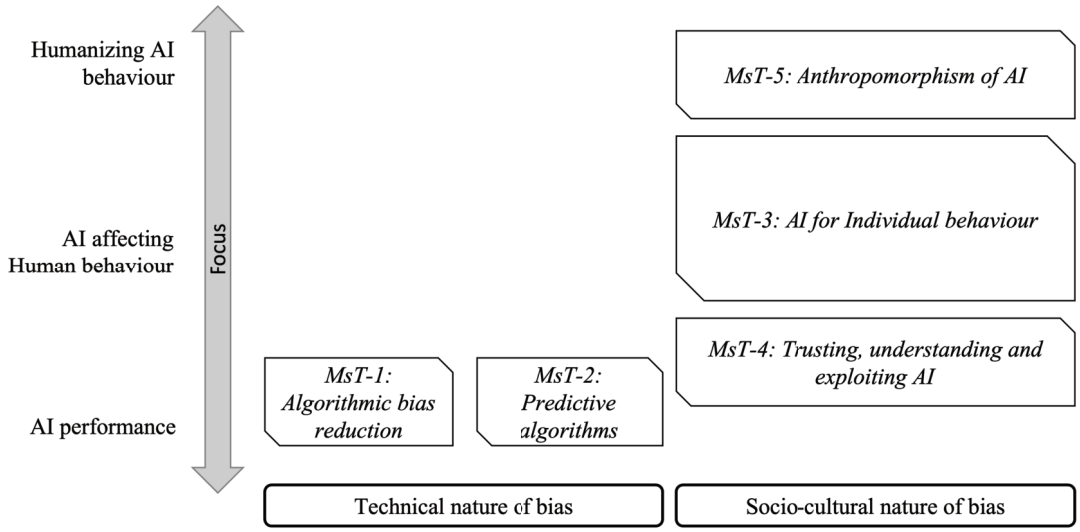

Figure 6. Meso-topics distributions according to the nature and focus of the study

Meso-topic 3 AI for Individual Behaviour. The topics in this cluster are 1) AI to predict customer intentions and the implications on privacy, 13) AI tools for monitoring social media, 14) AI & big data in finance, 15) ML to predict quality of service in healthcare. The contributions in this meso-topic focus on exploring the factors (e.g., emotions, knowledge) affecting the adoption, the use and the relationship with AI technology. The fields and the aspects investigated are manifold. A popular strand of research is related to the use of AI

tools to analyse the factors that are most impactful in predicting customer intentions and preferences, with a focus on the privacy issues that arise when using new technologies to perform this kind of research. Some of the most relevant papers on the topic study the consequences of the implementation of AI in the shopping experience and its effects on personalisation, with a focus on trust and privacy issues (Ameen et al., 2021, 2022). Other studies use neural network models to determine the most impactful factors that contribute to the use of mobile commerce (m-commerce) (Chong, 2013; Hew et al., 2016). Relevance is also given to voice based digital assistants (VBDA) such as Alexa and Google Assistant, trying to determine the factors that have an impact on consumers' privacy concerns (Vimalkumar et al., 2021). AI techniques are also used to explore the behaviour and intentions of companies and users on social media, investigating various settings. Pelau et al., (2021) and Hassan, Hudaefi and Caraka, (2022) conducted research on user opinion on cryptocurrencies and acceptance of AI technology from different consumer clusters, respectively. David et al., (2016) developed a classifier algorithm, trained with political parties' posts, to predict the political orientation of users. While Gunarathne, Rui and Seidmann, (2022) proved a case of racial bias by inspecting the behaviour of US airline employees interacting on social media. They found that employees were less likely to respond to African American customers when a complaint was presented. Another field inspected by the studies populating the meso-topic is the financial sector. More specifically, many contributions use AI to analyse people's behaviour and emotions and explore the effect that the application of financial AI tools has on investors. Eachempati and Srivastava, (2022) used NLP tools to analyse the role of investor sentiment in impacting asset prices by developing two different sentiment indices based on news and corporate disclosure. Other empirical studies explore the most important factors and bias of robot-advisory services in finance on a sample of Indian investors (Bhatia et al., 2021) and the use of ML algorithms to predict and prevent financial risk of enterprises conducted on a sample of 21 companies (B. Gao, 2022). In the meso-topic, AI and ML algorithms are also used to predict the quality of service in the healthcare sector, as well as to monitor parameters related to the improvement of human welfare. The majority of the papers examined construct ML algorithms to discover and assess the impact of the factors that contribute most to the quality of service inside hospitals and clinics by inspecting different cases (e.g., strokes, tuberculosis) (Ali et al., 2018; Cunnama et al., 2020). While Suleiman, Tight and Quinn, (2016) focus on the prediction of the levels of particulate matter in cities using ML algorithms.

Meso-topic 4 Trusting, understanding and exploiting AI. This group comprises the topics 4) Content analysis on AI tools, 6) Using digital technologies to overcome barriers inside enterprises, 8) Learning analytics in education, 12) Bright and dark side of AI in the public sector, 16) AI fairness in decision-making, 18) Explainable AI to build trust between human and machine. The aim of the meso-topic is to describe how AI is changing the sectors in which it is implemented and how to identify and address the issues that arise after its adoption. These studies use primary and secondary data with the aim of describing the performances of AI and ML tools. There are empirical studies that focus on the importance for hospitality firms of social media analytics to examine tourist experience and make marketing decisions (Lin et al., 2021) and on the evaluation of AI tools, such as artificial intelligence video interviews for employee recruitment by examining the opinion of the applicants (J. Y. Kim & Heo, 2022). There are also literature reviews focusing on artificial intelligence application and scope in different markets and industries (Simon, 2019), on the evolution and enhancement of sentiment analysis techniques (Qazi et al., 2017) and on the use and application of technology related to AI and digital transformation, such as RFID sensors (Raza, 2022). Many contributions are also involved in studies on the use of digital technologies to overcome institutional bias, operational issues and communication difficulties inside companies, while other research focuses on barriers such as security, performance and standardisations that appear when converting to a digital company (S. Li et al., 2022; Rathore et al., 2022). Wu and Kane, (2021) focus on the introduction of modern digital collaboration tools in a workplace and find that these tools ensure greater access to knowledge and expertise inside the firm, but, on the other hand, they favour people with determined characteristics more than others. Liming Xu, Mak and Brintrup, (2021) study the evolution of agent-based systems over the years by describing the current state of the art and the impact of such systems on companies. This topic also includes studies investigating the role of AI tools in the public and education sectors. High emphasis is given to learning analytics tools and the role they play in supporting teachers in their work. Some of the most relevant studies conduct empirical research with the aim of inspecting the results that higher explainability and an improved knowledge base have on teachers. All the studies conclude that these factors are of paramount importance to implementing the tools and enhancing trust in their users (Celik, 2023; Kitto & Knight, 2019; Marshall et al., 2022; Nazaretsky et al., 2022). Other contributions try to investigate the issues of AI applications in the public sector. More specifically, they try to describe, with practical examples, the pros and cons of AI adoption together with future implications. Langer and Cornelius

J. König, (2021) provide a description of opacity in Human Resource Management (HRM) algorithms. They explain the benefits and the issues derived from algorithm opacity and try to address the problem by giving recommendations to improve transparency in AI algorithms. On the same subject, Anastasiadou, Santos and Montargil, (2021) and Bracci, (2022) underline the need for transparency in AI algorithms in public sector operations and in accountability governance in order to provide a better service to the public. Importance is also given to the education sector and on how digital inequalities could represent a paramount issue in the future of labour market (Ozkazanc-Pan, 2021). The meso-topic also addresses a delicate theme, the fairness of algorithmic decision-making and the impact that it has on minorities. Feuerriegel, Dolata and Schwabe, (2020) investigate the challenges and opportunities of fair AI with respect to people, technology and organisations, underlining the unprecedented opportunities that it brings to society. Bolander, (2019) draws attention on the potential of AI in decision-making also highlighting the many potential weaknesses inside the field, as a warning for thoughtful implementation from companies. Other research focuses on empirical studies about AI fairness in different scenarios, also inspecting the issues that appear when efforts are made to enhance fairness in algorithms (Fu et al., 2022; Grote & Keeling, 2022; Köchling et al., 2021). Considerable importance is also given to how to address human lack of trust in AI applications in many sectors, such as financial and healthcare areas, where the decisions made by an AI application might have serious repercussions on people. More specifically, the discussion concerns how algorithms' explainability and domain knowledge can augment users' awareness of AI decisions and help them not to over-trust them (Dikmen & Burns, 2022). Furthermore, insights are given on the design and reporting of AI programs to enhance the efficacy of human-machine collaboration and build trust with end users (Alam & Mueller, 2022; Vössing et al., 2022). Finally, Friedrich et al., (2022) underline the importance of the study of human-AI relationships and how it contributes to shaping the relations between human beings and the surrounding world. They focus on the impact that black box AI has on physicians inside clinics.

Meso-topic 5 Anthropomorphism of AI. In this cluster, we find the topics 10) Machine anthropomorphism, 17) AI becomes human. The meso-topic focuses on the anthropomorphisation of machines. In particular, two themes are explored. The first, more practical and topical, concerns the effects that a machine mimicking human behaviour has on people. While the second refers to the possible consequences of the development of an AI with human-like capabilities. As for the first theme, studies explore

the concept of automation anthropomorphism and its impact on people. The most relevant research empirically studies the changes in trust and preference in users when they have to interact with a highly anthropomorphised agent. Results show that the more the machine reflects human behaviour and empathy, the more it inspires trust and preference in the users (de Visser et al., 2016; Peng et al., 2022; Schelble et al., 2022). The importance of training and communication in human-autonomy teams and how this leads to higher performance is also emphasised (C. J. Johnson et al., 2021). The second theme raises a relevant question, can AI be seen, perceived and treated like a human being? The most relevant papers in this cluster study this aspect. Swanepoel, (2021) investigates the possibility of giving moral agency status to AI. The author states that deliberate norm adherence (the capacity to act in accordance with a rule for reasons unrelated to those inherent in the rule) is fundamental to gaining such status. Li, (2021) discusses the importance that a "friendly AI" will have in Artificial General Intelligence (AGI) applications and identifies four issues related to its implementation in terms of human needs and moral development. Greene, (2017) studies moral intuitions and the case in which they are acquired by computational techniques such as ML methods. Finally, Sparrow, (2004) proposes a test to assess whether an AI can replace a human being in a moral dilemma. Interestingly, all the answers to the questions are negative, the AI cannot be compared to a human being yet, but these studies are of paramount importance to address possible issues in the future.

AIS List of Premier Journals articles inside the meso-topics. In Table 4, we have associated each article with its meso-topic. We selected a total of ten articles belonging to the list, three of them are associated with the meso-topic of technical nature (1 – 2) and the others with the socio-cultural ones (3 – 4). Similar to the other contributions used to describe the first two meso-topics, the articles of the list are focused on introducing analytical frameworks that use AI algorithms to investigate and predict various aspects in a more efficient way, by reducing bias and improving accuracy. Specifically, the articles associated with the meso-topic propose a text analytics framework used for inspecting the correlation between the host description and the trust perception and purchase behaviour of the guest on the platform Airbnb (Zhang et al., 2020). The other paper is focused instead on combining textual and numerical data to improve forecasts concerning stock returns (Geva & Zahavi, 2014). The last article of the list (Teodorescu et al., 2021) belonging to the second meso-topic seems to be an outlier . After an analysis, we concluded that the article could be a better fit for the fourth meso-topic. The reason behind its position could be attributed to the fact that inside the paper, there are sections

focused on describing the fairness criteria used in machine learning prediction algorithms. Such an aspect could easily mislead the LDA algorithm that is based on the statistical distribution of the keywords inside the documents. Moreover, the article is one of the less related to the topic according to the theta distribution. The contributions contained in Table 4, related to the third meso-topic, are instead focused on inspecting the impact of AI on people or in using AI to investigate the relationship between digital phenomena and people's attributes and perceptions. The topics within these papers are related to the use and improvement of AI algorithms for the study of online platforms. More specifically, some of the studies deal with discovering manipulations within online reviews (N. Kumar et al., 2018) or investigating which online review attributes people find most useful (Eslami et al., 2018), while others on measuring the impact a humanised chatbot has on people (Schanke et al., 2021). Finally, the articles embedded in the fourth meso-topic, discuss the topics of responsible AI and the dark side of AI. B. Kim et al. (2020) propose a framework that enhances the transparency and explicability of AI algorithms inside decision-making processes. The other papers describe the phenomena of ubiquitous IT and responsible AI. They focus on the issues that could arise from the use of AI, how the research community is trying to address these problems and underline the impact (positive and negative) that this technology is having in the IS field (Mikalef et al., 2022; Ransbotham et al., 2016). Both give avenues for future research, focusing on relevant areas related to the dark side of AI (e.g., fairness, accountability, transparency, bias, privacy).

Table 6. Meso-topics titles and possible future research opportunities

Topics	Tentative future research questions
Meso-topic 1 Algorithmic bias reduction	Although bias reduction from a technical point of view could make AI techniques more reliable, what are the implications of their use? For example, the case of NLP and image recognition tools used to make delicate decisions, such as diagnosis in healthcare. Could increasing the complexity of the AI system for making more accurate decisions, affect bias reduction?
Meso-topic 2 Predictive algorithms	Some studies investigate the comparison between the predictive outcomes conducted only by humans and other results achieved by AI-human augmentation. In these cases, which aspects should be evaluated besides the technical perspective? What are the possible implications for individuals and organisations of regulators' and policy makers' interventions concerning the development of ML models used for prediction and classification?
Meso-topic 3 AI for Individual behaviour	What are the main factors that impact user privacy and trust concerns when using AI technology? What are the main issues related to the adoption of a technology based on AI algorithms from the perspective of a customer (individual or organisation)?

Table 6. (Continued)

Meso-topic 4 Trusting, understanding, and exploiting AI	What are the consequences that algorithmic bias inside ML tools brings to the results of a firm's decision-making process? Is there a difference in the reaction towards and reception of new AI-based technology between employees having different roles? What will the implications be for employees to work side by side with AI? Are there any factors that might cause a disparity between employees within an organisation?
Meso-topic 5 Anthropomorphism of AI	What are the main attributes that influence the relationship between humans and machines when working together? In which cases can the machine be perceived as a colleague and not a tool, and what are the long-term implications of human-machine interaction? What are the legal implications of an AI that is comparable to a human being? How could policy makers address the regulatory challenges that arise from the development of AGI?
AIS List of Premier Journals articles	What are the processes for bias assessment in AI artefacts? How do they change according to the different kinds of bias? How the difficulty to define an ideal unbiased scenario in the reality could affect the identification, definition and mitigation of AI biases? What are the techniques or processes that, the organizations which develop AI artefacts, could follow to reduce the presence of different kinds of bias? What is the impact of an increased transparency of AI artefact in organizational settings? Is there a difference between decisional levels?

Conclusion

In this study, we performed a CLR to explore the evolution of the debate concerning AI bias, ethics and fairness in organisational and business studies. First, we performed an impact analysis evaluating the impact of journals and articles, then we conducted a content analysis using a topic modelling algorithm identifying 18 topics that are relevant and enough for clustering the papers of our dataset. We then manually analysed the most relevant papers within each topic in order to refine its title and description. Finally, combining the outcome of the automatic content analysis and the review performed by the authors, we identified the possibility of classifying the 18 topics into two main categories called macro-topics (e.g. technical and socio-cultural nature of the bias). Afterwards, looking in detail at the content of the papers belonging to each topic, we recognized five middle categories, called meso-topics. The results show coherence between impact and content analysis concerning the distinction between studies adopting a technical or a socio-cultural perspective in their discussion. Within the research, we have devoted specific sections to exploring the debate within the journals belonging to the AIS premier list. Their analysis has revealed the growing emphasis on the potential problems arising from the adoption and use of AI. Beyond that, some papers focus on the use of AI to analyse and mitigate ethical and trust issues.

From a theoretical perspective, we provided a rigorous and easy replicable method to conduct a computational literature review study. More specifically, we used the framework given by Antons et al. (2021) adding a further step concerning the in-depth analysis and description of each topic, thus making the

interpretation of results an iterative process to improve the inspection of the topic model output. CLR allows us to qualitatively analyse a large amount of data. As mentioned, the benefits are numerous. First of all, a very large number of articles can be examined, in terms of both content and impact. CLR also allows scholars to conduct research in a time-saving manner, as it is quite difficult to analyse thousands of documents manually in a limited amount of time. Moreover, thanks to CLR, it is easier to simultaneously analyse multi-domain literature to discover hidden patterns that would otherwise be difficult to find. In this contribution, we also provided a classification of literature concerning AI bias, ethics and fairness with a focus on organisational and business studies that could help other researchers have a better understanding of the domain and have an outlook of the different research strands. While previous studies focus on a limited set of papers by exploring the phenomenon in a specific field, we explored a broader set of articles, investigating different aspects concerning the same phenomenon debated in different fields of study. Furthermore, we presented a tentative set of research questions based on the content of each meso-topic that could provide opportunities for future research. Finally, we tried to describe in detail all the steps that allowed us to perform the CLR hoping that this article might be of support to scholars willing to carry out similar research.

The outcome of AI artefacts such as predictive algorithms and large language models are increasingly impressive and reliable thanks to the data produced by humankind. AI is capable of synthesising the data into useful information and accelerating cognitive processes (Brynjolfsson et al., 2023). Nowadays, this technology is highly supportive in many areas, but is it possible that we will reach a point where humans will stop producing data and be replaced by AI in this task, especially in contexts more suitable for automation? What will be the consequences for our cognitive capabilities? Although this scenario is very dystopian and remote, the oversimplification of many tasks due to the introduction of AI has profound consequences on our cognitive development (Sætra, 2023). It is, therefore, significant to develop AI tools in a way that their outcome is as interpretable and transparent as possible (Dignum, 2023; B. Kim et al., 2020). Consequently, it is crucial that organisations and policymakers act in the development and support of digital awareness campaigns to sensitise people on the importance of new technologies, such as AI. Differences in awareness could provide fertile ground for the emergence of different kinds of bias that could lead, for example, to social inequality problems (data is gathered and represents mostly age groups that are most likely to engage with new technologies or most developed countries where the majority of the people have access to the internet) (Madhavan et al., 2022;

Su Yiran et al., 2023). A similar discourse could be made with specific reference to the interaction between employees and AI. It is relevant for people to increase their awareness about the way AI solutions work (e.g., why an AI artefact produces a determined outcome in a particular context), aiming at using it as an augmentation powerful tool rather than a substitute (C. J. Johnson et al., 2021; Teodorescu et al., 2021).

Since AI is becoming an increasingly powerful and pervasive tool, concerns about its ethical and fair use are growing (Zuiderwijk et al., 2021). Governments are working together with academics and firms to address issues related to the development, testing and distribution of AI tools by seeking regulations that could be shared globally. Although progress has been made in identifying the ethical principles of AI and regulations for its development, they lack a common interpretation, delaying the creation of a set of rules to be observed by those developing this type of technology (Jobin et al., 2019; Martin et al., 2019). Moreover, the firms, need to better address and rethink their strategies concerning ethical and sociotechnical issues due to AI (D'Cruz et al., 2022). The article provides a clear distinction of the research strands related to AI bias, fairness and ethics that might be helpful to managers during the development, testing and adoption/distribution phases of an AI system. The different classifications help to identify the studies that address a particular issue: from the choice of a system used to predict the outcome of a specific project to the decision of implementing an AI application that closely resembles human beings or that is used by consumers. Moreover, focusing on AI ethical issues, this research provides some insights and references useful in promoting the responsible use of AI.

The contribution has some limitations. First of all, it could be interesting to explore more in-depth the theta distribution of the topic modelling output to study papers that are relevant to more than one topic. This analysis could allow a more precise description of hidden connections between topics and avoid biases related to the inspection of the most relevant articles only. Further studies could also concentrate on different kinds of sources, such as conference proceedings, thus providing the latest research trends in the domain. Furthermore, new NLP techniques could be used to explore the same datasets and compare the results. Nowadays, the proliferation of scientific literature increases the need for systematic, replicable and rigorous literature reviews, as well as the resources needed to conduct them (Badger et al., 2000). In parallel, digital technologies are becoming more and more pervasive in our lives. Therefore, it should be fundamental to understand these tools and the implications connected to their use from different perspectives.

CHAPTER 3

Exploring AI bias and mitigation challenges using a socio-technical perspective

Abstract. In an era where Artificial Intelligence (AI) increasingly permeates the global landscape, its potential to impact many people necessitates heightened vigilance against biases in AI-driven outputs. This research employs a grounded methodology, gathering insights through semi-structured interviews with AI professionals engaged in various stages of the AI lifecycle. Our objective is to develop a socio-technical framework that systematically identifies and mitigates AI bias, including various aspects of stakeholder participation. Initial findings from this study have helped us refine our interview protocols and expanded our understanding of overlooked AI biases. These insights offer a deeper understanding of the influence exerted by non-end-user stakeholders on the AI development lifecycle, providing a multi-dimensional perspective on the nuances of AI bias and its emergence and mitigation.

Keywords: Artificial intelligence, AI bias, Grounded theory, Bias identification and mitigation, AI lifecycle.

Introduction

The importance and relevance of artificial intelligence in the global landscape increases year by year (Makarius et al., 2020; Smacchia and Za, 2022). On an economic level, the market value is expected to grow to $2,000 million by 2030[13]. This technology is extremely multidisciplinary and is used to perform and support different processes within public and private organisations (D'Cruz et al., 2022). Attention towards AI is not only increasing from an academic point of view but is also finding the strongest response at a practitioner level (Stanford, 2023). Especially in recent years, AI-based artefacts are beginning to be adopted on a large scale, implying not only the utmost importance of this technology but also its ability to influence an ever-increasing number of people (Stanford, 2023). Due to this trend, there is a growing awareness of the need for regulation and shared practices in the AI lifecycle phases, ensuring compliance with ethical and social principles, avoiding discrimination, privacy reduction, and unfairness (Haenlein et al., 2022; Smacchia and Za, 2023). These aspects led numerous governments and institutions to develop comprehensive guidelines and frameworks,

[13] Next Move Strategy Consulting. (March 26, 2023). Artificial intelligence (AI) market size worldwide in 2021 with a forecast until 2030 (in million U.S. dollars) [Graph]. In Statista. Retrieved November the 2nd 2023, from https://www.statista.com/statistics/1365145/artificial-intelligence-market-size/.

resulting in a significant rise in regulations aimed at enhancing AI design, development, and deployment processes, with a particular focus on risk management and fairness compliance (Golpayegani et al., 2022; Smuha, 2021).

AI is increasingly becoming an indispensable technology inside organisational settings, and many organisations are using AI software not only to automate tasks but also as support for decision-making (Mikalef et al., 2022). Moreover, the outcomes of this technology have the power to influence an increasingly large number of people. Consequently, possible biases contained in AI artefacts can create or reinforce issues such as discrimination, social inequality, or lack of privacy, creating serious problems (Zuiderwijk et al., 2021). The presence of bias is not always negative, for instance, a biased dataset could be needed in developing a tool to diagnose breast cancer, a disease that is much more common in women than men[14]. However, in many cases, AI bias could lead to prejudice or preference toward a specific group or institution because of its characteristics (Mehrabi et al., 2021). For instance, it is reported that many voice recognition software are more accurate for male voices instead for female ones, raising concerns about the use of these applications in official contexts such as examinations (Bajorek, 2019). While this is only an example, there are numerous cases of fairness and ethical cases related to AI. These problems require increasing attention to the reduction of possible biases within AI artefacts and constantly seeking solutions that can improve their performance and accuracy (Ntoutsi et al., 2020).

While studies with a technical approach to mitigate and address AI bias have always been an important part of academic research, in the last few years, there has been increasing attention to social and ethical aspects related to this phenomenon (Smacchia & Za, 2023). To date, researchers have overlooked behavioural, organisational, and social implications, concentrating mostly on computational solutions to bias (Someh et al., 2019). While it cannot be denied that a technical approach is necessary to address AI bias, artefacts based on this technology are created in very complex contexts driven by various social dynamics (Dignum, 2023). Therefore, when studying this domain, it is necessary to adopt a socio-technical perspective that relates not only to the mathematical and computational constructs but also to the values and behaviours of those who design and develop the application, the environmental context, the organisational processes, and the

[14] Gerlinde Weger (October, 2023), Using Bias Intentionally in Artificial Intelligence. Retreived March 20th 2024 from: https://www.thefastmode.com/expert-opinion/33425-using-bias-intentionally-in-artificial-intelligence.

interconnections with the end users (Akter et al., 2021; Dignum, 2018; Schwartz et al., 2022; Someh et al., 2019). Organisational, individual, and environmental characteristics vary depending on the context in which an organisation operates. Different ideals, beliefs, and practices may lead actors involved in the design and development of an AI artefact can result in the inadvertent neglect or overconsideration of critical elements, potentially impacting the accuracy and fairness of its outcome. Similarly, the different tasks in which AI is applied have an impact on the definition and identification of bias (Kordzadeh & Ghasemaghaei, 2022). Hence, when inspecting how the biases emerge during the design, development, and deployment phases of AI artefacts, there is the need to consider multiple aspects that techno-solutionism alone cannot address (Schwartz et al., 2022). Furthermore, although numerous studies delve into this domain from a conceptual standpoint, there is a growing necessity for empirical research to grasp the ramifications of AI bias and to identify effective strategies for its mitigation (Kordzadeh & Ghasemaghaei, 2022; Mikalef et al., 2022).

This research aims to investigate the processes of identification and mitigation of AI bias inside companies that design, develop, and deploy AI artifacts using a grounded approach. More specifically, our objective is to develop a theoretical framework that delineates the process for identifying bias and, subsequently, mitigating it throughout the AI lifecycle, thereby helping practitioners and policymakers address this phenomenon. Since the nature and the contexts of AI bias are extremely varied, we have decided to adopt a grounded approach that allows the generation of a theory rooted in empirical evidence to increase relevance and applicability (Jaton, 2021). Through this study, we aim to answer the following questions:

How are biases identified by individuals engaged in AI projects during various phases of their lifecycle, including design, development, and deployment? What processes are employed to mitigate these biases?
Who are the key actors responsible for identifying and mitigating AI bias throughout the three critical phases of the AI lifecycle: design, development, and deployment?
What are the main challenges and opportunities concerning AI bias as perceived by those involved in AI projects?
What is the role that recent regulations and guidelines have inside the AI lifecycle phases?

The present research is organized as follows: First, the concept of AI bias is briefly introduced, followed by a concise literature review that provides context and motivation for our study and informs our sample selection.

Subsequently, the research protocol is outlined. The final sections present the results obtained from the preliminary data analysis and discuss the future directions of our research.

Theoretical Background

Investigation into AI bias is well-established; however, until recently, most inquiries have predominantly approached the issue from a technical standpoint. Recently, due to the increasing deployment of AI applications, practitioners and scholars began to underline the complexity of the phenomenon and the necessity to include behavioural, social, and ethical aspects in its examination (Mehrabi et al., 2021; Smacchia & Za, 2023; Someh et al., 2019). Similar to the bias definition, AI bias may be defined as a deviation in the results produced by a specific artefact from the ground truth, which disproportionately benefits or disadvantages certain groups or individuals over others (Mehrabi et al., 2021; Akter et al., 2021; Van Giffen et al., 2022). While this definition is commonly accepted in the literature, conceptualising AI bias is problematic, at least in some fields, due to the complexities (e.g., historical, geographical, ideological) involved in establishing a shared ground truth (Brickell, 2006). Many studies provide a definition along with a division of different kinds of bias and possible mitigation strategies. Kordzadeh and Ghasemaghaei, (2022) developed a framework based on eight theoretical concepts. They argue that algorithmic bias affects perceived fairness and, hence, the behavioural response of the users. This relation is influenced by contextual dimensions that concur with the emergence of bias. For example, individual and organisational factors could contribute to the possible under-representation of a determined characteristic in the training data. Van Giffen et al., (2022) conducted a literature review that identified eight groups of bias in machine learning. Based on this distinction, they provide several mitigation strategies by contextualising them inside real cases. Similarly, Mehrabi et al., (2021) differentiated among various forms of bias, outlining corresponding strategies for mitigation. These biases are categorised based on context, data, algorithm, and user interaction, which are interconnected in a continuous loop, perpetually exerting influence on one another. Another categorisation is provided in the realm of digital-driven innovation processes, in which (Akter et al., 2021) recognise three groups of bias, those deriving from the data (e.g., under-representation of the sample, inadequate data cleaning), those due to the used method (e.g., spurious correlations, flawed choice of the algorithm) and those coming from the society related to historical and institutional behaviour. The US National Institute of Standards and Technology (NIST) also offers

significant input by supplying a document aimed at offering recommendations for mitigating AI biases using a socio-technical perspective, in which multiple aspects are taken into consideration. Furthermore, they emphasise the necessity of a human-centric approach and the crucial role of involving domain experts in the design and development phases of AI artefacts (Schwartz et al., 2022). In their contribution, they provide an interesting division of AI bias into three different categories:

- Systemic bias: customs and beliefs surrounding a particular institution that may influence how AI is designed and developed (e.g., social/historical/institutional bias).
- Computational bias: issues in the underlying data/methods that affect the performance of the outcomes of an IA artefact (e.g., measurement/representation/label/feedback bias).
- Human bias: Human perceptions and heuristics impacting the use of an AI artefact (e.g., deployment/anchoring/availability/confirmation bias).

The three kinds of bias are deeply interconnected. While heuristics serve to expedite decision-making and thus reduce complexity, they can also lead to biased outcomes (Tversky & Kahneman, 1974). Phenomena like anchoring, availability, and confirmation bias stem from human ignorance, previous experiences, or convictions and are entwined with institutional, societal, and historical biases. Indeed, the influence is bidirectional: individuals shape society, which, in turn, affects individual behaviours (Yela, 2004; Banaji et al., 2021). Bias coming from individuals and society inevitably influences the way data, algorithms, and development procedures are carried out, creating possible computational or statistical issues (Kordzadeh & Ghasemaghaei, 2022; Schwartz et al., 2022).

AI regulations and guidelines

The discourse surrounding AI bias and the necessity for AI artefacts to adhere to ethical and fairness standards has gained significant relevance in governmental and institutional domains (Golpayegani et al., 2022; Smuha, 2021). Numerous governments and institutions are actively developing or have already established guidelines, frameworks, and regulations aimed at enhancing the design, development, and deployment processes of AI, with a strong emphasis on risk management and fairness compliance (Golpayegani et al., 2022; Novelli et al., 2023). From a governmental perspective, there has been a marked increase in the number of states adopting comprehensive or use-case-specific AI regulations and guidelines, rising from 25 in 2022 to 127 in 2023. For

example, the United States has introduced significant documents and frameworks to address issues such as discrimination, safety measures, and human rights in AI[15]. For instance, the National Institute of Standards and Technology (NIST) proposed the AI Risk Management Framework, which provides a four-step guideline for managing information security and privacy risks. This framework is relevant also for its adherence to the most recent U.S. laws and directives[16]. Additionally, the U.S. government has proposed the Blueprint for an AI Bill of Rights, which outlines five key principles focused on safety, privacy, rights, and awareness to address AI-related issues[17]. In December 2023, the European Parliament approved the AI Act, establishing the first comprehensive directive on AI within the European Union. This foundational legislation addresses potential issues arising from the rapid growth in AI use. The AI Act implements stringent restrictions and, in some cases, prohibitions on specific applications of AI, such as biometric identification, social scoring, and law enforcement activities. It categorizes AI applications based on varying levels of risk and delineates the responsibilities and obligations of AI providers (Novelli et al., 2023). Additionally, the Act specifies the documentation requirements for general-purpose AI systems, which can be utilized across multiple applications[18].

These regulations and frameworks are supported and complemented by international guidelines and recommendations that offer comprehensive direction on the execution of processes throughout the AI lifecycle. UNESCO's "Recommendation on the Ethics of Artificial Intelligence" outlines a comprehensive framework centred on four core principles: human rights and dignity, peace, diversity and inclusiveness, and environmental sustainability. This document emphasizes fundamental ethical principles essential for the responsible development of AI. Additionally, it offers policymakers guidance for AI regulation and provides implementation strategies for companies to ensure ethical AI practices[19]. The International Standards Organization (ISO) has intensified its focus on AI and the mitigation of AI bias. In 2022, ISO published a

[15] European Parlament, United States approach to artificial intelligence, retrieved June 3rd 2024 from: https://www.europarl.europa.eu/RegData/etudes/ATAG/2024/757605/EPRS_ATA(2024)757605_EN.pdf
[16] NIST, AI RMF Playbook, retrieved June 3rd 2024 from: https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook
[17] The White House Office of Science, Blueprint for an AI Bill of Rights, retrieved June 3rd 2024 from: https://www.whitehouse.gov/ostp/ai-bill-of-rights/
[18] EU Artificial Intelligence Act, High level summary of the AI Act, retrieved June 4th 2024 from: https://artificialintelligenceact.eu/high-level-summary/
[19] UNESCO, Ethics of artificial intelligence, retrieved June 3rd 2024 from: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

comprehensive overview of AI ethical and social concerns (ISO/IEC 24368:2022), encapsulating the key concepts underlying these issues. Subsequently, in 2023, ISO proposed the first certification for AI management systems. This certification focuses on each step of the AI lifecycle, providing detailed guidelines for the proper management of processes to ensure the responsible development and deployment of AI systems[20].

Methodology

To explore the processes concerning how AI bias is identified and mitigated across the AI lifecycle, and to craft a theory that describes and explains these methods, we employ a case-based approach, utilising grounded theory as the foundation of our research (Corbin & Strauss, 1990). The AI lifecycle represents the progression of an AI artefact through three different phases: design, development, and deployment, and it could be applicable to all AI solutions (D. De Silva & Alahakoon, 2022). In performing our research using a grounded approach, we adopt the research protocol described by Pozzebon et al., (2011), consisting of a five-stage analysis.

The first phase concerns the design of the research. According to the guidelines of Corbin and Strauss, (1990), we performed a literature review not with the objective of identifying a theoretical framework but to provide orientation concerning the phenomenon. This review served a dual purpose: it highlighted the significance and motivation behind our study, and it helped in our selection of cases to examine. The literature emphasised the value of a socio-technical viewpoint when investigating AI bias, noting the significance of individual, organisational, and environmental attributes in AI development. Consequently, rather than focusing on determined companies, our sample is composed of a diverse group of AI experts who have been involved in the phases of the AI lifecycle, including the design, development, and deployment phases (D. De Silva & Alahakoon, 2022). Our methodology prioritised diversity in participant selection, recognising the value of encompassing a range of backgrounds and organisational roles to capture a broad spectrum of perspectives and insights on the phenomenon under study. To capture a variety of professional viewpoints, we interviewed individuals from different roles within their organisations. This study serves as an initial investigation, in which we conducted preliminary interviews with AI experts actively engaged in both industrial and academic spheres.

[20] ISO, ISO/IEC 42001:2023, retrieved June 3rd 2024 from: https://www.iso.org/standard/81230.html

These initial discussions were necessary for refining and validating our interview questions, as well as providing an early exploration of the field, laying the groundwork for more comprehensive research. In the second phase of our study, we developed a research protocol concerning data collection and initial analysis. Opting for semi-structured interviews, we aimed to steer conversations with participants while allowing for flexibility in their responses. In particular, we performed five interviews that lasted, on average, 53 minutes. Table 1, provides an additional explanation about our preliminary sample.

Table 1. Information about the selected sample for the preliminary interviews.

Role	Interview length	AI product
Founder/Chief research officer	40 min	Machine translation tool based on ML
Founder/Designer	55 min	Tool for detecting dyslexia (B-LSTMs)
Founder/External role	65 min	Computer vision for damage detection
Founder/Chief Executive Officer	55 min	Computer vision tool for cashier-less shops
Employee/Cloud architecture	45 min	Unstructured to structured data converter

Our interview framework was divided into five distinct segments: the first part consisted of general inquiries about the respondents' projects and AI product details. The subsequent three segments were dedicated to exploring each phase of the AI product lifecycle—design, development, and deployment. The final segment delved into future challenges, the significance of various types of AI bias, and the importance of addressing them. Even though some participants were more focused on a specific phase of the AI lifecycle, we still asked everyone the same questions to avoid biases and to gather as many insights as possible. After the first interview iteration, we modified the interview script by refining some existing questions and adding a sixth segment concerning the impact of AI regulation and guidelines. During the third and fourth phases, we organised our data, categorising the interviews according to the interviewee's backgrounds and their years of experience in the AI field. Selecting AI experts from varied disciplines and with differing experience levels could improve the interviews' outcomes by providing diverse viewpoints, ensuring a balance between theoretical and practical insights, and facilitating the investigation of a complex and multifaced field such as AI. This will be followed by the data analysis process, conducted in three stages of coding: open, axial, and selective. In the open coding stage, we will analyse the text, coding segments based on emergent concepts directly derived from the data, avoiding any preconceived notions. These labels will then be grouped into broader categories through an iterative process involving frequent discussions among the authors. The axial coding phase will be centred on delineating the relationships between these broader categories and the more detailed labels. At this stage, we will try to develop a theoretical framework based on the collected data. Finally, in the selective coding stage,

we will iterate on another round of coding. This time, it will be conducted through the lens of the concepts and models developed in the preceding phases, allowing for a more nuanced and theoretically informed analysis. In the last phase, we will compare the obtained results with the literature. This comparison is aimed at identifying commonalities that would enable us to refine our theoretical model. By aligning our results with established scholarly work, we seek to ensure that our model is both grounded in empirical data and resonant with the broader academic discourse. Up to this point, we have started the open coding process based on the interviews we conducted. In the next section, we will present and examine the preliminary results that have emerged from this data. Furthermore, we will outline our strategy for continuing and completing our research. In Figure 1 we provide a graphical representation of the five-step methodology we have followed.

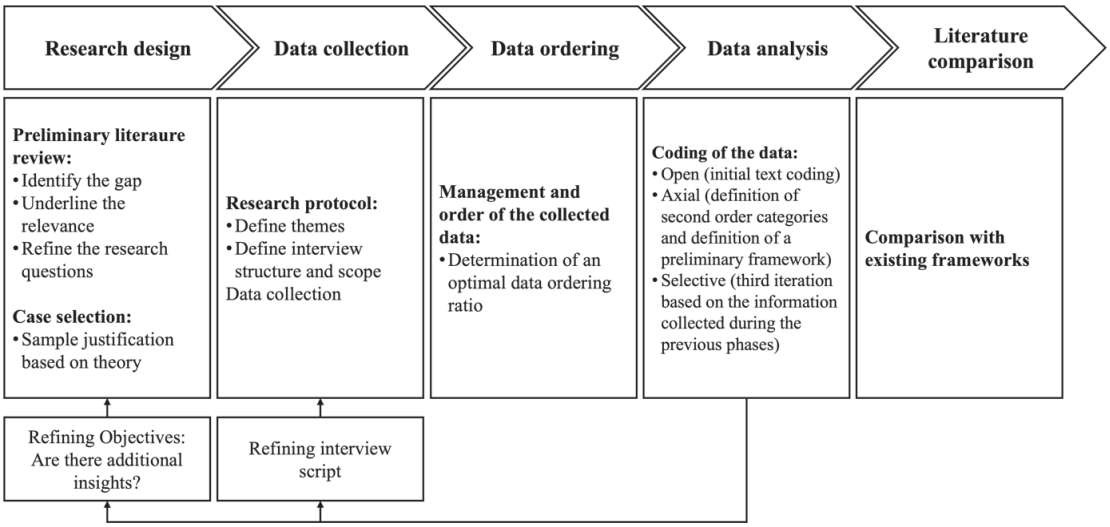

Figure 1. Graphical representation of the five-step methodology adopted

Results and discussion

The interview data provided valuable insights into the various stages of the AI lifecycle, particularly highlighting challenges in bias emergence and mitigation. In Figure 2, starting from the framework developed by NIST, we also represented the interconnection between the different kinds of AI bias that we encountered in the interviews. For each of them, we also provided some examples useful to better understand their nature. A critical issue identified during the design phase relates to data exploration and collection. Challenges here are multifaceted, often stemming from the need to review state-of-the-art literature and select appropriate datasets and models for artefact development. A significant concern arises when developers lack control over underlying data, potentially leading to biases. For example, using pre-trained AI models can unintentionally

embed bias, as they're often built on broad datasets lacking in diversity. In such cases, communicating these issues to data owners is important, alongside disclaiming potential biases in training datasets within the artefact's documentation. Another notable challenge is the limited availability of data. This issue is particularly acute for smaller firms, where the lack of partnerships or understanding with data proprietors can hinder access to comprehensive and representative datasets. While collecting additional, but not perfectly representative, data can aid in refining the training of artefacts, there is a risk of misalignment with the software's intended purpose, thereby introducing biased results. This aspect underscores the need for a more nuanced approach to data collection and usage in AI development, ensuring alignment with the intended application and mitigating bias risks. Another pivotal challenge in design lies in the imperative for a comprehensive grasp of environmental factors, extending beyond the identification of needs. Often, cognitive heuristics like availability bias may constrain the knowledge base of developers, leading to the oversight of critical elements in the design of AI artefacts. This oversight can lead to project failures. Consequently, the diversity in team backgrounds and cultures becomes crucial, not just beneficial, for the effective and accurate design of AI products, ensuring a more holistic and informed approach that accounts for a wider range of influencing factors. "*But the way to get into machine learning product, if you don't have the data, you need to be creative and get it from somewhere*".

In the development phase, the significance of ongoing feedback iterations, both within the internal team and between the team and the customer (typical of agile methodology) became evident. Challenges encountered in this phase are predominantly technical, encompassing tasks such as data correction to balance datasets and the application of advanced techniques to enhance model precision and efficiency or to verify the model's accuracy. A critical aspect of this phase is the continuous testing of the model in collaboration with the customer. This involves proactive consideration of potential biases, researching and applying cutting-edge metrics to detect any bias, and utilising diverse benchmarks for consistent comparison of outcomes. Such rigorous and collaborative testing is of paramount importance for ensuring the model's robustness and alignment with user needs and expectations. During this phase, we identified an additional type of bias that we did not take into account previously. This bias can originate from the hardware used to operate the artefact. For example, a bias may stem from the graphics card drivers employed in accelerating deep learning models. If one brand dominates the graphics card market, there's a risk that many models will be optimised exclusively

for that hardware, resulting in suboptimal performance or leading to possible errors when the hardware of another manufacturer is used. Further complications may arise when an AI tool is deployed across different platforms (e.g. tablet, PC, smartphone), each with its unique interaction modalities. A bias is likely to manifest if interacting with a particular platform proves more challenging, thereby disadvantaging the end user. A fitting example could be represented by an AI application designed for keyboard interaction rather than touchscreens. In scenarios where tablets or smartphones are utilized, the user experience and accessibility of the application will undoubtedly be diminished. This highlights the importance of considering hardware diversity and cross-platform compatibility in AI development to mitigate such biases. Another significant risk to consider in the development phase is the potential pitfalls of training AI models with data generated by other AI systems. This approach can lead to a perilous feedback loop where existing biases are not only perpetuated but potentially amplified. Such a loop can cause the model to reinforce and escalate any biases present in the initial AI-generated data, creating a cycle that continuously feeds and exacerbates these biases. "*You need to up very good metrics, set yourself up for some benchmarks or where you want to achieve. And those metrics should be compatible to what the customer is expecting*".

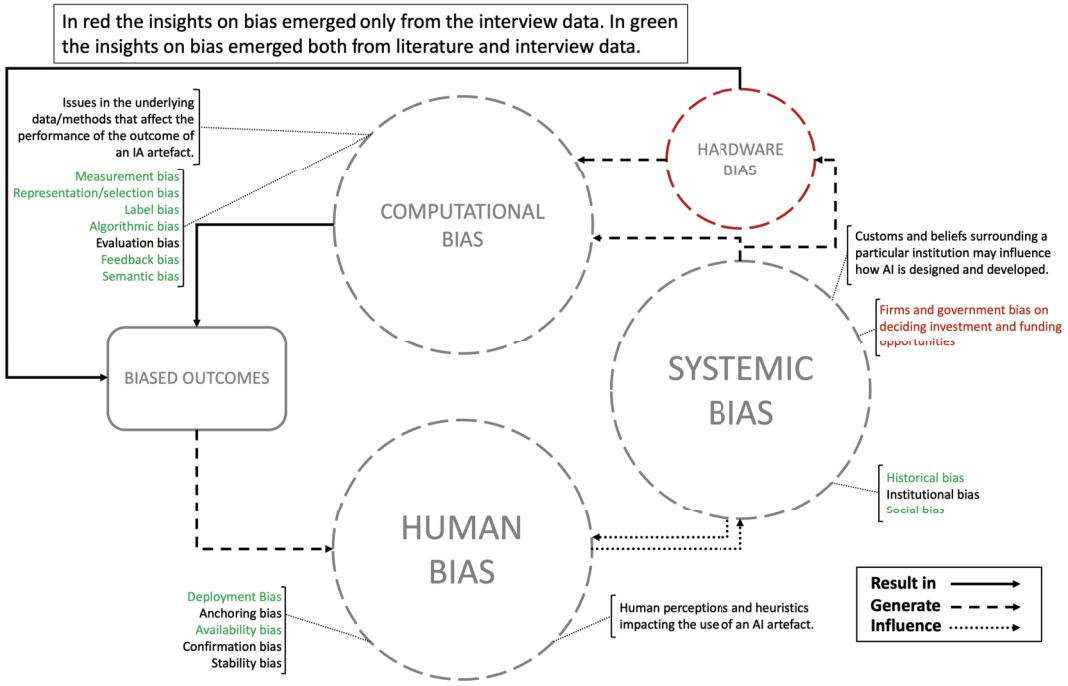

Figure 2. AI bias categories intersection and typologies

In the deployment phase, interviewees unanimously underscored the criticality of field testing the artefact in real-world scenarios, as it is in these contexts that hidden biases are most likely to surface. It is relevant for customers to actively report any issues with the artefact, although this isn't guaranteed, particularly with free solutions where users may tolerate certain flaws. Beyond engaging with end-users, initiating dialogues with other stakeholders involved in the project is crucial to understanding and addressing the root causes of problems in the AI artefacts. Many biases identified during this phase often align with cognitive biases impacting the training data. For instance, anchoring or availability bias might lead developers to neglect relevant data in model training, resulting in subpar performance in specific situations. Additionally, engaging in conversations with stakeholders other than the end-users is significant, as their influence can significantly impact decision-making processes. If these stakeholders are not adequately convinced of the AI solution's efficacy, it could be prone to failure due to resistance, highlighting the importance of broad stakeholder engagement in ensuring the successful deployment and adoption of AI solutions. Finally, the preliminary results highlighted issues concerning transparency in AI. Despite being a major concern among researchers and practitioners in the AI field, achieving full explainability of algorithms and completely eradicating bias remains a challenging endeavour. Consequently, a key responsibility of developers extends beyond technical development; it involves educating the public about the inherent risks of bias in AI systems and enhancing their understanding of what AI truly entails. This educational role is crucial in fostering a more informed and critical perspective among users, helping them to navigate and interact with AI technologies more effectively and responsibly. "*So, the most difficult thing to answer your question was basically to create a product that added value to each of these stakeholders. I mean, basically, to put all these people thinking in the same way or to make all these people agree was more difficult than doing the research itself, really*". In Table 2, we describe our interview script, highlighting the macro themes and the ratio behind them.

Table 2. Description of the interview protocol

General Theme	Description
Introduction & focus on the context	In this section, we first outline our research objectives and describe the methodology used for conducting the interview. Following this, we detail the interview structure, beginning with "ice-breaking" questions designed to gather general information about the interviewee's position, their years of experience with AI, and their tenure at the current company. Subsequently, we delve into specific questions regarding a particular AI solution the interviewee is involved with, which will serve as the primary focus of the interview.
Design	This section of the interview is dedicated to the design process in the AI lifecycle. If it applies, we ask questions about the main processes, the actors involved, the management of the communication, the main challenges how they are addressed.

Table 2. (Continued)

Development	After shifting to the next section, we proceed, if applicable, to ask questions similar to those in the design section. This section also includes inquiries about the metrics used to test and monitor the AI solution, ensuring a thorough understanding of its performance and reliability.
Deployment	This section delves into the post-deployment phase of the AI solution, examining what emerges once it is operational outside the company. In addition to exploring the main challenges and strategies for managing potential biases, this segment places a strong emphasis on the relationship with customers and their critical role in addressing and resolving issues that may arise (if applicable with the background and role of the respondent).
Regulation & Guidelines	After conducting the preliminary interviews, it became evident that regulations and guidelines, particularly the AI Act in the EU, are becoming increasingly important in the AI landscape. Consequently, we developed a new section to focus on the impact that both compulsory and voluntary laws and frameworks have on the processes throughout the AI lifecycle. This section examines how these regulations influence the design, development, deployment of AI solutions, seeking to gather insights from experts on the subject.
Future outlook	The final section concerned mostly general questions about AI bias and the future challenges that will arise in the AI domain. Before ending the interview, we usually ask if the respondent has some additional point to highlight.

Future work and expected results

In line with the approach detailed in our methodology section, we are presently involved in the data-gathering stage of our research. Up until now, we have successfully conducted five preliminary interviews. These initial discussions have been instrumental in refining our inquiry questions, gathering initial feedback on our representation of AI bias (Figure 2), and acquiring some preliminary insights. This ongoing expansion of our interview pool is a crucial step in enriching our understanding and ensuring a robust and diverse collection of perspectives on the subject. Upon completing the collection and systematic organisation of our data, we will embark on a detailed and iterative analysis through a threefold coding approach. Currently, we have initiated this process with open coding, which has yielded preliminary insights. Moreover, after the preliminary data analysis, we discovered interesting insights that allowed us to refine our interview protocol. More specifically, we adjusted some questions concerning the role of AI and the possible implications of the hardware used during the development of AI artefacts. Additionally, we expanded the discourse concerning AI regulations and guidelines to discover the impact that recent policies have on the AI lifecycle processes. We plan to collect data concerning these aspects in future interviews.

Our research objectives are multifaceted and aimed at advancing the field of AI bias. Firstly, we seek to gain a more exhaustive understanding of the AI bias domain. This involves refining the existing classification proposed by NIST, and that was further adapted (Figure 2), to better capture the nuances and complexities inherent in this area of study. Secondly, we are committed to identifying and suggesting potential directions for future research within this domain. This aspect of our work is crucial in guiding subsequent

academic inquiries and practical applications in addressing AI bias. Lastly, our goal is to develop a comprehensive framework that not only outlines the processes for identifying and mitigating AI bias but also incorporates the key stakeholders involved across various stages of the AI lifecycle. This framework shall encompass a wide array of viewpoints from individuals possessing varied experiences and backgrounds, especially those engaged in the development and implementation of diverse AI-based products and services. We posit that this inclusive methodology is imperative for acquiring a comprehensive and multifaceted comprehension of the phenomena associated with AI bias.

This research could have relevance in both theoretical and practical realms. Theoretically, the study is aimed at enhancing the existing body of literature by offering an in-depth and expansive analysis of AI bias, particularly focusing on its identification and mitigation processes. This analysis is grounded in empirical data, thereby enriching the theoretical understanding of the subject. Furthermore, the research is designed to propose new directions for future scholarly exploration in this area, potentially guiding subsequent academic endeavours. From a practical standpoint, the study is poised to benefit a range of professionals involved in the AI lifecycle. It aims to provide these practitioners with possible insights into the emergence and mitigation of bias within various stages of AI development, insights that are rooted in empirical findings gathered from field experts. Additionally, the adoption of a socio-technical perspective in our research is expected to raise awareness among practitioners and policymakers about the significance of addressing societal biases, which are often neglected in technical discussions (Schwartz et al., 2022; Someh et al., 2019).

CHAPTER IV

Does AI reflect Human Behaviour? Exploring the Presence of Gender Bias in AI Translation tools

Abstract. Natural language processing tools are becoming more and more important in our daily lives, enabling us to perform many tasks in a timely and efficient manner. However, as the utilisation of these tools grows, so does the risk of unexpected consequences due to the presence of bias. This study investigates the presence of gender bias within the most popular neural machine translation and large language model tools. We defined a set of Italian sentences concerning ten specific jobs where the gender of the subjects is not explicitly mentioned. Employing those AI tools, we translated the sentences from Italian to English, requiring the gender to be explicitly mentioned. Afterwards, we developed a survey to obtain human translations for the same sentences, allowing us to compare the differences between the responses generated by the tools and those from individuals. Results show a high presence of gender bias, especially for the jobs associated with a male gender, and demonstrate a consistency between the outcome obtained by the tools and the results of the survey. These findings serve as a starting point for exploring the origins of gender bias within natural language processing tools and how they reflect gender distributions in our society and human behaviour regarding job occupations.

Keywords: Natural Language Processing, Neural Machine Translation, Large Language Models, Gender Bias, Artificial Intelligence Bias.

Introduction

In the last few years, artificial intelligence (AI) has gained much attention from both scholars and practitioners, with a particular focus on societal aspects (Smacchia & Za, 2022). Within this context, natural language processing (NLP) tools, such as chatbots, translators and speech recognition, have witnessed a surge in popularity because they allow users to interact without using machine language (Khurana et al., 2023). NLP, a subfield of AI, encompasses the development of machine algorithms capable of understanding human language, consisting of two main branches: natural language understanding and natural language generation (Khurana et al., 2023; Young et al., 2017). The first branch is related to the study of linguistics and to the understanding of human language, while the second is focused on the process of producing such language

(Khurana et al., 2023). NLP finds application in various tasks, such as automatic summarization, discourse analysis, speech recognition, machine translations, and many others (ibid). In the last few years, this technology gained popularity due to the introduction of neural networks and large language models (LLM) that are able to solve many NLP tasks without having specific training data for a given job (zero-shot tasks) (Chowdhery et al., 2022; C. Qin et al., 2023; Stahlberg, 2020). Machine translation (MT) is one of the NLP fields significantly impacted by this transition (Stahlberg, 2020).

MT allows a machine to translate natural language sentences, providing an instant translation of words, sentences, documents and even speeches (Tan et al., 2020; Vieira et al., 2021). In the past, MT primarily relied on a set of predetermined translation rules and linguistic knowledge. However, nowadays, state-of-the-art MT systems implement neural networks that use a training dataset of parallel texts to translate an input sentence to a target output by using the information acquired during its training (Castilho et al., 2017). MT serve as a useful tool that simplifies the way we communicate with people who speak other languages by providing large volumes of translated text or speech in a few seconds (Doherty, 2016). As the industry value continues to grow, so does the number of tools available in the market that are used worldwide by companies and individuals for many tasks[21, 22]. Since this technology is broadly used, unexpected results such as mistranslations or biases could have severe implications. Vieira et al., (2021) conducted a review in which they explored MT bias across legal and medical fields. They reported interesting examples, such as the low accuracy of Google Translate for translating medical instructions to patients suffering from diabetes. Additionally, they discuss an incident in which Google Translate was used by a police officer to obtain consent during inspections, which later faced challenges during the court audience due to concerns about overcoming language barriers. In both cases, the translation tools outcomes can cause serious damage to the individuals involved, underscoring the paramount importance of addressing these kinds of issues.

In this study, we focus on examining the presence of gender bias within AI translation tools. Specifically, we selected five jobs mostly done by women and five jobs mostly done by men (on the basis of

[21] Global Market Insights. (April 6, 2017). Machine translation market size worldwide, from 2016 to 2024 (in million U.S. dollars) [Graph]. In Statista. Retrieved May 14th, 2023, from: https://www.statista.com/statistics/748358/worldwide-machine-translation-market-size/.

[22] Stanford University. (April 15, 2023). Number of commercial machine translation (MT) programs available globally from 2017 to 2022 [Graph]. In Statista. Retrieved May 14th, 2023, from: https://www.statista.com/statistics/1378793/mt-translation-programs-number/.

relevant reports). We then constructed sentences involving these jobs and translated them from Italian to English using the API of five different AI translation tools (Google Translate, Deepl, Microsoft Azure, GPT 3.5 and Text DaVinci). All the original sentences in Italian did not include a subject (the gender was not made explicit). Since in Italian, it is possible to omit the subject in a sentence, our experiment aims at measuring the gender bias in English translations where, on the other hand, the subject must always be made explicit. To compare the results obtained from an AI with those of human beings, we also developed a questionnaire that we submitted to a sample of Italian individuals.

The purpose of this study is to address the limitations identified in the literature by investigating the presence of gender bias in the outcome of five different AI tools when sentences without subject are presented. More specifically, we aim to answer the following questions:

To what extent does gender bias occur inside the most common AI translation tools?

Related to gender bias, is there a difference in the output of a translation depending on how the query is formulated?

Concerning the translation output, which jobs are most likely to be biased?

What are the results of a comparison between the results obtained from an AI with those obtained by Human beings?

The contribution of this study is structured as follows: In the next section, we provide a theoretical background about the gender bias inside MT, discuss the limitations detected, and outline the methodology we used to perform our research. Subsequently, we present the results obtained and discuss them. Finally, in the conclusion, we state the theoretical and practical contribution, along with the limitations and avenues for future research.

Theoretical Background

Inside the AI debate, research focusing on identifying, investigating, and addressing biases has become increasingly relevant (Smacchia & Za, 2023). Although AI and its subfields bring enormous potentialities and advantages in many domains, the presence of biases could bring dangerous and unpleasant consequences (Van Giffen et al., 2022). Since AI is used worldwide by private, public and governmental organisations to support the decision-making process of many activities, its outcome influences many people simultaneously

(Zuiderwijk et al., 2021b). Bias inside data and resources used to train the AI model or inside the algorithm could cause and amplify issues such as discrimination, inequality, and lack of privacy and trust (Cathy O'Neil, 2016; Hoffmann, 2019; Mehrabi et al., 2021; Sun et al., 2020; Van Giffen et al., 2022).

Among the possible biased outcomes concerning AI technology, the presence of gender bias inside neural MT tools has a significant relevance (Bernagozzi et al., 2021; Stanovsky et al., 2019; Tomalin et al., 2021). Gender bias occurs when one gender is preferred or prejudiced with respect to the others. NLP algorithms or training datasets on which machine translation tools are based could reinforce or amplify damaging gender stereotypes (Sun et al., 2020). There are many real-world cases related to this issue, from resume filtering giving preference to male candidates to credit applications refused to women even if they have a higher score than a male applicant (Kelley et al., 2022; Sun et al., 2020).

Gender bias in MT tools based on AI could be attributed to different factors. For example, behaviours or habits embedded in society described by the training data could affect the outcome of a translator. Technical issues in the data or the algorithm used to develop the artefact could lead to some forms of gender bias. Finally, since AI tools are capable of learning, different use in various contexts could influence the behaviour of the MT (Savoldi et al., 2021).

The presence of gender bias inside neural MT tools is not a new topic. There are many press articles that underline sexist behaviour in online translators[23, 24]. In addition, several articles also tried to inspect and address this issue. (Stanovsky et al., 2019) propose an evaluation method for gender bias inside MT tools based on morphological analysis. They analysed six MT tools, both industrial and academic, and detected the presence of translation errors related to gender. Another study focuses on reducing gender bias by training a neural MT tool and applying several automated techniques to address the issue. Results indicate that when these techniques are used to reduce the gender bias in a neural MT tool, the performances are worse with respect to the baseline system. Among the methods used, only domain adaptation provides better results after a model is trained (Tomalin et al., 2021). (Bernagozzi et al., 2021), used a method proposed by (Srivastava &

[23] Parmy Olson, The Algorithm That Helped Google Translate Become Sexist, Forbes. Retrieved May 14th, 2023 from: at: https://www.forbes.com/sites/parmyolson/2018/02/15/the-algorithm-that-helped-google-translate-become-sexist/.
[24] Shivali Best, Is Google Translate SEXIST? Users report biased results when translating gender-neutral languages into English, Daily Mail. Retrieved May 14th, 2023 from: https://www.dailymail.co.uk/sciencetech/article-5136607/Is-Google-Translate-SEXIST.html

Rossi, 2018) to evaluate the presence of gender bias in MT tools that could be used by third parties. Moreover, through a questionnaire, they inspect the impact that such biases have on people and how this issue is perceived. (Schiebinger, 2014) underlines the necessity to analyse gender bias inside MT tools as she notes their trend to translate with a default masculine form.

Although this phenomenon has been studied several times, there are some aspects that still need to be addressed. The language source of multiple datasets used to perform the research is English (Bernagozzi et al., 2021; Stanovsky et al., 2019). Many studies have not considered the fact that in some languages sentences without a subject are valid while, in other languages, the subject must be outlined (Srivastava & Rossi, 2018). Moreover, some contributions use a limited number of tools to make comparisons and simple sentences (Bernagozzi et al., 2021; Srivastava & Rossi, 2018). To carry out the research, we adapted the framework of (Srivastava & Rossi, 2018) to assess the presence of gender bias inside MT tools without having access to their source code or training data.

In the next section, we describe the process adopted to conduct our research. Starting from studies on gendered language concerning the reinforcement of gender stereotypes in job titles (Bigler & Leaper, 2015; Liben et al., 2002) and from statistics about the occupations mostly performed by a specific gender, we developed a set of sentences to translate. Every phrase is related to a specific job that could be used to identify both men and women (gendered-language free). We avoided referring to occupations that could be related specifically to masculine or feminine forms to avoid possible bias due to gendered language.

Methodology

This contribution adopts an exploratory design as the main aim is to seek new insights into the evaluation of gender bias inside AI translation tools (Makri & Neely, 2021). Our primary goal was to assess the presence of bias in translators when a sentence with no subject (a common practice in some languages) is translated into a sentence in which the subject must be mandatorily made explicit. To this end, we have developed a set of sentences in Italian in which the subject has not been made explicit and which can be attributed to either a male or a female, and then we translated them into English. Each sentence focuses on ten particular jobs, which

we chose by searching within Eurostat databases[25, 26], in World Bank Group[27] and U.S. Bureau of Labor Statistics[28] reports. More specifically, we selected the five jobs mostly done by women and just as many held mostly by men. In order to be selected, the name of the job had to be impersonal in Italian (not attributable to a male or female audience). We developed three sentences for each job; the first one was quite simple, while the other two were more elaborate. Since all sentences have been translated into English, we arranged for the first sentence to contain only personal pronouns, while the other two contained both personal pronouns and possessive adjectives. As an example, the following three sentences are those already translated in English and referring to carer (in bold the words concerning the gender):

- She said she works as a carer in the home of an elderly person.
- She started her job as a carer after an elderly person requested assistance.
- After losing her job as a carer, she thought of changing life.

In the next step, we selected the tools used to perform our study. We chose them according to two parameters: the popularity of the translator and the availability of the APIs. After a screening, we chose three of the most used neural MT tools in the market: Google Translate, DeepL, and Microsoft Azure. Since LLM models are becoming more and more relevant inside the debate and are also trained on much bigger datasets than translators, we decided to add to the current list, the models GPT 3.5 turbo and Text-DaVinci 003 on which Chat GPT is based (King, 2023; C. Qin et al., 2023). We referred to the API documentation of these MT tools to write a Python script in which we translated the 30 sentences in three different ways, first, we translated the entire text corpus (all sentences at once), then we translated all the corpus from an external document and finally, we translated the sentences one by one. The results, consisting of 90 translations per tool, were listed in an Excel sheet and labelled according to the gender that was made explicit in the English translation (M for male, F for female and N for neutral) and job. Depending on the output of the translation we classified the

[25] European Jobs Monitor 2021: Gender gaps and the employment structure. Retrieved May 16th, 2023: https://www.aranagenzia.it/attachments/article/12440/European%20Jobs%20Monitor%202021-Gender%20gaps%20and%20the%20employment%20structure.pdf

[26] Jobs still split along gender lines. Retrieved May 16th, 2023: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/EDN-20180307-1

[27] Gender-based Employment Segregation: Understanding Causes and Policy Interventions. Retrieved May 16th, 2023: https://documents1.worldbank.org/curated/en/483621554129720460/pdf/Gender-Based-Employment-Segregation-Understanding-Causes-and-Policy-Interventions.pdf

[28] Labor Force Statistics from the Current Population Survey. Retrieved May 16th, 2023: https://www.bls.gov/cps/cpsaat39.htm.

results as biased when the output reflects the underlying data used to develop the sentences, converging biased when the output is influenced by the previous translations (underlying data) and converges towards the same result and unbiased when the output is translated using in a neutral way or the translation path doesn't reflect a biased behaviour.

We then analysed the results measuring the presence of bias in the translation, considering each tool individually and in aggregated form. Then, we compared the results with those obtained from the questionnaire for each job. Lastly, we considered the translation trend between the different responses to the last question concerning the aim of the study.

To better explain the processes on which the research is based, we graphically represented our research protocol (Figure 1) that we adapted from the model of (Srivastava & Rossi, 2018) to assess the presence of bias inside MT tools by third parties. Moreover, to assess the presence of bias, we take into account the statistics concerning the gender gap inside the selected jobs. To make the model clearer and more reproducible, we used the business process modelling notation (BPMN).

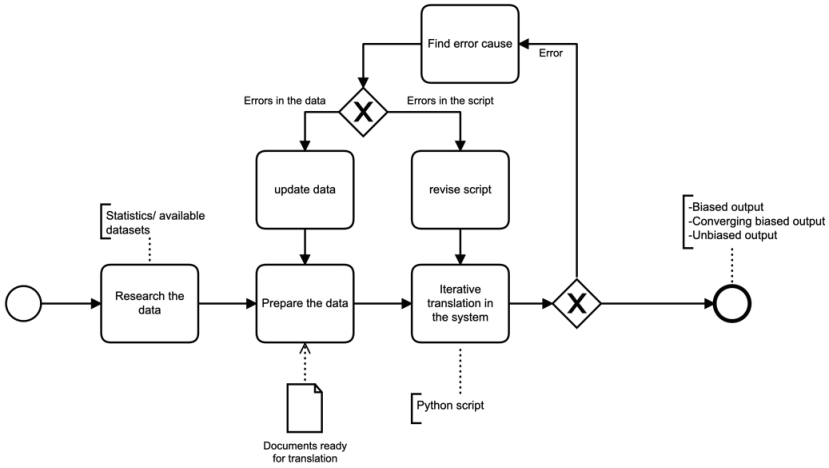

Figure 1. Research protocol

In this study, we also developed a questionnaire to compare the results obtained by AI tools with those obtained by humans. More specifically, our aim is to understand whether gender bias is still present in the way people think in our society.

The questionnaire reported the 30 sentences in Italian translated into English, to which all references regarding gender (pronouns and adjectives) were appropriately removed. The respondents were asked to complete the sentences and give them a proper meaning (fill-in-the-blank questions). All the sentences were

presented in the same order they were submitted to the AI tools. At the beginning and at the end of the form, respondents were also asked demographic questions. The first question was related to the English level to assess the familiarity of the respondents with the language. After completing the 30 questions related to the translated sentences, respondents were asked the gender they identify with, their Italian region of origin, their age, and their educational qualification. All the demographic questions, although mandatory, had the option "I prefer not to answer" or "other". Finally, we added a further question asking the respondent to explain in a few words what was, in her/his opinion, the main aim of this study. After the questionnaire was completed, we provided information about the objective of the research.

We decided to use the fill-in-the-blank structure for various reasons. First, letting people translate all the sentences would have gone beyond our purposes, leading to possible errors and a difficult analysis of the outcomes, caused by an excessive amount of noise within the dataset. Second, we wanted to make the questionnaire available to people who are not proficient in the English language, reducing the possibility of mistakes due to their lack of expertise.

We submitted the questionnaire to a random sample of 100 Italian speakers. The respondents were 53. Of these responses, 11 were discarded because they were inconsistent with what was required within the questionnaire. The data concerning the 42 questionnaires, was collected from late April to early June 2023. To develop the survey, we used the tool Jotform. Further information about the sample is discussed in paragraph 4.2.

Results

In this section we present the results obtained by the analysis conducted on the translation tools first and then the outcome of the questionnaire.

Analysis of AI translation tools

We analyse the results given by the MT tools considering first the outcome of every single tool, concerning the presence of bias, taking into account the statistics that we used to select the jobs to develop the sentences (Figure 2). For each AI tool, we provide the percentage of female, male and neutral translations both in aggregate (statistics for the five jobs done mostly by men and by women) and individual form (statistics for every single job selected). Afterwards, we summarise those results in a combined diagram.

Microsoft Azure. The tool is consistent between the three methods of translation. We obtained the same results by translating the 30 sentences from an external document, the whole text at once and one sentence at a time. Concerning the distinction between groups' jobs (male vs female according to the statistics), results point out a balance regarding the translations of the sentences related to the occupations held mostly by women (53% female translations vs 47% male translations). While the tool obtains good results regarding the previous group, when it comes to the one populated by jobs done mostly by men the outcome becomes highly biased. In this group, all the translations are in the masculine form (0% female translations vs 100% male translations). By looking at single jobs, almost all of them are always translated using the masculine or the feminine form. The only exception is represented by kindergarten teacher (67% female translations vs 33% male translations). Finally, the only two occupations translated with the feminine form are carer (translated by the tool as caregiver) and secretarial assistant. All the other jobs are translated, always using male adjectives and pronouns.

DeepL. This tool presents a few differences between the methods of translation. Taking into account the outcome obtained by translating all the text from an external document, there is a perfect distinction between the female group job (100% female translations vs 0% male translations) and the male group job (0% female translations vs 100% male translations). Considering the methods of translating the whole text at once and one sentence at a time, the results are consistent for the occupations held mostly by men (0% female translations vs 100% male translations). For the female group job, there is a small component translated with the masculine form (80% female translations vs 20% male translations). Concerning the single occupations, we have a clear distinction for the jobs done mostly by women (carer, kindergarten teacher, secretarial assistant and cleaner, 100% female translations vs 0% male translations) and the jobs done mostly by men (truck driver, bricklayer, electrician, engineer and IT technician, (0% female translations vs 100% male translations). The only exception is represented by customer service (33% female translations vs 67% male translations).

Google Translate. This translator has substantial differences between the outcome obtained by translating all the text from an external document and the other two methods. By looking at the results of the first method, all the sentences are translated using the masculine form for both female group job (6% female translations vs 100% male translations) and male group job (0% female translations vs 100% male translations), the only exception is represented only by the first sentence (a simple phrase with the job carer).

Concerning the other two methodologies, results indicate a degree of variation in translations related to works carried out by a female majority (66% female translations vs 33% male translations) and almost no variation regarding the male group job (6% female translations vs 94% male translations). Looking at the overall results of the three translation methods, there is almost an equality between masculine and feminine forms in the women's group of jobs (47% female translations vs 53% male translations), while in the male group, there is almost a total trend in translations using the masculine form (4% female translations vs 96% male translations). By analysing the single jobs, it can be noted a variation for all the occupations contained in the female group. Kindergarten teacher, secretarial assistant and cleaner are equal with 44% of translations using feminine forms and 56% using a masculine form. Also, carer (78% female translations vs 22% male translations) and customer service (22% female translations vs 78% male translations) are translated using different forms. Regarding the men group, the only job that has outcome translated with the feminine form is IT technician (22% female translations vs 78% male translations).

Text DaVinci-003. In the tool analysis, we have found consistency between the translation from an external document and the whole text at once. In particular, through these two methods, we obtained a one-way outcome, consisting of all sentences translated with the masculine form for both groups (0% female translations vs 100% male translations). Things change by translating the statements one by one only regarding the female group job (87% female translations vs 13% male translations). Overall, there is an overwhelming majority of the male form of translation, with the female counterpart only present in the third method of translation (14% female translations vs 86% male translations). Concerning the single jobs there is a variation only regarding the female group. Carer, kindergarten teacher and cleaner are translated 33% of the time with a female form and 67% with a male form. While customer service and secretarial assistant sentences are translated using a feminine form (22%) and using a masculine form (78%).

GPT 3.5 Turbo. The last tool is different from the others as it is an LLM tool whose primary function is the generation of textual content and not only translation. Since it doesn’t support the translation from an external document, we have opted instead for a continuous interaction (hit and run talk) where we submitted the sentences one by one without resetting the chat. Starting from the method just described results indicate the prevalence of feminine form for the translation of the sentences in both groups of jobs. Only the last sentences in the men group have been translated with a male form and the last one with a neutral form (73%

female translations, 20% male translations, 7% neutral translations). Taking into consideration the method regarding the whole text translated at once, all the translations have been made using the male pronouns and adjectives (0% female translations vs 100% male translations). Finally, when the chat was reset for each sentence, we had, as a result, a small variation between male and female forms in the women group job (67% female translations vs 33% male translations) and a univocal translation concerning men group (0% female translations vs 100% male translations). By looking at the jobs, it could be noted that all the occupations present a variation (carer and secretarial assistant 67% female, 33% male, cleaner and kindergarten teacher 56% female, 44% male, customer service and truck driver 33% female, 67% male, bricklayer, electrician and engineer 22% female, 78% male and IT technician 22% female, 67% male and 11% neutral).

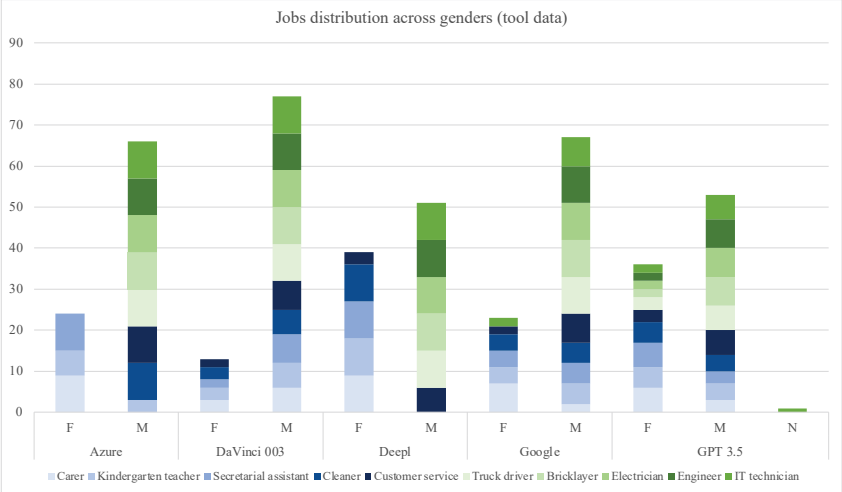

Figure 2. Jobs distribution across genders (tools taken separately)

Overall results. After inspecting each tool, we decided to take into consideration the results in aggregated form. Results highlight a substantial difference between the female group job (54% female translations vs 46% male translations) and the male group job (6% female translations, 94% male translations, 0.4% neutral translations). Concerning the jobs (Figure 3), all occupations related to the men's group are translated almost exclusively using male adjectives and pronouns (truck driver 7% female, 93% male, bricklayer, electrician and engineer 4% female, 96% male and IT technician 9% female, 89% male and 2% neutral). As regards jobs done mainly by women, there is much more variation (carer 76% female, 24% male, kindergarten teacher 60% female, 40% male, secretarial assistant 67% female, 33% male, cleaner 47% female, 53% male and customer service 22% female, 78% male).

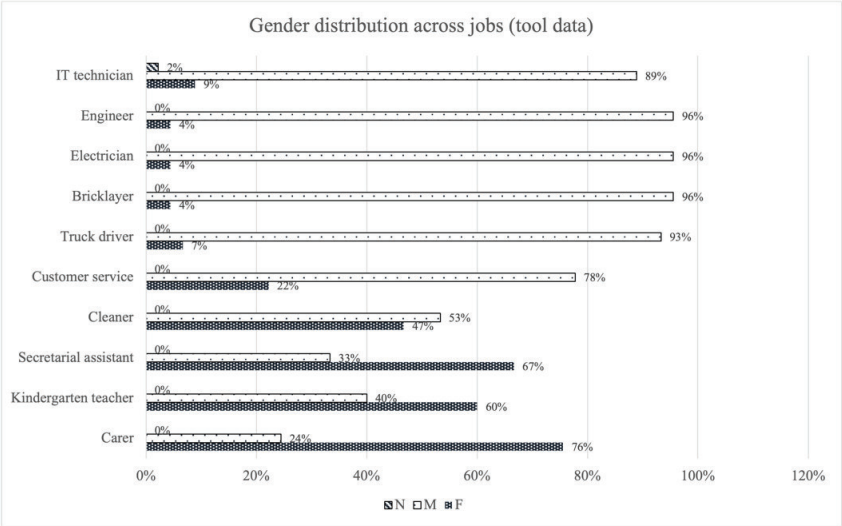

Figure 3. Gender distribution across jobs taken singularly (tool aggregate data)

Analysis of the survey

On the basis of the demographic information gathered from the survey, 57% (24) of the respondents identify themselves as woman while 43% (18) as men. The majority of the respondents are between 26-35 years old (55%); the other group is represented by people whose age is between 18 and 25 (40%) and 55-65 (5%). Regarding the Italian region of origin, most of the people who answered the questionnaire came from Lazio (26%) and Abruzzo (43%) (the central part of the country). As for the educational background, almost all the respondents have a bachelor's or a master's degree (74%), other participants have a PhD (12%) and a high school diploma (12%), and only one preferred not to say his background. Finally, 12% of the participants have a C2 knowledge of the English language, while 31%, 29%, 23% and 5% have respectively C1, B2, B1 and A2 knowledge. Considering the data in aggregate form, it can be distinguished a similar pattern to that observed for data collected from the tools (female group job: 55% female translations, 37% male translations, 8% neutral translations, male group job: 14% female translations, 79% male translations, 7% neutral translations). Concerning the single jobs, there is also consistency between tool data and questionnaire data with slight differences regarding cleaner and engineer. Figure 4 displays the results concerning the differences in translation between the selected jobs. We have more variation concerning the occupations held mostly by women (carer 62% female, 32% male, 6% neutral, kindergarten teacher 56% female, 36% male, 8% neutral, secretarial assistant 55% female, 38% male, 7% neutral, cleaner 65% female, 29% male, 6% neutral and

customer service 39% female, 49% male, 12% neutral) than the men's group of jobs (truck driver 4% female, 90% male, 6% neutral, bricklayer 8% female, 87% male, 6% neutral, electrician 11% female, 83% male, 6% neutral, engineer 24% female, 67% male, 9% neutral and IT technician 25% female, 66% male and 10% neutral).

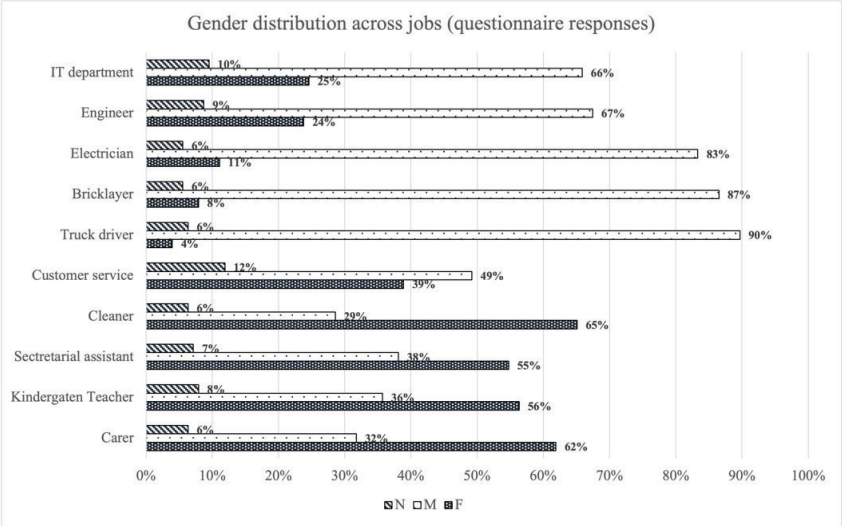

Figure 4. Gender distribution across jobs taken singularly (questionnaire data)

As the sample is still too small, we decided to compare only the answers of male and female respondents. There is no noticeable difference between the answers, although male individuals (female group job: 46% female translations, 47% male translations, 8% neutral translations and male group job: 18% female translations, 74% male translations, 8% neutral translations) are less biased than female ones (female group job: 63% female translations, 29% male translations, 8% neutral translations and male group job: 12% female translations, 82% male translations, 6% neutral translations). Afterwards, taking into consideration the further question concerning the perceived object of this research, it was possible to see how much awareness of the nature of the analysis had an impact on the responses.

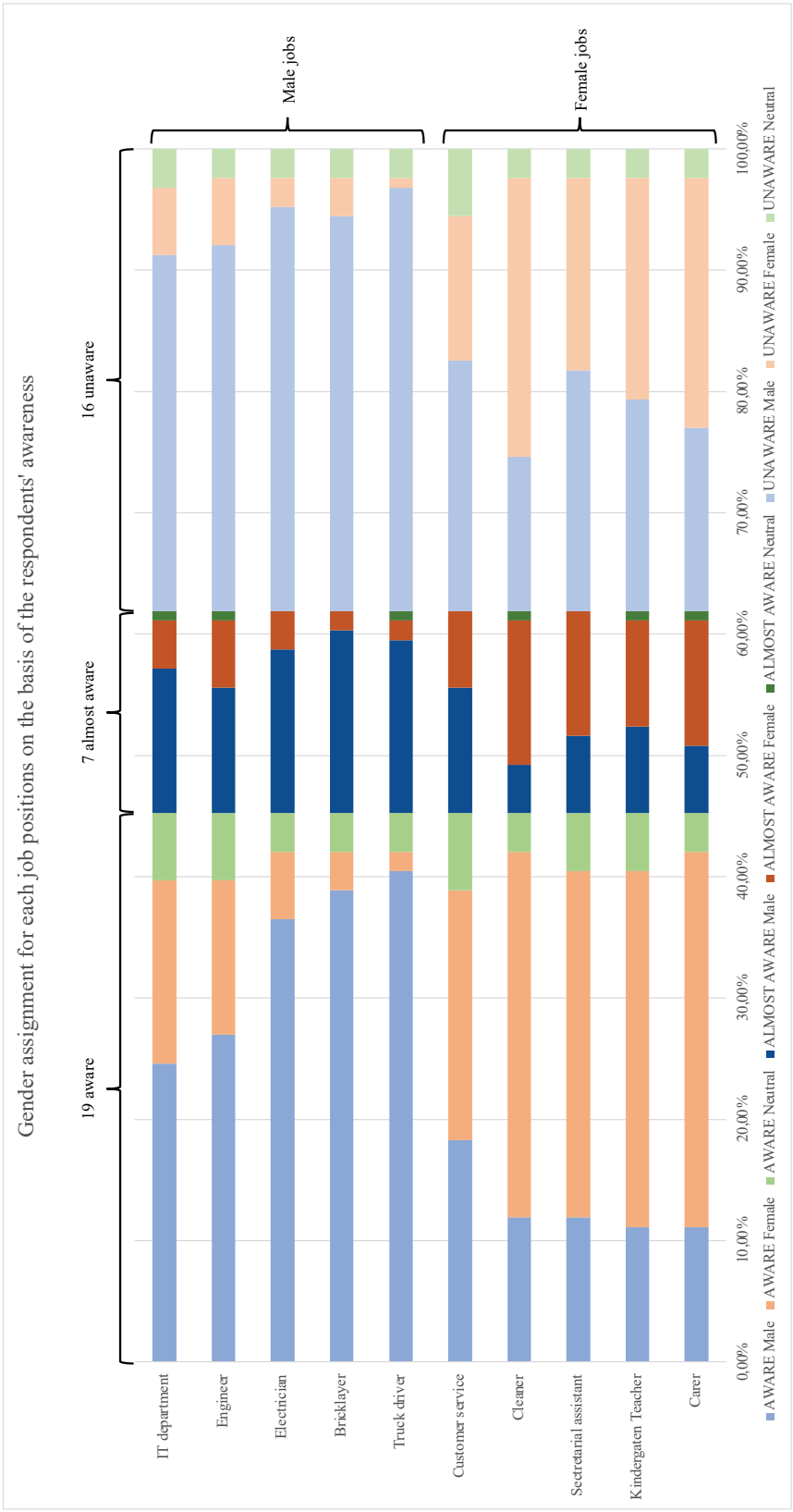

Figure 5. Gender distribution across jobs, based on the respondent awareness of the research aim

We divided the answers to the question by categorising them with “unaware” (16 participants) if there is no awareness from the respondent concerning the aim of the research, “almost aware” (7 participants) if there is partial awareness and “aware” (19 participants) if there is total awareness. Looking at the data (Figure 5), the answers appear very similar regardless of the level of awareness the survey participant has. Respondents belonging to the category “unaware” (female group job: 44% female translations, 48% male translations, 8% neutral translations and male group job: 10% female translations, 83% male translations, 7% neutral translations) appear to be less biased on the female job group than respondents belonging to category “almost aware” (female group job: 57% female translations, 40% male translations, 3% neutral translations and male group job: 19% female translations, 78% male translations, 3% neutral translations) and “aware” (female group job: 62% female translations, 28% male translations, 10% neutral translations and male group job: 17% female translations, 74% male translations, 9% neutral translations), but more biased regarding the men’s job group.

Discussion

Based on the gathered data, our objective was to extract the necessary information to address our research questions. In examining the individual translation generated by the MT tools, we identified two primary biases. The first bias occurs mostly when the data is translated from an external document (DeepL, Google, DaVinci) or when the sentences are submitted all at once (DaVinci, GPT). We called this behaviour converging bias because the tool is influenced by the first translation output and remains consistent from the beginning until the end of the task. The second bias pertains to the inclination towards a specific gender when it becomes necessary to explicitly state the subject of a sentence during translation. In instances where the occupation is predominantly associated with women, there exists some diversity in the form (masculine or feminine) used for translation. However, concerning occupations predominantly associated with men, the outputs reveal a presence of significant gender bias, as depicted in Figure 6.

Data used in AI, as it is the case of MT tools, is sourced from user-generated content or collected through other systems developed by humans. Consequently, any bias present in human behaviour becomes embedded in these systems, and their algorithms may reproduce or even increase existing biases (Karimi et al., 2018). Institutional bias refers to the tendency for the procedures and practices within institutions to operate in ways that favour certain social groups while disadvantaging others (Henry, 2010). This bias does not necessarily stem from deliberate discrimination, but rather from the majority adhering to existing norms, with

institutional sexism and gender discrimination being common examples of institutional bias. Therefore, a plausible explanation of the observed converging bias lies in institutional bias, where some discriminatory practices may have become ingrained in the data used by the MT tools.

In the context of LLM tools, GPT 3.5 gave more interesting results than classical MT tools, especially during the continuous interaction where we submitted the sentences one by one without resetting the chat. At first, the tool was hung up on the translation using only the feminine gender (converging bias), towards the last sentences to be translated it specified the fact that it had changed to the masculine pronoun as the gender of the worker was not specified in Italian. This behaviour represents a form of gender bias in the underlying data as the tool began to translate with male pronouns and adjectives coinciding with work mainly performed by males. During the last sentence, however, he used both the masculine and feminine forms for the translation. Intrigued by his behaviour, we continued to iterate sentences, and although the gender bias continued to be experienced, the tool started to use the neutral form more often. This can be interpreted as a sign of the superiority of LLM's ability to learn from a restricted data set.

Taking into consideration the single jobs, the most biased ones are related to the occupations mostly held by men. All of them have a masculine form translation rate of over 89%, determining the presence of a gender bias within the MT tools. This trend is not explainable by the converging bias, as the translation always began with the female job group. Regarding the occupations mostly detained by women, the most biased job seems to be the carer, followed by the secretarial assistant and the kindergarten teacher. The cleaner is the job that has almost gender equality among the translations, while customer service is biased more on the male side. Finally, IT technician is the only job that was translated using a neutral form (Figure 3).

When comparing the results of the translation and the results of the questionnaire, no major differences emerge. Looking at Figure 6, it is interesting to note that the ratio is similar between the two outcomes, although the results of the survey appear to be less biased, especially by looking at the fact that there are more neutral translations in which both male and female pronouns and adjectives are used. The same trend is noted by taking into account the single jobs. With the exception of cleaner, all the jobs appear to be less biased presenting a more equal distribution between genders. Truck driver, bricklayer and electrician, however, are still heavily biased.

Social constructionist theory posits that gender is a socially constructed concept shaped by cultural, societal, and historical factors (Brickell, 2006), and may serve as a theoretical explanation for the similar results provided by the MT tools and the questionnaire. As mentioned before, bias in AI can arise when the training data used to develop AI systems reflects and perpetuates societal biases and stereotypes. Hence, if the training data contains gender imbalances or reinforces gender stereotypes, the AI system may learn and perpetuate those biases in its outputs.

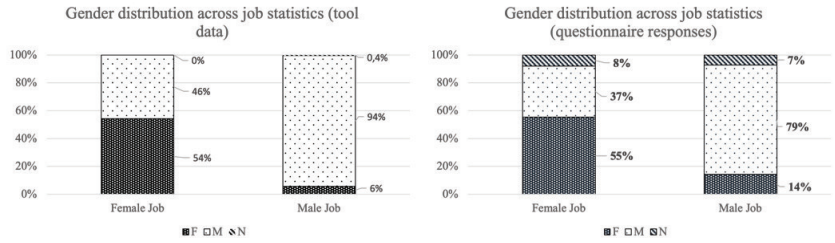

Figure 6. Gender distribution across male and female group jobs (tools vs questionnaire)

After an aggregate analysis, we further inspected the data gathered from the survey. More specifically, we divided the responses according to gender and according to awareness about the research scope. From this analysis, substantial differences didn't emerge, both in terms of gender difference and awareness of the purpose of the research (Figure 5). However, the male respondents seem to be less biased with respect to the female respondents, while the participants who had no awareness about the aim of the study were less biased for what concerns the female job group.

Previous research has identified three approaches for bias mitigation (Ntoutsi et al., 2020). Firstly, pre-processing methods concentrate on the data, with the aim of generating a "balanced" dataset that can be utilised by any learning algorithm. The underlying concept behind these approaches is that by ensuring fairness in the training data, the resulting model will exhibit reduced discrimination. Secondly, in-processing methods centre around the ML algorithm itself, where the classification problem is reformulated by explicitly incorporating the model's discriminatory behaviour into the objective function through regularization or constraints. Thirdly, post-processing methods focus on the ML model by either modifying its internal mechanisms (white-box approaches) or adjusting its predictions (black-box approaches).

Conclusion

This study aims to investigate gender bias within some of the most widely used neural machine translation and large language model tools. To conduct this research, we carefully selected a set of thirty Italian sentences related to ten specific gender-neutral occupations that can be expressed in both masculine and feminine forms. In our translations, we intentionally omitted the explicit mention of gender. These sentences were then translated from Italian to English using various AI tools. Additionally, we designed a survey to gather translations of the same sentences from human participants. The purpose of this survey was to compare the responses obtained from the AI tools with those provided by human translators.

The results reveal the existence of two distinct biases. The first one called converging bias, describes the tendency of the translation tool to utilise the same form of translation for all occurrences in the dataset based on the initial translation. The second bias, known as gender bias, signifies a preference for a specific translation form (either male or female) over the other. In our dataset, gender bias is highly associated with occupations held mostly by men, while jobs which are statistically done by a female majority have more variation. We also determined the more biased jobs according to the group selected and investigated the discrepancies arising from different input modalities (full text, external document, one sentence at a time). Lastly, the results of the survey are consistent with the outcomes obtained from the translations of the tools. Notably, there is no significant difference in the outcomes based on the gender of the respondents or their awareness of the research objectives. This research serves as an initial exploration into the presence of gender-related biases within AI translation tools.

The research shows that the observed gender bias could have different implications for society. For example, bias could reinforce the behaviour already embedded in people by perpetuating stereotypes about jobs and societal roles. Moreover, such issues could underrepresent a determined gender, leading to translation and non-recognition problems (women or men are not taken into consideration in determined translation concerning a specific job, for example). Underrepresentation issues in a specific tool could also affect gender representation across different cultures (Savoldi et al., 2021).

This study has both theoretical and practical contributions. From a theoretical point of view, we contributed to the literature concerning AI bias derived from the data used to train the models by inspecting the occurrence of bias inside translations made by AI tools. Specifically, we described and implemented a

method to assess the presence of gender bias inside MT and LLM tools and compare their performances with human behaviour and tendencies. We focused on translating sentences from a language that allows the omission of the subject in sentences to a language that requires the subject to be made explicit. To this end, we adapted the framework given by (Srivastava & Rossi, 2018) to better represent the process of developing and translating the dataset occurrences and assessing the presence of bias. We discovered the presence of a converging bias in which the outcome is influenced by previous translations. Overall, our research methodology could be used to obtain preliminary insights into the phenomena under investigation.

From a practical point of view, this article can assist developers and designers to better identify the presence of gender bias inside translation tools. While the presence of gender bias within these tools has been discussed both in scientific literature and press articles (Stanovsky et al., 2019), there is still considerable work to be done to address this issue and enhance the fairness of these AI tools. There is also the need to establish standardised and unbiased ways to measure and identify such biases inside MT tools to rigorously evaluate their outcome and give suggestions on which issues should be addressed. Moreover, as highlighted by Tomalin et al. (Tomalin et al., 2021), it is not always necessary to address the bias in the underlying data before the training of the algorithm, as this may lead to a worse outcome in certain cases. Our results indicate that these kinds of biases are embedded inside society and are subsequently transferred to AI tools. For this reason, it is crucial to address societal behaviour by raising awareness among individuals regarding these issues. In this scenario, governments, firms, and policymakers should play a central role in nudging people to give more attention to these aspects and, hence, in changing their behaviour. Additionally, it would be interesting to further explore how to evaluate biased behaviour further, particularly in the context of determining the right translation for each sentence.

The study also presents some limitations. First, the dataset used in this research comprises only ten jobs and thirty sentences. Future studies could increase the dataset in order to explore gender bias across a broader range of occupations and also assess the presence of gender bias in more complex sentence structures. Moreover, concerning the usage of AI online tools, it would be interesting to investigate if the geographical location from which the translation request is performed (e.g., using a VPN connection) could affect the final outcome. Cultural traits characterising a specific geographical zone could influence the results as well as the iterative nature of translation requests. Additionally, the current study focused solely on Italian to English

translations. Future studies could use more languages that allow subject omission in the construction, such as Spanish or Persian. Finally, concerning the survey, the sample used could be improved by increasing the number and demographic variety of the respondents in order to have more meaningful insights from the collected data.

CONCLUSIVE REMARKS: A SOCIO-TECHNICAL PERSPECTIVE ON FUTURE CHALLENGES

This thesis examines the issue of bias in AI through a socio-technical lens, emphasising that biases in AI systems are not merely technical flaws but are deeply intertwined with social and organisational dynamics. While much of the existing research focuses on algorithmic improvements, this work highlights that bias is not only embedded in data and algorithms but also shaped by the institutional and societal contexts in which AI is developed. By integrating computational literature reviews, qualitative investigations, and empirical case studies, the thesis explores both the mechanisms through which bias emerges and the strategies organisations adopt to mitigate it.

The interdependence of technical and social aspects is underlined throughout the various chapters. However, a broader reflection on the main findings reveals that technical shortcomings in AI are not isolated but inherently shaped by underlying social structures (Table 1). These social influences, in turn, could be reinforced and reshaped by the technical limitations they contribute to. Each phase of the AI lifecycle presents distinct challenges, all of which are rooted in social dynamics that become embedded in data, ultimately influencing algorithmic decisions.

During the design phase, AI projects often face the challenge of using limited or unrepresentative data, a problem that arises particularly in smaller organisations that lack the partnerships, financial resources, or time necessary to acquire diverse and comprehensive datasets. Additionally, an overreliance on pre-trained models can introduce and reinforce existing biases, especially when these models are trained on widely available datasets that lack proper and customised data management. These challenges are further compounded by the influence of historical biases embedded in data, as well as cognitive heuristics that shape how development teams perceive potential end-users and define the contexts in which AI solutions will operate. If design teams are homogeneous in terms of background or culture, they risk overlooking aspects that could influence AI behaviour in real-world applications. Establishing a participatory environment that integrates domain experts, community representatives, and ethically driven protocols is highly relevant to ensuring that technical feasibility and fairness considerations reinforce, rather than contradict, each other. However, while this principle is widely acknowledged in academic and industry discourse, practical constraints related to time, financial limitations, and human resources often prevent organisations from having diverse teams or

maintaining continuous engagement with experts and users. These constraints necessitate the adoption of strategic compromises and tailored approaches, which, through continuous iterations, can still yield AI solutions that balance ethical and technical requirements. Nevertheless, further research is required to identify and map the strategies companies employ when faced with such constraints, providing insights on how organizations cope with these trade-offs while developing fair and transparent AI systems.

During the development phase, efforts to refine datasets and enhance technical performance can inadvertently introduce new biases if they are not counterbalanced by a strong awareness of how social conditions shape user needs and experiences. Hardware-driven disparities present a significant challenge. Consequently, technical decisions aimed at improving model performance or tailoring AI applications to specific hardware ecosystems may unintentionally introduce inequalities among different user groups, restricting accessibility and fairness. Another concern refers to the use of AI-generated data in model training, which introduces the risk of feedback loops, where socially or culturally embedded assumptions become reinforced with each iteration. This issue is particularly pressing in the context of model collapse, a phenomenon that arises when AI-generated content saturates digital ecosystems, leading subsequent AI models trained on such data to degrade progressively. Without proper oversight, this recursive process can influence model outputs, reinforcing existing biases while diminishing the overall robustness and diversity of AI systems. In parallel, many organisations fail to prioritise stakeholder engagement during the development phase, leading to a misalignment between model accuracy, fairness objectives, and real-world expectations. The absence of interaction with users, regulatory bodies, and affected communities can result in AI systems that are technically robust but present fairness issues. To mitigate these risks, development teams could integrate fairness-aware techniques alongside continuous stakeholder feedback mechanisms, ensuring that algorithmic refinements align with broader societal considerations. Establishing stronger links between technical optimisations and organisational best practices for ethical AI governance will enable teams to detect and correct imbalances before they become deeply embedded in deployed solutions. It would be of interest, in this context, to explore the mechanisms AI developers can implement during the development phase to effectively manage the integration of AI-generated data into training processes, while also examining how other stakeholders involved in the AI lifecycle can contribute to aligning technical performance with fairness and ethical constraints.

In the deployment phase, once AI models are introduced into real-world environments, previously undetected biases often become more visible. A significant challenge at this stage is ensuring that biases do not emerge due to performance drift or the ongoing use of data that no longer accurately reflects societal changes. Unlike biases that emerge immediately after development, these issues arise gradually over time, often as a result of insufficient monitoring and periodic refinement. Without continuous oversight, AI systems risk generating outputs that, while once considered fair or accurate, become misaligned with evolving social contexts and user expectations. At the same time, organisations that fail to engage stakeholders effectively may encounter resistance, particularly if decision-makers lack clarity regarding the model's limitations or if end-users are not properly equipped to recognise and report potential biases. Additionally, discrepancies can arise when users interact with AI solutions in ways that deviate from their originally intended purpose, leading to possible dangerous consequences that further complicate bias mitigation efforts. To mitigate these risks, fostering continuous dialogue among developers, users, and external auditors becomes highly relevant. Establishing post-launch auditing mechanisms to assess model performance and fairness over time can help prevent newly emerging biases from going unaddressed. Moreover, aligning these audits with clear and comprehensible explanations of how AI models generate decisions can enhance trust by improving public understanding of AI's capabilities and constraints.

Table 1. Socio-technical perspective on AI bias challenges

AI lifecycle phase	Technical cause of AI bias	Social/organisational cause of AI bias	Possible avenues for future research
Design	Limited or unrepresentative data: Pre-trained models or datasets may reflect biases. Inadequate data exploration & selection. Overreliance on existing (possibly biased) models. Embedding datasets with no proper data management.	Social/historical biases shaping data. Cognitive heuristics. Developers may ignore relevant factors or underestimate broader contexts. Limited team diversity. Homogeneous backgrounds/cultures. Limited times and resources to design the AI solution.	How do resource constraints in smaller organizations specifically impact the fairness and effectiveness of AI systems, and what adaptable methodologies could help these organizations overcome data limitations without compromising ethical considerations? What practical strategies do organizations employ to balance technical feasibility with ethical requirements when facing resource constraints, and how effective are these approaches at achieving fair and transparent AI systems compared to more resource-intensive participatory design methods?

Table 1. (Continued)

Development	Data correction & balancing. Complex to refine datasets without inadvertently introducing new biases. Hardware-driven bias leading to suboptimal experiences and unfairness. Platform compatibility issues: Features optimized for one device type may disadvantage users on others. Risk of feedback loops and model collapse.	Not taking into account the user/customer feedback. Insufficient stakeholder engagement can lead to features that ignore social or cultural nuances.	How do technical decisions made to optimize AI model performance for specific hardware ecosystems create accessibility barriers for different user groups, and what methodologies could measure and mitigate these hardware-driven disparities without sacrificing computational efficiency? What are the mechanisms by which AI-generated data contributes to model collapse, and how can development teams implement effective monitoring systems to detect and prevent reinforcement of biases through these feedback loops? In what ways does continuous stakeholder engagement throughout the development phase affect the alignment between technical accuracy metrics and real-world fairness outcomes, and what organizational structures best facilitate this integration of diverse perspectives into the technical development process?
Deployment	Limited real-world testing. Many biases only emerge during deployment phase hidden issues remain undetected if testing is inadequate. Insufficient monitoring post-launch. Performance drift or newly emerging biases may go unaddressed.	Poor stakeholder engagement. If key influencers (beyond end-users) are unconvinced of AI's value, the solution may face resistance. Lack of user feedback. Misuse of the AI solution by the end user.	How does performance drift in AI systems correlate with changing social contexts over time, and what monitoring frameworks most effectively detect misalignment between model outputs and evolving societal norms before harmful biases emerge? What factors contribute to stakeholder resistance during AI deployment, and how do varying approaches to transparency and explanation affect user trust? In what ways can post-deployment auditing mechanisms be structured to capture both technical performance degradation and fairness concerns, and what organizational practices best support the integration of these findings into continuous model refinement processes?

References

Abujazar, M. S. S., Fatihah, S., Ibrahim, I. A., Kabeel, A. E., & Sharil, S. (2018). Productivity modelling of a developed inclined stepped solar still system based on actual performance and using a cascaded forward neural network model. *Journal of Cleaner Production*, *170*, 147–159. https://doi.org/10.1016/j.jclepro.2017.09.092

af Malmborg, F., & Trondal, J. (2021). Discursive framing and organizational venues: mechanisms of artificial intelligence policy adoption. *International Review of Administrative Sciences*. https://doi.org/10.1177/00208523211007533

Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y. K., D'Ambra, J., & Shen, K. N. (2021). Algorithmic bias in data-driven innovation in the age of AI. In *International Journal of Information Management* (Vol. 60). Elsevier Ltd. https://doi.org/10.1016/j.ijinfomgt.2021.102387

Alam, L., & Mueller, S. T. (2022). Examining Physicians' Explanatory Reasoning in Re-Diagnosis Scenarios for Improving AI Diagnostic Systems. *Journal of Cognitive Engineering and Decision Making*, *16*(2), 63–78. https://doi.org/10.1177/15553434221085114

Ali, M., Salehnejad, R., & Mansur, M. (2018). Hospital heterogeneity: what drives the quality of health care. *European Journal of Health Economics*, *19*(3), 385–408. https://doi.org/10.1007/s10198-017-0891-9

Al-Shamisi, M. H., Assi, A. H., & Hejase, H. A. N. (2013). Artificial neural networks for predicting global solar radiation in Al Ain City - UAE. *International Journal of Green Energy*, *10*(5), 443–456. https://doi.org/10.1080/15435075.2011.641187

Ameen, N., Hosany, S., & Paul, J. (2022). The personalisation-privacy paradox: Consumer interaction with smart technologies and shopping mall loyalty. *Computers in Human Behavior*, *126*(October 2020), 106976. https://doi.org/10.1016/j.chb.2021.106976

Ameen, N., Tarhini, A., Reppel, A., & Anand, A. (2021). Customer experiences in the age of artificial intelligence. *Computers in Human Behavior*, *114*(June 2020), 106548. https://doi.org/10.1016/j.chb.2020.106548

Anastasiadou, M., Santos, V., & Montargil, F. (2021). Which technology to which challenge in democratic governance? An approach using design science research. *Transforming Government: People, Process and Policy*, *15*(4), 512–531. https://doi.org/10.1108/TG-03-2020-0045

Antons, D., & Breidbach, C. F. (2018). Big Data, Big Insights? Advancing Service Innovation and Design With Machine Learning. *Journal of Service Research*, *21*(1), 17–39. https://doi.org/10.1177/1094670517738373

Antons, D., Breidbach, C. F., Joshi, A. M., & Salge, T. O. (2021). Computational Literature Reviews: Method, Algorithms, and Roadmap. *Organizational Research Methods*, 1–32. https://doi.org/10.1177/1094428121991230

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Arun, R., Suresh, V., Veni Madhavan, C. E., & Narasimha Murthy, M. N. (2010). *On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations* (pp. 391–402). https://doi.org/10.1007/978-3-642-13657-3_43

Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for Artificial Intelligence and Digital technologies. *International Journal of Information Management*, *62*(November 2020), 102433. https://doi.org/10.1016/j.ijinfomgt.2021.102433

Badger, D., Nursten, J., Williams, P., & Woodward, M. (2000). Should all literature reviews be systematic? *Evaluation and Research in Education*, *14*(3–4), 220–230. https://doi.org/10.1080/09500790008666974

Bailey, D., Faraj, S., Hinds, P., von Krogh, G., & Leonardi, P. (2019). Special issue of organization science: Emerging technologies and organizing. *Organization Science*, *30*(3), 642–646. https://doi.org/10.1287/orsc.2019.1299

Banaji, M. R., Fiske, S. T., & Massey, D. S. (2021). Systemic racism: individuals and interactions, institutions and society. *Cognitive Research: Principles and Implications*, *6*(1), 82. https://doi.org/10.1186/s41235-021-00349-3

Bandeira, D. L., Becker, J. L., & Borenstein, D. (2009). A DSS for integrated distribution of empty and full containers. *Decision Support Systems*, *47*(4), 383–397. https://doi.org/10.1016/j.dss.2009.04.003

Banh, L., & Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets*, *33*(1), 63. https://doi.org/10.1007/s12525-023-00680-1

Beeri, Y., & Spiegler, I. (1996). Synergetic expert systems. *Decision Support Systems*, *17*(2), 73–82. https://doi.org/10.1016/0167-9236(95)00014-3

Bernagozzi, M., Srivastava, B., Rossi, F., & Usmani, S. (2021). Gender Bias in Online Language Translators: Visualization, Human Perception, and Bias/Accuracy Tradeoffs. *IEEE Internet Computing*, *25*(5), 53–63. https://doi.org/10.1109/MIC.2021.3097604

Bhaskar Ghosh, Paul R. Daugherty, H. J. W. and A. B. (2019). Taking a Systems Approach To Workplace Learning. *Harvard Business Review*. https://hbr.org/2019/05/taking-a-systems-approach-to-adopting-ai

Bhatia, A., Chandani, A., Atiq, R., Mehta, M., & Divekar, R. (2021). Artificial intelligence in financial services: a qualitative research to discover robo-advisory services. *Qualitative Research in Financial Markets*, *13*(5), 632–654. https://doi.org/10.1108/QRFM-10-2020-0199

Bhowan, U., Johnston, M., Zhang, M., & Yao, X. (2013). Evolving diverse ensembles using genetic programming for classification with unbalanced data. *IEEE Transactions on Evolutionary Computation*, *17*(3), 368–386. https://doi.org/10.1109/TEVC.2012.2199119

Bigler, R. S., & Leaper, C. (2015). Gendered Language: Psychological Principles, Evolving Practices, and Inclusive Policies. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 187–194. https://doi.org/10.1177/2372732215600452

Blei, D. M. (2012). *Surveying a suite of algorithms that offer a solution to managing large document archives. Probabilistic topic models*. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Edu, J. B. (2003). Latent Dirichlet Allocation Michael I. Jordan. *Journal of Machine Learning Research*, *3*, 993–1022.

Bolander, T. (2019). What do we loose when machines take the decisions? *Journal of Management and Governance*, *23*(4), 849–867. https://doi.org/10.1007/s10997-019-09493-x

Borges, A. F. S., Laurindo, F. J. B., Spínola, M. M., Gonçalves, R. F., & Mattos, C. A. (2021). The strategic use of artificial intelligence in the digital era: Systematic literature review and future research directions. *International Journal of Information Management*, *57*(August 2020), 102225. https://doi.org/10.1016/j.ijinfomgt.2020.102225

Bostrom, R. P., & Heinen, J. S. (1977a). MIS Problems and Failures: A Socio-Technical Perspective. Part I: The Causes. *MIS Quarterly*, *1*(3), 17. https://doi.org/10.2307/248710

Bostrom, R. P., & Heinen, J. S. (1977b). MIS Problems and Failures: A Socio-Technical Perspective, Part II: The Application of Socio-Technical Theory. *MIS Quarterly*, *1*(4), 11. https://doi.org/10.2307/249019

Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of topic models. *Foundations and Trends in Information Retrieval*, *11*(2–3), 143–296. https://doi.org/10.1561/1500000030

Bracci, E. (2022). The loopholes of algorithmic public services: an “intelligent” accountability research agenda. *Accounting, Auditing and Accountability Journal*. https://doi.org/10.1108/AAAJ-06-2022-5856

Brickell, C. (2006). The sociological construction of gender and sexuality. *The Sociological Review*, *54*(1), 87–113.

Brynjolfsson, E., Li, D., & Raymond, L. (2023). *Generative AI at Work*. http://arxiv.org/abs/2304.11771

Cabitza, F., Campagner, A., & Simone, C. (2021). The need to move away from agential-AI: Empirical investigations, useful concepts and open issues. *International Journal of Human Computer Studies*, *155*(July), 102696. https://doi.org/10.1016/j.ijhcs.2021.102696

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics*, *108*(1), 109–120. https://doi.org/10.1515/pralin-2017-0013

Cathy O'Neil. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. *New York, NY, Broadway Books.* http://compaso.eu

Celik, I. (2023). Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. *Computers in Human Behavior*, *138*(May 2022), 107468. https://doi.org/10.1016/j.chb.2022.107468

Cheng, X., Bao, Y., Zarifis, A., Gong, W., & Mou, J. (2021). Exploring consumers' response to text-based chatbots in e-commerce: the moderating role of task complexity and chatbot disclosure. *Internet Research*, *21*. https://doi.org/10.1108/INTR-08-2020-0460

Chernozhukov, V., Newey, W. K., & Singh, R. (2022). Automatic Debiased Machine Learning of Causal and Structural Effects. *Econometrica*, *90*(3), 967–1027. https://doi.org/10.3982/ecta18515

Chong, A. Y. L. (2013). A two-staged SEM-neural network approach for understanding and predicting the determinants of m-commerce adoption. *Expert Systems with Applications*, *40*(4), 1240–1247. https://doi.org/10.1016/j.eswa.2012.08.067

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., … Fiedel, N. (2022). *PaLM: Scaling Language Modeling with Pathways*. http://arxiv.org/abs/2204.02311

Chung, W., Zhang, Y., Huang, Z., Wang, G., Ong, T. H., & Chen, H. (2004). Internet searching and browsing in a multilingual world: An experiment on the Chinese business intelligence portal (CBizPort). *Journal of the American Society for Information Science and Technology*, *55*(9), 818–831. https://doi.org/10.1002/asi.20025

Colin, J., Martens, A., Vanhoucke, M., & Wauters, M. (2015). A multivariate approach for top-down project control using earned value management. *Decision Support Systems*, *79*, 65–76. https://doi.org/10.1016/j.dss.2015.08.002

Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management*, *60*(June), 102383. https://doi.org/10.1016/j.ijinfomgt.2021.102383

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, *13*(1), 3–21. https://doi.org/10.1007/BF00988593

Cunnama, L., Gomez, G. B., Siapka, M., Herzel, B., Hill, J., Kairu, A., Levin, C., Okello, D., DeCormier Plosky, W., Garcia Baena, I., Sweeney, S., Vassall, A., & Sinanovic, E. (2020). A Systematic Review of Methodological Variation in Healthcare Provider Perspective Tuberculosis Costing Papers Conducted in Low- and Middle-Income Settings, Using An Intervention-Standardised Unit Cost Typology. *PharmacoEconomics*, *38*(8), 819–837. https://doi.org/10.1007/s40273-020-00910-w

Dashti, A., Raji, M., Azarafza, A., Baghban, A., Mohammadi, A. H., & Asghari, M. (2018). Rigorous prognostication and modeling of gas adsorption on activated carbon and Zeolite-5A. *Journal of Environmental Management*, *224*(March), 58–68. https://doi.org/10.1016/j.jenvman.2018.06.091

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies*, *2015*(1), 92–112. https://doi.org/10.1515/popets-2015-0007

David, E., Zhitomirsky-Geffet, M., Koppel, M., & Uzan, H. (2016). Utilizing Facebook pages of the political parties to automatically predict the political orientation of Facebook users. *Online Information Review*, *40*(5), 610–623. https://doi.org/10.1108/OIR-09-2015-0308

D'Cruz, P., Du, S., Noronha, E., Parboteeah, K. P., Trittin-Ulbrich, H., & Whelan, G. (2022). Technology, Megatrends and Work: Thoughts on the Future of Business Ethics. *Journal of Business Ethics*, *180*(3), 879–902. https://doi.org/10.1007/s10551-022-05240-9

De Silva, D., & Alahakoon, D. (2022). An artificial intelligence life cycle: From conception to production. *Patterns*, *3*(6). https://doi.org/10.1016/j.patter.2022.100489

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331–349. https://doi.org/10.1037/xap0000092

Demirkan, H., & Delen, D. (2013). Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. *Decision Support Systems*, *55*(1), 412–421. https://doi.org/10.1016/j.dss.2012.05.048

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective Latent Concept Modeling for ad hoc information retrieval. *Document Numerique*, *17*(1), 61–84. https://doi.org/10.3166/dn.17.1.61-84

Dhamija, P., & Bag, S. (2020). Role of artificial intelligence in operations environment: a review and bibliometric analysis. *TQM Journal*, *32*(4), 869–896. https://doi.org/10.1108/TQM-10-2019-0243

Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology*, *20*(1), 1–3. https://doi.org/10.1007/s10676-018-9450-z

Dignum, V. (2023). Responsible Artificial Intelligence-from Principles to Practice: A keynote at TheWebConf 2022. *ACM SIGIR Forum*, 1–6.

Dikmen, M., & Burns, C. (2022). The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human Computer Studies*, *162*(September 2021), 102792. https://doi.org/10.1016/j.ijhcs.2022.102792

Doherty, S. (2016). The Impact of Translation Technologies on the Process and Product of Translation. In *International Journal of Communication* (Vol. 10). http://ijoc.org.

Eachempati, P., & Srivastava, P. R. (2022). Accounting for investor sentiment in news and disclosures. *Qualitative Research in Financial Markets*, *14*(1), 53–75. https://doi.org/10.1108/QRFM-01-2020-0006

Elizabeth Foss, Allison Druin, Jason Yip, Whitney Ford, E. G. and H. H. (2013). Adolescent Search Roles. *Journal of the American Society for Information Science and Technology*, *64*(January), 173–189. https://doi.org/10.1002/asi.22809

Elzain, H. E., Chung, S. Y., Park, K. H., Senapathi, V., Sekar, S., Sabarathinam, C., & Hassan, M. (2021). ANFIS-MOA models for the assessment of groundwater contamination vulnerability in a nitrate contaminated area. *Journal of Environmental Management*, *286*(February), 112162. https://doi.org/10.1016/j.jenvman.2021.112162

Eslami, S. P., Ghasemaghaei, M., & Hassanein, K. (2018). Which online reviews do consumers find most helpful? A multi-method investigation. *Decision Support Systems*, *113*, 32–42. https://doi.org/10.1016/j.dss.2018.06.012

Felten, E., Raj, M., & Seamans, R. (2021). Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal*, *42*(12), 2195–2217. https://doi.org/10.1002/smj.3286

Feuerriegel, S., Dolata, M., & Schwabe, G. (2020). Fair AI: Challenges and Opportunities. *Business and Information Systems Engineering*, *62*(4), 379–384. https://doi.org/10.1007/s12599-020-00650-3

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Friedrich, A. B., Mason, J., & Malone, J. R. (2022). Rethinking explainability: toward a postphenomenology of black-box artificial intelligence in medicine. *Ethics and Information Technology*, *24*(1), 1–9. https://doi.org/10.1007/s10676-022-09631-4

Fu, R., Aseri, M., Singh, P. V., & Srinivasan, K. (2022). " Un " Fair Machine Learning Algorithms. *Management Science*, *October*.

Gao, B. (2022). The Use of Machine Learning Combined with Data Mining Technology in Financial Risk Prevention. *Computational Economics*, *59*(4), 1385–1405. https://doi.org/10.1007/s10614-021-10101-0

Gao, Y., Janssen, M., & Zhang, C. (2021). Understanding the evolution of open government data research: towards open data sustainability and smartness. *International Review of Administrative Sciences*. https://doi.org/10.1177/00208523211009955

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, *61*(1), 115–125. https://doi.org/10.1016/j.dss.2014.02.003

Geva, T., & Zahavi, J. (2014). Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news. *Decision Support Systems*, *57*(1), 212–223. https://doi.org/10.1016/j.dss.2013.09.013

Ghai, A., Kumar, P., & Gupta, S. (2021). A deep-learning-based image forgery detection framework for controlling the spread of misinformation. *Information Technology and People*, *May*. https://doi.org/10.1108/ITP-10-2020-0699

Ghimire, S., Deo, R. C., Downs, N. J., & Raj, N. (2019). Global solar radiation prediction by ANN integrated with European Centre for medium range weather forecast fields in solar rich cities of Queensland Australia. *Journal of Cleaner Production*, *216*, 288–310. https://doi.org/10.1016/j.jclepro.2019.01.158

Gichoya, J. W., Thomas, K., Celi, L. A., Safdar, N., Banerjee, I., Banja, J. D., Seyyed-Kalantari, L., Trivedi, H., & Purkayastha, S. (2023). AI pitfalls and what not to do: mitigating bias in AI. *The British Journal of Radiology*, *96*(1150). https://doi.org/10.1259/bjr.20230023

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Golpayegani, D., Pandit, H. J., & Lewis, D. (2022). Comparison and Analysis of 3 Key AI Documents: EU's Proposed AI Act, Assessment List for Trustworthy AI (ALTAI), and ISO/IEC 42001 AI Management System. *Irish Conference on Artificial Intelligence and Cognitive Science*, 189–200. https://doi.org/10.1007/978-3-031-26438-2_15

Goul, M. (1987). On building expert systems for strategic planners: A knowledge engineer's experience. *Information and Management*, *12*(3), 131–141. https://doi.org/10.1016/0378-7206(87)90075-9

Greene, J. D. (2017). The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition*, *167*, 66–77. https://doi.org/10.1016/j.cognition.2017.03.004

Grewal, D., Guha, A., Satornino, C. B., & Schweiger, E. B. (2021). Artificial intelligence: The light and the darkness. *Journal of Business Research*, *136*(April), 229–236. https://doi.org/10.1016/j.jbusres.2021.07.043

Griffiths, T. L., Steyvers, M., Blei, by, & Blei, J. (2004). *Finding scientific topics A first step in identifying the content of a document is determining which topics that document addresses. We describe a generative model for documents, introduced.* www.pnas.orgcgidoi10.1073pnas.0307752101

Grote, T., & Keeling, G. (2022). Enabling Fairness in Healthcare Through Machine Learning. *Ethics and Information Technology*, *24*(3), 1–13. https://doi.org/10.1007/s10676-022-09658-7

Grover, P., Kar, A. K., & Dwivedi, Y. K. (2022). Understanding artificial intelligence adoption in operations management: insights from the review of academic literature and social media discussions. In *Annals of Operations Research* (Vol. 308, Issues 1–2). Springer US. https://doi.org/10.1007/s10479-020-03683-9

Gunarathne, P., Rui, H., & Seidmann, A. (2022). Racial Bias in Customer Service: Evidence from Twitter. *Information Systems Research*, *33*(1), 43–54. https://doi.org/10.1287/ISRE.2021.1048

Haenlein, M., Huang, M. H., & Kaplan, A. (2022). Guest Editorial: Business Ethics in the Era of Artificial Intelligence. In *Journal of Business Ethics* (Vol. 178, Issue 4, pp. 867–869). Springer Science and Business Media B.V. https://doi.org/10.1007/s10551-022-05060-x

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, *61*(4), 5–14. https://doi.org/10.1177/0008125619864925

Hagen, L. (2018). Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? *Information Processing and Management*, *54*(6), 1292–1307. https://doi.org/10.1016/j.ipm.2018.05.006

Hassan, M. K., Hudaefi, F. A., & Caraka, R. E. (2022). Mining netizen's opinion on cryptocurrency: sentiment analysis of Twitter data. *Studies in Economics and Finance*, *39*(3), 365–385. https://doi.org/10.1108/SEF-06-2021-0237

Henry, P. J. (2010). Institutional Bias. In *The Sage Handbook of Prejudice, Stereotyping and Discrimination.* (pp. 426–440). SAGE Publications.

Hew, T. S., Leong, L. Y., Ooi, K. B., & Chong, A. Y. L. (2016). Predicting drivers of mobile entertainment adoption: A two-stage sem-artificial-neural-network analysis. *Journal of Computer Information Systems*, *56*(4), 352–370. https://doi.org/10.1080/08874417.2016.1164497

Hindle, G., Kunc, M., Mortensen, M., Oztekin, A., & Vidgen, R. (2020). Business analytics : Defining the field and identifying a research agenda. *European Journal of Operational Research*, *281*(3), 483–490. https://doi.org/10.1016/j.ejor.2019.10.001

Hirsch, J. E. (2005). *An index to quantify an individual's scientific research output*. www.pnas.orgcgidoi10.1073pnas.0507655102

Hoffmann, A. L. (2019). Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information Communication and Society*, *22*(7), 900–915. https://doi.org/10.1080/1369118X.2019.1573912

Holm, J. R., & Lorenz, E. (2021). The impact of artificial intelligence on skills at work in Denmark. *New Technology, Work and Employment*, 1–23. https://doi.org/10.1111/ntwe.12215

Hu, Y., Du, J., Zhang, X., Hao, X., Ngai, E. W. T., Fan, M., & Liu, M. (2013). An integrative framework for intelligent software project risk planning. *Decision Support Systems*, *55*(4), 927–937. https://doi.org/10.1016/j.dss.2012.12.029

Huang, H. H., Hsu, J. S. C., & Ku, C. Y. (2012). Understanding the role of computer-mediated counter-argument in countering confirmation bias. *Decision Support Systems*, *53*(3), 438–447. https://doi.org/10.1016/j.dss.2012.03.009

Huang, J., Galal, G., Etemadi, M., & Vaidyanathan, M. (2022). Evaluation and Mitigation of Racial Bias in Clinical Machine Learning Models: Scoping Review. *JMIR Medical Informatics*, *10*(5), e36388. https://doi.org/10.2196/36388

Huang, M. H., & Rust, R. T. (2018). Artificial Intelligence in Service. *Journal of Service Research*, *21*(2), 155–172. https://doi.org/10.1177/1094670517752459

Huang, Y., Fields, K. G., & Ma, Y. (2022). A tutorial on generative adversarial networks with application to classification of imbalanced data. *Statistical Analysis and Data Mining*, *15*(5), 543–552. https://doi.org/10.1002/sam.11570

Hugh J. Watson. (2017). Preparing for the Cognitive Generation of Decision Support. *MIS Quarterly Executive*, *16*(3), 153–169.

Hunkenschroer, A. L., & Luetge, C. (2022). Ethics of AI-Enabled Recruiting and Selection: A Review and Research Agenda. In *Journal of Business Ethics* (Vol. 178, Issue 4, pp. 977–1007). Springer Science and Business Media B.V. https://doi.org/10.1007/s10551-022-05049-6

Jaton, F. (2021). Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application. *Big Data & Society*, *8*(1), 205395172110135. https://doi.org/10.1177/20539517211013569

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Jialun Qin, Yilu Zhou, M. C. and H. C. (2006). Multilingual Web Retrieval: An Experiment in English–Chinese Business Intelligence Jialun. *Journal of the American Society for Information Science and Technology*, *57*(March), 671–683. https://doi.org/10.1002/asi.20329

Joan Palmiter Bajorek. (2019). Voice Recognition Still Has Significant Race and Gender Biases. *Harvard Business Review*. https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johnson, C. D., Bauer, B. C., & Niederman, F. (2021). The automation of management and business science. *Academy of Management Perspectives*, *35*(2), 292–309. https://doi.org/10.5465/AMP.2017.0159

Johnson, C. J., Demir, M., McNeese, N. J., Gorman, J. C., Wolff, A. T., & Cooke, N. J. (2021). The Impact of Training on Human–Autonomy Team Communications and Trust Calibration. *Human Factors*. https://doi.org/10.1177/00187208211047323

Kaplan, A., & Haenlein, M. (2019a). Rulers of the world, unite! The challenges and opportunities of artificial intelligence. *Business Horizons*, *63*(1), 37–50. https://doi.org/10.1016/j.bushor.2019.09.003

Kaplan, A., & Haenlein, M. (2019b). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Karimi, F., Génois, M., Wagner, C., Singer, P., & Strohmaier, M. (2018). Homophily influences ranking of minorities in social networks. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-29405-7

Kelley, S., Ovchinnikov, A., Hardoon, D. R., & Heinrich, A. (2022). Antidiscrimination Laws, Artificial Intelligence, and Gender Bias: A Case Study in Nonmortgage Fintech Lending. *Manufacturing and Service Operations Management*, *24*(6), 3039–3059. https://doi.org/10.1287/msom.2022.1108

Khan, A. A., Badshah, S., Liang, P., Waseem, M., Khan, B., Ahmad, A., Fahmideh, M., Niazi, M., & Akbar, M. A. (2022). Ethics of AI: A Systematic Literature Review of Principles and Challenges. *ACM International Conference Proceeding Series*, 383–392. https://doi.org/10.1145/3530019.3531329

Khatib, T., Mohamed, A., Mahmoud, M., & Sopian, K. (2011). Modeling of daily solar energy on a horizontal surface for five main sites in Malaysia. *International Journal of Green Energy*, *8*(8), 795–819. https://doi.org/10.1080/15435075.2011.602156

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Kim, B., Park, J., & Suh, J. (2020). Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems*, *134*. https://doi.org/10.1016/j.dss.2020.113302

Kim, J. Y., & Heo, W. G. (2022). Artificial intelligence video interviewing for employment: perspectives from applicants, companies, developer and academicians. *Information Technology and People*, *35*(3), 861–878. https://doi.org/10.1108/ITP-04-2019-0173

King, M. R. (2023). A Conversation on Artificial Intelligence, Chatbots, and Plagiarism in Higher Education. In *Cellular and Molecular Bioengineering* (Vol. 16, Issue 1, pp. 1–2). Springer. https://doi.org/10.1007/s12195-022-00754-8

Kitsios, F., & Kamariotou, M. (2021). Artificial intelligence and business strategy towards digital transformation: A research agenda. *Sustainability (Switzerland)*, *13*(4), 1–16. https://doi.org/10.3390/su13042025

Kitto, K., & Knight, S. (2019). Practical ethics for building learning analytics. *British Journal of Educational Technology*, *50*(6), 2855–2870. https://doi.org/10.1111/bjet.12868

Köchling, A., Riazy, S., Wehner, M. C., & Simbeck, K. (2021). Highly Accurate, But Still Discriminatory: A Fairness Evaluation of Algorithmic Video Analysis in the Recruitment

Context. *Business and Information Systems Engineering*, *63*(1), 39–54. https://doi.org/10.1007/s12599-020-00673-w

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409. https://doi.org/10.1080/0960085X.2021.1927212

Korteling, J. E. (Hans), van de Boer-Visschedijk, G. C., Blankendaal, R. A. M., Boonekamp, R. C., & Eikelboom, A. R. (2021). Human- versus Artificial Intelligence. *Frontiers in Artificial Intelligence*, *4*(March), 1–13. https://doi.org/10.3389/frai.2021.622364

Kumar, G., Singh, U. P., & Jain, S. (2021). Swarm Intelligence Based Hybrid Neural Network Approach for Stock Price Forecasting. In *Computational Economics* (Vol. 60, Issue 3). Springer US. https://doi.org/10.1007/s10614-021-10176-9

Kumar, N., Venugopal, D., Qiu, L., & Kumar, S. (2018). Detecting Review Manipulation on Online Platforms with Hierarchical Supervised Learning. *Journal of Management Information Systems*, *35*(1), 350–380. https://doi.org/10.1080/07421222.2018.1440758

Laitila, P., & Virtanen, K. (2022). Portraying probabilistic relationships of continuous nodes in Bayesian networks with ranked nodes method. *Decision Support Systems*, *154*(October 2021), 113709. https://doi.org/10.1016/j.dss.2021.113709

Lamboglia, R., Lavorato, D., Scornavacca, E., & Za, S. (2020). Exploring the relationship between audit and technology. A bibliometric analysis. *Meditari Accountancy Research*. https://doi.org/10.1108/MEDAR-03-2020-0836

Langer, M., & König, C. J. (2021a). Introducing a multi-stakeholder perspective on opacity, transparency and strategies to reduce opacity in algorithm-based human resource management. *Human Resource Management Review*, *xxxx*, 100881. https://doi.org/10.1016/j.hrmr.2021.100881

Langer, M., & König, C. J. (2021b). Trust in Artificial Intelligence: Comparing trust processes between human and automated trustees in light of unfair bias. *Journal of Business and Psychology*, *January*, 1–52. https://doi.org/10.1007/s10869-022-09829-9

Lappas, T., Sabnis, G., & Valkanas, G. (2016). The impact of fake reviews on online visibility: A vulnerability assessment of the hotel industry. *Information Systems Research*, *27*(4), 940–961. https://doi.org/10.1287/isre.2016.0674

Larson, L., & DeChurch, L. A. (2020). Leading teams in the digital age: Four perspectives on technology and what they mean for leading teams. *Leadership Quarterly*, *31*(1), 101377. https://doi.org/10.1016/j.leaqua.2019.101377

Lee, K. C., Han, I., & Kwon, Y. (1996). Hybrid neural network models for bankruptcy predictions. *Decision Support Systems*, *18*(1 SPEC. ISS.), 63–72. https://doi.org/10.1016/0167-9236(96)00018-8

Lee, K., Park, J., Kim, I., & Choi, Y. (2018). Predicting movie success with machine learning techniques: ways to improve accuracy. *Information Systems Frontiers*, *20*(3), 577–588. https://doi.org/10.1007/s10796-016-9689-z

Li, O. (2021). Problems with "Friendly AI." *Ethics and Information Technology*, *23*(3), 543–550. https://doi.org/10.1007/s10676-021-09595-x

Li, S., Iqbal, M., & Saxena, N. (2022). Future Industry Internet of Things with Zero-trust Security. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10199-5

Liben, L. S., Bigler, R. S., & Krogh, H. R. (2002). Language at Work: Children's Gendered Interpretations of Occupational Titles. *Child Development*, *73*(3), 810–828. https://doi.org/10.1111/1467-8624.00440

Lin, M. S., Liang, Y., Xue, J. X., Pan, B., & Schroeder, A. (2021). Destination image through social media analytics and survey method. *International Journal of Contemporary Hospitality Management*, *33*(6), 2219–2238. https://doi.org/10.1108/IJCHM-08-2020-0861

Lind, M. R., & Sulek, J. M. (2000). A methodology for forecasting knowledge work projects. *Computers and Operations Research*, *27*(11–12), 1153–1169. https://doi.org/10.1016/S0305-0548(99)00150-1

Liu, J., Gong, M., Xiao, L., Zhang, W., & Liu, F. (2022). Evolving Connections in Group of Neurons for Robust Learning. *IEEE Transactions on Cybernetics*, *52*(5), 3069–3082. https://doi.org/10.1109/TCYB.2020.3022673

Lundvall, B. Å. (2017). Is there a technological fix for the current global stagnation?: A response to Daniele Archibugi, Blade Runner economics: Will innovation lead the economic recovery? *Research Policy*, *46*(3), 544–549. https://doi.org/10.1016/j.respol.2016.06.011

Madhavan, M., Wangtueai, S., Sharafuddin, M. A., & Chaichana, T. (2022). The Precipitative Effects of Pandemic on Open Innovation of SMEs: A Scientometrics and Systematic Review of Industry 4.0 and Industry 5.0. In *Journal of Open Innovation: Technology, Market, and Complexity* (Vol. 8, Issue 3). MDPI. https://doi.org/10.3390/joitmc8030152

Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research*, *120*, 262–273. https://doi.org/10.1016/J.JBUSRES.2020.07.045

Makri, C., & Neely, A. (2021). Grounded Theory: A Guide for Exploratory Studies in Management Research. *International Journal of Qualitative Methods*, *20*. https://doi.org/10.1177/16094069211013654

Marshall, R., Pardo, A., Smith, D., & Watson, T. (2022). Implementing next generation privacy and ethics research in education technology. *British Journal of Educational Technology*, *53*(4), 737–755. https://doi.org/10.1111/bjet.13224

Martin, K., Shilton, K., & Smith, J. (2019). Business and the Ethical Implications of Technology: Introduction to the Symposium. In *Journal of Business Ethics* (Vol. 160, Issue 2, pp. 307–317). Springer. https://doi.org/10.1007/s10551-019-04213-9

Martin Kunca, M. J. M. and R. V. (2018). A computational literature review of the field of System Dynamics from 1974 to 2017. *Journal of Simulation*, *12*(2), 115–117. https://doi.org/10.1080/17477778.2018.1468950

McLean, G., Osei-Frimpong, K., & Barhorst, J. (2021). Alexa, do voice assistants influence consumer brand engagement? – Examining the role of AI powered voice assistants in influencing consumer brand engagement. *Journal of Business Research*, *124*(April 2019), 312–328. https://doi.org/10.1016/j.jbusres.2020.11.045

Meek, T., Barham, H., Beltaif, N., Kaadoor, A., & Akhter, T. (2017). Managing the ethical and risk implications of rapid advances in artificial intelligence: A literature review. *PICMET 2016 - Portland International Conference on Management of Engineering and Technology: Technology Management For Social Innovation, Proceedings*, 682–693. https://doi.org/10.1109/PICMET.2016.7806752

Meghan Rimol. (2021). Gartner Identifies Four Trends Driving Near-Term Artificial Intelligence Innovation. *Gartner*, 1–3.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6). https://doi.org/10.1145/3457607

Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side' of AI. In *European Journal of Information Systems* (Vol. 31, Issue 3, pp. 257–268). Taylor and Francis Ltd. https://doi.org/10.1080/0960085X.2022.2026621

Mortenson, M. J., & Vidgen, R. (2016). International Journal of Information Management A computational literature review of the technology acceptance model. *International Journal of Information Management*, *36*(6), 1248–1259. https://doi.org/10.1016/j.ijinfomgt.2016.07.007

Mousavi, R., Raghu, T. S., & Frey, K. (2020). Harnessing Artificial Intelligence to Improve the Quality of Answers in Online Question-answering Health Forums. *Journal of Management Information Systems*, *37*(4), 1073–1098. https://doi.org/10.1080/07421222.2020.1831775

Mulgan, T. (2019). Corporate Agency and Possible Futures. *Journal of Business Ethics*, *154*(4), 901–916. https://doi.org/10.1007/s10551-018-3887-1

Nadj, M., Maedche, A., & Schieder, C. (2020). The effect of interactive analytical dashboard features on situation awareness and task performance. *Decision Support Systems*, *135*(May), 113322. https://doi.org/10.1016/j.dss.2020.113322

Nagel, L. (2020). The influence of the COVID-19 pandemic on the digital transformation of work. *International Journal of Sociology and Social Policy*, *40*(9–10), 861–875. https://doi.org/10.1108/IJSSP-07-2020-0323

Nazaretsky, T., Ariely, M., Cukurova, M., & Alexandron, G. (2022). Teachers' trust in AI-powered educational technology and a professional development program to improve it. *British Journal of Educational Technology*, *53*(4), 914–931. https://doi.org/10.1111/bjet.13232

Nishant, R., Kennedy, M., & Corbett, J. (2020). Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda. *International Journal of Information Management*, *53*(April), 102104. https://doi.org/10.1016/j.ijinfomgt.2020.102104

Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., & Floridi, L. (2023). Taking AI risks seriously: a new assessment model for the AI Act. *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01723-z

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., … Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), 1–14. https://doi.org/10.1002/widm.1356

Okoli, C., & Schabram, K. (2010). Working Papers on Information Systems A Guide to Conducting a Systematic Literature Review of Information Systems Research. *Working Papers on Information Systems*, *10*(2010). https://doi.org/10.2139/ssrn.1954824

Ozkazanc-Pan, B. (2021). Diversity and future of work: inequality abound or opportunities for all? *Management Decision*, *59*(11), 2645–2659. https://doi.org/10.1108/MD-02-2019-0244

Pelau, C., Pop, M. I., Ene, I., & Lazar, L. (2021). Clusters of skeptical consumers based on technology and ai acceptance, perception of social media information and celebrity trend setter.

Journal of Theoretical and Applied Electronic Commerce Research, *16*(5), 1231–1247. https://doi.org/10.3390/jtaer16050069

Pendharkar, P. C. (2007). The theory and experiments of designing cooperative intelligent systems. *Decision Support Systems*, *43*(3), 1014–1030. https://doi.org/10.1016/j.dss.2005.05.028

Peng, C., Merat, N., Romano, R., Hajiseyedjavadi, F., Paschalidis, E., Wei, C., Radhakrishnan, V., Solernou, A., Forster, D., & Boer, E. (2022). Drivers' Evaluation of Different Automated Driving Styles: Is It Both Comfortable and Natural? *Human Factors*, *44*(0), 1–38. https://doi.org/10.1177/00187208221113448

Pereira, V., Hadjielias, E., Christofi, M., & Vrontis, D. (2021). A systematic literature review on the impact of artificial intelligence on workplace outcomes: A multi-process perspective. *Human Resource Management Review*, *September*, 100857. https://doi.org/10.1016/j.hrmr.2021.100857

Peter Luhn. (1958). The automatic creation of literature abstracts. *IBM Journal of Research*.

Piri, S., Delen, D., & Liu, T. (2018). A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets. *Decision Support Systems*, *106*, 15–29. https://doi.org/10.1016/j.dss.2017.11.006

Pozzebon, M., Petrini, M., de Mello, R. B., & Garreau, L. (2011). Unpacking researchers' creativity and imagination in grounded theorizing: An exemplar from IS research. *Information and Organization*, *21*(4), 177–193. https://doi.org/10.1016/j.infoandorg.2011.09.001

Pütz, L., & Werner, A. (2024). Absorptive capacity in family firms: a systematic literature review. *Review of Managerial Science*, *18*(2), 577–632. https://doi.org/10.1007/s11846-023-00651-y

Qazi, A., Raj, R. G., Hardaker, G., & Standing, C. (2017). A systematic literature review on opinion types and sentiment analysis techniques: Tasks and challenges. *Internet Research*, *27*(3), 608–630. https://doi.org/10.1108/IntR-04-2016-0086

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). Is ChatGPT a General-Purpose Natural Language Processing Task Solver? *ArXiv Preprint ArXiv:2302.06476*. http://arxiv.org/abs/2302.06476

Qin, Y., Xu, Z., Wang, X., & Skare, M. (2024). Artificial Intelligence and Economic Development: An Evolutionary Investigation and Systematic Review. *Journal of the Knowledge Economy*, *15*(1), 1736–1770. https://doi.org/10.1007/s13132-023-01183-2

Ransbotham, S., Fichman, R. G., Gopal, R., & Gupta, A. (2016). Special section introduction: Ubiquitous IT and digital vulnerabilities. *Information Systems Research*, *27*(4), 834–847. https://doi.org/10.1287/isre.2016.0683

Rathore, B., Gupta, R., Biswas, B., Srivastava, A., & Gupta, S. (2022). Identification and analysis of adoption barriers of disruptive technologies in the logistics industry. *International Journal of Logistics Management*, *33*(5), 136–169. https://doi.org/10.1108/IJLM-07-2021-0352

Raza, S. A. (2022). A systematic literature review of RFID in supply chain management. *Journal of Enterprise Information Management*, *35*(2), 617–649. https://doi.org/10.1108/JEIM-08-2020-0322

Reis, J., Santo, P. E., & Melão, N. (2019). Artificial Intelligence in Government Services: A Systematic Literature Review. *Advances in Intelligent Systems and Computing*, *930*, 241–252. https://doi.org/10.1007/978-3-030-16181-1_23

Riazy, S., Simbeck, K., & Schreck, V. (2021). Systematic Literature Review of Fairness in Learning Analytics and Application of Insights in a Case Study. In *Communications in Computer and Information Science: Vol. 1473 CCIS*. Springer International Publishing. https://doi.org/10.1007/978-3-030-86439-2_22

Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev. Online*, *15*(94).

Rowe, F. (2014). What literature review is not: Diversity, boundaries and recommendations. *European Journal of Information Systems*, *23*(3), 241–255. https://doi.org/10.1057/ejis.2014.7

Sabherwal, R., & Chan, Y. E. (2001). Alignment between Business and IS Strategies: A Study of Prospectors, Analyzers, and Defenders. *Information Systems Research*, *12*(1), 11–33. https://doi.org/10.1287/isre.12.1.11.9714

Saeed, U., Jan, S. U., Lee, Y. D., & Koo, I. (2021). Fault diagnosis based on extremely randomized trees in wireless sensor networks. *Reliability Engineering and System Safety*, *205*(October 2020), 107284. https://doi.org/10.1016/j.ress.2020.107284

Sætra, H. S. (2023). Generative AI: Here to stay, but for good? *Technology in Society*, *75*. https://doi.org/10.1016/j.techsoc.2023.102372

Sarker, S., Chatterjee, S., Xiao, X., & Elbanna, A. (2019). The Sociotechnical Axis of Cohesion for the IS Discipline: Its Historical Legacy and its Continued Relevance. *MIS Quarterly*, *43*(3), 695–719. https://doi.org/10.25300/MISQ/2019/13747

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, *9*, 845–874. https://doi.org/10.1162/tacl

Schanke, S., Burtch, G., & Ray, G. (2021). Estimating the impact of “humanizing” customer service chatbots. *Information Systems Research*, *32*(3), 736–751. https://doi.org/10.1287/ISRE.2021.1015

Schelble, B. G., Lopez, J., Textor, C., Zhang, R., McNeese, N. J., Pak, R., & Freeman, G. (2022). Towards Ethical AI: Empirically Investigating Dimensions of AI Ethics, Trust Repair, and Performance in Human-AI Teaming. *Human Factors*. https://doi.org/10.1177/00187208221116952

Schiebinger, L. (2014). Scientific research must take gender into account. *Nature*, *507*(7490), 9–9. https://doi.org/10.1038/507009a

Schwab, K. (2016). The Fourth Industrial Revolution: what it means and how to respond. *World Economic Forum*, 1–7.

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a standard for identifying and managing bias in artificial intelligence. *US Department of Commerce, National Institute of Standards and Technology*, *3*. https://doi.org/10.6028/NIST.SP.1270

Semenova, V., & Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *Econometrics Journal*, *24*(2), 264–289. https://doi.org/10.1093/ectj/utaa027

Shadikur Rahman, Syeda Sumbul Hossain, Md. Shohel Arman, Lamisha Rawshan, Tapushe Rabaya Toma, Fatama Binta Rafiq, and K. B. Md. B. (2020). Assessing the Effectiveness of Topic Modeling Algorithms in Discovering Generic Label with Description. In *Advances in Intelligent Systems and Computing: Vol. 1130 AISC*. https://doi.org/10.1007/978-3-030-39442-4_5

Shobana, V., & Kumar, N. (2015). Big data - A review. *International Journal of Applied Engineering Research*, *10*(55), 1294–1298. https://doi.org/10.26634/jit.6.1.13507

Sievert, C., & Shirley, K. (2015). *LDAvis: A method for visualizing and interpreting topics*. 63–70. https://doi.org/10.3115/v1/w14-3110

Silva, J. C. Da, Saxena, A., Balaban, E., & Goebel, K. (2012). A knowledge-based system approach for sensor fault modeling, detection and mitigation. *Expert Systems with Applications*, *39*(12), 10977–10989. https://doi.org/10.1016/j.eswa.2012.03.026

Simon, J. P. (2019). Artificial intelligence: scope, players, markets and geography. *Digital Policy, Regulation and Governance* , *21*(3), 208–237. https://doi.org/10.1108/DPRG-08-2018-0039

Sjödin, D., Parida, V., Palmié, M., & Wincent, J. (2021). How AI capabilities enable business model innovation: Scaling AI through co-evolutionary processes and feedback loops. *Journal of Business Research*, *134*(June), 574–587. https://doi.org/10.1016/j.jbusres.2021.05.009

Smacchia, M., & Za, S. (2022). Artificial Intelligence in Organisation and Managerial Studies: A Computational Literature Review. *ICIS 2022 Proceedings. 6*, 0–17. https://aisel.aisnet.org/icis2022/soc_impact_is/soc_impact_is/6

Smacchia, M., & Za, S. (2023). Exploring Artificial Intelligence Bias, Fairness and Ethics in Organisation and Managerial Studies. *ECIS 2023 Research Papers*, 362. https://aisel.aisnet.org/ecis2023_rp/362

Smith, G. (2020). Data mining fool's gold. *Journal of Information Technology*, *35*(3), 182–194. https://doi.org/10.1177/0268396220915600

Smiti, S., & Soui, M. (2020). Bankruptcy Prediction Using Deep Learning Approach Based on Borderline SMOTE. *Information Systems Frontiers*, *22*(5), 1067–1083. https://doi.org/10.1007/s10796-020-10031-6

Smuha, N. A. (2021). From a 'race to AI' to a 'race to AI regulation': regulatory competition for artificial intelligence. *Law, Innovation and Technology*, *13*(1), 57–84. https://doi.org/10.1080/17579961.2021.1898300

Someh, I., Davern, M., Breidbach, C. F., & Shanks, G. (2019). Ethical issues in big data analytics: A stakeholder perspective. *Communications of the Association for Information Systems*, *44*(1), 718–747. https://doi.org/10.17705/1CAIS.04434

Sparrow, R. (2004). The turing triage test. *Ethics and Information Technology*, *6*(4), 203–213. https://doi.org/10.1007/s10676-004-6491-2

Srivastava, B., & Rossi, F. (2018). Towards Composable Bias Rating of AI Services. *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 284–289. https://doi.org/10.1145/3278721.3278744

Stahl, B. C., Timmermans, J., & Mittelstadt, B. D. (2016). The ethics of computing: A survey of the computing-oriented literature. *ACM Computing Surveys*, *48*(4). https://doi.org/10.1145/2871196

Stahlberg, F. (2020). Neural Machine Translation: A Review. In *Journal of Artificial Intelligence Research* (Vol. 69). https://scholar.google.com/scholar?q=%22neural+machine+

Stanford, H. A. I. (2023). *Artificial Intelligence Index Report 2023*. https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf

Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating Gender Bias in Machine Translation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1679–1684. http://arxiv.org/abs/1906.00591

Su Yiran, Wang Emma, & Berthon Pierre. (2023). Ethical Marketing AI? A Structured Literature Review of the Ethical Challenges Posed by Artificial Intelligence in the Domains of Marketing

and Consumer Behavior. *56th Hawaii International Conference on System Sciences (HICCS), Hyatt Regency Maui, January 3-6, 2023*.

Suleiman, A., Tight, M. R., & Quinn, A. D. (2016). Assessment and prediction of the impact of road transport on ambient concentrations of particulate matter PM10. *Transportation Research Part D: Transport and Environment*, *49*, 301–312. https://doi.org/10.1016/j.trd.2016.10.010

Sun, T., Gaut, A., Tang, S., Huang, Y., Elsherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2020). Mitigating Gender Bias in Natural Language Processing: Literature Review. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 1630–1640.

Suzuki, Y. (2009). A decision support system of dynamic vehicle refueling. *Decision Support Systems*, *46*(2), 522–531. https://doi.org/10.1016/j.dss.2008.09.005

Suzuki, Y. (2012). A decision support system of vehicle routing and refueling for motor carriers with time-sensitive demands. *Decision Support Systems*, *54*(1), 758–767. https://doi.org/10.1016/j.dss.2012.09.004

Swanepoel, D. (2021). The possibility of deliberate norm-adherence in AI. *Ethics and Information Technology*, *23*(2), 157–163. https://doi.org/10.1007/s10676-020-09535-1

Tako, A. A., & Robinson, S. (2012). The application of discrete event simulation and system dynamics in the logistics and supply chain context. *Decision Support Systems*, *52*(4), 802–815. https://doi.org/10.1016/j.dss.2011.11.015

Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., & Liu, Y. (2020). Neural machine translation: A review of methods, resources, and tools. In *AI Open* (Vol. 1, pp. 5–21). Elsevier B.V. https://doi.org/10.1016/j.aiopen.2020.11.001

Teodorescu, M. H. M., Morse, L., Awwad, Y., & Kane, G. C. (2021). Failures of fairness in automation require a deeper understanding of human–ml augmentation. *MIS Quarterly: Management Information Systems*, *45*(3), 1483–1499. https://doi.org/10.25300/MISQ/2021/16535

Tiku, N. (2022). The Google engineer who thinks the company's AI has come to life. *The Washington Post*, 1–12. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

Tomalin, M., Byrne, B., Concannon, S., Saunders, D., & Ullmann, S. (2021). The practical ethics of bias reduction in machine translation: why domain adaptation is better than data debiasing. *Ethics and Information Technology*, *23*(3), 419–433. https://doi.org/10.1007/s10676-021-09583-1

Toorajipour, R., Sohrabpour, V., Nazarpour, A., Oghazi, P., & Fischl, M. (2021). Artificial intelligence in supply chain management: A systematic literature review. *Journal of Business Research*, *122*(September 2020), 502–517. https://doi.org/10.1016/j.jbusres.2020.09.009

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. *British Journal of Management*, *14*(3), 207–222. https://doi.org/10.1111/1467-8551.00375

Trunk, A., Birkel, H., & Hartmann, E. (2020). On the current state of combining human and artificial intelligence for strategic organizational decision making. *Business Research*, *13*(3), 875–919. https://doi.org/10.1007/s40685-020-00133-x

Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. (2015). Big data analytics: a survey. *Journal of Big Data*, *2*(1), 1–32. https://doi.org/10.1186/s40537-015-0030-3

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Upadhyay, N., Upadhyay, S., & Dwivedi, Y. K. (2021). Theorizing artificial intelligence acceptance and digital entrepreneurship model. *International Journal of Entrepreneurial Behaviour and Research*. https://doi.org/10.1108/IJEBR-01-2021-0052

Vakkuri, V., Kemell, K. K., & Abrahamsson, P. (2019). Implementing Ethics in AI: Initial Results of an Industrial Multiple Case Study. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11915 LNCS*, 331–338. https://doi.org/10.1007/978-3-030-35333-9_24

Van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, *144*(January), 93–106. https://doi.org/10.1016/j.jbusres.2022.01.076

van Wezel, M., & Potharst, R. (2007). Improved customer choice predictions using ensemble methods. *European Journal of Operational Research*, *181*(1), 436–452. https://doi.org/10.1016/j.ejor.2006.05.029

Verma, S., Sharma, R., Deb, S., & Maitra, D. (2021). Artificial intelligence in marketing: Systematic review and future research direction. *International Journal of Information Management Data Insights*, *1*(1), 100002. https://doi.org/10.1016/j.jjimei.2020.100002

Vieira, L. N., O'Hagan, M., & O'Sullivan, C. (2021). Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. In *Information Communication and Society* (Vol. 24, Issue 11, pp. 1515–1532). Routledge. https://doi.org/10.1080/1369118X.2020.1776370

Vimalkumar, M., Sharma, S. K., Singh, J. B., & Dwivedi, Y. K. (2021). 'Okay google, what about my privacy?': User's privacy perceptions and acceptance of voice based digital assistants. *Computers in Human Behavior*, *120*(February), 106763. https://doi.org/10.1016/j.chb.2021.106763

Vogl, T. M., Seidelin, C., Ganesh, B., & Bright, J. (2020). Smart Technology and the Emergence of Algorithmic Bureaucracy: Artificial Intelligence in UK Local Authorities. *Public Administration Review*, *80*(6), 946–961. https://doi.org/10.1111/puar.13286

Vössing, M., Kühl, N., Lind, M., & Satzger, G. (2022). Designing Transparency for Effective Human-AI Collaboration. *Information Systems Frontiers*, *May*, 877–895. https://doi.org/10.1007/s10796-022-10284-3

Vu, H. L., Ng, K. T. W., Richter, A., & An, C. (2022). Analysis of input set characteristics and variances on k-fold cross validation for a Recurrent Neural Network model on waste disposal rate estimation. *Journal of Environmental Management*, *311*(October 2021), 114869. https://doi.org/10.1016/j.jenvman.2022.114869

Wagner, G., Lukyanenko, R., & Paré, G. (2022). Artificial intelligence and the conduct of literature reviews. *Journal of Information Technology*, *37*(2), 209–226. https://doi.org/10.1177/02683962211048201

Wang, F., Liang, X., Xu, L., & Lin, L. (2022). Unifying Relational Sentence Generation and Retrieval for Medical Image Report Composition. *IEEE Transactions on Cybernetics*, *52*(6), 5015–5025. https://doi.org/10.1109/TCYB.2020.3026098

Wang, Q., Li, B., & Singh, P. V. (2018). Copycats vs. original mobile apps: A machine learning copycat-detection method and empirical analysis. *Information Systems Research*, *29*(2), 273–291. https://doi.org/10.1287/isre.2017.0735

West, D., & Blackman, D. (2015). Performance Management in the Public Sector. *Australian Journal of Public Administration*, *74*(1), 73–81. https://doi.org/10.1111/1467-8500.12130

Wilson, C. (2022). Public engagement and AI: A values analysis of national strategies. *Government Information Quarterly*, *39*(1), 101652. https://doi.org/10.1016/j.giq.2021.101652

Wu, L., & Kane, G. C. (2021). Network-biased technical change: How modern digital collaboration tools overcome some biases but exacerbate others. *Organization Science*, *32*(2), 273–292. https://doi.org/10.1287/orsc.2020.1368

Xiaoyan Cai, W. L. (2011). Enhancing Sentence-Level Clustering With Integrated and Interactive Frameworks for Theme-Based Summarization. *Journal of the American Society for Information Science and Technology*, *62*(October), 2067–2082. https://doi.org/10.1002/asi.21593

Xie, J., & Jain, T. (2024). Unpacking micro-CSR through a computational literature review: An identity heterogeneity view of internal stakeholders. *Journal of Business Research*, *172*, 114451. https://doi.org/10.1016/j.jbusres.2023.114451

Xie, W., Shen, L., & Duan, J. (2021). Adaptive Weighting of Handcrafted Feature Losses for Facial Expression Recognition. *IEEE Transactions on Cybernetics*, *51*(5), 2787–2800. https://doi.org/10.1109/TCYB.2019.2925095

Xu, L., Mak, S., & Brintrup, A. (2021). Will bots take over the supply chain? Revisiting agent-based supply chain automation. *International Journal of Production Economics*, *241*(February), 108279. https://doi.org/10.1016/j.ijpe.2021.108279

Yela, C. (2004). Interpersonal Attraction. In *Encyclopedia of Applied Psychology* (pp. 379–385). Elsevier. https://doi.org/10.1016/B0-12-657410-3/00057-X

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2017). Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Computational Intelligence Magazine*, *13*(3), 55–75. http://arxiv.org/abs/1708.02709

Zach, T. H. and L. (2012). Use of Microblogging for Collective Sense-Making During Violent Crises: A Study of Three Campus Shootings. *Journal of the American Society for Information Science and Technology*, *63*(January), 34–47. https://doi.org/10.1002/asi.21685

Zhang, L., Yan, Q., & Zhang, L. (2018). A computational framework for understanding antecedents of guests' perceived trust towards hosts on Airbnb. *Decision Support Systems*, *115*, 105–116. https://doi.org/10.1016/j.dss.2018.10.002

Zhang, L., Yan, Q., & Zhang, L. (2020). A text analytics framework for understanding the relationships among host self-description, trust perception and purchase behavior on Airbnb. *Decision Support Systems*, *133*. https://doi.org/10.1016/j.dss.2020.113288

Zhou, C., Huang, B., & Fränti, P. (2022). A review of motion planning algorithms for intelligent robots. *Journal of Intelligent Manufacturing*, *33*(2), 387–424. https://doi.org/10.1007/s10845-021-01867-z

Zhu, Y. Q., Corbett, J. ueline, & Chiu, Y. Te. (2021). Understanding employees' responses to artificial intelligence. *Organizational Dynamics*, *50*(2), 100786. https://doi.org/10.1016/j.orgdyn.2020.100786

Zuiderwijk, A., Chen, Y. C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*, *38*(3), 101577. https://doi.org/10.1016/j.giq.2021.101577