AI AND DIVERSITY IN A DATAFIED WORLD OF WORK

INTERNATIONAL PERSPECTIVES ON EQUALITY, DIVERSITY AND INCLUSION

Series Editor: Mustafa F. Özbilgin

Recent Volumes:

Volume 1: *Practical and Theoretical Implications of Successfully Doing Difference in Organizations,* Edited by Donnalyn Pompper

Volume 2: *Gender, Careers and Inequalities in Medicine and Medical Education: International Perspectives,* Edited by Maria Tsouroufli

Volume 3: *Management and Diversity: Perspectives from Different National Contexts,* Edited by Mustafa F. Özbilgin and Jean-François Chanlat

Volume 4: *Management and Diversity: Thematic Approaches,* Edited by Mustafa F. Özbilgin and Jean-François Chanlat

Volume 5: *The Strength of Difference: Itineraries of Atypical Bosses,* Edited by Norbert Alter

Volume 6: *Race Discrimination and Management of Ethnic Diversity and Migration at Work: European Countries' Perspectives,* Edited by Joana Vassilopoulou, Julienne Brabet and Victoria Showunmi

Volume 7: *Lived Experiences of Exclusion in the Workplace: Psychological and Behavioural Effects,* Edited by Kurt April, Babar Dharani and Amanda Ronita April

Volume 8: *Management and Organizational Studies on Blue & Grey Collar Workers: Diversity of Collars,* Edited by Joanna Paliszkiewicz and Demet Varoğlu

Volume 9: *Contemporary Approaches in Equality, Diversity and Inclusion: Strategic and Technological Perspectives,* Edited by Berk Kucukaltan

Volume 10: *Care and Compassion in Capitalism,* Edited by Cagri Yalkin and Mustafa F. Özbilgin

Volume 11A: *Future Workscapes: Strategic Insights and Innovations in Human Resources and Organizational Development,* Edited by Joanna Paliszkiewicz, Demet Varoğlu and Olena Kulykovets

Volume 11B: *Future Workscapes: Emerging Business Trends and Innovations,* Edited by Joanna Paliszkiewicz, Demet Varoğlu and Olena Kulykovets

INTERNATIONAL PERSPECTIVES ON EQUALITY,
DIVERSITY AND INCLUSION VOLUME 12

AI AND DIVERSITY IN A DATAFIED WORLD OF WORK: WILL THE FUTURE OF WORK BE INCLUSIVE?

EDITED BY

JOANA VASSILOPOULOU
Brunel University of London, UK

AND

OLIVIA KYRIAKIDOU
The American College of Greece, Greece

United Kingdom – North America – Japan
India – Malaysia – China

Emerald Publishing Limited
Emerald Publishing, Floor 5, Northspring, 21-23 Wellington Street, Leeds LS1 4DL.

First edition 2025

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-1-83797-297-5 (Print)
ISBN: 978-1-83797-296-8 (Online)
ISBN: 978-1-83797-298-2 (Epub)

ISSN: 2051-2333 (Series)

CONTENTS

ABOUT THE EDITORS

Joana Vassilopoulou is a Professor in Equality, Diversity, and Inclusion & HRM at Brunel University of London. She has over 50 peer-reviewed publications in books and journals, such as *Human Resource Management Journal, Journal of World Business,* and *Work, Employment and Society*. She has successfully acquired grants with a total value of over £2.5 million, such as for the OECD, the European Academy of Management, EU Horizon2020, and Erasmus+. She is a certified trainer in Managing Gender and diversity and HRM and has developed and delivered training for organisations internationally. She is also the co-founder of CEFI, the Centre for Inclusion at Work (https://www.inclusionat- work.eu), offering consulting and training services to organisations on diversity, equity, and inclusion at work. She has a PhD from the University of East Anglia, UK, and a Sociology degree from the University of Duisburg-Essen, Germany.

Olivia Kyriakidou is an Organisational Psychologist, Assistant Professor in Psychology, and Coordinator of the MS in Organisational Psychology at Deree-ACG. She is also the Co-founder of CEFI, the Centre for Inclusion at Work, offering consulting and training services to organisations on issues of diversity, equity, and inclusion at work. Her current research interests are focussed on the field of diversity, equity, and inclusion in the workplace with a special interest in the impact of artificial intelligence and automation processes on employees and their employment.

This page intentionally left blank

ABOUT THE CONTRIBUTORS

Kurt April is the Allan Gray Chair, an Endowed Professorship, specialising in Leadership, Diversity, and Inclusion, and is Adjunct Faculty of Saïd Business School, University of Oxford (UK). His most recent diversity, equity, and inclusion (DEI) books were *Lived Experiences of Exclusion in the Workplace: Psychological and Behavioural Effects* (Emerald Publishing) and *12 Lenses Into Diversity in South Africa* (KR Publishing). Outside of academia, he is Managing Partner of LICM Consulting, Director (DEI) at Oxford Acuity, Director of the Achievement Awards Group and bountiXP, as well as Ambassador of the Global Unashamedly Ethical Movement. He is a Series Editorial Board Member of Palgrave (Springer) on EDI and Indigenization in Business (Switzerland), Associate Editor for *Equality, Diversity and Inclusion: An International Journal* (UK), Associate Editor for *Design Science* (UK), and Editorial Board Member of the *European Management Review* (UK).

Nergiz Çetin-Kılıç is a researcher at Özyeğin University and a PhD candidate in Management and Organisation at Marmara University, Istanbul, Turkey. Her research primarily focuses on organisational behaviour, with a particular interest in work–family dynamics and gender equality in the workplace. Recently, she has expanded her work to explore the role of artificial intelligence in shaping diversity, equity, and inclusion within organisations.

Preeya Daya is an educator, researcher, and consultant in the fields of leadership, people management, and organisational behaviour. She is committed to building the awareness of how values-orientated, conscious leaders can enable thriving, purpose-led organisations where people are engaged and deliver high performance. She has a PhD from the University of Cape Town and is a Certified Integral Coach. In addition to her role at the Centre for Values-Based Leadership, she is an Academic Director and Board Member at the Achievement Awards Group. She serves on the Board of the Institute of Natural Resources and as a Council Member for the Durban Chamber of Commerce, where she heads the Human Resources Chapter. She is a Co-Editor of the book *12 Lenses Into Diversity in South Africa*, serves as an Associate Editor, and has authored numerous peer-reviewed publications.

Cihat Erbil is an Associate Professor in the Department of Business Administration at Ankara Haci Bayram Veli University, Turkey. His research focuses on organisational sociology and critical management studies. Through his work, he aims to amplify the voices of 'others' and increase their visibility.

Theodoros Fouskas is an Assistant Professor in 'Sociology with an emphasis on Migration and Public Health' at the Department of Public Health Policy, School of Public Health, University of West Attica (Greece). His research and teaching interests combine the sociological study of migration, migrant/refugee health and access to health care, health inequalities and racial and ethnic disparities, precarious employment, low-status/low-paid work, domestic work and caregiving, migrant community associations, solidarity and pro/anti-migrant mobilisations, social integration and exclusion of third-country nationals, and migration and health policies. His latest authored monograph is *2021 Precarious Lives of Maids, Nannies and Caregivers in Greece*, Nova Publishers and his latest edited publication is *2021 Immigrants, Asylum Seekers and Refugees in Times of Crises*, EPLO (2 Volumes). Since 2005, he has participated in European and national research projects on migrants and other vulnerable social groups.

Bessie Goumas recently earned a Master of Science in Organisational Psychology. With a background in professional and academic administration, she is focussed on advancing workplace inclusion and diversity. She is especially interested in how artificial intelligence can be leveraged to drive equitable hiring practices and support the career growth of marginalised groups, contributing to fairer and more inclusive work environments.

Nur Gundogdu is an Assistant Professor of Strategy and Sustainability at Birmingham Business School, University of Birmingham, UK. Previously, she worked at Brunel University of London and Aston University. Her main research interests include artificial intelligence and digital transformation for sustainability, sustainable, responsible, and inclusive business, well-being and mental health, gender equality, and diversity at work. She co-authored the chapter titled 'Artificial Intelligence, Gig Economy, and Precarity' in the book *Diversity and Precarious Work During Socio-Economic Upheaval*, published by Cambridge University Press. Throughout her career, she has carried out academic and professional careers together and collaborated with academia, government, industry, and non-governmental organisations in a national and international context. Her research and consultancy work has received funding from prominent organisations, including EU Horizon 2020, UKRI ESRC, UKRI Research England, UKRI EPSRC, and the British Council.

Tally Hatzakis is a dedicated researcher with over 25 years of experience at the intersection of technology and society, especially focussed on human flourishing. Her work spans high-impact European Union-funded projects and mentorship of emerging scholars, consistently pushing boundaries in both academia and industry. Her expertise covers foresight research, strategic analysis, user research, and co-design methodologies. In the past seven years, she has focussed on artificial intelligence in diverse domains and actively advocated for inclusivity in tech, championing the involvement of persons with disabilities in shaping emerging technologies. Her contributions are reflected in an extensive publication record.

Lefteris Kretsos is a Senior Lecturer in Business and Management at Brunel University of London. He has contributed to over 15 international research projects funded by bodies like the European Commission and the ESRC, with numerous consultancy projects and publications addressing labour market regulation and precarious employment. Previously, as Greece's Cabinet Minister for Digital Policy and Media, he led initiatives positioning Greece as a global film production hub through incentives such as cash rebates and tax deductions, significantly advancing the country's creative industries.

Maria Manteli holds an MSc in Organisational Psychology, and in her many years of vocational experience in multinational companies in Greece and abroad, she has a passion for enhancing employee well-being and promoting diversity, equity, and inclusion strategies. She is currently a corporate wellness advisor at The Healing Tree Project co-op and a mentor at Women On Top, an organisation working for the economic and professional empowerment of women and for gender equality and inclusion in the workplace. She also leads the circle 'Wellbeing at Work' at Lean In, a global community dedicated to helping foster leadership, advancement, and inclusion for women in the workplace and co-leads the circle 'Expat Women Returning to Greece', aiming to support Greece Brain Gain. She also participates in initiatives that include artificial intelligence (AI) and possible ways AI can be utilised to promote employee well-being, inclusive workplaces, and in designing training and development strategies for the corporate future.

Aybike Mergen is a social sciences scholar who is extremely curious about leadership, decision-making mechanisms, artificial intelligence, and ethics. She is currently an Assistant Professor of Management at Özyeğin University. She obtained her PhD from the Graduate School of Business, Koç University, and she was a Visiting Scholar at the Schulich School of Business, York University, during her PhD studies. She obtained two master's degrees, an MSc in Economics from Tilburg University and an MA in Political Science from Sabancı University. She has two years of management consultancy experience in Ernst & Young and Deloitte.

Mustafa F. Özbilgin is a Professor of Organisational Behaviour at Brunel University of London. His research delves into workplace equality, diversity, and inclusion from comparative and relational perspectives, focussing on transforming policies and practices. An advocate for workplace democracy and humanisation, he has held numerous key roles, including Editor-in-Chief of the *British Journal of Management* and *European Management Review*. He's also the founder of the Equality, Diversity and Inclusion: an International Conference (EDI). With over 30 books and 200 papers in top-tier journals, his influence extends globally.

Doris Schneeberger is a Postdoctoral University Assistant at the Institute for Change Management and Management Development at the Vienna University of Economics and Business, Austria. Her current research focuses on the role of nonhuman animals in organisations from a business ethics perspective. She is

the author of the forthcoming book *Envisioning a Better Future for Nonhuman Animals: Towards Future Animal Rights Declarations* and has published her research in *Business Ethics, the Environment and Responsibility*, and *The Oxford Handbook of Animal Organization Studies*, among many other publications.

Evi Tasoula is a human resource (HR) professional with expertise in talent acquisition, employee experience, and HR operations. She holds a master's in Organisational Psychology and a bachelor's in Accounting and Finance. As an HR Operations Specialist at Printec Group of Companies, she excels in streamlining recruitment, improving onboarding, and driving employee engagement. Previously, she worked as an Assistant Accountant, contributing to the company's digital transformation. Passionate about how technology can enhance people's lives, she combines her tech-savviness with a commitment to continuous learning, always looking for ways to improve organisational processes and culture.

Laura Traavik, PhD, is an Associate Professor in the Department of Leadership and Organization at Kristiania University College, Norway, where she also serves as the Program Director for the Master of Leadership. Additionally, she leads the research group on Workplace Inclusion and Diversity. Her teaching and research interests include inclusion, diversity, team dynamics, negotiation, and digital mindsets. She is also an active speaker in Norway's private and public sectors, focussing on inclusion, diversity, and negotiation topics.

George Varvatsoulis is a doctoral student in Science and Technology Studies at York University, Canada. His academic work centres on the intersections of technology, media, and society, with a focus on how digital innovation shapes modern workplaces. In his research, he has developed the concept of Organisational Technoscience, a field that examines the influence of emerging technologies on organisational structures and practices. In addition to his research, he is a dedicated educator, earning multiple award nominations for his teaching. He holds master's degrees in Media Production and Digital Media from Toronto Metropolitan University, Business Administration from the University of Essex, and Organisational Psychology from the American College of Greece. Drawing on both academic and professional experience, he integrates practical insights from his leadership roles into his teaching, preparing students to thrive in complex, evolving environments.

Emily Yarrow is a Senior Lecturer in Management and Organisations, in the Leadership, Work, and Organisation Subject Group at Newcastle University Business School, UK. She is a Senior Fellow of the Higher Education Academy (SFHEA) and Academic Member of the CIPD (MCIPD). Her scholarly work focuses on and contributes to contemporary understandings of gendered organisational behaviour, inequality regimes, women's experiences of organisational life, and the (inclusive) future of work. Her current work is focussed on algorithmic bias perpetuation in hiring platforms, exploring how subjective preferences that reflect societal biases, which are inscribed in algorithms, may lead to unfair or discriminatory hiring outcomes.

Patrizia Zanoni is a Full Professor at the School of Social Sciences of Hasselt University, Belgium, of which she is one of the founders. Drawing on critical theories including contemporary strands of Marxism and critical discourse analysis, her research investigates difference, work, and value under capitalism and beyond. She has written extensively on the unequal fragmentation of labour (or 'diversity'), social reproduction, digital technologies reshaping work, employment and labour, the platform economy, control and resistance, and the possibility of organising the economy and society otherwise. She is currently principal investigator (PI) in two large European projects (Humans in Digital Logistics and INSPIRE) and Co-Editor-in-Chief of *Organization*: the critical journal of organisation, theory, and society. In 2022, she received an Emma Goldman award from the Flax Foundation for her work on inequality.

Jannes Zwaenepoel is postdoctoral researcher at the School for Social Sciences at Hasselt University, Belgium. His research investigates how advanced digital technologies such as machine learning algorithms affect work and workers. Questions of how people experience work during times of digital transition but also how this transition affects policy are central to his research. He frequently employs critical theories to understand how innovations can be used for multiple agendas.

This page intentionally left blank

PREFACE

As artificial intelligence (AI) rapidly transforms industries, reshaping decision-making and redefining efficiency, it simultaneously brings forth profound ethical questions, particularly in the context of diversity, equity, and inclusion (DEI) in the workplace. The technology heralds exciting potential: AI can automate mundane tasks, streamline recruitment, personalise employee experiences, and support data-driven insights for better workforce management. Yet, beneath these promises lies an urgent need to critically examine AI's impacts on marginalised communities and the fabric of organisational life.

This book grew from recognising that AI's integration into our daily work is not a neutral process; instead, it is one intricately bound to the biases and inequalities embedded in our social structures. Algorithms may be programmed to be objective, yet they inherit and often amplify the biases from the data that train them, the environments where they are deployed, and the societies surrounding them. This book addresses the gaps between AI's potential benefits and the risks of reinforcing structural inequities, offering readers a nuanced understanding of how AI affects DEI in far-reaching and, at times, unsettling ways.

The impact of AI on DEI is inherently interdisciplinary, intersecting fields such as sociology, ethics, human resources (HR), technology studies, law, and organisational psychology. As such, this book brings together a diverse group of scholars, practitioners, and thought leaders, each offering unique insights into AI's influence on workplace equity. Through an interdisciplinary lens, this book explores topics ranging from the theoretical foundations of AI bias to real-world applications in human resource management (HRM), migration, disability inclusion, and trade unions, to name a few. Together, these contributions build a comprehensive picture of the landscape of AI in today's workplaces and beyond.

THEMATIC OVERVIEW OF THIS BOOK

The chapters within this volume navigate the complex and layered challenges that AI presents for DEI in contemporary organisations.

The opening chapters set the stage by investigating AI bias's theoretical and historical roots, examining how human prejudices can become encoded in algorithms through socio-technical systems. Drawing on theories of intersectionality and systemic discrimination, these chapters reveal that biases in AI are not incidental flaws but reflections of deep-seated social inequities that AI can inadvertently entrench if left unchecked.

Subsequent chapters delve into AI's practical applications in the workplace, particularly in HRM. For instance, this book explores how algorithmic

management tools in HR risk reinforcing biases that disproportionately disadvantage marginalised groups despite their utility in streamlining recruitment and employee evaluation. Case studies of notable incidents – such as the failure of Amazon's AI hiring tool to avoid gender bias – highlight the unintended consequences that can arise when AI systems operate without adequate oversight or accountability. Through critical analysis, this book underscores the need for frameworks like fairness, accountability, and transparency (FAccT) to serve as ethical pillars guiding AI's deployment in high-stakes fields.

Chapters dedicated to the experiences of marginalised communities reveal AI's varied impact on people with disabilities, gig economy workers, migrant populations, and even vegans and vegetarians in the workplace. These groups face unique challenges in an AI-driven workforce, from issues of accessibility and inclusion to new forms of workplace surveillance and assessment that may deepen existing vulnerabilities. Through these perspectives, readers gain an understanding of the nuanced ways AI shapes individual lives and professional opportunities, often along pre-existing fault lines of social inequity.

NAVIGATING ETHICAL AI: A FRAMEWORK FOR THE FUTURE

This book does not merely critique AI's current shortcomings and aims to equip readers with frameworks and strategies to build more inclusive and ethical AI systems. A key message across chapters is that while AI technology is complex, achieving fairness and accountability requires a collective effort that transcends technical solutions. It calls for continuous reflection on how AI systems are developed, deployed, and regulated and for including diverse voices in the conversation to mitigate risks of bias and discrimination.

Several chapters offer practical recommendations for 'coding for equity' by embedding ethical considerations directly into the AI development process. By advocating for transparency in algorithmic processes, regular bias audits, and ongoing stakeholder engagement, this book presents actionable steps towards creating AI that works for everyone – not just for those who already hold positions of power and privilege.

INTERDISCIPLINARY COLLABORATION AND THE ROAD AHEAD

One of this book's most pressing calls to action is the need for interdisciplinary collaboration in AI development and governance. As AI continues to evolve, so must our approaches to understanding and mitigating its impact on diverse communities. Experts from ethics, technology, law, and social sciences must come together to inform and guide AI's trajectory, ensuring that it aligns with principles of social justice and human dignity. By involving people from various fields and backgrounds, particularly those from underrepresented communities, we can build AI systems that respect and reflect our society's diversity.

CONCLUSION

As we step further into an era where AI is woven into the fabric of our workplaces and institutions, the urgency of understanding its DEI implications has never been greater. This book seeks to illuminate the complexities of AI's impact on work and inclusion, advocating for a proactive and ethically grounded approach to AI governance. We hope that the perspectives shared within these pages inspire readers – technologists, HR professionals, policymakers, or simply concerned citizens – to engage with AI critically and champion a vision of the future where technology serves all of humanity equitably.

In the end, AI's impact on society is not predetermined. It is shaped by the choices we make today – the frameworks we adopt, the voices we include, and the values we prioritise. This book is both a reflection on the present and a roadmap for the future, urging readers to envision a world where AI augments our abilities and upholds the principles of justice, fairness, and inclusion that define our humanity.

This page intentionally left blank

CHAPTER 1

INTRODUCTION: AI AND DIVERSITY IN A DATAFIED WORLD OF WORK – WILL THE FUTURE OF WORK BE INCLUSIVE?

Joana Vassilopoulou[a] and Olivia Kyriakidou[b]

[a]*Brunel Business School, Brunel University of London, UK*
[b]*American College of Greece, Greece*

INTRODUCTION

This introductory chapter sets the stage for this edited volume, outlining the rationale for an interdisciplinary approach to studying the impact of artificial intelligence (AI) on diversity, equality, and inclusion (DEI). It provides an overview of the key themes and questions in the subsequent chapters, highlighting the importance of collaboration across different fields to develop holistic solutions to AI's challenges.

The rapid advancement of AI technology is reshaping organisational operations and decision-making processes. This edited volume explores AI's multifaceted impacts on DEI within the workplace, emphasising the importance of interdisciplinary research to understand and address these effects fully. By bringing together scholars and practitioners from diverse fields, including sociology, organisational psychology, human resource management (HRM), DEI, ethics, and technology, this edited volume builds bridges between disciplines to offer a comprehensive perspective on AI's implications.

This edited volume examines AI and the future of workplace DEI in a datafied world of work to provide novel insights into the extant knowledge and practices

AI and Diversity in a Datafied World of Work: Will the Future of Work be Inclusive?
International Perspectives on Equality, Diversity and Inclusion, Volume 12, 1–15

ISSN: 2051-2333/doi:10.1108/S2051-233320250000012002

in the domain. AI is becoming an increasingly important consideration for organisations (Vassilopoulou et al., 2022). For example, algorithms are now widely used by human resource (HR) professionals and other key stakeholders within organisations. Algorithms are defined as 'precise recipes that specify the exact sequence of steps required to solve a problem' (MacCormick, 2012, p. 2). Examples of their use include the employment of such devices to 'enhance' various HR decisions (Langer et al., 2019) and to save costs and increase productivity (McDonald et al., 2017). While there is indeed utility to using algorithms in HRM (Cheng & Hackett, 2021), this does not come without its problems (Ong, 2019), such as a lack of regulatory measures (Ajunwa, 2020) or 'good' and transparent (historically unbiased) employment data (Citron & Pasquale, 2014), but also a recognition of the implications of AI decision-making on employee control, surveillance, ethics, and discrimination, raising questions around the governance of such phenomena (Ajunwa, 2020). Moreover, it has highlighted how human biases can be inscribed into the code of algorithms, embedding and sustaining inequalities while assuming objectivity (Raghavan et al., 2020). On the back of such evidence, critical scholars have called for an understanding of the processes through which algorithms may mask inequality and discrimination, reinforce and replicate social and organisational inequalities, and, in some instances, even amplify human bias (Kellogg et al., 2020; Köchling et al., 2020).

Moreover, AI has the potential to automate many tasks, which could lead to job displacement on a large scale. Only recently did Buzzfeed announce plans to replace human journalists with ChatGPT (from AI firm OpenAI) robots. Tech website CNET has reportedly been using an AI tool to generate articles that human editors later scan for accuracy before publication. Contrastingly, AI can also create new jobs and improve productivity, leading to overall economic growth, or it can free people from labour and provide more recreation time. In this sense, AI reshapes the working environment in ways that we have not seen before and makes critical scholars wonder about the future of work in a datafied world. It is important for society to proactively manage the impact of AI on employment and reskill workers for future jobs while also managing those who are managed out of positions taken over by AI. But what are these future jobs, and what does it mean for workers that AI will replace? Who is set to gain, and who is set to lose from the AI revolution? Does AI empower or disempower workers, perpetuating the existing marginalisation of already disadvantaged groups while giving the impression that it empowers them? What does all these mean for DEI at work? More importantly, how should the datafied future of work be, and for whom and how can we pursue it?

AI is now widely used in organisations, and it is predicted that AI will replace a variety of professions in the near future. However, many professionals, especially in the HRM and DEI fields, do not understand the implications of AI in the workplace or are unaware of how their organisation implements AI systems at work. Moreover, little research has been conducted to date investigating and discussing the possible implications of AI for DEI in a datafied world of work. The edited volume's purpose is to explore various topics and questions relevant to understanding and mitigating the possible impact of AI on DEI.

More specifically, it will try to explore (a) how AI is currently used in organisations by a variety of management and HRM processes and how it changes the existing working environment, (b) the challenges that it imposes for the inclusion of all employees and specifically for already marginalised communities, and (c) the opportunities it offers if developed and managed responsibly and ethically. To further the discussion, we welcomed theoretically and empirically informed chapter contributions from various disciplines and geographical areas. Accordingly, the chapters in this edited volume hold great potential to advance the current knowledge and practices by exploring the lifecycle of AI development, the propagation of historical biases through AI, and the challenges of aligning AI with diverse human values while filling this gap mentioned above by adopting an interdisciplinary approach.

THEORETICAL FOUNDATIONS OF BIAS IN AI

Socio-technical Systems and the Construction of Bias

The conceptualisation of bias within AI is grounded in recognising that technology and society are fundamentally intertwined through socio-technical systems. Scholars such as Bruno Latour (2007) and Thomas Hughes (1983) have articulated how technological development does not occur in isolation but is shaped by the social, economic, and cultural contexts that inform its creation and implementation. AI systems, particularly large language models (LLMs), represent quintessential examples of socio-technical constructs in which societal biases and inequalities are reflected and reinforced.

From this perspective, biases in AI are not incidental flaws of data or algorithmic structures; instead, they are symptomatic of the broader social inequalities embedded in the datasets themselves (Eubanks, 2018). These systems are constructed and influenced by the cultural and institutional frameworks of the societies that produce them, often leading to systemic biases within AI technologies (O'Neil, 2016). Understanding the intersection between socio-technical systems and AI development highlights the need for a more inclusive approach to AI design, accounting for the broader societal impacts of these technologies.

The Role of LLMs and the Stochastic Parrot Analogy

The development of LLMs such as OpenAI's GPT, Google's Gemini, and Anthropic's Claude illuminates the complexities of bias within AI. These models, trained on vast datasets derived from diverse human-generated texts, exemplify AI systems' potential and limitations. The 'stochastic parrot' analogy, introduced by Bender et al. (2021), is particularly illustrative of the operational mechanics of LLMs. These models predict word sequences based on observed statistical patterns in training data but lack genuine comprehension or ethical reasoning (Bender & Koller, 2020). Consequently, LLMs are predisposed to replicate and amplify the biases present within their data.

The analogy underscores that LLMs lack human capacities for emotional sensitivity, contextual awareness, and ethical discernment despite their apparent coherence and sophistication. This limitation becomes particularly concerning when these models are deployed in high-stakes applications, such as employment, where biased outputs may disproportionately affect marginalised communities (Bender et al., 2021).

Intersectionality and the Complexities of Bias in AI

Intersectionality, a term introduced by Kimberlé Crenshaw (1989), provides a critical framework for analysing the multidimensional nature of bias within AI. Intersectionality emphasises how overlapping aspects of identity – such as race, gender, age, and socioeconomic status – interact to produce unique experiences of marginalisation and discrimination. When developed without an intersectional approach, AI systems risk exacerbating these compounded biases, disproportionately harming already marginalised groups (Buolamwini & Gebru, 2018).

Explicit and implicit biases can operate across multiple levels, complicating efforts to identify and mitigate their impact. For instance, AI systems deployed in healthcare may fail to account for the intersection of race and gender, leading to diagnostic disparities that disproportionately affect women of colour (Buolamwini & Gebru, 2018). Therefore, traditional approaches to bias mitigation, which often isolate specific identity markers, fall short. A holistic and intersectional approach is necessary to develop AI systems that are genuinely inclusive (Noble, 2018).

The Morphogenetic Approach and the Emergence of Bias

Margaret Archer's (1995) morphogenetic approach provides a valuable framework for understanding how biases in AI systems emerge and evolve. This approach emphasises the dynamic interplay between social structures and human agency, suggesting that social phenomena, including biases, are not static but are continuously shaped and reshaped through interaction. In the context of AI, the biases embedded in training data are not fixed; they evolve as AI systems interact with the social environments in which they are deployed. This process is evident in the feedback loops that can occur when AI systems are implemented in decision-making processes (Christin, 2020). For instance, in the criminal justice system, algorithms like COMPAS, which are used to predict recidivism, can perpetuate racial biases that already exist in the data (Angwin et al., 2016). As these systems are used over time, they can create a cycle where biased outcomes reinforce the data used to train the models, leading to increasingly entrenched disparities (Benjamin, 2019).

Archer's framework helps to explain why efforts to mitigate bias in AI must go beyond simply correcting flawed data or algorithms. It requires continuous reflection and adjustment, considering how AI systems interact with and influence the social systems in which they operate. This approach also underscores the importance of interdisciplinary collaboration in AI development, bringing

together expertise from sociology, ethics, and computer science to address the complex and evolving nature of bias in AI.

The Challenge of Transparency and the 'Black Box' Nature of AI

One of the significant challenges in addressing bias in AI is the 'black box' nature of many AI systems, particularly deep learning models (Lipton, 2016). These systems are often highly complex, with decision-making processes that humans do not interpret easily (Doshi-Velez & Kim, 2017). This lack of transparency makes it difficult to diagnose and correct biases, as it is not always clear how the model arrived at a particular decision or outcome. With a clear understanding of how AI models process data and make decisions, it is easier to ensure these systems operate fairly and ethically. The opacity of AI systems not only complicates efforts to mitigate bias but also undermines public trust in AI technologies, particularly when these technologies are used in critical areas such as healthcare and employment (O'Neil, 2016).

Addressing this challenge requires developing new tools and methods for making AI systems more transparent and interpretable. This includes efforts to create more explainable AI (XAI) models that can provide insights into AI systems' decision-making processes (Miller, 2019). However, transparency alone is insufficient; it must be accompanied by broader efforts to embed ethical considerations into developing and deploying AI technologies (Selbst et al., 2019).

ETHICAL CONSIDERATIONS IN AI DESIGN AND THE ROLE OF DIVERSITY

Ethical Imperatives in AI Development

As AI becomes increasingly integrated into critical aspects of human life, the ethical implications of its design, deployment, and impact have come to the forefront of academic and public discourse. The ethical challenges associated with AI are multifaceted, encompassing fairness, accountability, transparency, and inclusivity issues. These challenges are not merely technical problems to be solved but are deeply rooted in broader societal and moral considerations. Dignum (2017) and Gabriel (2020) argue that the development of AI systems must be guided by a commitment to ethical principles that align with human values, particularly those related to justice, equity, and the protection of fundamental rights.

From an ethical standpoint, the central concern in AI design is ensuring that these technologies do not perpetuate or exacerbate existing social inequalities. The ethical design of AI involves more than just mitigating biases in algorithms; it requires a fundamental reevaluation of the values and assumptions that underpin AI development (Jobin et al., 2019). This includes considering who is involved in the design process, whose values are being prioritised, and how decisions are made about the goals and functions of AI systems. Ethical AI design must be proactive rather than reactive, anticipating potential harms

and actively working to prevent them through thoughtful and inclusive design practices (Vallor, 2016).

The Role of Diversity in Ethical AI Design

One of the most significant factors influencing the ethical outcomes of AI systems is the diversity – or lack thereof – within the teams that design, develop, and deploy these technologies (Hanna et al., 2020). The field of AI is currently dominated by a relatively homogeneous group of individuals, particularly in terms of race, gender, and socioeconomic background (West et al., 2019). This lack of diversity has profound implications for the development of AI systems, as it can lead to the reinforcement of narrow perspectives and the exclusion of essential voices in the design process.

Research has shown that diverse teams are better equipped to identify and address biases in AI systems (Schiebinger, 2014). This is because individuals from different backgrounds bring various perspectives, experiences, and insights, which can help uncover potential issues that might otherwise go unnoticed. For example, a team that includes members from marginalised communities may be more attuned to how AI systems could disproportionately impact these groups, leading to more thoughtful and equitable design choices. Vassilopoulou et al. (2022) also discuss the concept of 'algorithmic hygiene', which refers to the need for systematic efforts to ensure that AI systems are free from biases that can perpetuate discrimination. This concept underscores the importance of diversity in AI development teams, as diverse perspectives are crucial for maintaining the 'hygiene' of algorithms – ensuring that they are clean from the biases that could otherwise contaminate their outputs.

Moreover, the involvement of diverse stakeholders is essential for identifying and mitigating biases and ensuring that AI systems are designed to meet the needs of all users, not just those who fit a particular demographic profile (Burnett et al., 2015). This means that AI development must be inclusive, involving voices from a wide range of communities, particularly those that have historically been marginalised or underrepresented in technology development.

Intersectionality and Ethical AI Design

The concept of intersectionality is crucial to understanding the ethical implications of AI design. For instance, an AI system used in hiring might be biased not only against women or people of colour individually but might also disproportionately disadvantage women of colour due to the intersection of race and gender (Buolamwini & Gebru, 2018). These compounded biases can lead to more severe and pervasive discrimination, so AI design processes must take an intersectional approach (Benjamin, 2019). This approach requires technical solutions to identify and mitigate biases and a deep understanding of how different forms of identity intersect and influence experiences of discrimination.

Ethical AI design, therefore, must incorporate an intersectional perspective, ensuring that the systems developed are sensitive to the diverse ways in which

people experience discrimination (Birhane, 2021). This involves not only including diverse voices in the design process but also developing methods to test AI systems for intersectional biases, ensuring that the outputs do not disproportionately harm those already marginalised in multiple ways.

Fairness, Accountability, and Transparency as Ethical Pillars

The fairness, accountability, and transparency (FAccT) framework is a cornerstone in the ethical design of AI systems. This framework provides a set of principles that guide the development of AI systems towards more ethical outcomes (Friedler et al., 2019). Fairness ensures that AI systems do not discriminate against any group of people and promote equitable outcomes (Binns, 2018). Accountability refers to the ability to hold those who design and deploy AI systems responsible for their impacts (Diakopoulos, 2016). Transparency is about making the processes and decisions of AI systems understandable and accessible to all stakeholders.

The FAccT framework is critical in the context of diversity because it emphasises the need for AI systems to be designed and evaluated in ways that are inclusive of diverse perspectives. For instance, fairness in AI can only be achieved if the system is tested against a wide range of demographic variables, ensuring that it performs equitably across all groups (Ananny & Crawford, 2018). Similarly, accountability is more easily maintained when the development process includes diverse voices who can offer insights into potential risks and harms that might not be immediately apparent to a more homogeneous group of developers.

Transparency, too, is closely linked to diversity. Making AI systems transparent requires technical efforts to open the 'black box' of AI decision-making and clear communication accessible to people from all backgrounds (Gebru et al., 2021). This includes explaining how AI systems work in ways that are understandable to non-experts and ensuring that the impacts of these systems are communicated to all stakeholders, including those from marginalised communities.

The European Union's AI Act

The European Union's AI Act represents a pioneering effort to regulate AI systems within a comprehensive ethical framework (European Commission, 2021). The AI Act categorises AI systems based on their risk to public safety and fundamental rights, with particular attention to high-risk sectors such as healthcare, law enforcement, and employment. The Act requires that AI systems in these sectors meet stringent standards for accuracy, security, and data protection to prevent harm and discrimination.

One key aspect of the AI Act is its emphasis on bias mitigation throughout the AI development lifecycle (Veale & Zuiderveen Borgesius, 2021). This includes requirements for transparency, such as detailed documentation and traceability of AI decision-making processes, as well as accountability measures that hold developers and operators responsible for the impacts of their systems. By mandating these standards, the AI Act seeks to align AI development with broader

societal values, ensuring that AI technologies contribute to the public good rather than exacerbating existing inequalities.

The AI Act also highlights the importance of human oversight in AI systems, particularly in high-stakes domains (AlgorithmWatch, 2021). This oversight is crucial for ensuring that AI systems do not operate autonomously in ways that could harm individuals or communities. By requiring human involvement in AI decision-making processes, the Act aims to safeguard against the potential risks of AI, particularly those related to bias and discrimination.

Bias Detection and Mitigation Techniques

Effective bias mitigation in AI involves a multifaceted approach that addresses potential biases at every stage of the AI development lifecycle, from data collection and preprocessing to model development, validation, and deployment. Preprocessing techniques involve carefully selecting and preparing data before it is used to train AI models (Barocas et al., 2019). One of the most critical steps in bias mitigation is ensuring that training data represent the diverse populations the AI system will interact with. This includes addressing issues like sampling bias, where particular groups may be underrepresented in the data, and measurement bias, where data collection methods may systematically favour specific outcomes. Data curation is essential for creating balanced datasets that accurately reflect the real world's diversity. Strategies for achieving this include balancing demographic representations, removing toxic language disproportionately affecting marginalised groups, and prioritising modern, inclusive data sources over outdated ones.

In-processing techniques involve directly incorporating fairness objectives into the AI model during training. This can be done by embedding constraints or regularisation techniques that ensure the model learns to make accurate and fair decisions. For example, adversarial debiasing, where the model is trained to be antagonistic to its own biases, can help reduce discrimination in the final model outputs. Companies like Anthropic are developing AI systems with built-in ethical guidelines – called 'constitutional principles' – which guide the model's behaviour towards fairness and ethicality. These principles are enforced through reinforcement learning techniques that refine the model's outputs based on how well they align with these ethical objectives.

Post-processing techniques are applied after the AI model has been trained to adjust its outputs and ensure they do not perpetuate bias (Mehrabi et al., 2021). This can involve re-weighting, where the outputs are adjusted to reflect fairer distributions, or filtering, where outputs that could reinforce stereotypes or harmful biases are removed or modified (Feldman et al., 2015). For example, AI image generation tools like DALL-E have been adjusted to provide more diverse representations of professions and roles, countering historical biases by depicting women and people of colour in positions of authority and expertise. Similarly, Google's Gemini AI has implemented filters to prevent the generation of images that reinforce racial or gender stereotypes, showcasing the importance of post-processing in bias mitigation.