Essays on Responsible Artificial Intelligence

Mengran Xiong

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Social Sciences
Management School

May 2024

Abstract

Artificial intelligence (AI) investments by organisations across sectors have skyrocketed in recent years, driven by their potential to enhance business performance and marketing effectiveness through improved decision quality in task augmentation and automation within competitive environments. However, the black-box nature of many modern AI systems, powered by machine learning (ML) algorithms, raises significant ethical and societal challenges, making *responsible AI* a critical and widely discussed topic. Its scope includes ethical and responsible approaches to designing, deploying, and implementing AI systems that uphold individual rights and promote social good. In response to the growing discourse, this thesis, comprising three empirical essays, adopts a pragmatic philosophical approach to examine the multifaceted implications of AI adoption across organisational, managerial, and individual levels using a multi-method research design. Essay 1, based on a practice-based view (PBV), develops and validates a conceptual model informed by corporate social responsibility (CSR) and information technology (IT) business value to examine how responsible AI practices create business value, supported by multiple case studies for empirical analysis. Essay 2 advances theoretical understanding of managing AI's evolving frontiers (e.g., autonomy, learning, and inscrutability) through a qualitative, interpretive study. It introduces a grounded model that integrates technical and social aspects of responsible AI design and governance, connecting instrumental and humanistic outcomes to highlight the multifaceted nature of responsible AI implementation in contemporary organisations. Essay 3 focuses on the consumer side, investigating how generative AI (GenAI) impacts consumer intentions and trust in healthcare marketing. Through an experimental design, it integrates construal level theory (CLT), the accessibility-diagnosticity model (ADM), and cue utilisation theory (CUT) to reveal how goal- versus duty-oriented messaging drives the GenAI-powered healthcare service adoption across three domains: prevention, diagnosis, and treatment, offering insights for digital healthcare marketing communications. Collectively, this thesis contributes novel insights into responsible AI, advancing theoretical research while providing practical strategies for organisations to implement AI ethically and effectively at multiple levels.

Acknowledgements

The work presented in this thesis would not have been possible without the support of many exceptional people. I am deeply grateful to all those who, even if not named here, supported me through my struggles and celebrated my small victories along the way. This has been a transformative journey, and finally, I have made it to the finish line.

My heartfelt thanks go to my supervisors, Dr Yichuan Wang and Professor Hossein Olya, for their invaluable guidance, unwavering support, insightful challenges, and constant encouragement throughout this journey. From the early stages of refining research proposals to the final submission of the thesis, their patience, wealth of wisdom, and ability to enlighten have been vital to my academic growth. I am truly grateful for the immeasurable contributions they have made to my development, helping me achieve goals I once thought were beyond my reach.

My sincere thanks are extended to Dr Sabrina Thornton, the principal investigator of a research project I worked on at Sheffield University Management School (SUMS), and to my co-authors: Dr Long Chen, Dr Jiao Ji, Dr Si Li, Dr Weisha Wang, Dr Minhao Zhang, and Renxian Zuo (listed in alphabetical order). Collaborating with them on journal publications has been an incredibly rewarding experience. Their academic expertise and dedication have been a constant source of motivation and inspiration.

I am especially grateful to my thesis examiners, Dr Sabrina Thornton and Dr Dongmei Cao, for their time, effort, invaluable comments, and constructive suggestions, which greatly enhanced the quality and presentation of this thesis. I also wish to thank my first-year confirmation review panel members, Professor Fraser McLeay and Professor Don Webber, for their helpful insights into my initial work. Additionally, I thank all other SUMS staff members for their friendly support.

A special thank you goes to all the interviewees who generously shared their thoughts and personal experiences during our conversations. Their openness allowed me to gain a deeper understanding of the research topic, adding significant meaning and depth to my work.

I am grateful to my fellows, Dr Xiaoyu Gan, Dr Jingshan Yang, Peilin Jiang, and many others at SUMS and other institutions, for their kindness, support, and understanding, which have made my journey so much more enjoyable. I would also like to thank my friends in Manchester for the wonderful moments we shared.

I own my deepest gratitude to my parents, Wei Xiong and Guangmin Li, for their love, encouragement, and unwavering belief in me. Without them, I would not have had the courage to begin this journey. Look how far I have come. To my brother, Yilin Xiong, thank you for the confidence in me throughout this long process; it kept me positive and motivated. I am deeply grateful to my parents-in-law, Zhongping Wu and Yirui Xu, for their unconditional love and support. The dedication of both my family and my husband's family continues to inspire me as I reflect on my PhD journey.

A greatest and most special acknowledgement goes to my husband, Dr Hao Wu, who has supported me without hesitation in this and every endeavour, and to our children, Elaine Wu and Ethan Wu. Thank you, Hao, for your unreserved love and for being an inspiration for me to pursue my dreams. Your support throughout the highs and lows of this journey, and the care and sacrifices you made so I could focus on writing, have been truly invaluable. Thank you, Elaine and Ethan, for your little smiles that lifted me during tough times, and for your understanding when I was not fully present for you. Having you all in my life has made me realise that I am stronger and better than ever before.

Declaration

I, the author, confirm that the thesis I have submitted for examination for the PhD degree at the University of Sheffield is my own work, except where it includes contributions from collaborative manuscripts. I am aware of the University's Guidance on the Use of Unfair Means (www.sheffield.ac.uk/ssid/unfair-means). This work has not been previously presented for an award at this, or any other, university. The contributions of the author and co-authors have been explicitly indicated below. Appropriate credit has been given within the thesis where reference has been made to the work of others.

The author conceived, designed, and conducted the research presented in Chapters II, III and IV under the supervision of Dr Yichuan Wang and Professor Hossein Olya. The initial drafts of the manuscripts were prepared by the author and subsequently reviewed, revised, and refined with input from the co-authors.

A revised version of the essay developed in collaboration with co-authors outside the PhD supervision team (i.e., Dr. Jiao Ji, Renxian Zuo) in Chapter II is currently under revision for a peer-reviewed journal, with further details discussed in Chapter I. Additionally, a version of the essay in Chapter III has been accepted for publication as a book chapter, and a version of the essay in Chapter IV is planned for submission to a peer-reviewed journal.

Research Output

Publications

These manuscripts are being developed in collaboration with co-authors, including contributions beyond the scope of my PhD research.

Wang, W., Chen, L., **Xiong, M.**, & Wang, Y. (2023). Accelerating AI adoption with responsible AI signals and employee engagement mechanisms in health care. *Information Systems Frontiers.*

Li, S., **Xiong, M.**, Wang, Y., & Zhang, M. (2022). How does product-celebrity congruence and content sponsorship affect perceived altruism among consumers?. *Resources, Conservation & Recycling.*

Wang, Y., **Xiong, M.**, & Olya, H. (2020). Towards an understanding of responsible artificial intelligence practices. *Proceedings of the 53rd Hawaii International Conference on System Sciences.*

Work-in-progress

These manuscripts are being developed in collaboration with co-authors, building upon the contributions of my PhD research.

Xiong, M., Ji, J., Wang, Y., Olya, H., & Zuo, R. (under revision). Responsible artificial intelligence (AI) attention and firm innovation: An Attention-based view. *Journal of Product Innovation Management.*

Xiong, M., Wang, Y., & Olya, H. (in press). 'Real-ising' the benefits of responsible artificial intelligence (AI) management. In H. Olya (Ed.), *Responsible service management.* Emerald Publishing.

Xiong, M., Wang, Y., & Olya, H. (in preparation). Goal versus duty: Marketing communications for generative artificial intelligence (GenAI) healthcare service.

Conferences

Wang, Y., **Xiong, M.**, & Olya, H. (2020). Toward an understanding of responsible artificial intelligence practices. Paper presented at the *53rd Hawaii International Conference on System Sciences,* Maui, HI, USA.

Xiong, M., Wang Y., & Olya, H. (2020). From the darkness cometh the light: Actualising responsible artificial intelligence (AI) in driving business value. Paper presented at the *4th International Conference of Marketing, Strategy and Policy - Conference/Doctoral Colloquium,* London, UK.

Table of Contents

List of Figures

List of Tables

Acronyms

AI	Artificial Intelligence
ABV	Attention-Based View
ADM	Accessibility-Diagnosticity Model
ANOVA	Analysis of Variance
APPI	Act on the Protection of Personal Information
AR	Augmented Reality
CDO	Chief Data Officer
CIO	Chief Information Officer
CTO	Chief Technology Officer
CLT	Construal Level Theory
CRAiO	Chief Responsible AI Officer
CSR	Corporate Social Responsibility
CUT	Cue Utilisation Theory
EAD	Ethically Aligned Design
ES	Enterprise Systems
ESG	Environmental, Social, and Governance
EU	European Union
GANs	Generative Adversarial Networks
GenAI	Generative Artificial Intelligence
GDPR	General Data Protection Regulation
GTM	Grounded Theory Method
IS	Information Systems
IT	Information Technology
LLM	Large Language Model
ML	Machine Learning
MLCC	Machine Learning Crash Course
NLP	Natural Language Processing
PBV	Practice-Based View
RBV	Resource-Based View
SAP	Strategy-as-Practice
SOA	Service-Oriented Architecture
TMT	Top Management Team
TPB	Theory of Planned Behaviour
UK	United Kingdom
US	United States
VRIN	Valuable, Rare, Inimitable, and Non-Substitutable

Chapter I

Thesis Introduction

1. Research Background and Rationale

Artificial intelligence (AI), an evolving frontier of emerging computing capabilities (i.e., autonomy, learning, and inscrutability; Berente et al., 2021), simulates human intelligence to perform tasks such as decision-making and content creation (Huang & Rust, 2021, 2024). AI autonomously learns from environmental changes, enabling it to tackle complex issues beyond the capabilities of earlier computational technologies (Rai et al., 2019; Verganti et al., 2020). With rapidly advancing capabilities such as increased autonomy and deeper learning (Baird & Maruping, 2021; Berente et al., 2021), AI is poised to transform the global economy. This impact is underscored by substantial investments in AI solutions, with worldwide business spending estimated at $154 billion in 2023—a figure projected to continue growing (Thormundsson, 2024).

Public and private organisations across diverse industries have rapidly embraced AI to automate and assist with business operations, decision-making processes, and interactive capabilities, yielding significant benefits in efficiency, consistency, and personalisation (Agrawal et al., 2019; Asatiani et al., 2021). AI adoption spans a broad range of applications, including physical robots for manufacturing (Huang & Rust, 2021), predictive analytics for hiring (Hoffman et al., 2018) and medical diagnostics (Jussupow et al., 2021), virtual agents for customer service (Schanke et al., 2021), and augmented intelligence for coworking (Jain et al., 2021). Given that AI capabilities are evolving with time, its applications will continue to grow at a remarkable speed and be adopted in an astounding variety of problem domains across industries (Berente et al., 2021). In particular, recent advancements in generative AI (GenAI), which enables the creation of novel content, have

revolutionised how we create, innovate, and interact with technology (Huang & Rust, 2023).

As AI continues to evolve at an unprecedented pace, its adoption is set to accelerate, reshaping industries while introducing both new opportunities and challenges for organisations worldwide. While AI's role in improving marketing effectiveness and business performance has become increasingly critical; however, it also presents significant concerns, including privacy invasions, data leaks, biases, and misinformation (Mikalef et al., 2022). Firms that fail to adequately address these issues when implementing AI may lose public trust in their commitment to ethical practices, suffer damage to their brand reputation, and incur substantial financial penalties and legal expenses (Jobin et al., 2019). A notable example is Clearview AI, a facial recognition company that faced severe repercussions for its unethical use of AI technology. The company was fined $10 million and banned from selling its technology in certain regions for collecting the faces of UK citizens from websites and social media without their consent (Heikkilä, 2022).

As concerns about AI misuse and unintended consequences grow, interest in responsible AI is rapidly increasing. An Accenture survey (Eitel-Porter & Grosskopf, 2022) reports that 80% of companies plan to increase investment in responsible AI, and 77% view AI regulation as a priority. Firms engage in responsible AI initiatives by adopting ethical guidelines and right-based values that enforce transparency, explainability, and fairness (Arrieta et al., 2020; Shneiderman, 2021). Despite the growing importance of responsible AI, academic investigation into its strategic implications is still rare and lacking in detail (Huang & Rust, 2021). Existing studies tend to focus more on the technical aspects of responsible AI development (Díaz-

Rodríguez et al., 2023), such as algorithmic design and bias detection, rather than the broader strategic considerations that firms must navigate when integrating AI responsibly into their operations like governance frameworks and managerial accountability.

Responsible AI has gained increasing scholarly attention and activated growing conceptual debates about its meanings and applications (Arrieta et al., 2020; Díaz-Rodríguez et al., 2023; Mikalef et al., 2022). It is described as the prospect of implementing AI technology guided by ethical principles and regulatory initiatives to enhance human capabilities and empower businesses in a way that aligned with ethical and societal values (Mikalef et al., 2022; Siala & Wang, 2022). Unlike traditional AI applications, which focus primarily on efficiency and cost reduction, responsible AI aims to balance technological advancements with ethical considerations and societal values (Mikalef et al., 2022). It encompasses various principles aimed at ensuring that AI systems are ethical, transparent, and beneficial to individuals and society. The literature review identifies several key concepts underlying responsible AI, including explainability, transparency, human-centered design, human-AI augmentation, consideration of human uniqueness, and privacy and data security (see Table I.1).

Table I.1 Responsible AI and its underlaying concepts in the literature

Responsible AI	Definitions	Sources
Explainability	An AI system is considered to be explainable if it supplies clear evidence, support, or reasoning related to an outcome from or a process, particularly in complex or opaque "black box" models.	Arrieta et al. (2020); Bauer et al. (2023); Rai (2020)
Transparency	An AI system is considered transparent if it is, by itself, understandable and accessible to users, stakeholders, or regulators in terms of its	Díaz-Rodríguez et al. (2023); Siala & Wang (2022)

	workings, decision-making processes, underlying algorithms, and any potential biases or limitations.	
Human-centred AI	An AI system is designed and developed to be human-centric, intuitive, and responsive to the needs, values, and well-being of individuals and society.	Bankins & Formosa (2023); Liu-Thompkins et al. (2022)
Human-AI augmentation	An AI system is designed to augment human abilities by enhancing performance in data processing and automation, while humans contribute oversight, creativity, and intuition, creating a synergistic partnership for improved decision-making and productivity.	Jussupow et al. (2021); Raisch & Krakowski (2021); Tschang & Almirall (2021)
Uniqueness consideration	An AI system is designed to respect and adapt to human uniqueness by recognising and responding to individual characteristics, preferences, and needs, ensuring personalised and effective interactions.	Longoni et al. (2019)
Privacy and data security	An AI system is designed to prioritise privacy and data security by securely handling personal information and protecting it from unauthorised access or breaches, while adhering to privacy regulations.	Thomaz et al. (2020)

Table I.2 summarises key studies on responsible AI, exploring its implementation across diverse contexts and reinforcing the importance of aligning AI systems with human values while mitigating ethical and societal risks. These studies provide essential insights into the three focal areas of this thesis, that is, responsible AI at the organisational, managerial, and consumer levels, while also illustrating the intricate and multidimensional challenges associated with AI design, governance, management, and communication. The existing literature underscores the need for a holistic approach to responsible AI adoption, ensuring AI systems are ethically sound, strategically integrated, and effectively communicated to foster trust and promote long-term sustainability. However, significant gaps remain in translating responsible AI principles into actionable frameworks that generate business value while addressing regulatory, managerial, and consumer concerns. This thesis

advances the discourse on responsible AI by proposing structured frameworks that operationalise and manage AI in practical, value-driven ways across organisational and managerial domains, while also examining the role of responsible marketing communications in fostering consumer trust and adoption in GenAI-enabled digital healthcare.

At the organisational level, responsible AI is increasingly recognised as a strategic imperative, requiring organisations to integrate responsible AI practices with their corporate objectives, regulatory compliance, and business performance. While AI ethics principles offer essential guidance, their practical implementation into actionable governance mechanisms remains challenging. For instance, Bauer et al. (2023) illustrate the paradox of AI explainability, suggesting that while transparent AI explanations enhance user comprehension, they may also inadvertently reinforce cognitive biases, ultimately compromising decision-making quality. This complexity underscores the need for structured AI governance frameworks that strike a balance between enhancing transparency and safeguarding against cognitive distortions. Identifying the most successful cases of responsible AI implementation can provide valuable insights into addressing this challenge. This thesis, particularly Essay 1 in Chapter II, extends the practice-based view (PBV) (Bromiley & Rau, 2014) by conceptualising responsible AI as an evolving, industry-wide practice shaped by regulatory mandates and stakeholder expectations. By analysing best and transferrable practices from industry leaders, this research identifies effective strategies for institutionalising responsible AI. The findings demonstrate that firms can move beyond viewing responsible AI as a compliance obligation, instead leveraging it as a strategic driver that enhances business value,

strengthens competitive differentiation, and fosters long-term organisational resilience.

At the managerial level, responsible AI presents complex sociotechnical challenges, necessitating adaptive AI design and governance frameworks that balance AI autonomy with human oversight. For instance, research on AI decision-making in high-stakes environments (e.g., Jussupow et al., 2021) highlights the risks of over-reliance on AI without adequate human supervision, particularly in domains such as healthcare, where AI can improve diagnostic accuracy but may also introduce new vulnerabilities. This aligns with the sociotechnical perspective (Berente et al., 2021), which emphasises that responsible AI adoption is not merely a technical or regulatory challenge but an organisational and managerial one that requires multi-stakeholder involvement. This thesis, particularly Essay 2 in Chapter III, extends these discussions by advocating for multi-level AI management frameworks that engage policymakers, industry leaders, organisational decision-makers, and end-users, moving beyond technocentric approaches that focused solely on algorithmic fairness and technical transparency (Díaz-Rodríguez et al., 2023). Instead, it highlights the critical role of organisational leadership in guiding AI design (i.e., integrate ethical considerations like explainability, privacy, and fairness) and governance (i.e., ensure oversight, risk management, and accountability) while embedding responsible AI principles into corporate decision-making and risk management throughout AI lifecycle.

At the consumer level, responsible AI is pivotal in shaping trust and driving user engagement, particularly in sectors like healthcare and finance, where AI-driven decisions significantly impact individuals. Studies on human-centred AI (e.g., Liu-

Thompkins et al., 2022; Longoni et al., 2019) reveal that consumers often resist AI due to perceived deficiencies in empathy and personalisation, fearing that AI systems may compromise autonomy, uniqueness, and fairness. These insights are directly relevant to Essay 3, Chapter IV of this thesis, which explores responsible AI communication strategies through theoretical lens of construal level theory (CLT; Trope & Liberman, 2010), the accessibility-diagnosticity model (ADM, Feldman & Lynch, 1988) and cue utilisation theory (CUT, Richardson et al., 1994). This study argues that combining communication strategies (message orientation: goal, duty × message construal: how, why) can effectively foster positive consumer responses and reduce algorithm aversion. Consumers often rely on extrinsic cues—such as message framing—to assess AI credibility and risk, highlighting the importance of transparent and explainable communication in shaping consumer perceptions of AI-powered services. By integrating artificial empathy and human-centred design into AI solutions, organisations can enhance trust and engagement, ensuring AI is perceived as a responsible and reliable tool rather than an impersonal or opaque system.

This thesis advances the understanding of responsible AI by addressing research gaps across organisational, managerial, and consumer levels. Through conceptual and grounded modelling, qualitative inquiry, and experimental research, this thesis provides actionable insights for organisations to effectively implement, manage, and communicate responsible AI practices that align with consumer, corporate, and societal expectations—driving business value, strengthening governance, and fostering consumer engagement.

Table I.2 Literature review on responsible AI

Study	RAI considered	Outcome/s	Mechanisms/Boundary conditions	Level of analysis	Study design	Study context	Key findings
Bauer et al. (2023)	Explainability	Users' sense making of information and mental model adjustments	Task domain, user expertise, and explanation presentation *(boundary conditions)*	Individual level	Cross-sectional *(experiment)*	Real estate industry	Feature-based AI explanations reshape users' decision-making but can reinforce confirmation bias, leading to persistent misconceptions, biased decisions, and behavioural risks.
Fügener et al. (2022)	Human-AI augmentation	Classification accuracy, delegation rate, certainty	The existence of complementarities, the recognition of complementarities, and the execution of efficient delegation rules *(boundary conditions)*	Individual level	Cross-sectional *(experiment)*	Image classification	AI-human collaboration works best when AI delegates tasks, but humans' poor self-assessment (lack of metaknowledge) hinders effective delegation, limiting collaboration success.
Gnewuch et al. (2024)	Human-AI augmentation; Transparency *(human involvement disclosure)*	Adoption a more human-oriented communication style, employee workload	Impression management concerns *(mechanism)*	Individual level	Cross-sectional *(experiment)*	Tele-communication	Disclosing human involvement makes customers adopt a more human-centred communication style, increasing employee workload, driven by customers' impression management concerns.
Huang & Rust (2024)	Human-centred AI *(feeling AI for customer care)*	Emotional connection	N/A	Individual level	Cross-sectional *(interview)*	Customer service	Gen AI has potential in customer care, but technological challenges remain in improving emotion recognition and response, with marketing strategies needed to ensure effective emotional connection and customer value.
Jia et al. (2024)	Human-AI augmentation	Increased employee creativity and sales performance	Redistribution of job tasks *(mechanism)* Employee skill level (*Boundary condition*)	Individual *(employee) level*	Cross-sectional *(mixed methods)*	Telemarketing	AI assistance enhances employee creativity, particularly among higher-skilled employees. This AI-augmented creativity is skill-biased

							and leads to increased sales performance. AI redefines job design, intensifying interactions with serious customers and enabling employees to generate innovative solutions.
Jussupow et al. (2021)	Human-AI augmentation	Improved decision-making accuracy and cognitive load management	N/A	Individual *(physician level)*	Mixed methods	Medical diagnostics	AI systems can effectively augment medical decision-making by reducing error rates and cognitive load, especially when physicians are trained to critically evaluate AI advice using metacognitive strategies. Reliance on AI without sufficient scrutiny can lead to errors, especially when AI advice is incorrect.
Liu-Thompkins et al. (2022)	Human-centred AI *(artificial empathy)*	Relationship satisfaction, customer well-being, customer loyalty, customer equity	Human-AI gap in affective/social customer experience quality/activation *(mechanism)*	Interaction between AI systems and individuals	Cross-sectional *(qualitative study and one pilot experimental study)*	Customer service interaction	Artificial empathy can significantly improve the emotional and social dimensions of customer experience by making AI interactions feel more human-like and responsive to individual emotional states, thereby enhancing customer satisfaction and potentially fostering greater customer loyalty.
Longoni et al. (2019)	Uniqueness consideration	Choice, willingness to pay, relative preference, partworth utility, propensity to follow	Uniqueness neglect *(mechanism)* Personalisation, recipient of care, AI role, stakes, other domains *(boundary conditions)*	Individual level	Cross-sectional	Healthcare	Consumers resist AI healthcare due to concerns about individual uniqueness, but this decreases when AI is framed as personalised, supports human decisions, or applies to others.
Longoni & Cian (2022)	Uniqueness consideration,	Preference for recommendations	Competence perceptions, hedonic/utilitarian perceptions; unique	Individual level	Cross-sectional	Various: beauty, real	AI recommenders are preferred for utilitarian tasks due to perceived competence, but human

	Human-AI augmentation		preference matching, augmented intelligence *(boundary conditions)*			estate, food, fashion	recommenders are favoured for hedonic tasks. This effect is moderated by the nature of the recommendation and can be attenuated with interventions.
Luo et al. (2019)	Transparency	Customer purchases	Perceived knowledge; Empathy of chatbots *(mechanisms)*	Individual level	Cross-sectional	Financial service company	Undisclosed AI chatbots are as effective as skilled workers, but early disclosure reduces purchase rates by 79.7% due to perceptions of lower knowledge and empathy. Delaying disclosure or using prior AI experience mitigates this effect.
Padigar et al. (2022)	Transparency *(announcements of AI-embedded new product innovations)*	Investor response	Innovation stage, route to innovation, innovation complexity *(boundary conditions)*	Firm level	Longitudinal	Various (S&P 500 firms)	Investors respond favourably to AI-integrated product innovations when firms have strong marketing departments, especially for later-stage, complex innovations or those involving external assets.
Thomaz et al. (2020)	Privacy and data security	N/A	N/A	Individual level	Conceptual	Marketing	• Predicts a shift towards more private and controlled web environments, influencing marketing strategies. • Marketers can effectively use conversational agents to navigate this new landscape by adapting to consumer privacy preferences and managing to engage both privacy-sensitive (Ghosts) and less concerned (Buffs) consumer segments.

2. Thesis Structure

This thesis consists of three empirical essays presented in Chapters II, III, and IV, each offering a comprehensive examination of responsible AI from organisational, managerial, and consumer perspectives. Chapters I and V serve as the overall introduction and conclusion, respectively.

Chapter II offers an initial understanding of the 'responsible AI' phenomenon by exploring the potential business value generated through the implementation of responsible AI practices in organisations. This study integrates corporate social responsibility (CSR), enterprise systems (ES) benefits, and the practice-based view (PBV) to develop a conceptual model linking responsible AI practices to business value. Using case materials, the study analyses AI implementation projects and assess their business value potential through a proposed conceptual model based on a 'cause-and-effect' structure, a widely used analytical approach in the information systems (IS) field (e.g., Mueller et al., 2010; Shang & Seddon, 2002; Wang et al., 2018). By examining 51 AI implementations and industry insights, this study identifies six key practices that enhance regulatory compliance and business benefits. The findings contribute both theoretically and practically, offering a framework for responsible and value-driven AI implementation.

As our investigation progresses, it becomes evident that businesses often struggle to adhere to abstract AI ethical principles and to execute AI initiatives in an ethically and socially responsible manner. This realisation shifts the focus beyond merely recognising the business value of responsible AI implementation to understanding the challenges associated with its effective management. Thus, in Chapter III, the research aims to provide empirical evidence to support organisational decision-

makers in navigating these complexities. Relying solely on secondary data, such as company reports, may fail to capture the nuances of AI management practices, particularly decision-making processes and operational challenges, which are rarely disclosed publicly. To bridge this gap, Chapter III adopts an interpretive research approach, gathering primary data through semi-structured interviews with elite informants who share their experiences, insights, and perceptions of responsible AI practices. To enhance the robustness of our findings, these interview results are triangulated with archival data.

The outcome of Chapter III is an empirically grounded model that offers an integrative perspective on responsible AI management, focusing on AI design and governance aspects through a sociotechnical lens. By integrating empirical findings with theoretical frameworks, this study provides a nuanced understanding of AI management practices, highlighting key dimensions of responsibility (i.e., evidence, epistemic, and outcome), governance mechanisms (i.e., structural, procedural, and relational), and organisational outcomes (i.e., instrumental and humanistic). This approach ensures academic rigour and practical relevance, providing a holistic view of responsible AI design, governance, and management.

While Chapters II and III examine responsible AI from organisational and managerial perspectives, limited research exists on how these initiatives influence consumer behaviour, particularly in relation to trust and the adoption of AI-driven products and services. The healthcare sector has been a focal point for AI applications (Jussupow et al., 2021), driven by the rising demand for digital healthcare solutions, including AI-enabled disease identification and medical

interventions for mental health conditions. According to a study by BCG[1], GenAI is projected to expand more rapidly in healthcare than in any other sector, with an estimated compound annual growth rate of 85% until 2027. However, many MedTech companies adopt a cautious 'wait-and-see' approach to GenAI development due to stringent regulatory requirements and persistent consumer resistance to medical AI (Longoni et al., 2019). Given this hesitation, marketing scholars are encouraged to explore intervention strategies—such as persuasive messaging—to enhance consumer acceptance of medical AI, enabling companies to effectively create and capture value from GenAI applications.

In Chapter IV, the research investigates marketing communication strategies for GenAI healthcare applications (i.e., prevention, diagnosis, and treatment) to integrate responsible AI insights into a forward-looking healthcare inquiry. Specifically, the study examines the effectiveness of goal-oriented versus duty-oriented messaging, drawing on real-world observations that suggest their potential significance and addressing the gap in existing scientific evidence through an empirics-first approach (Golder et al., 2023). Experimental findings reveal that duty-oriented messages work best for prevention, while goal-oriented messages are most effective paired with "why" explanation for prevention and "how" explanation for diagnosis, driven by information diagnosticity influencing consumer responses to GenAI healthcare. This study explores how construal level effects and accessibility-diagnosticity effects shape consumer engagement with GenAI healthcare services,

[1] https://www.bcg.com/publications/2023/generative-ai-in-medtech

offering evidence-based strategies to optimise marketing communication and enhance GenAI adoption.

The three essays presented in this thesis are closely related to key sub-fields within marketing, particularly marketing strategy and consumer behaviour[2]. From a strategic perspective, this thesis examines how responsible AI implementation influences firm-level decision-making, resource investment, and competitive advantage. Essay 1 investigates how organisations strategically adopt CSR-oriented responsible AI practices to create business value, reinforcing responsible AI as a strategic asset that aligns with corporate governance and innovation management. Essay 2 extends this perspective by examining managerial challenges in scaling responsible AI practices, focusing on key considerations in marketing strategy such as AI design tactics, AI governance mechanisms, decision-making structures, and cross-functional collaboration. From a consumer behaviour perspective, this thesis investigates how responsible GenAI communication strategies shape consumer trust and adoption intentions. By integrating CLT, ADM, and CUT, Essay 3 explores how goal- and duty-oriented responsible AI marketing messages, combined with psychological framing (e.g., why vs. how explanations), align with consumer mindsets and influence their engagement with GenAI healthcare services.

By addressing both strategic and behavioural marketing dimensions, this thesis positions responsible AI as a critical enabler of both ethical integrity and business

[2] Marketing strategy examines how firms interact with stakeholders, emphasising managerial decision-making, resource allocation, and competitive positioning. Consumer behaviour explores the psychological processes that influence how individuals think, feel, and make decisions, particularly when evaluating and choosing between alternatives (Kotler et al., 2019).

performance. Through both theoretical development and empirical evidence, it contributes to marketing literature by illustrating how organisations can leverage responsible AI practices and narratives to enhance competitive positioning while improving consumer acceptance of GenAI-enabled services, particularly in digital healthcare contexts.

3. Publication Status and Author Contribution

This thesis follows a publication-based format, consisting of a collection of three essays structured for publication in peer-reviewed journals and a book chapter, aligning with academic conventions in responsible Ai research. These studies are integrated into the thesis alongside traditional components, the introduction and conclusion, to form a coherent and substantial body of research. The thesis makes an original contribution to the field of responsible AI, with each study contributing novel insights into responsible AI business value, responsible AI management, and responsible AI marketing communication.

The status of these studies in terms of publication and submission is as follows: Chapter II, which develops a conceptual model linking responsible AI practices to business value, has been submitted to the *Journal of Product Innovation Management* and is currently under third-round review. This study has been refined to better align with the special issue on "Artificial Intelligence, Stakeholder Engagement, and Innovation Value". The revised manuscript applies the attention-based view (ABV) to examine how top management team (TMT) attention to responsible AI influences firm innovation, addressing gaps in understanding responsible AI's strategic role at the firm level. While existing research focuses on individual user perceptions and

technical ethics, this study conceptualises responsible AI attention as managerial awareness that integrates ethical principles into AI-driven processes to foster innovation.

Using a responsible AI dictionary developed from 527 case descriptions across 30 organisations, the study employs natural language processing (NLP) and large language model (LLM) techniques to analyse earnings call transcripts from 2,670 firm-year observations spanning S&P 500 firms from 2011 to 2021. The results show that TMT's responsible AI attention positively correlates with firm innovation, particularly in low-tech industries and firms with short-term-focused shareholders. However, the presence of a Chief Technology Officer (CTO) was not found to significantly amplify this effect. These findings contribute to the AI and innovation management literatures by providing theoretical and practical insights into how firms can leverage responsible AI as a strategic driver of innovation.

Chapter III presents the empirical investigation into responsible AI management, which has been accepted for publication as a book chapter in *Responsible Service Management* and is forthcoming in 2025. Meanwhile, Chapter IV examines AI marketing communication strategies and consumer trust and adoption of generative AI (GenAI) healthcare services. This study is currently being prepared for submission to the *Journal of Service Research* in 2025.

The research presented in this thesis is primarily the work of the author, with significant contributions to the conceptual development, data collection, analysis, and writing. For Chapter II, the author was responsible for the theoretical framework, data collection, analysis, and manuscript preparation, while co-authors

contributed to methodological refinements and additional theoretical discussions. In Chapter III, the author led the empirical investigation, conducted qualitative data collection and analysis, and drafted the manuscript, with co-authors providing support by refining the analytical approach and reviewing the manuscript. For Chapter IV, the author designed the experimental study, developed the theoretical positioning, and conducted data analysis, with co-authors contributing input on experimental design. While collaboration with supervisors and co-authors has been invaluable in refining and strengthening the research, the core intellectual contributions, research execution, and writing of this thesis are predominantly the author's own work. This thesis represents the author's original contribution to the field of responsible AI, bridging insights from responsible AI business value, responsible AI management, and responsible AI marketing communication.

Chapter I References

Agrawal, A., Gans, J., & Goldfarb, A. (2019). *The economics of artificial intelligence: An agenda*. Chicago: University of Chicago Press.

Arrieta, A. B., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58,* 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Asatiani, A., Malo, P., Nagbøl, R., & Penttinen, E. (2021). Sociotechnical envelopment of artificial intelligence: An approach to organizational deployment of inscrutable artificial intelligence systems. *Journal of the Association for Information Systems, 22*(2), 325–352. https://doi.org/10.17705/1jais.00664

Baird, A., & Maruping, L. M. (2021). The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. *MIS Quarterly, 45*(1), 315–341. https://doi.org/10.25300/MISQ/2021/15882

Bankins, S., & Formosa, P. (2023). The ethical implications of artificial intelligence (AI) for meaningful work. *Journal of Business Ethics, 185*(4), 725–740. https://doi.org/10.1007/s10551-023-05339-7

Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research, 34*(4), 1582–1602. https://doi.org/10.1287/isre.2023.1199

Bromiley, P., & Rau, D. (2014). Towards a practice-based view of strategy. *Strategic Management Journal, 35*(8), 1249–1256. https://doi.org/10.1002/smj.2238

Díaz-Rodríguez, N., et al. (2023). Connecting the dots in trustworthy artificial intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion, 99,* 101896. https://doi.org/10.1016/j.inffus.2023.101896

Eitel-Porter, R., & Grosskopf, U. (2022). *From AI compliance to competitive advantage: Becoming responsible by design*. Accenture. https://www.accenture.com/us-en/insights/artificial-intelligence/ai-compliance-competitive-advantage

Feldman, J. M., & Lynch, J. G. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology, 73*(3), 421–435. https://doi.org/10.1037/0021-9010.73.3.421

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information Systems Research, 33*(2), 678–696. https://doi.org/10.1287/isre.2021.1079

Golder, P. N., et al. (2023). Learning from data: An empirics-first approach to relevant knowledge generation. *Journal of Marketing, 87*(3), 319–336. https://doi.org/10.1177/00222429221129200

Gnewuch, U., Morana, S., Hinz, O., Kellner, R., & Maedche, A. (2024). More than a bot? The impact of disclosing human involvement on customer interactions

with hybrid service agents. *Information Systems Research*, 35(3), 936–955. https://doi.org/10.1287/isre.2022.0152

Hoffman, M., Kahn, L., & Li, D. (2018). Discretion in hiring. *The Quarterly Journal of Economics, 133*(2), 765–800. https://doi.org/10.1093/qje/qjx042

Huang, M. H., & Rust, R. T. (2021). A strategic framework for artificial intelligence in marketing. *Journal of the Academy of Marketing Science, 49*(1), 30–50. https://doi.org/10.1007/s11747-020-00749-9

Huang, M. H., & Rust, R. T. (2024). The caring machine: Feeling AI for customer care. *Journal of Marketing, 88*(5), 1–23. https://doi.org/10.1177/00222429231224748

Jain, H., et al. (2021). Editorial for the special section on humans, algorithms, and augmented intelligence: The future of work, organizations, and society. *Information Systems Research, 32*(3), 675–687. https://doi.org/10.1287/isre.2021.1046

Jia, N., Luo, X., Fang, Z., & Liao, C. (2024). When and how artificial intelligence augments employee creativity. *Academy of Management Journal, 67*(1), 5–32. https://doi.org/10.5465/amj.2022.0426

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

Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research, 32*(3), 713–735. https://doi.org/10.1287/isre.2020.0980

Kotler, P., Keller, K. L., Brady, M., Goodman, M., & Hansen, T. (2019). *Marketing management* (4th European ed.). Pearson.

Liu-Thompkins, Y., Okazaki, S., & Li, H. (2022). Artificial empathy in marketing interactions: Bridging the human-AI gap in affective and social customer experience. *Journal of the Academy of Marketing Science, 50*(6), 1198–1218. https://doi.org/10.1007/s11747-022-00892-5

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Longoni, C., & Cian, L. (2022). Artificial intelligence in utilitarian vs. hedonic contexts: The "word-of-machine" effect. *Journal of Marketing, 86*(1), 91–108. https://doi.org/10.1177/0022242920957347

Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science, 38*(6), 937–947. https://doi.org/10.1287/mksc.2019.1192

Mikalef, P., et al. (2022). Thinking responsibly about responsible AI and "the dark side" of AI. *European Journal of Information Systems, 31*(3), 257–268. https://doi.org/10.1080/0960085X.2022.2026621

Padigar, M., Pupovac, L., Sinha, A., & Srivastava, R. (2022). The effect of marketing department power on investor responses to announcements of AI-embedded new product innovations. *Journal of the Academy of Marketing Science, 50*(6), 1277–1298. https://doi.org/10.1007/s11747-022-00873-8

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review, 46*(1), 192–210. https://doi.org/10.5465/amr.2018.0072

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

Shang, S., & Seddon, P. B. (2002). Assessing and managing the benefits of enterprise systems: The business manager's perspective. *Information Systems Journal, 12*(4), 271–299. https://doi.org/10.1046/j.1365-2575.2002.00132.x

Shneiderman, B. (2021). Responsible AI: Bridging from ethics to practice. *Communications of the ACM*, 64(8), 32-35. https://doi.org/10.1145/3445973

Siala, H., & Wang, Y. (2022). SHIFTing artificial intelligence to be responsible in healthcare: A systematic review. *Social Science & Medicine, 296,* 114782. https://doi.org/10.1016/j.socscimed.2022.114782

Thormundsson, B. (2024). *Global spending on AI 2023, by industry*. Statista. https://www.statista.com/statistics/1446052/worldwide-spending-on-ai-by-industry/

Verganti, R., Vendraminelli, L., & Iansiti, M. (2020). Innovation and design in the age of artificial intelligence. *Journal of Product Innovation Management, 37*(3), 212–227. https://doi.org/10.1111/jpim.12523

Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change, 126,* 3–13. https://doi.org/10.1016/j.techfore.2015.12.019

Chapter II

From the Darkness Cometh the Light: Actualising Responsible Artificial Intelligence (AI) in Driving Business Value

Abstract

As organisations across industries increasingly invest in artificial intelligence (AI) for business operations, ethical and societal concerns surrounding its adoption have prompted a growing strategic emphasis on responsible AI practices. While high-level principles of responsible AI are well established, there remains limited understanding of its operationalisation and its role in generating business value. This study addresses this gap by integrating corporate social responsibility (CSR) theory (i.e., economic, ethical, and corporate citizenship dimensions), the enterprise systems (ES) benefit framework (i.e., operational, managerial, strategic, information technology [IT] infrastructure, and organisational benefits), and the practice-based view (PBV) to develop a conceptual model of the business value of responsible AI. Drawing on recent information systems (IS) literature on AI ethics and IT value, as well as an analysis of 51 publicly available AI implementations from 13 in-depth case studies, the model identifies six responsible AI practices and their dominant linkages to business value creation. To ensure practical relevance, an applicability check was conducted through interviews with industry professionals. The findings advance theoretical understanding of responsible AI by extending CSR and IS business value perspectives to AI governance, demonstrating how ethical AI practices drive tangible business benefits while ensuring corporate responsibility and regulatory compliance. Practically, this study provides organisations with a framework to implement responsible AI and maximise its business value.

Keywords: responsible artificial intelligence (AI), business ethics, business value, corporate social responsibility (CSR), practice-based view (PBV)

1. Introduction

Artificial intelligence (AI) is the development of machines capable of simulating human intelligence to perform complex tasks, including decision-making and predictive analysis. AI systems learn from their environments and autonomously adapt to changes without direct human intervention, enabling businesses to tackle challenges beyond the capabilities of traditional computational technologies (Rai et al., 2019; Verganti et al., 2020). Given its rapid advancements, AI is transforming industries worldwide and influencing an astounding variety of business activities (Davenport et al., 2020; Dwivedi et al., 2021). Companies across diverse sectors, including automotive (e.g., Audi AG), finance (e.g., Capital One, USAA), and information technology (IT; e.g., Google, IBM), are integrating AI to optimise business models and deliver innovative customer experiences (Chen et al., 2012).

However, AI proliferation also comes with certain ethical and societal concerns, such as privacy invasions, data leakages, algorithmic biases, and misinformation (Jobin et al., 2019). These concerns are particularly pronounced with recent generative AI (GenAI) technologies, which generate human-like content and engage in complex decision-making but pose substantial risks despite their proven popularity. Companies that fail to address these challenges risk reputational damage, regulatory scrutiny, and consumer distrust. For example, Microsoft's controversial dissolution of its AI ethics and society team has heightened concerns about unchecked AI development and corporate responsibility (Schiffer & Newton, 2023).

Stakeholders—including customers, regulators, investors, and employees—are demanding greater accountability in AI development and deployment. Researchers

and policymakers have cautioned against the indiscriminate adoption of AI without ethical safeguards, citing risks such as discriminatory outputs and unintended societal consequences (Dwivedi et al., 2023; Stokel-Walker & Van Noorden, 2023). AI experts have even called for a temporary halt to the development of powerful AI systems, such as ChatGPT, until safety protocols are in place to ensure responsible implementation (Osborne, 2023). Despite these concerns, many organisations lack structured policies and governance frameworks for implementing AI responsibly (Ramaswamy et al., 2018). There remains an urgent need for businesses to operationalise responsible AI in ways that align with stakeholder expectations, regulatory requirements, and ethical standards while maintaining competitive advantages.

A growing body of research has explored AI ethics, focusing on high-level principles guiding AI design, development, and deployment (e.g., Floridi et al., 2018; Jobin et al., 2019), and examining system design embedded with ethical implications (e.g., Peters et al., 2020; Vakkuri et al., 2019; Wearn et al., 2019). However, these principled approaches often remain abstract and disconnected from practical implementation, leaving a critical gap in defining responsible AI and understanding how its practices contribute to business value. *Business value*, broadly defined, encompasses the tangible and intangible benefits that companies derive from investments and strategic initiatives. It extends beyond financial performance to include operational efficiency, innovation, customer satisfaction, brand reputation, and long-term sustainability (Melville et al., 2004). In the digital era, companies increasingly recognise that business value is multidimensional, shaped by both economic returns and broader societal impacts.

In the context of AI-driven businesses, responsible AI implementation has the potential to create business value by mitigating risks, enhancing decision-making, fostering consumer trust, and improving brand reputation through compliance with legal, societal, and organisational standards. However, many organisations struggle to operationalise responsible AI in a way that delivers measurable business value, as existing research primarily focuses on high-level AI principles with limited guidance on practical implementation. This gap highlights the need for a more structured approach to responsible AI that translates ethical considerations into business practices, enabling organisations to balance responsible AI adoption with performance-driven outcomes.

To examine the business value of responsible AI, we draw on the well-established ES benefits framework (Shang & Seddon, 2002) and the IT business value literature (e.g., Kohli & Grover, 2008; Schryen, 2013), which provides a robust theoretical foundation for understanding how organisations derive value from IT investments beyond financial benefits but extend to strategic, operational, and organisational improvements (Gregor et al., 2006). The IT business value perspective is particularly relevant because AI is fundamentally an IT innovation, and its value creation mechanisms align with prior research on ES (Mueller et al., 2010), data analytics (Wang et al., 2018), and digital transformation (Bharadwaj et al., 2013; Mithas et al., 2011). Applying IT business value perspectives enables a structured analysis of how responsible AI practices enhance performance and stakeholder relationships, linking ethical AI to measurable business outcomes. This approach ensures responsible AI is not merely compliance-driven but a source of sustainable competitive advantage.

Our study seeks to address the aforementioned gap by conceptualising responsible AI practices and investigating their business value implications through empirical case studies. Specifically, we build on the practice-based view (PBV; Bromiley & Rau, 2016), a strategic management perspective that extends the resource-based view (RBV) by emphasising organisational practices as key drivers of competitive advantage, and corporate social responsibility (CSR) perspectives while drawing from information systems (IS) research on IT business value to develop a conceptual model. This model allows us to synthesise business value from responsible AI practices and explore the following research question: *how can companies effectively implement responsible to drive business value?* Our research objectives are to: (1) conceptualise responsible AI practices; (2) identify the business benefits of responsible AI; (3) establish a framework linking responsible AI to business value; and (4) develop effective implementation strategies for responsible AI. These objectives provide a comprehensive understanding of responsible AI and actionable insights for integrating it into corporate strategy.

We conducted a qualitative analysis of 13 real-world responsible AI implementation cases across different industries, identifying key responsible AI practices and their impact on business value. This multi-industry approach enabled us to capture diverse AI governance strategies and the varying ways organisations integrate responsible AI into their operations. Additionally, we performed an applicability check (i.e., importance, accessibility, and suitability; Rosemann & Vessey, 2008) through interviews with industry practitioners to assess the practical relevance and feasibility of our findings. This step ensured that our conceptual model aligns with

real-world organisational challenges and strategies, strengthening its applicability for businesses seeking to implement responsible AI effectively.

This research contributes to both theory and practice. Theoretically, it offers a more holistic view of responsible AI by incorporating CSR dimensions and introducing novel constructs that link responsible AI to business value. It also extends the understanding of IT business value in AI contexts, using an ES benefits framework to explain how responsible AI influences organisational performance through a practice-based approach. Practically, our conceptual model offers actionable insights for companies, helping them implement responsible AI practices that not only mitigate ethical risks but also enhance business value creation.

The remainder of this study is structured as follows. The second section provides the theoretical background, defining responsible AI practices and their potential business benefits, leading to the development of our conceptual model. The third section outlines the research methodology. The fourth section presents findings from the multiple case study analysis and applicability check. Finally, the fifth section discusses theoretical and practical implications, study limitations, and future research directions.

2. Theoretical Background and Research Model

2.1 Conceptualising responsible AI through the lens of corporate social responsibility

Over the past decade, the IS literature has increasingly highlighted ethical concerns stemming from digital transformation (Culnan & Williams, 2009; Gerlach et al., 2019; Xu et al., 2011). However, only a limited number of IS studies have explored

the nexus between ethical issues arising from digital technology use and CSR (e.g., Flyverbom et al., 2019; Newell & Marabelli, 2015). While AI offers significant opportunities for processing large-scale data and enabling data-driven decision-making (Brynjolfsson & Mitchell, 2017), its responsible adoption presents ethical, legal, and organisational challenges. Algorithmic bias, transparency issues, and data privacy risks threaten fairness and accountability, particularly in high-stakes sectors such as healthcare, finance, and law enforcement (Danks & London, 2017; Martin & Murphy, 2017; Mehrabi et al., 2021; Mittelstadt et al., 2016). Many organisations face governance gaps, struggling to balance profit-driven AI adoption with ethical responsibility (Dwivedi et al., 2021). The complexities lie in balancing technological advancements with ethical, legal, and societal considerations.

Newell and Marabelli (2015) have therefore called on IS scholars to explore how companies' irresponsible use of AI-driven analytics affects individuals, companies, and societies. Some recent studies have responded to this call. For instance, Wright and Schultz (2018) proposed an ethical AI framework by integrating stakeholder theory[3] (Donaldson & Preston, 1995) with social contract theory[4] (Donaldson & Dunfee, 1994), outlining best practices to mitigate AI-driven ethical risks such as acknowledging the transition, minimising disruptions, and reducing social inequalities. Similarly, Flyverbom et al. (2019) explored the intersection between digital transformation and responsible business, emphasising the need for

[3] *Stakeholder theory* posits that businesses should consider the interests of all stakeholders—not just shareholders—when making decisions. Stakeholders include anyone affected by a company's operations, such as employees, customers, suppliers, communities, and investors.
[4] *Social contract theory* is a philosophical framework that posits that individuals voluntarily consent, either explicitly or implicitly, to form a society and abide by its rules in exchange for protection, security, and social order.

companies to extend beyond profit maximisation and proactively address ethical and societal challenges when adopting digital technologies such as AI.

As companies increasingly recognise CSR's role in fostering sustainable business growth—encompassing economic, environmental, and societal dimensions—the integration of responsible AI into forward-thinking business strategies has gained prominence (Baskentli et al., 2019; Sen & Bhattacharya, 2001). Research shows that CSR initiatives enhance organisational performance by improving stakeholder relationships and corporate reputation (e.g., Margolis & Walsh, 2003; Öberseder et al., 2013; Peloza & Shang, 2011; Weber, 2008). Simultaneously, the rise of digitalisation as a megatrend in the global economy has prompted companies to adopt data-driven technologies such as AI to predict customer needs, personalise products and services, and maintain a competitive advantage (Dwivedi et al., 2021; Huang & Rust, 2018). However, the ethical and societal risks associated with AI adoption necessitate responsible implementation aligned with CSR principles (Bocquet et al., 2013). Failure to integrate CSR into digital business strategies may hinder AI adoption from reaching its full potential, exposing firms to reputational risks, regulatory scrutiny, and decreased consumer trust (Bernal-Conesa et al., 2017; Martin & Murphy, 2017). Regulatory frameworks and public awareness of ethical AI issues have also pushed firms to take greater responsibility for AI implementations (Osburg & Lohrmann, 2017). Despite these efforts, a structured conceptualisation of responsible AI remains underdeveloped in the literature, particularly concerning its operationalisation in business settings and its role in value creation.

Recent discussions on responsible AI (Arrieta et al., 2020; Dignum, 2019; Peters et al., 2020) highlight the importance of integrating ethical AI principles into business practices. However, these studies primarily focus on the theoretical prospects of responsible AI without establishing a clear theoretical framework for its practical implementation. To bridge this gap, we propose a conceptualisation of responsible AI using the three CSR conceptions outlined by Windsor (2016): economic, ethical, and corporate citizenship. This framework extends prior CSR literature to the AI context, offering an analytical lens through which responsible AI practices can be understood in organisational settings. By applying CSR as a guiding framework, this study contributes to AI ethics and responsible AI literature by demonstrating how responsible AI practices align with corporate objectives and generate business value while advancing social good.

2.1.1 The economic conception of responsible AI

The economic conception of CSR prioritises the creation of corporate value and market wealth, aligning CSR initiatives with fiduciary responsibility, minimalist public policy, and customary business ethics (Moon et al., 2005; Windsor, 2006). This perspective resonates with the classical shareholder view, where businesses' primary responsibility is to maximise long-term financial returns for their owners (Van Marrewijk, 2003). Within this paradigm, AI is regarded as a technological enabler embedded in business operations to enhance operational efficiency, drive product and process innovation, and generate sustained economic benefits. Companies adopting AI from an economic CSR perspective focus on profitability and productivity, with minimal emphasis on broader societal or ethical concerns.

Applying this perspective to responsible AI, we define the *economic conception of responsible AI* as "the practice of deploying AI systems to optimise business operations, enhance competitiveness, and create economic value while ensuring responsible use in product and service offerings". This requires companies to implement AI in ways that align with financial objectives while mitigating risks that could undermine trust and long-term viability.

2.1.2 The ethical conception of responsible AI

The ethical conception of CSR has emerged in response to rising ethical concerns and evolving data regulations, focusing on moral business conduct (Windsor, 2006). This perspective advocates impartial moral reflection (e.g., unbiased evaluation of ethical issues), self-regulation, and altruism duties, encouraging businesses to go beyond legal compliance and uphold stakeholder rights. Van Marrewijk (2003) further suggested that companies should not only be accountable to shareholders but also balance the interests of stakeholders—including employees, customers, regulators, and society at large. In line with this approach, Someh et al. (2019) built upon stakeholder theory and discourse ethics to highlight key ethical concerns in big data analytics, such as data privacy, algorithmic fairness, and transparency.

Extending this to the context of AI implementation, AI can have unintended ethical consequences—such as reinforcing biases, making opaque decisions, or breaching user privacy—necessitating their alignment with ethical principles. Ethical AI governance must incorporate fairness, accountability, and transparency, ensuring that AI applications are designed and deployed in ways that respect individual rights and stakeholder interests. However, ethical CSR may appear ambiguous when

viewed as rigid compliance rather than a dynamic organisational process (Windsor, 2006). Thus, deploying human-AI hybrids (Rai et al., 2019)—where AI complements human judgment—can help companies navigate ethical complexities and enhance stakeholder trust. Building on these insights, we define the *ethical conception of responsible AI* as "the practice of ethically integrating AI into business operations, ensuring fairness, transparency, and accountability to protect stakeholder interests".

2.1.3 The corporate citizenship conception of responsible AI

The corporate citizenship perspective of CSR extends beyond economic and ethical responsibilities to encompass companies' voluntary contributions to societal well-being (Windsor, 2006). It views philanthropy as a strategic tool to enhance corporate reputation and acknowledges businesses as both private and public actors that influence economic and social systems. This approach aligns with Van Marrewijk's (2003) societal CSR approach, which emphasises public legitimacy, regulatory cooperation, and long-term societal impact. Someh et al. (2019) further identified key societal concepts in ethical big data analytics, such as power imbalance, surveillance, coercion, and principles and guidelines. They address the equitable use of big data analytics, the necessity of regulatory measures to protect individuals from potential harm, privacy concerns, and the balance between individual autonomy and reliance on data contributions in big data analytics services.

In the context of responsible AI, corporate citizenship goes beyond compliance and ethical design to actively leveraging AI for social good (Jobin et al., 2019). For

example, H&M's Body Scan Jeans project[5] utilises AI to enhance personalised fashion recommendations while reducing waste from returns, aligning AI use with environmental sustainability and inclusivity. This demonstrates how AI can serve as a catalyst for social fairness, diversity, and sustainability, reinforcing companies' role as responsible corporate citizens. Thus, we define the *corporate citizenship conception of responsible AI* as "the practice of leveraging AI to promote socially responsible initiatives, advancing equity, inclusiveness, and sustainability".

2.2 Conceptualising the business value of responsible AI using

IT business value refers to the impact of IT investments on organisational performance and has been widely explored in IS literature, often linked to IS capabilities and firm-level benefits (Devaraj & Kohli, 2003; Langdon, 2006; Melville et al., 2004). As AI advancements enhance IT capabilities, enabling both the augmentation of existing IS systems and the development of AI-powered solutions (Rzepka & Berger, 2018), it becomes critical to understand how responsible AI adoption contributes to business value. While traditional IT business value research focuses on efficiency gains, cost reductions, and competitive advantage derived from IT investments, responsible AI business value extends beyond financial gains, integrating ethical, social, and regulatory considerations while accounting for stakeholder expectations crucial for long-term organisational sustainability. This positions responsible AI as a driver of sustainable business growth.

[5] https://hmgroup.com/our-stories/responsible-ai-is-better-ai/

To conceptualise responsible AI business value, we adapt and extend Shang and Seddon's (2002) multidimensional ES benefit framework (see Table II.1), which categorises IT benefits into five dimensions: operational, managerial, strategic, IT infrastructure, and organisational benefits, to map potential responsible AI benefits and capture the unique business value AI offers to organisations. This framework is well-suited for evaluating the business value of responsible AI for three key reasons. First, it has been empirically validated in prior studies assessing specific IS benefits (e.g., service-oriented architecture [SOA], big data analytics) in organisational contexts (Mueller et al., 2010; Wang et al., 2018), making it a robust foundation for analysing AI-enabled systems. Second, it provides a more structured approach to capturing diverse benefits beyond financial returns using 21 sub-dimensions of IS benefits, allowing us to assess responsible AI practices from multiple organisational perspectives. Third, three key benefit dimensions—IT infrastructure, operational, and strategic benefits—have been acknowledged in AI ethics research (Peters et al., 2020; Sambasivan & Holbrook, 2018), reinforcing the relevance of this framework in bridging AI ethics with IS benefits in organisational contexts.

Table II.1 Responsible AI business value dimensions

Dimensions	Subdimensions (all consequences of responsible AI use)
Operational	Responsible AI streamlines processes and automates tasks, enhancing operational efficiency while maintaining ethical safeguards. AI-driven automation reduces costs, minimises human errors, and improves service delivery. Key business value include: • *cost reduction* through AI-enabled process automation and resource optimisation • *cycle time reduction* via real-time data processing and predictive analytics • *productivity improvement* by augmenting human capabilities and reducing repetitive tasks

	• *quality improvement* by augmenting human capabilities and reducing repetitive tasks • *customer service improvement* via AI chatbots, personalised recommendations, and proactive issue resolution
Managerial	Responsible AI supports better resource allocation, enhances decision-making, and ensures ethical AI governance. By integrating AI insights with human judgment, organisations can optimise resource management and improve business planning. Key business value include: • *better resource management* through AI-driven forecasting and optimisation models • *improved decision making and planning* by leveraging AI-powered analytics for data-driven strategies • *performance improvement* via AI-powered monitoring systems that track key performance indicators and detect inefficiencies
Strategic	AI creates sustained competitive advantages by enabling innovation, supporting business growth, and facilitating external collaborations. Responsible AI ensures that these advantages are ethically grounded and aligned with corporate social responsibility goals. Key business value include: • *support for business growth* by leveraging AI for market expansion and operational scalability • *support for business alliance* through AI-enabled data-sharing and collaborative platforms • *building business innovations* by enabling AI-driven research, development, and new product creation • *building cost leadership* through AI-driven efficiencies and intelligent automation • *generating product differentiation* via AI-enhanced personalisation and adaptive offerings • *building external linkages* through responsible AI partnerships and ethical AI deployment
IT infrastructure	AI-driven infrastructure improvements lay the foundation for digital transformation while ensuring responsible implementation. Organisations gain enhanced IT flexibility and scalability while mitigating risks associated with AI deployment. Key business value include: • *building business flexibility* to accommodate AI-driven advancements and regulatory changes • *reducing IT costs* by optimising infrastructure for AI-driven automation and cloud-based AI services • *enhancing capabilities for responsible AI implementation,* ensuring compliance with ethical standards, governance frameworks, and data protection policies
Organisational	AI reshapes organisational structures by fostering knowledge-sharing, employee empowerment, and customer-centric approaches. Responsible AI aligns technological advancements with social responsibility, ensuring that AI adoption benefits all stakeholders. Key business value include:

	• *changing work patterns* by enabling AI-assisted decision-making and adaptive workflows • *facilitating organisational learning* through AI-driven knowledge management and training systems • *empowering employees* by integrating AI as a collaborative tool rather than a replacement • *building a common vision* by aligning AI strategies with corporate values and stakeholder expectations

Note. Adapted from Shang and Seddon (2002)

2.3 Conceptual model from the practice-based view

Early research drawing on the resource-based view (RBV) emphasised the importance of acquiring IT resources that are valuable, rare, inimitable, and non-substitutable (VRIN) to gain a competitive edge in the marketplace (Barney, 2001). Within the IT business value domain, extensive studies have examined the synergetic effects of IT resources. For instance, Melville et al. (2004) proposed an integrated model suggesting that IT business value is enhanced when technological IT, human IT, and complementary organisational resources are effectively bundled. Tanriverdi (2006) further demonstrated that corporate performance benefits arise from combining IT infrastructure, IT strategy-making, IT human resource management, and IT vendor management resources. Additionally, several studies have highlighted the role of complementary factors, such as organisational structure, culture, policies, and external conditions, in amplifying IT's impact on business performance (Nevo & Wade, 2010; Wang et al., 2019). A recent study by Mikalef and Krogstie (2020) explored how big data analytics resources interact with contextual factors to drive business process innovation. Collectively, these studies extend RBV by specifying how IT resources, when combined with complementary assets and business processes, enhance organisational performance.

However, while RBV remains a dominant framework in IS literature, it may not be the most suitable lens for understanding responsible AI practices. RBV primarily focuses on securing unique, firm-specific resources to outcompete rivals, emphasising the exclusivity and inimitability of resources as sources of competitive advantage. In contrast, responsible AI is not confined within a firm but relies on shared knowledge, industry-wide best practices, and evolving ethical standards that transcend firm boundaries. Competitive advantage in responsible AI does not stem from possessing inimitable AI resources but rather from effectively adopting, operationalising, and sustaining responsible AI practices in alignment with regulatory, societal, and stakeholder expectations.

To better capture the practice-oriented nature of responsible AI, we adopt the PBV, which offers an alternative theoretical perspective to RBV. PBV, proposed by Bromiley and Rau (2014), shifts the focus from exclusive, inimitable resources to publicly available, imitable practices that can be transferred across companies to enhance performance. It defines practices as "a defined activity or set of activities that a variety of companies might execute" (p. 1249), with variations in execution leading to differences in organisational performance (Bromiley & Rau, 2014). Further, Russo-Spena and Mele (2012) described practices as incorporating subjects, actions, tools, and contexts, conceptualising innovation as a collection of co-creation practices.

Applying PBV to responsible AI, we argue that responsible AI practices represent a transferable set of best practices that organisations can adopt to ethically integrate AI and generate business value. Unlike RBV, which would frame responsible AI as a firm-specific, proprietary asset and a source of competitive differentiation, PBV

provides a dynamic framework for evaluating how organisations institutionalise responsible AI governance, balance ethical considerations with business objectives, and adapt AI practices in response to evolving societal and regulatory expectations. This perspective also aligns with the "strategy-as-practice" (SAP) approach (Vaara & Whittington, 2012) in strategic management literature, which has been widely applied in IS research to explore how IT-enabled practices shape organisational capabilities and performance (Huang et al., 2014).

We develop a conceptual model of responsible AI business value (see Figure II.1) through the lens of PBV, providing a practice-oriented framework to examine how organisations implement responsible AI, navigate challenges, and align adoption with strategic objectives. By framing responsible AI as a transferable practice rather than mere regulatory compliance, the model categorises responsible AI practices into three CSR-based dimensions: economic, ethical, and corporate citizenship. These dimensions align with PBV's focus on practices as defined, actionable activities that organisations can execute to achieve specific outcomes. The model further integrates responsible AI practices with the multidimensional ES benefit framework, categorising their business value into operational, managerial, strategic, IT infrastructure, and organisational dimensions. This integration allows for a structured assessment of how responsible AI practices enhance organisational performance, build stakeholder trust, and support sustainable business growth (Kohli & Grover, 2008). By linking PBV, CSR, and ES benefit frameworks, the model offers a comprehensive perspective on systematically implementing responsible AI to achieve both economic value, and broader ethical and societal impact.

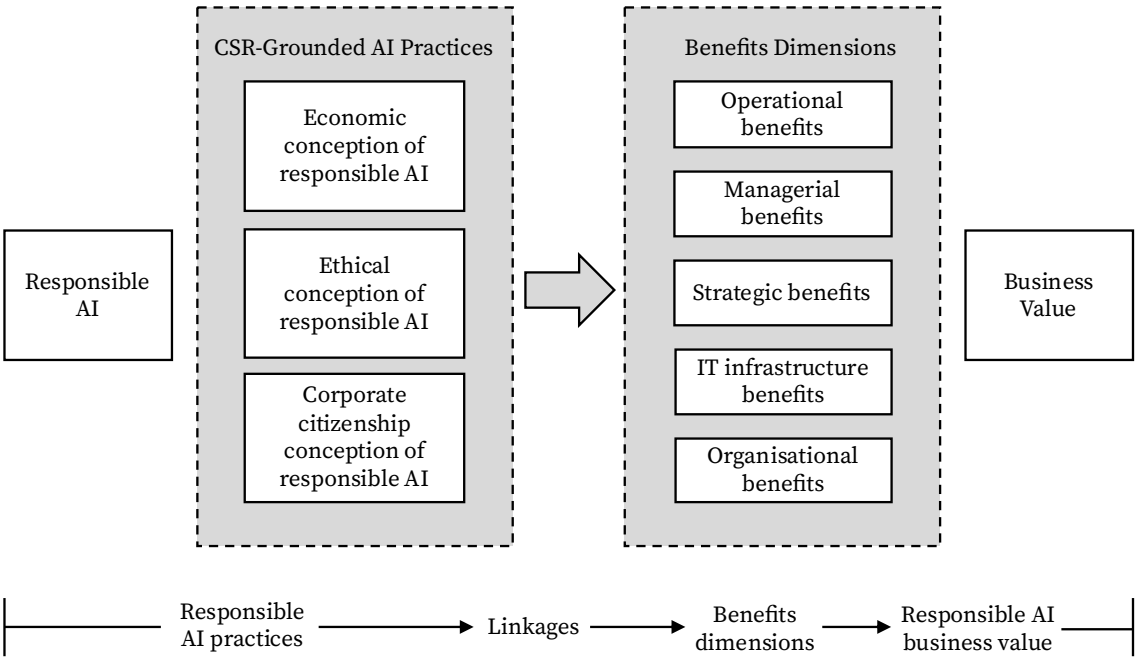


Figure II.1 Conceptual model of responsible AI business value

3. Method

3.1 Research design and approach

Our study aimed to identify key responsible AI practices and examine their business value potential. Given the nascent state of knowledge on this topic, we adopted an interpretivist paradigm, which acknowledges that multiple realities are shaped by human experiences and subjective interpretations (Walsham, 2006). Since responsible AI is still an emerging concept, valuable insights can be derived from real-world industry cases that illustrate how organisations across different sectors implement AI responsibly from a PBV. To achieve this, we employed a multiple case study approach, a well-established method for conducting in-depth, multi-faceted explorations of complex, real-life phenomena (Eisenhardt & Graebner, 2007; Yin, 2014). This approach is particularly well-suited for exploratory, practice-based

research, as it allows for a systematic investigation of the relationship between responsible AI practices and business value creation (Kohli & Grover, 2008).

Previous studies on IT business value have effectively used this method to examine emerging technologies such as SOA and big data analytics (e.g., Mueller et al., 2010; Wang et al., 2018), developing theoretical constructs and propositions on how IT and IS implementations contribute to organisational performance (e.g., economic potential). By applying this approach to responsible AI, we extend IT business value research by generating novel insights into how CSR-grounded AI practices drive business value within organisations. In our interpretive case study approach, we systematically decomposed the documented statements on real-world responsible AI implementations using our proposed conceptual model (see Figure II.1). This structured analysis enabled us to identify recurring patterns linking responsible AI practices to IT benefits that drive business value creation. Specifically, we analysed responsible AI practices through the lens of CSR, categorised business benefits using the ES benefits framework, and mapped the interconnections between these elements. This integrative approach provided a systematic framework for understanding how organisations operationalise responsible AI, linking ethical considerations to tangible business outcomes. By aligning these dimensions, we provided empirical evidence supporting the business value potential of responsible AI, reinforcing PBV as a relevant theoretical foundation for understanding how responsible AI practices contribute to organisational performance and societal impact.

To further validate our findings and ensure their practical relevance, we incorporated an *applicability check*, a recommended approach in IS research for

evaluating the real-world significance of conceptual models (Rosemann & Vessey, 2008). We conducted eight expert interviews with AI professionals from diverse industries to assess the model's relevance across three key dimensions: importance, accessibility, and suitability. The *importance* dimension examined the critical role of responsible AI practices in shaping ethical AI adoption, ensuring that the model aligns with industry needs. The *accessibility* dimension assessed whether the conceptual framework was presented in a clear, structured, and actionable manner for industry practitioners. Lastly, the *suitability* dimension explored the extent to which the identified responsible AI practices can be effectively applied in real-world AI implementations.

Findings from these expert interviews confirmed that responsible AI practices are essential for ethical AI deployment and that a structured framework can support companies in navigating AI governance. Several experts highlighted that the visual representation of responsible AI practices within our model made it easier for businesses to identify key areas for AI implementation. Others noted that the framework could be adapted into practical tools, such as responsible AI handbooks or training guidelines, to support AI governance efforts. This validation process strengthens both the practical significance and theoretical robustness of our responsible AI business value model, reinforcing its potential as a strategic tool for organisations seeking to balance AI ethics with business performance.

3.2 Data collection

Given the emergent nature of responsible AI practice, selecting organisations with demonstrated success in ethical AI design and governance was essential for

deriving meaningful insights into responsible AI business value. Following established case study methodologies (Eisenhardt & Graebner, 2007; Yin, 2014), we employed a purposeful sampling strategy to identify cases that offered revelatory insights into the intersection of AI ethics, governance, and business performance. This approach ensured that case selection was not ad hoc but systematically guided, strengthening the study's explanatory depth and theoretical contribution.

To ensure empirical rigour and relevance, we established systematic case selection criteria prior to data collection. *First,* the cases had to demonstrate the empirical implementation of responsible AI, with documented evidence of ethical AI design and governance frameworks in organisational contexts. This criterion ensured that our analysis centred on tangible business impacts rather than theoretical discussions or speculative claims. *Second,* we prioritised industry and geographical diversity, selecting cases from multiple sectors and regions to capture a broad spectrum of responsible AI challenges, opportunities, and contextual variations shaping AI strategies and business value outcomes. AI design and governance are industry-specific, with distinct challenges across sectors. Regional differences also influence AI governance; for instance, the European Union (EU)' General Data Protection Regulation (GDPR), mandates strict data privacy and accountability, whereas the United States (US) relies on industry self-regulation. By incorporating cross-sectoral cases from diverse regulatory environments, we captured a comprehensive view of responsible AI practices and business value potential, while accounting for variations in governance models, regulatory constraints, and ethical considerations. *Third,* we selected only cases where responsible AI initiatives had progressed beyond the conceptual or planning stages and had reached execution or

completion. This ensured that our study focused on realised business outcomes rather than aspirational commitments. By grounding our findings in observed business value rather than hypothetical benefits, this approach strengthened the validity of our analysis.

The sampling frame comprised organisations widely recognised as early adopters of responsible AI, identified through systematic searches of corporate AI governance disclosures, CSR reports, and media coverage of AI ethics initiatives. In 2020, we collected 51 case materials covering 13 organisations actively implementing responsible AI, spanning North America and Europe. To ensure methodological rigour, case selection was not arbitrary but strategically focused on organisations that had institutionalised responsible AI practices, offering empirical insights into its business value. These early adopters provided rich contextual data for examining how responsible AI is operationalised, making them revelatory cases that illuminate key dynamics within this evolving field (Yin, 2014). Following case selection, we employed a multi-source data collection strategy to enhance the robustness of findings through triangulation. As outlined in Table II.2, data sources included academic research articles, industry reports, CSR disclosures, vendor case studies, consultant analyses, and media coverage from newspapers and magazines discussing AI adoption, ethical risks, and business implications related to the case companies. The integration of diverse sources strengthened the credibility of our findings, ensuring a well-rounded analysis of responsible AI implementation and its business value.

A critical concern in case study research is the risk of selection bias (Hernán et al., 2004), particularly when replying on company reports and publicly available

sources, as firms may selectively disclose information that supports positive narratives while downplaying challenges or ethical dilemmas. Additionally, discrepancies in the availability and depth of materials across cases may introduce biases in comparative analysis. To mitigate these risks, we employed a two-pronged approach. First, we critically evaluated the credibility of each case by cross-referencing industry reports, regulatory findings, and independent analyses to corroborate organisational claims about responsible AI implementation. Second, we conducted an applicability check, following Rosemann and Vessey's (2008) guidelines, to assess the importance, accessibility, and suitability of our findings. This step was reinforced through expert interviews, providing external validation and deeper insights into responsible AI practices and their business value potential beyond self-reported corporate narratives.

Moreover, the composition of our sample significantly shaped our findings, as it focused on early adopters of responsible AI, primarily in North America and Europe, where regulatory and corporate attention to AI ethics is more pronounced. While this approach captures leading practices, it may not reflect the experiences of organisations with less mature AI governance or those in regions with different regulatory landscapes. This suggest that while our study provides robust insights into responsible AI's business value, future research should expand to diverse geographic and organisational contexts, leveraging direct organisational data to capture a more comprehensive view of responsible AI adoption and its evolving business impact.

Table II.2 Summary of data sources and analysis

Data Type	Data Source (quantity)	Responsible AI implementation cases	Use in Analysis
Research papers	2 papers (107 pages)	• Alder Hey Children's Hospital *(Healthcare, UK)* • Audi AG *(Automobile manufacturing, Germany)* • Capital One *(Financial & banking, US)* • Equinor ASA *(Petroleum & energy, Norway)* • Ernst & Young *(EY Global, Professional services, UK)* • Google *(Software, US)* • H&M *(Clothing retail, Sweden)* • IBM *(Software, US)* • IESO Digital Health *(Healthcare, UK)* • NVIDIA Corporation *(Computing, US)* • PwC *(Professional services, UK)* • Quantcast *(Software, US)* • Sage Group *(Software, UK)*	• Provide detailed understanding of responsible AI practices through case studies • Examines corporate strategies, disclosures, as well as third-party and media perspectives on responsible AI • Identify patterns linking responsible AI practices and corresponding potential benefits for business value creation
Company reports (including CSR reports)	16 reports (511 pages)		
Vendor implementation cases, analyst & consultant studies	6 websites (72 pages) 5 documents (121 pages)		
Newspaper, magazine articles	22 articles (148 pages)		

3.3 Data analysis

To systematically analyse the case materials and generate meaningful insights, we adopted a structured yet flexible approach to qualitative content analysis (Elo & Kyngäs, 2008). This method integrates theoretical grounding (i.e., CSR, ES benefits) with empirical observations, allowing for the iterative development of a conceptual model of responsible AI business value. Given that responsible AI remains an emerging and underexplored domain, our approach facilitated the identification of key themes and patterns from real-world implementations without rigidly imposing predefined theoretical constructs. Instead, it enabled abductive reasoning, allowing insights from empirical data to inform conceptual frameworks. Following Elo and Kyngäs' (2008) three-phase coding process: preparation, organising, and reporting,

we systematically extracted, categorised, and synthesised data to construct an empirically grounded yet theoretically informed model of responsible AI business value.

Preparing for the coding process. We first defined three key aspects of the coding process: the unit of analysis, the level of analysis, and the purpose of evaluation (Elo & Kyngäs, 2008). Coding units were identified as themes, which could appear as sentences, paragraphs, or entire sections, as long as they conveyed meaningful insights into responsible AI practices and their implications (Minichiello et al., 1990). At the level of analysis, we examined responsible AI implementation across multiple industries to ensure a broader dataset and avoid sector-specific limitations. This cross-industry perspective allowed the identification of both commonalities and variations in AI governance models, ethical considerations, and business value realisation. The primary purpose of evaluation was to uncover recurring patterns in responsible AI practices and their direct and indirect contributions to business value, offering empirical insights into how organisations integrate ethical AI while achieving economic and societal benefits.

To enhance coding reliability and coder alignment, we developed a comprehensive coding instruction based on Krippendorff's (2018) recommendations for content analysis. The instruction provided definitions of the three CSR-based dimensions of responsible AI (i.e., economic, ethical, and corporate citizenship) and descriptions of the five business value dimensions (i.e., operational, managerial, strategic, IT infrastructure, and organisational benefits). Two coders from the author team were trained using these predefined constructs and provided with detailed coding

procedures, illustrative examples, and data management protocols to ensure a consistent and rigorous analytical approach.

Organising the coding process. We conducted our qualitative analysis through a three-step coding process: open, axial, and selective coding (Strauss & Corbin, 1990; Wiesche et al., 2017), enabling an abductive refinement of theoretical constructs based on empirical observations. In the *open coding* phase, one coder initially reviewed the case materials to identify statements explaining how responsible AI practices contribute to business value. These statements were provisionally categorised under two broad themes: responsible AI practices and corresponding business value. The coding was then verified by a second coder to ensure consistency. Subsequently, both coders analysed and classified these statements into the predefined CSR-grounded responsible AI practices and the five responsible AI business value dimensions (Table II.1). They then reviewed these classifications, reclassifying where necessary to enhance analytical rigour. In the *axial coding* phase, coders further refined the data by identifying sub-elements within responsible AI practices and their business value outcomes. This stage involved extracting key attributes of responsible AI and linking them to specific benefits within the business value framework. Additionally, coders identified linkages between responsible AI elements and business value dimensions, refining the theoretical structure through iterative comparison and abstraction (Marshall & Rossman, 2014; Urquhart et al., 2010). In the *selective coding* phase, coders conducted a comparative analysis across similar coded elements to ensure consistency and eliminate redundancy. This stage focused on finalising core categories and relationships between responsible AI practices and business value dimensions. The coding process ensured that only the

most significant linkages were retained, resulting in a structured model for understanding how responsible AI contributes to business value creation.

To ensure interrater reliability, each coded statement was initially labelled by one coder and independently reviewed by a second coder (Schilling, 2006). Any discrepancies were resolved through discussion, resulting in an inter-coder agreement rate of 87%. The final dataset identified 66 linkages between responsible AI practices and business value dimensions.

Reporting the coding process and the results. The final results were presented in a cross-tabulation matrix (see Table II.3), mapping responsible AI practices to corresponding business value outcomes. This matrix provided a structured framework for analysing how responsible AI generates business value. Additionally, the results were visualised in the research model (see Figure II.2 in the finding section), highlighting the most prominent linkages. Appendix A (Tables II.A1 and II.A2) includes illustrative examples demonstrating how case materials were distilled into meaningful insights through this structured coding process.

Table II.3 Linkages between responsible AI practices and business benefits

Responsible AI practices	Potential business benefits				
	Operational	Managerial	Strategic	IT infrastructure	Organisational
Data governance	**12**	2	**6**	**6**	2
Risk control	2	1	0	3	0
Ethically designed solutions	3	0	1	**10**	1
Human and AI coordination	1	2	1	2	0
Training and education	0	1	3	2	1
Social acceptance and sustainability	0	0	3	0	1

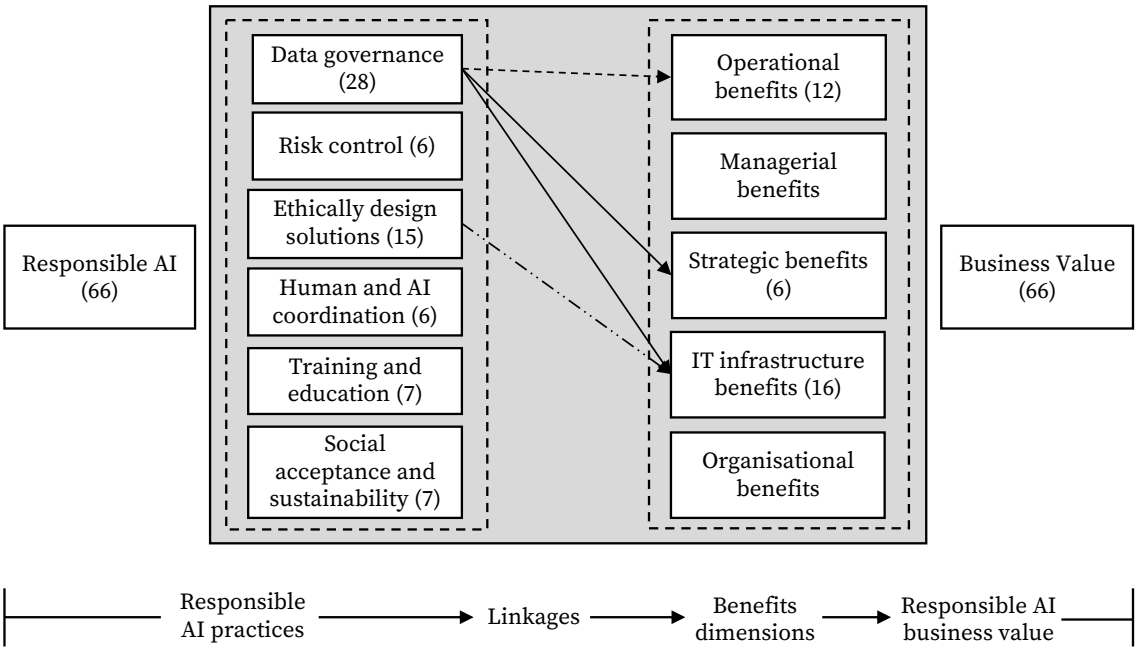
4. Results and Discussion

Building on the theoretical foundations of CSR, IT business value (e.g., ES benefits framework), and PBV, our conceptualisations (Figure II.1) inform the development of a responsible AI business value model (Figure II.2). This model, derived from the analysis of 51 case materials and 66 identified linkages, provides empirical insights into how responsible AI practices generate business value. Our discussion focuses on three key aspects: the occurrences of responsible AI practices and business value elements, the linkages between responsible AI practices and business value, and the propositions emerging from the four most prominent linkages.

Occurrences in our dataset indicate instances where a responsible AI practice was explicitly linked to a business benefit within a case document. To ensure a balanced analysis, some linkages appeared in multiple documents within the same case company, highlighting deeply embedded practices, while others spanned multiple

cases, reinforcing their cross-industry relevance. The nature of these occurrences carries distinct implications. Multi-document, single-case occurrences suggest a core organisational strategy, consistently integrated and documented. Multi-case occurrences indicate widely applicable responsible AI strategies across industries. Isolated occurrences may reflect emerging trends or underreported but potentially significant initiatives. These distinctions help assess the strength and transferability of responsible AI practices in generating business value.

Certain elements in our model, such as training and education, social acceptance and sustainability, and organisational and managerial benefits, did not exhibit strong direct linkages. However, their lower occurrence does not indicate irrelevance. Instead, these factors likely play more indirect roles in shaping responsible AI adoption and its long-term impact. Training and education foster AI literacy and ethical awareness within organisations but may not yield immediate, quantifiable business returns. Similarly, social acceptance and sustainability, though crucial for long-term AI adoption, may not yet be prioritised due to challenges in measuring their financial impact. Organisational and managerial benefits likely emerge as secondary effects, with improved decision-making and workforce engagement stemming from AI transparency and reliability but not always explicitly documented. As AI governance frameworks mature and regulatory pressures increase, these factors may become more prominent in driving responsible AI's business value.



Note: (#) represents number of occurrences this element was coded in the cases analysed
- - - - - ▶ represents the highest linkages (12)
—··—··▶ represents the second highest linkages (10)
——▶ represents the third highest linkages (6)

Figure II.2 Results of the responsible AI business value model

4.1 Synthesis of responsible AI practice

Our case analysis identified six primary responsible AI practices, comprising 15 underlying elements, all structured within the three conceptual dimensions of responsible AI: economic, ethical, and corporate citizenship (see Table II.4). The findings highlight variations in how organisations prioritise and implement responsible AI, which may also be influenced by sampling choices and resource availability in supporting these initiatives.

The *economic conception of responsible AI* emerged as the most dominant, with 34 coded occurrences. This suggests that organisations primarily adopt responsible AI to achieve tangible business benefits, such as improved operational efficiency, competitive advantage, and measurable returns on investment. The strong

presence of this dimension underscores the practical business incentives driving responsible AI adoption. However, this dominance may also reflect selection bias in our case materials, as many organisations are more likely to report AI initiatives that align with economic benefits rather than those aimed at ethical or societal impact. Given that companies often disclose success stories in public reports, responsible AI projects with clear economic returns may be more prominently featured than those with broader, long-term ethical or societal objectives.

The *ethical conception of responsible AI*, with 28 coded occurrences, also holds substantial significance, indicating an increasing organisational awareness of AI's ethical implications. Organisations recognise the necessity of embedding fairness, transparency, explainability, and accountability in AI governance, reflecting the growing regulatory and societal pressures to ensure AI is deployed responsibly. However, the prevalence of ethical AI initiatives may also be influenced by the regulatory environments of the sampled organisations. For instance, companies operating in jurisdictions with stricter AI regulations (e.g., GDPR in Europe) are more likely to prioritise and report ethical AI practices. Additionally, organisations with greater financial and human resources may have dedicated AI ethics teams or governance structures that facilitate responsible AI adoption, further influencing the presence of ethical AI initiatives in our dataset.

The *corporate citizenship conception of responsible AI* appeared the least frequently, with only four occurrences. This may suggest that organisations struggle to operationalise AI's broader societal contributions or perceive them as long-term objectives rather than immediate priorities. Alternatively, businesses may prioritise economic and ethical concerns over societal impact due to the challenges of

measuring social value in AI-driven initiatives. However, the limited representation of corporate citizenship AI initiatives in our sample may also reflect resource constraints and industry-specific variations. Organisations with greater financial flexibility and public visibility (e.g., large tech firms) may have a higher capacity to engage in AI initiatives that promote social good, whereas resource-limited ones may focus on compliance and direct economic benefits. Furthermore, industries with strong societal influence, such as healthcare or financial services, may be more inclined to adopt AI for social impact compared to cost-efficiency-driven-sectors.

Table II.4 Responsible AI practices

Elements of responsible AI practice		Sub-elements	Number of occurrences	
Economic conception of responsible AI	Data governance	• Explainability – Provide meaningful and personalised explanations of AI-generated results	10	34
		• Transparency – Ensure AI usage is transparent to individuals, external constituencies, and communities	12	
		• Reliability – Institutionalise data use and management practices to build trust among users	6	
	Risk control	• Conduct comprehensive assessments of security risks in data use	2	
		• Formulate rules and mechanisms to mitigate economic and performance risks	4	
Ethical conception of responsible AI	Ethically designed solutions	• Foster an ethical mindset and culture within companies and among employees	3	28
		• Develop ethical AI algorithms through collaboration with engineers, scientists, universities, and other stakeholders	10	
		• Use ethical principles to guide socially significant decision-making	2	
	Human and AI coordination	• Integrate human intelligence and skills with AI to facilitate human-AI collaboration	3	

		• Use AI applications to augment human capabilities	1	
		• Train AI agents to correspond with human competencies	2	
	Training and education	• Initiate organisational learning about AI ethics	2	
		• Provide training courses to guide the ethical use of AI for wider stakeholders	5	
Corporate citizenship conception of responsible AI	Social acceptance and sustainability	• Apply AI to promote social sustainability concerning human well-being	3	4
		• Address AI's social acceptance challenges in society	1	
Total			66	

4.1.1 Economic responsible AI practices

Our analysis highlights *data governance* (28 occurrences; hereafter, only the number will be noted) as the most prominent element of economic responsible AI practices, particularly in ensuring transparency (12), explainability (10), and reliability (6) in AI applications. This aligns with Bryson and Winfield's (2017) argument that AI systems must be transparent to stakeholders, enabling them to understand how AI processes data and reaches decisions. Several responsible AI initiatives in our case materials reinforce this priority. For example, IBM's AI Supplier's Declarations of Conformity (SDoCs) function as fact sheets detailing AI products' development, outputs, and performance testing to enhance explainability (Arnold et al., 2019). Interestingly, despite its critical role in AI governance, reliability received less attention in the case materials. A potential explanation is that reliability is perceived as a baseline requirement for emerging technologies, whereas responsible AI discussions may prioritise AI-specific challenges, such as opacity and interpretability. This suggests that while transparency and explainability are seen

as differentiators in responsible AI implementation, reliability may be implicitly assumed rather than actively emphasised.

Moreover, *risk control* (6) emerges as a secondary economic element, emphasising the need for security risk assessment for data use (2) and mechanisms for managing economic and performance risks (4). The relatively lower emphasis on risk control in responsible AI case materials does not necessarily indicate a lack of organisational awareness or action. Instead, it may reflect the embedded nature of risk control within existing governance structures, the challenge of quantifying AI-specific risks, and the evolving regulatory environment. Additionally, organisations may selectively prioritise highlighting AI's economic benefits and ethical considerations over risk governance in public disclosures.

4.1.2 Ethical responsible AI practices

The most prevalent ethical responsible AI practice in our case materials is *ethically designed solutions* (15), underscoring that ethical considerations should not be reactive interventions but embedded in AI development from the outset. This is followed by *training and education* (7) and *human-AI coordination* (6), reflecting the growing recognition of the need to integrate ethical principles across AI's lifecycle.

Among the sub-elements of ethically designed solutions, developing ethical algorithms (10) is the most frequently mentioned, suggesting a proactive approach to mitigating ethical risks in AI deployment. Additionally, cases highlight the importance of fostering an ethical mindset and culture within companies (3) and applying ethical principles to socially significant decisions (2). Ensuring AI ethics

involves both designing and governing ethical AI; thus, guiding principles must be defined to inform the ethical deployment of AI systems. However, as Taddeo and Floridi (2018) argue, defining universal AI ethics principles is inherently complex due to variations in cultural contexts and application domains. An example of translating ethical AI principles into actionable practices is H&M Group's responsible AI initiatives, which include a 30-question ethical AI checklist for new and ongoing projects. This structured approach demonstrates how organisations can institutionalise ethical AI practices, moving beyond abstract principles to concrete implementation strategies.

In the training and education domain, responsible AI initiatives focus on initiating organisational learning (2) and providing AI ethics training for wider stakeholders (5). This suggests that while AI ethics training is gaining traction, it remains a developing area that requires broader adoption across industries. The human-AI coordination element encompasses integrating human expertise with AI systems (3), augmenting human capabilities through AI applications (1), and training AI agents to align with human competencies (2). These findings reinforce the idea that AI should be designed to complement rather than replace human intelligence, ensuring ethical and responsible human-AI interactions.

4.1.3 Corporate citizenship conception responsible AI practices

Our case analysis indicates that corporate citizenship responsible AI practices are the least represented, with only four occurrences. The identified sub-elements include *applying AI for social sustainability* (3) and *addressing AI's social acceptance challenges* (1). Despite academic interest in balancing AI's economic and social

sustainability impacts, real-world implementation appears limited. For one instance, NVIDIA Clara, an AI-powered medical imaging platform, was designed to bridge the gap between legacy imaging systems and modern AI-driven diagnostics (NVIDIA, 2019). However, such initiatives remain relatively rare in our case materials.

While the observed trends provide valuable insights, we acknowledge that the distribution of responsible AI practices in our findings may reflect selection bias in case materials rather than the representation of industry-wide adoption. The dominance of economic and ethical responsible AI practices in our dataset could be influenced by the availability of publicly documentation and companies' strategic emphasis on market-driven and regulatory priorities. The scarcity of corporate citizenship AI cases may reflect limited real-world adoption, underrepresentation of such initiatives in publicly available data, or the general underdevelopment of AI for social good initiatives, possibly driven by market-driven priorities and a lack of immediate business incentives. Moreover, many corporate citizenship AI projects remain fragmented and confined to pilot programmes rather than fully integrated business strategies. While tech firms like Microsoft and Google have publicly committed to AI for humanitarian and sustainability causes, most companies—especially outside the tech sector—have yet to embed such initiatives into their core operations. This may stem from challenges in measuring social impact and competing resource allocations, where organisations prioritise profit-driven AI applications over long-term societal benefits.

4.2 Synthesis of responsible AI business value

A crucial aspect of evaluating responsible AI in a business context is determining how these practices translate into tangible business value. To address this, we examined the linkages between responsible AI practices and their potential business value (see Table II.5). Our analysis highlights IT infrastructure enhancement (23) as the primary business value of responsible AI. In an increasingly digital landscape, responsible AI strengthens IT systems by mitigating data biases and algorithmic errors (10) and ensuring data accuracy (7) for reliable management. Operational benefits (18) also play a key role, particularly in optimising personalised customer service (12) while maintaining ethical integrity. Strategic benefits (14) emerge through improved customer trust (6) and competitive positioning. However, managerial (6) and organisational benefits (5) appear less frequently, indicating that responsible AI adoption currently emphasises IT infrastructure and operational efficiency over broader structural and managerial transformation. At this stage, firms prioritise AI-driven customer engagement and risk mitigation, integrating responsible AI primarily within IT and operational frameworks rather than embedding it as a catalyst for systemic organisational change.

Table II.5 Responsible AI business value

Elements of responsible AI business value	Sub-elements	Number of occurrences	
IT infrastructure benefits	• Mitigate biases and errors in data and algorithms	10	23
	• Reduce IT task complexity	4	
	• Track data effectively	2	
	• Maintain data accuracy	7	
Operational benefits	• Transform customer engagement in a way that creates a dialogic environment	2	18
	• Improve operational efficiency	4	
	• Optimise personalised customer service	12	
Managerial benefits	• Improve speed and quality of managerial decisions	4	6
	• Gain insights quickly about changing market demand	2	
Strategic benefits	• Develop close collaborations with stakeholders	4	14
	• Build strong brand image	4	
	• Improve customer trust	6	
Organisational benefits	• Facilitate organisational learning	2	5
	• Improve employee well being	3	
Total		66	

4.2.1 Dominant linkages: The role of data governance

Our analysis highlights four dominant linkages. Notably, three of these stem from data governance, each aligning with one of the three most frequently mentioned business benefits: 12 out of 66 occurrences (hereafter, only the number will be noted) are linked to operational benefits, 6 to IT infrastructure benefits, and another 6 to strategic benefits. These findings underscore the pivotal role of robust AI data governance in enhancing operational efficiency, minimising security risks, and fostering trust and collaboration.

Companies that prioritise AI data governance are better positioned to streamline operations and mitigate risks while strengthening their competitive position by building a reputation for transparency and reliability. Specifically, transparent AI governance can reduce security breaches, prevent algorithmic biases, and ensure that AI-driven decision-making aligns with business objectives and ethical standards. Given the increasing regulatory scrutiny surrounding AI, businesses that establish clear and accountable AI data governance structures not only reduce legal risks but also enhance credibility with stakeholders, including customers, regulators, and investors. Additionally, the strong link between data governance and strategic benefits suggests that AI systems built on robust governance structures can foster stronger stakeholder collaborations. When AI applications operate transparently and responsibly, they contribute to higher consumer trust, increased regulatory acceptance, and improved cross-sector partnerships, positioning firms as leaders in responsible AI adoption.

4.2.2 IT infrastructure benefits: Ethically designed AI

Another key finding is that ethically designed AI solutions predominantly contribute to IT infrastructure benefits (10). This suggests that ethical AI extends beyond compliance to directly improve technical robustness, system reliability, and efficiency. Ethical AI design ensures that systems are free from biases, operate with fairness, and uphold data privacy standards, which in turn enhances the reliability of AI models. Companies that invest in ethical AI from the early stages of system development benefit from more accurate, robust, and dependable AI-driven decision-making. For example, integrating explainability features in AI systems can

foster transparency that aligns with growing regulatory and consumer demands. Moreover, ethical AI design supports long-term sustainability in AI adoption by preventing issues related to algorithmic discrimination, reputational damage, and legal liabilities. Companies that proactively integrate ethical considerations into AI development reduce the risk of systemic failures and public backlash, thereby safeguarding investments and reinforcing the business case for responsible AI.

4.2.3 Operational benefits: AI personalisation and customer engagement

Among the most prominent business benefits, operational efficiency emerges as a key driver of responsible AI adoption (18), with optimised personalised customer service accounting for 12 of these. This reflects the growing industry trend toward AI-driven customer engagement, where companies leverage AI responsibly to tailor products, services, and interactions while adhering to ethical guidelines. AI personalisation has long been a key competitive advantage, yet concerns over consumer data privacy and algorithmic fairness have led firms to adopt more responsible approaches. The strong linkage between responsible AI practices and dialogic customer engagement highlights that firms increasingly focus on creating transparent, two-way interactions rather than simply leveraging AI for efficiency gains. This shift reflects a growing recognition that ethical AI is not only a regulatory obligation but also a value-generating business strategy. For example, responsible AI practices ensure that recommendation systems do not reinforce discriminatory biases and that AI-driven marketing communications are truthful, inclusive, and non-intrusive. Companies that align AI-driven customer engagement with ethical

standards are more likely to differentiate themselves in competitive markets, securing both market share and consumer goodwill.

4.2.4 Managerial and organisational benefits: Areas for future growth

While our findings show that managerial (6) and organisational (5) benefits are less frequently mentioned compared to IT infrastructure, operational, and strategic benefits, they hold significant untapped potential. As responsible AI adoption matures, firms are likely to increasingly leverage AI to enhance decision-making, workforce productivity, and corporate learning cultures. For managerial benefits, responsible AI can improve the speed and quality of decision-making (4) by delivering accurate, unbiased, and explainable insights. AI-driven analytics enable executives to monitor market trends, anticipate customer needs, and refine business strategies. However, our case analysis suggests that many firms are still in the early stages of integrating AI into managerial workflows in ways that align with responsible AI principles. Similarly, responsible AI can foster organisational learning (2) and enhance employee well-being (3). AI-powered tools can facilitate knowledge sharing, support ethical leadership, and promote inclusive decision-making. For instance, AI applications can be designed to reduce employee burnout through workload balancing could contribute to a more sustainable AI-powered work environment. However, these applications remain underexplored in our dataset, suggesting a need for further research and broader adoption in practice.

To conclude, our findings confirm that responsible AI is both an ethical duty and a strategic asset that enhances business value by improving system reliability, efficiency, and competitiveness through consumer trust and regulatory alignment.

Strong linkages to IT infrastructure, operational, and strategic benefits suggest that firms prioritising responsible AI gain a competitive advantage while mitigating risks. However, the limited focus on corporate citizenship responsible AI indicates that organisations have yet to fully harness AI for broader societal impact. While AI-driven sustainability and social initiatives hold promise, they remain secondary to economic and ethical benefits, highlighting the need for further research and innovation in AI governance models that balance corporate responsibility with profitability. A more holistic approach embedding responsible AI across managerial, organisational, and societal dimensions will be essential for maximising its long-term value.

4.3 Proposition development

Building on our case analysis, we formulated four propositions based on the four most dominant linkages between responsible AI practices and business benefits. These linkages, identified through systematic coding of case materials, represent the most frequently observed patterns in responsible AI implementation at the time of data collection—specifically, how responsible AI practices enhance business value through data governance, ethical design, and transparency—reinforcing their tangible business value. The propositions serve both a confirmatory role, validating the responsible AI business value model, and an interpretative function, providing empirical grounding for theoretical development.

Proposition 1: Transparent personalised customer service strengthens operational benefits of data governance.

Transparency in AI operations is essential for fostering trust and enhancing customer engagement (Bryson & Winfield, 2017). Effective AI strategies depend on robust data governance, especially under regulatory frameworks like the GDPR[6], which mandate explicit consumer consent for data use. While such regulations may initially appear restrictive—limiting behavioural targeting and personalised marketing—they can, when approached with the right data policies, enhance consumer confidence in data sharing (Vayena et al., 2018). A study by the Data & Marketing Association (2018) found that 80% of consumers are willing to share personal data if they understand its intended use. This underscores the potential of transparent AI policies to transform regulatory compliance into a competitive advantage. Capital One, for instance, has made their credit card criteria system transparent and enhances customer trust by providing detailed explanations of algorithmic decisions on credit card applications (Knight, 2017). Also, Alder Hey Children's Hospital's AI-powered Alder Play app improves healthcare quality and patient engagement by offering transparent access to online medical records (Alderheycharity, 2017; de Fine Licht & de Fine Licht, 2020). These initiatives highlight the preference of AI users for transparent data use, which in turn help businesses in improving the quality of their customer service. As Shang and Seddon (2002) assert, process quality improvements yield operational benefits, reinforcing the need for transparency in AI-driven services.

[6] https://gdpr-info.eu/

Proposition 2: Developing ethical algorithms through stakeholder collaboration mitigates data biases and algorithmic errors in AI systems.

Minimising ethical concerns in AI requires proactive design measures to address algorithmic bias, privacy risks, and decision-making opacity (Ågerfalk, 2020). Ensuring ethical AI implementation requires human oversight to evaluate system functionalities and scrutinise underlying algorithms, enhancing accountability and trust. Google, for example, integrates ethics considerations into AI development through collaborative stakeholder engagement, ensuring socially responsible AI applications (Pichai, 2018). They have integrated responsible AI principles into its research and product development, designing user-centred AI systems aligned with best practices for software development. By prioritising fairness and transparency, Google aims to mitigate biases and proactively detect potential errors before they arise. The company has further institutionalised responsible AI governance by forming a 'responsible innovation team' of domain experts to conduct preliminary ethical assessments. Additionally, Google has appointed a senior executive council to oversee critical AI-related decisions and established an external advisory board for independent guidance on AI governance (Walker, 2018; Gershgorn, 2018). Similarly, Quantcast, a firm specialising in AI-driven marketing, applies machine learning (ML) to protect customers' brand safety and combat misinformation in real-time. Embedding ethical principles into AI algorithms ensures AI outcomes align with corporate responsibility standards (Kearns & Roth, 2019). Thus, ethical AI design is not merely a compliance obligation but a crucial safeguard against reputational and regulatory risks.

Proposition 3: Data governance based on AI reliability contributes to data accuracy. Reliable AI requires high-quality, consented data to prevent biases and inaccuracies (Ramaswamy et al., 2018). Effective AI governance relies on high-quality, reliable data, requiring rigorous verification of data sources, recognition of data limitations, and the establishment of clear data management protocols. For example, PwC partnered with H2O.ai to develop GL.ai[7], an AI-driven auditing bot that replicates the decision-making processes of human expert auditors, detecting financial anomalies and ensuring transaction accuracy. While AI-driven data analytics enhance decision-making, they also pose risks such as privacy breaches, fraud, and identity theft (Cohen, 2018; Martin & Murphy, 2017). Accurate AI outcomes depend on access to personal information, including online transaction records, yet the widespread availability of such data can have detrimental effects on individuals, businesses, and society (Bryson & Winfield, 2017). Growing distrust from data leaks further challenges businesses in data collection. Regulatory frameworks in many countries and regions, including GDPR and Japan's Act on the Protection of Personal Information (APPI)[8], mandate transparency and consumer control over personal data, reinforcing the need for robust AI governance (Ramaswamy et al., 2018). Organisations must institutionalise responsible practices that secure explicit user consent, ensure transparent data usage, and effectively communicate AI decision-making processes to enhance user trust. Prioritising AI reliability fosters consumer confidence, facilitates more accurate data collection, and ultimately improves overall AI performance.

[7] https://www.pwc.com/sk/en/current-press-releases/harnessing-the-power-of-ai-to-transform-the-detection-of-fraud-and-error.html
[8] https://www.ppc.go.jp/files/pdf/Act_on_the_Protection_of_Personal_Information.pdf

Proposition 4: Explainability in AI governance fosters user trust.

Providing clear, tailored explanations of AI decisions can reduce uncertainty and strengthen consumer trust (Rai, 2020; Ramaswamy et al., 2018). For instance, PwC's Responsible AI Toolkit exemplifies this by guiding businesses in adopting AI responsibly while offering personalised advisory services (PwC, 2019). Similarly, Alder Hey Children's Hospital leverages IBM Watson-powered analytics to provide explainable AI-driven healthcare services, improving patient experiences through interactive digital engagement and tailored health services (Alderheycharity, 2017). Features like virtual hospital tours, AI-based health tracking, real-time interactions with the virtual assistant 'Ask Oli', and augmented reality (AR)-powered interactive educational and entertainment tools on the Alder Play app (Ustwo, 2019) enhance healthcare delivery, promote transparency, and build trust in AI-driven medical services. Organisations that embed explainability into AI systems are better positioned to cultivate user confidence and reduce resistance to AI adoption, particularly in high-stakes domains such as healthcare.

To conclude, these propositions affirm that responsible AI is not just an ethical and regulatory obligation but a strategic driver of business value. Transparent data policies, ethical AI development, and strong governance frameworks enhance operational efficiency, mitigate risks, and foster long-term stakeholder trust. While the conceptual model offers a broad theoretical view of responsible AI business value, incorporating elements such as training and education, social acceptance and sustainability, and managerial and organisational benefits, these were less prominent in our dataset. In contrast, the four propositions emphasise the most empirically supported linkages shaping responsible AI's business impact. This

discrepancy highlights the need to prioritise robust, observable relationships while recognising that other elements may gain relevance as AI governance matures. By differentiating dominant from secondary linkages, we provide both confirmatory evidence for established theoretical perspectives and new insights into how responsible AI is operationalised in practice.

4.4 Validating results through applicability check

Following the methodological guidelines of Rosemann and Vessey (2008), an applicability check (i.e., importance, accessibility, and suitability) was conducted to evaluate the practical relevance of the research findings. Eight participants based in the United Kingdom (UK) with substantial expertise in AI were recruited from five professional AI-focused LinkedIn groups, and invitations were sent through LinkedIn's messaging platform. The final sample included five senior executives from different AI software companies and three experts with extensive experience in designing AI solutions. Table II.B1 in Appendix B provides detailed participant profiles.

Before the interviews, participants were provided with a structured summary of the conceptual model and research findings, including the key elements and linkages that emerged during the study. The discussions aimed to critically assess the importance, accessibility, and suitability of the findings within real-world AI implementation contexts. The interviews also sought to elicit participants' insights and professional perspectives on the practical challenges and opportunities of implementing responsible AI. The interviews were audio recorded, transcribed, and analysed, generating 37 pages of single-spaced text. This process ensured that

the research findings were not only theoretically robust but also validated in terms of their practical applicability. The diverse professional backgrounds of the participants added significant depth and critical perspectives, enabling a nuanced understanding of the business value potential of responsible AI practices in the real world.

Importance. All experts underscored the critical role of identifying responsible AI practices to support the implementation of AI ethics. The CEO of an AI start-up highlighted this significance, stating, "*To fully use AI ethically, this month we are recruiting a director who will be in charge of supervising AI developers and monitoring all AI projects running through the essential steps of AI ethics… It is a brand new position for our company…The recognition of responsible AI practice from your study would help us to understand what we can expect from this new role*". Another interviewee reinforced this view, noting, "*Responsible AI is imperative, and we need to care about it … As AI is everywhere, we need to get it right… The identification of responsible AI practice can be a good starting point for finding better ways to implement it*."

Accessibility. All participants agreed that the responsible AI practices were presented in a clear, visually appealing, and accessible format, offering actionable insights into how AI can be operated responsibly. Two AI solution designers observed, "*We actually need a clear and comprehensive picture on the development of ethical algorithms… It is easier to understand from your findings".* The majority of interviewees concluded*, "The evaluation of the importance associated with responsible AI practices [by frequency count] helps us prioritise the key areas to be implemented in our AI projects*."

Suitability. Experts expressed strong interest in the outcomes derived from the responsible AI model, agreeing that the identified practices and benefits are applicable and actionable for their AI implementations. A technical director of an AI software company commended the model, stating, "*Insights generated from this work would be very useful as it combines theory and practice specifically via identifying responsible AI practices directly from a set of successful AI implementation cases… I think it could further develop into a toolkit/handbook guiding us to deliver AI with ethical value.*" The sales director of another AI company echoed this sentiment, adding, "*It is challenging to convince our clients to emphasise on AI ethics since it comes with extra costs… The results of this study show the tremendous benefits when considering AI ethics. This would offer great momentum to our clients, and we could integrate this knowledge into our sales plan.*"

5. Conclusion

5.1 Theoretical contributions

Theoretically, our study contributes to the literature in three key areas. First, this research advances the theoretical discourse on responsible AI. By integrating CSR (Windsor, 2006) and IS business value (Shang & Seddon, 2002; Melville et al., 2004) perspectives through a PBV lens (Bromiley & Rau, 2014), we move beyond normative discussions of AI ethics to frame responsible AI as both an ethical and governmental imperative, and provide an empirically grounded framework that connects responsible AI practice with organisational business performance. Prior studies across different disciplines have primarily conceptualised responsible AI as a moral obligation via ethical principles (Jobin et al., 2019; Floridi et al., 2018). Yet,

our findings demonstrate its strategic relevance by conceptualising responsible AI through economic, ethical, and corporate citizenship dimensions (Windsor, 2006). thereby operationalising and extending CSR theory into the realm of AI governance while reinforcing both regulatory compliance and long-term value creation. This approach operationalises and extends CSR theory into the realm of AI governance, providing a structured framework that connects AI ethics with business strategy.

Moreover, applying the IS benefits framework (Shang & Seddon, 2002) offers a novel perspective on responsible AI business value, moving beyond abstract principles to practical, actionable implementation strategies and measurable business outcomes like enhanced operational efficiency, IT resilience, and stakeholder trust. By doing so, we reinforce both regulatory compliance and long-term business value creation, contributing to the theoretical maturation of responsible AI at its nascent stage. This demonstrates that responsible AI is not merely an ethical obligation but a strategic driver of competitive advantage, stakeholder trust, and sustainable innovation. Furthermore, our findings underscore the need for integrated AI design and governance approaches that align ethical responsibility with economic and societal sustainability.

Second, our study advances the understanding of IS business value by examining how responsible AI practices generate benefits through established IT value sources (ES benefit framework; Shang & Seddon, 2002). By applying this framework, we extend the IS business value discourse to responsible AI, integrating ethical and societal considerations with traditional performance metrics. We highlight that effective data governance (via explainability, transparency, and data accuracy) and ethically designed AI solutions (focusing on algorithmic fairness and biases

mitigation) enable AI-driven decision-making that fosters user trust, optimises product/service personalisation, and strengthens AI system robustness and data reliability. These, in turn, contribute to IT infrastructure resilience, operational efficiency, and strategic positioning, reinforcing responsible AI as both an ethical necessity and a business enabler.

At its core, responsible AI governance prioritises data protection and ethical information use, positioning IT infrastructure as the foundation for broader business value creation. This aligns with Newell and Marabelli's (2015) call to assess the societal impact of digital technology, demonstrating how responsible AI adoption benefits companies, consumers, and society. By unpacking the business value dimensions of responsible AI, our study extends IS business value research into this emerging domain, offering insights into how ethical AI adoption can serve as both a regulatory safeguard and a competitive advantage.

Third, despite growing interest in AI ethics, existing research remains fragmented, often limited to technical discussions, conceptual analyses, or single case studies that fail to capture responsible AI's broader business value across industries (Floridi et al., 2018; Mittelstadt, 2019). This study addresses this gap by systematically identifying responsible AI practices from multiple cases, providing an empirically grounded framework for their implementation. In doing so, we contribute to PBV (Bromiley & Rau, 2014) by demonstrating how responsible AI operates as a set of publicly known, imitable activities that are shaped by regulatory and societal expectations rather than a mere compliance requirement, and that can be transferred across organisations and refined over time. Unlike the RBV, which focuses on firm-specific resources, PBV better explains responsible AI's evolution

as an industry-driven practice. This aligns with SAP research, where organisations gain a competitive edge through effective execution of shared practices rather than exclusive resource ownership. Our applicability check (Rosemann & Vessey, 2008) through expert interviews validates the responsible AI business value model, ensuring its empirical grounding and practical relevance. Ultimately, this study advances responsible AI as a dynamic, practice-oriented concept that drives strategic business value.

5.2 Practical implications

Based on our case studies and expert interviews, we propose five strategic recommendations for embedding responsible AI into organisational practices while maximising business value. These strategies offer a structured approach to ethical AI implementation, aligning with regulations, stakeholder expectations, and corporate goals. Beyond risk mitigation, responsible AI can enhance efficiency, trust, and competitive advantage, driving long-term sustainability. Integrating ethical considerations into AI governance enables organisations to address emerging challenges while leveraging AI's potential for both economic and societal impact.

Strategy 1: Establishing AI ethics training and education for internal and external stakeholders. Our findings underscore the critical role of AI ethics training in fostering responsible AI practices. Effective training programmes equip managers and employees with the knowledge and skills needed to navigate ethical challenges in AI development and deployment. For instance, the IEEE's Global Initiative on

Ethics of Autonomous and Intelligent Systems[9] promotes ethical AI development through its Ethically Aligned Design (EAD) framework and P7000 standards. They offer certification programmes for AI designers, empowering them to prioritise ethical considerations in their work with training (Bryson & Winfield, 2017), and for AI systems that demonstrate adherence to ethical, human-centric principles such as transparency, accountability, and fairness and ensure that AI technologies align with organisational and societal values. Beyond technical training, organisations should adopt a multifaceted approach to AI ethics education, incorporating mentoring, cross-functional team workshops, and self-directed learning. Such initiatives not only cultivate an ethical AI culture but also embed ethical reasoning into decision-making processes. By investing in comprehensive AI ethics training, organisations can enhance internal governance, build stakeholder trust, and create sustainable business value through responsible AI implementation.

Google's Machine Learning Crash Course (MLCC)[10], another example, exemplifies AI ethics training by equipping university computer science faculties and professionals with foundational ML knowledge. To address AI bias, MLCC includes a fairness module available in 11 languages, ensuring global accessibility for staff training (Walker, 2018). Additionally, Google Cloud AutoML, leveraging techniques like 'learning2learn' and 'transfer learning' empowers users with limited ML expertise to develop high-quality predictive models (Li & Li, 2018). This initiative enhances productivity, fosters interdisciplinary collaboration, and promotes

[9] https://standards.ieee.org/industry-connections/ecpais/
[10] https://developers.google.com/machine-learning/crash-course

responsible AI deployment (Pichai, 2018), reinforcing the importance of sustainable AI education.

Strategy 2: Developing human-centric AI for risk management. Implementing responsible AI requires a structured risk management framework that spans the entire AI lifecycle, from design to deployment, and evaluation. Companies must proactively address key risks while investing in AI solutions, including security threats (e.g., cyber intrusions, vulnerabilities in open-source AI models), economic risks (e.g., job displacement), and performance risks (e.g., biases, inaccuracies, unintended consequences). To mitigate potential AI risks, organisations should establish clear, ethical governance frameworks, integrating predefined risk control protocols with measurable objectives and performance benchmarks.

A critical component of responsible AI is robust data governance, ensuring AI systems are trained on high-quality, unbiased datasets. Companies should systematically review both internally and externally sourced data to detect potential biases, inconsistencies, and security vulnerabilities. By implementing proactive data protocols—covering data collection, storage, processing, and application—organisations can enhance AI reliability while safeguarding against ethical and regulatory concerns. Beyond risk mitigation, human-AI collaboration is fundamental to ensuring AI augments, rather than replaces, human decision-making (Pavlou, 2018). A human-centric AI approach prioritises transparency, interpretability, and oversight, ensuring that AI remains aligned with human values and societal needs. Effectively managing AI risks not only enhances organisational resilience but also strengthens consumer and stakeholder trust, reinforcing the long-term business value of responsible AI practices.

Strategy 3: The Emergence of Chief Responsible AI Officers (CRAiO). As AI adoption accelerates, companies are increasingly expected to align AI deployment with CSR goals and ethical standards. Beyond its role in enhancing customer insights and operational efficiency, AI—when deployed responsibly—serves as a strategic asset that can enhance business reputation, foster consumer trust, and strengthen brand equity. However, a PwC survey[11] found that only 25% of about 250 surveyed companies took ethical concerns into account before investing in AI, underscoring a widespread immaturity in responsible AI practices.

To bridge this gap, organisations should establish the role of a Chief Responsible AI Officer (CRAiO)—a senior executive responsible for developing, overseeing, and implementing responsible AI strategies. The CRAiO would ensure that AI systems align with ethical guidelines, regulatory compliance, and stakeholder expectations, integrating responsible AI practices across business functions while fostering a culture of AI accountability. Given the cross-functional nature of AI governance, appointing a CRAiO requires an assessment of organisational resources, capabilities, and existing leadership structures to facilitate seamless collaboration across departments. Alternatively, companies can establish a multidisciplinary AI ethics advisory board to support executive leadership (Cobey & Boillet, 2018). Such a board can provide independent oversight, risk assessment, and strategic guidance, ensuring that AI governance aligns with both corporate and societal interests. Whether through a dedicated CRAiO or an advisory board, embedding

[11] https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai/responsible-ai-practical-guide.pdf

responsible AI leadership within organisations is essential for maximising AI's business value while mitigating ethical, legal, and reputational risks.

Strategy 4: Balancing economic and social sustainability in AI use. As AI adoption continues to grow rapidly, balancing economic benefits with social sustainability has become a critical priority for both academics and practitioners. While AI presents significant opportunities for business growth, efficiency, and competitive advantage, its deployment must also consider potential ethical, social, and environmental risks, including concerns over human rights, privacy, and employment displacement. Companies that fail to account for these broader societal impacts risk undermining product credibility, eroding consumer trust, and damaging brand reputation.

The ecological modernisation theory (Spaargaren & Mol, 1992) argues that sustainable development arises from synergising economic progress with social responsibility. In the context of AI, this means companies must develop AI-driven solutions that uphold ethical integrity while driving business value. This requires embedding responsible AI governance into corporate strategies, addressing ethical concerns throughout the AI lifecycle—from design to post-launch implementation, and integrating AI ethics into broader CSR initiatives. For example, Audi AG's Beyond AI Initiative[12] demonstrates a commitment to responsible AI by prioritising societal well-being alongside technological advancement. Their initiative tackles challenges such as public trust in autonomous driving and the future of work in the AI era, illustrating how AI innovation can align with long-term social sustainability

[12] https://www.audi-mediacenter.com/en/and-audi-initiative-8950

goals. By proactively addressing ethical concerns, businesses can enhance AI acceptance, build stakeholder confidence, and achieve sustainable profitability.

Strategy 5: Carrot-and-stick mechanism to regulate AI usage. The carrot-and-stick approach (i.e., rewards/incentives, punishments), widely applied IS research to regulate IT usage (Liang et al., 2013), provides a useful lens for managing AI ethics. Establishing mechanisms that encourage ethical AI behaviour while deterring misuse is essential for fostering responsible AI practices. Floridi et al. (2018) proposed several organisational strategies to promote ethical AI adoption. First, fostering cross-functional collaboration can facilitate discussions on AI's ethical and societal implications. For example, H&M introduced an Ethical AI Debate Club[13], enabling employees to explore hypothetical AI-related ethical dilemmas in the fashion industry. Such initiatives raise awareness and integrate ethical considerations into AI development and deployment. Second, developing a structured framework to align ethics, policies, and innovation ensures that AI-driven advancements are socially responsible. By embedding ethical governance into AI strategy, organisations can proactively mitigate risks while promoting AI-driven innovation. Third, deterrence mechanisms, including monitoring, auditing, and enforcement policies, play a crucial role in ensuring compliance. Implementing strict oversight and imposing consequences for unethical AI practices reinforces accountability and minimises AI-related harms. A combination of incentives and regulatory controls can drive ethical AI use, balancing organisational interests with broader societal responsibilities.

[13] https://hmgroup.com/our-stories/responsible-ai-is-better-ai/

5.3 Limitation and future research

The rapid advancement of AI presents opportunities for social good but also raises ethical risks and challenges. Despite increasing awareness, responsible AI adoption remains limited as organisations struggle to align AI implementation with both organisational and ethical, societal expectations. While this study advances AI and business ethics research by identifying responsible AI practices, business value, and recommendations, its limitations highlight areas for future exploration.

First, responsible AI remains an emerging concept, and its academic exploration is still in its early stages. Due to the limited availability of established research, much of the data in this study was drawn from publicly available documentation, which may introduce bias, as organisations often highlight successes while minimising challenges (Wang et al., 2018). Additionally, PBV suggests that a comprehensive understanding of responsible AI requires both qualitative and quantitative approaches (Bromiley & Rau, 2014, 2016). Future research should incorporate primary data from multiple stakeholders—including consumers, managers, and policymakers—to gain a more holistic perspective on responsible AI adoption at different levels.

Second, this study primarily focuses on responsible AI at the organisational level. While our findings underscore the role of consumer trust in AI adoption, a deeper examination of individual-level impacts is needed. Future research could explore how cognitive, emotional, and behavioural responses shape consumer acceptance of responsible AI. Experimental and survey-based studies could assess how issues such as algorithmic bias, privacy breaches, and AI-generated misinformation affect

consumer perceptions and engagement, further refining our understanding of responsible AI's societal implications.

Third, while this study presents a conceptual model linking CSR-grounded responsible AI practices to IS business value, further research is needed to examine contextual factors influencing responsible AI implementation. Specifically, future studies could investigate who within organisations drives responsible AI adoption, when these practices yield the most significant benefits, and what external forces—such as regulatory pressures, societal legitimacy, and market competitiveness—shape AI governance (Bromiley & Rau, 2016). Longitudinal case studies or empirical analyses could provide deeper insights into how responsible AI evolves over time and how companies integrate these practices into their strategic planning.

Fourth, our findings reflect a PBV perspective, which views responsible AI as a set of transferable, industry-wide practices shaped by regulatory and societal expectations. An RBV approach, however, would produce significantly different results, conceptualising responsible AI as a firm-specific capability that provides competitive advantage through proprietary AI governance mechanisms. Had we adopted RBV, our study would have focused on why some firms excel in responsible AI adoption while others lag, linking success to internal capabilities like AI talent acquisition and unique governance structures. PBV explores responsible AI's diffusion and institutionalisation, while RBV examines firms' development and protection of proprietary AI strategies. Thus, future research should bridge these perspectives, investigating how responsible AI fosters both industry-wide ethical norms and firm-level differentiation.

Finally, this study examines responsible AI within the current technological landscape, where AI primarily functions as an augmentative tool rather than a fully autonomous decision-maker. As AI systems gain greater autonomy, the nature of human-AI collaboration will shift, presenting new challenges in maintaining ethical oversight. Future research should investigate governance mechanisms that ensure responsible decision-making in increasingly autonomous AI systems. This includes exploring accountability frameworks, human-in-the-loop safeguards, and adaptive regulatory policies that balance efficiency with ethical responsibility. Empirical studies on real-world autonomous AI applications would provide valuable insights into sustaining ethical oversight as AI takes on greater decision-making authority.

Chapter II References

Ågerfalk, P. J. (2020). Artificial intelligence as digital agency. *European Journal of Information Systems, 29*(1), 1–8. https://doi.org/10.1080/0960085X.2020.1721947

Alderheycharity. (2017). *Download our brilliant new app now*. Alder Hey Children's Charity. https://www.alderheycharity.org/news/latest-news/the-alder-play-app-has-launched/

Arnold, M. (2018). *Factsheets for AI services*. IBM. https://www.ibm.com/blogs/research/2018/08/factsheets-ai/

Arnold, M., et al. (2019). FactSheets: Increasing trust in AI services through Supplier's Declarations of Conformity. *arXiv preprint.* https://doi.org/10.48550/arXiv.1808.07261

Arrieta, A. B., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Barney, J. B. (1991). Firm resources and sustained competitive advantage. *Journal of Management, 17*(1), 99–120. https://doi.org/10.1177/014920639101700108

Barney, J. B. (2001). Is the resource-based "view" a useful perspective for strategic management research? Yes. *Academy of Management Review, 26*(1), 41–56. https://doi.org/10.2307/259393

Baskentli, S., et al. (2019). Consumer reactions to corporate social responsibility: The role of CSR domains. *Journal of Business Research, 95*, 502–513. https://doi.org/10.1016/j.jbusres.2018.07.046

Bernal-Conesa, J. A., de Nieves Nieto, C., & Briones-Peñalver, A. J. (2017). CSR strategy in technology companies: Its influence on performance, competitiveness and sustainability. *Corporate Social Responsibility and Environmental Management, 26*, 96–107. https://doi.org/10.1002/csr.1393

Bharadwaj, A., El Sawy, O. A., Pavlou, P. A., & Venkatraman, N. (2013). Digital Business Strategy: Toward a Next Generation of Insights. *MIS Quarterly, 37*(2), 471–482. https://doi.org/10.25300/MISQ/2013/37:2.3

Bocquet, R., et al. (2013). Are firms with different CSR profiles equally innovative? Empirical analysis with survey data. *European Management Journal, 31*, 642–654. https://doi.org/10.1016/j.emj.2012.07.001

Bromiley, P., & Rau, D. (2014). Towards a practice-based view of strategy. *Strategic Management Journal, 35*(8), 1249–1256. https://doi.org/10.1002/smj.2238

Bromiley, P., & Rau, D. (2016). Operations management and the resource-based view: Another view. *Journal of Operations Management, 41*, 95–106. https://doi.org/10.1016/j.jom.2015.11.003

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science, 358*(6370), 1530–1534. https://doi.org/10.1126/science.aap8062

Bryson, J., & Winfield, A. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. *Computer,* 116–119. https://doi.org/10.1109/MC.2017.154

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165–1188. https://doi.org/10.2307/41703503

Cobey, C., & Boillet, J. (2018). *How do you teach AI the value of trust?* Ernst & Young. https://www.ey.com/en_uk/digital/how-do-you-teach-ai-the-value-of-trust

Cohen, M. C. (2018). Big data and service operations. *Production and Operations Management, 27*(9), 1709–1723. https://doi.org/10.1111/poms.12832

Culnan, M. J., & Williams, C. C. (2009). How ethics can enhance organizational privacy: Lessons from the ChoicePoint and TJX data breaches. *MIS Quarterly, 33*(4), 673–687. https://doi.org/10.2307/20650322

Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 4691–4697.

Davenport, T., et al. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science, 48*, 24–52. https://doi.org/10.1007/s11747-019-00696-0

Devaraj, S., & Kohli, R. (2003). Performance impacts of information technology: Is actual usage the missing link? *Management Science, 49*(3), 273–289. https://doi.org/10.1287/mnsc.49.3.273.12736

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way.* Springer (Artificial Intelligence: Foundations, Theory, and Algorithms). https://doi.org/10.1007/978-3-030-30371-6

DMA. (2018). *GDPR: A consumer perspective.* Data & Marketing Association. https://dma.org.uk/uploads/misc/5af5497c03984-gdpr-consumer-perspective-2018-v1_5af5497c038ea.pdf

Donaldson, T. & Dunfee, T.W. (1994). Toward a unified conception of business ethics: Integrative social contracts theory. *Academy of management review*, 19(2), 252-284. https://doi.org/10.5465/amr.1994.9410210749

Donaldson, T. & Preston, L.E. (1995). The stakeholder theory of the corporation: Concepts, evidence, and implications. *Academy of management Review*, 20(1), 65-91. https://doi.org/10.5465/amr.1995.9503271992

Dwivedi, Y. K., et al. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice, and policy. *International Journal of Information Management, 57*, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Dwivedi, Y. K., et al. (2023). Opinion paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges, and implications of generative conversational AI for research, practice, and policy. *International Journal of Information Management, 71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Eisenhardt, K. M., & Graebner, M. E. (2007). Theory building from cases: Opportunities and challenges. *The Academy of Management Journal, 50*(1), 25–32. https://doi.org/10.5465/amj.2007.24160888

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing, 62*(1), 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Farmaki, A., et al. (2022). Hotel CSR and job satisfaction: A chaordic perspective. *Tourism Management, 91*, 104526. https://doi.org/10.1016/j.tourman.2022.104526

de Fine Licht, K., & de Fine Licht, J. (2020). Artificial intelligence, transparency, and public decision-making. *AI & Society, 35*, 917–926. https://doi.org/10.1007/s00146-020-00960-w

Floridi, L., et al. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*, 689–707. https://doi.org/10.1007/s11023-018-9482-5

Flyverbom, M., Deibert, R., & Matten, D. (2019). The governance of digital technology, big data, and the internet: New roles and responsibilities for business. *Business & Society, 58*(1), 3–19. https://doi.org/10.1177/0007650317727540

Gerlach, J. P., et al. (2019). Flamingos on a slackline: Companies' challenges of balancing the competing demands of handling customer information and privacy. *Information Systems Journal, 29*(2), 548–575. https://doi.org/10.1111/isj.12222

Gershgorn, D. (2018). *Google created a 'responsible innovation team' to check if its AI is ethical*. Quartz. https://qz.com/1501998/google-created-a-responsible-innovation-team-to-check-if-its-ai-is-ethical/

Gregor, S., Martin, M., Fernandez, W., Stern, S., & Vitale, M. (2006). The transformational dimension in the realization of business value from information technology. *The Journal of Strategic Information Systems, 15*(3), 249–270. https://doi.org/10.1016/j.jsis.2006.04.001

Hernán, M.A., Hernández-Díaz, S., & Robins, J.M. (2004). A structural approach to selection bias. *Epidemiology, 15*(5),615-625. https://doi.org/10.1097/01.ede.0000135174.63482.43

Huang, J., et al. (2014). Site-shifting as the source of ambidexterity: Empirical insights from the field of ticketing. *The Journal of Strategic Information Systems, 23*(1), 29–44. https://doi.org/10.1016/j.jsis.2014.01.001

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

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

Kearns, M., & Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design.* New York: Oxford University Press.

Knight, W. (2017). *The financial world wants to open AI's black boxes*. MIT Technology Review. https://www.technologyreview.com/2017/04/13/152590/the-financial-world-wants-to-open-ais-black-boxes/

Kohli, R., & Grover, V. (2008). Business value of IT: An essay on expanding research directions to keep up with the times. *Journal of the Association for Information Systems, 9*(1), 23–39. https://doi.org/10.17705/1jais.00147

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology.* Sage Publications.

Langdon, C. S. (2006). Designing information systems capabilities to create business value: A theoretical conceptualization of the role of flexibility and integration. *Journal of Database Management, 17*(3), 1–18. https://doi.org/10.4018/978-1-60566-058-5.ch049

Li, F. F., & Li, J. (2018). *Cloud AutoML: Making AI accessible to every business.* Google Cloud. https://cloud.google.com/blog/topics/inside-google-cloud/cloud-automl-making-ai-accessible-every-business

Liang, H., Xue, Y., & Wu, L. (2013). Ensuring employees' IT compliance: Carrot or stick? *Information Systems Research, 24*(2), 279–294. https://doi.org/10.1287/isre.1120.0427

Margolis, J. D., & Walsh, J. P. (2003). Misery loves companies: Rethinking social initiatives by business. *Administrative Science Quarterly, 48*(2), 268–305. https://doi.org/10.2307/3556659

Marshall, C., & Rossman, G. B. (2014). *Designing qualitative research.* Sage Publications.

Martin, K. D., & Murphy, P. E. (2017). The role of data privacy in marketing. *Journal of the Academy of Marketing Science, 45*, 135–155. https://doi.org/10.1007/s11747-016-0495-4

Martínez-López, F. J., & Casillas, J. (2013). Artificial intelligence-based systems applied in industrial marketing: An historical overview, current and future insights. *Industrial Marketing Management, 42*(4), 489–495. https://doi.org/10.1016/j.indmarman.2013.03.001

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR), 54*(6),1-35. https://doi.org/10.1145/3457607

Melville, N., Kraemer, K., & Gurbaxani, V. (2004). Information technology and organizational performance: An integrative model of IT business value. *MIS Quarterly, 28*(2), 283–322. https://doi.org/10.2307/25148636

Mikalef, P., & Krogstie, J. (2020). Examining the interplay between big data analytics and contextual factors in driving process innovation capabilities. *European Journal of Information Systems, 29*(3), 260–287. https://doi.org/10.1080/0960085X.2020.1740618

Minichiello, V., et al. (1990). *In-depth interviewing: Researching people.* Melbourne, AU: Longman Cheshire.

Mithas, S., Ramasubbu, N., & Sambamurthy, V. (2011). How Information Management Capability Influences Firm Performance. *MIS Quarterly, 35*(1), 237–256. https://doi.org/10.2307/23043496

Moon, J., Crane, A., & Matten, D. (2005). Can corporations be citizens? Corporate citizenship as a metaphor for business participation in society. *Business Ethics Quarterly, 15*(3), 429–453. https://doi.org/10.5840/beq200515329

Morley, J., et al. (2020). From what to how: An initial review of publicly available AI ethics tools, methods, and research to translate principles into practices. *Science and Engineering Ethics, 26*, 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Mueller, B., et al. (2010). Understanding the economic potential of service-oriented architecture. *Journal of Management Information Systems, 26*(4), 145–180. https://doi.org/10.2753/MIS0742-1222260406

Nevo, S., & Wade, M. R. (2010). The formation and value of IT-enabled resources: Antecedents and consequences of synergistic relationships. *MIS Quarterly, 34*(1), 163–183. https://doi.org/10.2307/20721419

Newell, S., & Marabelli, M. (2020). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of "datification." *The Journal of Strategic Information Systems, 24*(1), 3-14. https://doi.org/10.1016/j.jsis.2015.02.001

NVIDIA. (2019). *2019 NVIDIA corporate social responsibility report.* https://www.nvidia.com/content/dam/en-zz/Solutions/documents/FY2019-NVIDIA-CSR-Social-Responsibility.pdf

Öberseder, M., Schlegelmilch, B. B., & Murphy, P. E. (2013). CSR practices and consumer perceptions. *Journal of Business Research, 66*(10), 1839–1851. https://doi.org/10.1016/j.jbusres.2013.02.005

O'Leary, D. E. (2013). Artificial intelligence and big data. *IEEE Intelligent Systems, 28*(2), 96–99. https://doi.org/10.1109/MIS.2013.39

Osborne, S. (2023). *Elon Musk and experts say AI development should be paused immediately.* Sky News. https://news.sky.com/story/elon-musk-and-others-sign-open-letter-calling-for-pause-on-ai-development

Osburg, T., & Lohrmann, C. (2017). *Sustainability in a digital world: New opportunities through new technologies.* Springer (CSR, Sustainability, Ethics & Governance (CSEG)). https://doi.org/10.1007/978-3-319-54603-2

Pavlou, P. A. (2018). Internet of things—Will humans be replaced or augmented? *NIM Marketing Intelligence Review, 10*(2), 42–47. https://doi.org/10.2478/gfkmir-2018-0017

Peloza, J., & Shang, J. (2011). How can corporate social responsibility activities create value for stakeholders? A systematic review. *Journal of the Academy of Marketing Science, 39*, 117–135. https://doi.org/10.1007/s11747-010-0213-6

Peters, D., et al. (2020). Responsible AI—Two frameworks for ethical design practice. *IEEE Transactions on Technology and Society, 1*(1), 34–47. https://doi.org/10.17863/CAM.49394

Pichai, S. (2018). *AI at Google: Our principles.* Google. https://blog.google/technology/ai/ai-principles/

PwC. (2019). *A practical guide to responsible artificial intelligence (AI).* https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai/responsible-ai-practical-guide.pdf

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science, 48*, 137–141. https://doi.org/10.1007/s11747-019-00710-5

Rai, A., Constantinides, P., & Sarker, S. (2019). Next-generation digital platforms: Toward Human–AI hybrids. *MIS Quarterly, 43*(1), iii–ix.

Ramaswamy, P., Jeude, J., & Smith, J. A. (2018). *Making AI responsible and effective.* Cognizant. https://www.thecognizant.com/site/assets/files/1940/making-ai-responsible-and-effective-codex3974.pdf

Rosemann, M., & Vessey, I. (2008). Toward improving the relevance of information systems research to practice: The role of applicability checks. *MIS Quarterly, 32*(1), 1–22. https://doi.org/10.2307/25148826

Russell, S., et al. (2015). Robotics: Ethics of artificial intelligence. *Nature, 521*(7553), 415–418. https://doi.org/10.1038/521415a

Russell, S., & Norvig, P. (2016). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Russo-Spena, T., & Mele, C. (2012). "Five Co-s" in innovating: A practice-based view. *Journal of Service Management, 23*(4), 527–553. https://doi.org/10.1108/09564231211260404

Rzepka, C., & Berger, B. (2018). User interaction with AI-enabled systems: A systematic review of IS research. *39th International Conference on Information Systems.* International Conference on Information Systems, San Francisco.

Schiffer, Z., & Newton, C. (2023). *Microsoft lays off team that taught employees how to make AI tools responsibly.* The Verge. https://www.theverge.com/2023/3/13/23638823/microsoft-ethics-society-team-responsible-ai-layoffs

Schilling, J. (2006). On the pragmatics of qualitative assessment. *European Journal of Psychological Assessment, 22*(1), 28–37. https://doi.org/10.1027/1015-5759.22.1.28

Schryen, G. (2013). Revisiting IS business value research: What we already know, what we still need to know, and how we can get there. *European Journal of Information Systems, 22*(2), 139–169. https://doi.org/10.1057/ejis.2012.45

Sen, S., & Bhattacharya, C. B. (2001). Does doing good always lead to doing better? Consumer reactions to corporate social responsibility. *Journal of Marketing Research, 38*(2), 225–243. https://doi.org/10.1509/jmkr.38.2.225.1883

Shang, S., & Seddon, P. B. (2002). Assessing and managing the benefits of enterprise systems: The business manager's perspective. *Information Systems Journal, 12*(4), 271–299. https://doi.org/10.1046/j.1365-2575.2002.00132.x

Shneiderman, B. (2021). Responsible AI: Bridging from ethics to practice. *Communications of the ACM,* 32–35.

Someh, I., et al. (2019). Ethical issues in big data analytics: A stakeholder perspective. *Communications of the Association for Information Systems, 44*, 718–747. https://doi.org/10.17705/1CAIS.04434

Spaargaren, G., & Mol, A. P. J. (1992). Sociology, environment, and modernity: Ecological modernization as a theory of social change. *Society & Natural Resources, 5*(4), 323–344. https://doi.org/10.1080/08941929209380797

Strauss, A., & Corbin, J. M. (1990). *Basics of qualitative research: Grounded theory procedures and techniques.* Thousand Oaks, CA: Sage.

Stokel-Walker, C., & Van Noorden, R. (2023). The promise and peril of generative AI. *Nature, 614,* 214–216. https://doi.org/10.1038/d41586-023-00340-6

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science, 361*(6404), 751–752. https://doi.org/10.1126/science.aat5991

Tanriverdi, H. (2006). Performance effects of information technology synergies in multibusiness companies. *MIS Quarterly, 30*(1), 57–77. https://doi.org/10.2307/25148717

Torresen, J. (2018). A review of future and ethical perspectives of robotics and AI. *Frontiers in Robotics and AI, 4*(75), 1–10. https://doi.org/10.3389/frobt.2017.00075

Urquhart, C., Lehmann, H., & Myers, M. D. (2010). Putting the "theory" back into grounded theory: Guidelines for grounded theory studies in information systems. *Information Systems Journal, 20*(4), 357–381. https://doi.org/10.1111/j.1365-2575.2009.00328.x

Ustwo. (2019). *Alder Play: Revolutionising patient care for children and their families.* https://www.ustwo.com/work/alder-play

Vaara, E. & Whittington, R. (2012). Strategy-as-practice: Taking social practices seriously. *Academy of Management Annals, 6*(1), 285-336. https://doi.org/10.5465/19416520.2012.672039

Vakkuri, V., Kemell, K. K., & Abrahamsson, P. (2019). Implementing ethics in AI: Initial results of an industrial multiple case study. In *Product-Focused Software Process Improvement. PROFES 2019*, 331–338.

Van Marrewijk, M. (2003). Concepts and definitions of CSR and corporate sustainability: Between agency and communion. *Journal of Business Ethics, 44*, 95–105. https://doi.org/10.1023/A:1023331212247

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine, 15*(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689

Walker, K. (2018). *Google AI Principles updates, six months in.* Google. https://www.blog.google/technology/ai/google-ai-principles-updates-six-months/

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong.* New York: Oxford University Press.

Walsham, G. (1995). Interpretive case studies in IS research: Nature and method. *European Journal of Information Systems, 4*(2), 74–81. https://doi.org/10.1057/ejis.1995.9

Walsham, G. (2006). Doing interpretive research. *European Journal of Information Systems, 15*(3), 320–330. https://doi.org/10.1057/palgrave.ejis.3000589

Wang, Y., et al. (2019). Leveraging big data analytics to improve quality of care in healthcare organizations: A configurational perspective. *British Journal of Management, 30*(2), 362–388. https://doi.org/10.1111/1467-8551.12332

Wang, Y., & Hajli, N. (2017). Exploring the path to big data analytics success in healthcare. *Journal of Business Research, 70*, 287–299. https://doi.org/10.1016/j.jbusres.2016.08.002

Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change, 126*, 3–13. https://doi.org/10.1016/j.techfore.2015.12.019

Wearn, O. R., Freeman, R., & Jacoby, D. M. P. (2019). Responsible AI for conservation. *Nature Machine Intelligence, 1*, 72–73. https://doi.org/10.1038/s42256-019-0022-7

Weber, M. (2008). The business case for corporate social responsibility: A company-level measurement approach for CSR. *European Management Journal, 26*(4), 247–261. https://doi.org/10.1016/j.emj.2008.01.006

Whittlestone, J., et al. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. *2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES),* 195–200.

Wiesche, M., et al. (2017). Grounded theory methodology in information systems research. *MIS Quarterly, 41*(3), 685–702. https://doi.org/10.25300/MISQ/2017/41.3.02

Windsor, D. (2006). Corporate social responsibility: Three key approaches. *Journal of Management Studies, 43*(1), 93–114. https://doi.org/10.1111/j.1467-6486.2006.00584.x

Wright, S. A., & Schultz, A. E. (2018). The rising tide of artificial intelligence and business automation: Developing an ethical framework. *Business Horizons, 61*(6), 823–832. https://doi.org/10.1016/j.bushor.2018.07.001

Xu, H., et al. (2011). Information privacy concerns: Linking individual perceptions with institutional privacy assurances. *Journal of the Association for Information Systems, 12*(12). https://doi.org/10.17705/1jais.00281

Yin, R. K. (2014). *Case study research: Design and methods* (2nd ed.). Sage Publications.

Chapter II Appendix A: Coding Examples

Table II.A1 Coding examples

Statement A	Open (underlined) and axial (*italic*) coding	Linkages confirmed by selective coding
IESO Digital Health (Case #9) "To protect the privacy and autonomy of users, make transparent the use of mental health data and ensure secure storage. Online mental health therapy applications that collect, store, and make use of personal data raise several important concerns around privacy, which in turn can pose risks to user autonomy. In particular, because of the kind of personal data that is now available to be collected (e.g., biometrics, location, and online behaviour) combined with advances in machine learning that make it possible to infer personal attributes from collected data companies are increasingly able to tailor messages and services to specific individuals or groups. This means that the more personal information a company has about someone, the more effectively they can target interventions in an attempt to influence them which may present new risks to patient autonomy." (Peters *et al.*, 2020, p.40)	Economic conception of responsible AI, Data governance *Transparency – The use of AI is transparent to individuals and other external constituencies and communities* Potential benefits, Operational benefits *Optimise personalised customer service*	• Comparing this statement to other similar codes (i.e., transparency and operational benefit) • A discrepancy on operational benefit occurred between two coders. The first coder coded it as optimising customer service in the first place, while the second coder argued that transparency helps IESO digital health tailor messages which could transform customer engagement and coded it as transforming customer engagement in a way that creates a dialogic environment • After discussion and debate, both coders agreed on transparency of AI as a specific responsible AI practice can lead to personalisation of customer service. • Recoding this statement as one of the linkages: Transparency → operational benefits

Table II.A2 Coding examples (continued)

Statement B	Open (underlined) and axial (*italic*) coding	Linkages confirmed by selective coding
NVIDIA Corporation (Case #10) "We share the widespread concerns about bias and maintaining ethical practices in AI, and they help guide the work of our data scientists. We host seminars at our GTCs around the world to train other engineers in upholding best practices. In our engagements with governments, we emphasize the importance of supporting positive uses of AI while considering issues around its development. Deployed properly, AI can remove bias in areas that people have struggled to address for decades. We hold ourselves to the highest standards in our research and support and encourage developers to work with the most accurate data sets to improve the algorithms in their work." (NVIDIA, 2019, p.39)	Ethical conception of responsible AI, Ethically designed solutions *Develop ethical algorithms by collaborating with engineers, scientists, universities, and other stakeholders* Potential benefits, IT infrastructure benefits *Mitigate biases and errors in data and algorithms*	• Comparing this statement to other similar codes (i.e., ethically design solutions and IT infrastructure benefits) • Both coders agreed that designing ethical AI solutions by collaborating with wider stakeholders can help NVIDIA mitigate biases in their algorithms • Recoding this statement as one of the linkages: Ethically design solutions → IT infrastructure benefits

Chapter II Appendix B: List of Interviews

Table II.B1 The profiles of the interviewees

Interviewees	Positions	Key duties in their organisations
Interviewee #1	CEO of an AI start-up company	Set up strategic direction and business strategy and manage team to scale all aspects of the company
Interviewee #2	AI software sales director	Execute and monitor the consultative sales plans and lead sales team to explore market opportunities
Interviewee #3	Technical director	Take responsibility for the maintenance and management of AI projects for their clients
Interviewee #4	Senior consultant	Undertakes consulting activities regarding AI development
Interviewee #5	Senior consultant	
Interviewee #6	Business analyst	Harness data and perform analysis to support business strategies
Interviewee #7	AI solution designer/architect	Design AI solutions specifically for their customer needs
Interviewee #8	AI solution designer/architect	

Chapter III

“Real-ising” the Benefits of Responsible Artificial Intelligence (AI) Management: An Interpretive Study

Abstract

Organisational decision-makers need to manage the ever-evolving frontiers (e.g., autonomy, learning, inscrutability) of artificial intelligence (AI) responsibly to navigate ethics-related challenges and realise organisational goals. Yet, much of the extant research on AI appears siloed and spans multiple disciplines, making it challenging to learn about management practices at the frontiers of AI that can support complex decision-making. The evolving expectations for managing AI necessitate theoretical advancements in understanding well-developed, evidence-based management practices. Our research approaches this challenge by situating responsible AI management practices within the design and governance of AI through a sociotechnical lens. Drawing on qualitative data (i.e., interviews, archival materials) and inspiration from diverse research, we develop an empirically grounded model for managing AI through an interpretive study. The model 1) delineates multifaceted responsibility (i.e., evidence, epistemic, outcome) related to AI design and identifies associated design tactics for each dimension; 2) specifies AI governance mechanisms (i.e., structural, procedural, relational) for enacting responsible AI management practices; and 3) identifies organisational performance outcomes (i.e., instrumental, humanistic) of management practices. Our findings contribute to AI ethics, governance, and management literature, offering researchers and practitioners an empirical exposition of managing AI with actionable guidance. By turning to theory for a guide, researchers can approach managerial issues concerning the future frontiers of AI with theoretical foundations, and practitioners can plan initiatives that effectively manage emerging ethics challenges.

Keywords: responsible artificial intelligence (AI), managing AI, responsible AI design, responsible AI governance, AI ethics, sociotechnical perspective

1. Introduction

Artificial intelligence (AI) has recently gained significant momentum (Dwivedi et al., 2021; Kopalle et al., 2022) in its integration into organisations (Asatiani et al., 2021; Keding, 2021). Unlike 'intelligent' information technology (IT) artifacts in the past, AI is more of an evolving frontier of emerging computational advancements (Berente et al., 2021) that has greater autonomy, deeper learning capability, and increased inscrutability (Baird & Maruping, 2021). Current AI technologies are capable of emulating human cognition and communication, presenting great possibilities for businesses to invent business models, devise product/service offerings, and reshape the nature of work (Davenport et al., 2020). Consequently, organisations across various fields have rapidly adopted AI to automate operations, support decision-making, enable dynamic interactions, and generate content, resulting in increased efficiency, consistency, and personalisation (Asatiani et al., 2021; Huang & Rust, 2021; 2023).

Given the rapidly advancing frontiers (e.g., autonomy, learning, and inscrutability) of AI, its applications are growing at a remarkable speed across diverse problem domains (Berente et al., 2021). However, these advancements also bring complex challenges related to data privacy and security, ethics and human rights, human-AI interfaces, and business strategies (Benbya et al., 2021). This has fuelled ongoing conceptual debates, resulting in calls for responsible AI and content-wise similar expressions such as trustworthy AI, which emphasise the critical need to manage AI in line with processes and rules that prioritise ethics and ensure societal benefits (Mikalef et al., 2022).

An unprecedented wave of regulatory initiatives has emerged, from organisation best-practice advocacy to national and international policymaking (Schiff et al., 2021), to reflect the understandable tensions between the urgency to implement AI and the need for it to be done so responsibly (Mittelstadt, 2019; Wang et al., 2020). They are working to determine where and how AI fits into existing data, technology regulations and policies, and human rights law (Schiff et al., 2021), and updating and creating corresponding rules for AI.

While scholarly reviews reveal prevailing thematic trends such as accountability and fairness from principled documents (Fjeld et al., 2020; Jobin et al., 2019), they often focus on societal and regulatory perspectives. Many organisations still struggle to translate abstract principles into practical applications (Mäntymäki et al., 2022) and identify effective AI governance mechanisms that yield desirable outcomes (Dennehy et al., 2023). Moreover, it remains unclear whether current regulatory instruments are sufficient for AI management (Mittelstadt, 2019), as the number of AI systems exhibiting ethical pitfalls continues to grow with the rapid algorithmic development (Benbya et al., 2020), often outpacing regulation development (Berente et al., 2021; Schneider et al., 2023).

Consequently, regulators and supervisory agencies tend to react to AI risks rather than anticipate them (Butler et al., 2023). This reactive approach, combined with business board oversight lags toward AI implementation, neglects AI's negative affordances on organisational stakeholders and society (Mikalef et al., 2022). The concept of AI governance has emerged (Gahnberg, 2021), but it has yet to explicitly define practices for managing AI within organisations (Schneider et al., 2023). Organisations struggling with AI oversight can benefit from understanding

potential regulatory mechanisms and actions institutionalised at organisational and policy levels (Mäntymäki et al., 2022) that define how to attain the needs dictated by responsible AI principles.

Managing AI involves more than the ticking of ethical 'boxes'; it requires addressing managerial uncertainties with theoretical advancements that go beyond principled approaches. Recent research has explored design practices to promote AI responsibility (e.g., Madaio et al., 2022; Rakova et al., 2021). Still, AI designers face challenges stemming from bias influenced by their personal perspectives and experiences, despite acknowledging the importance of the intervention of training datasets (Denton et al., 2021) and algorithmic mitigations (Agarwal et al., 2018). While publicly available toolkits have been developed for additional guidance and support (Morley et al., 2020), they fall short of explicitly explaining how to implement the recommendations (Deng et al., 2022) and are primarily designed for specific types of technical work (Wang et al., 2023). Thus, research needs to work towards a framework that supports designers in their 'responsible work' during the design of AI systems, ensuring that both technological and social elements are adequately addressed.

With its sociotechnical roots, the Information Systems (IS) field "considers both the technical artifacts and the individuals and collectives that develop and use these artifacts in a social context" (Sarker et al., 2019, p.696). This perspective has brought greater prominence to managing AI within the IS discourse (Ågerfalk, 2020; Berente et al., 2019), which involves communicating, leading, coordinating, and controlling the frontiers of AI to address complex decision-making problems. Despite growing interest in investigating and addressing challenges associated with managing AI

(e.g., Lebovitz et al., 2021; Teodorescu et al., 2021), most efforts target a different set of critical challenges. While substantial knowledge on managing IT exists in the IS field (Jarvenpaa & Ives, 1991), adapting this knowledge to AI for driving positive organisational results requires empirical and theoretical work. Therefore, IS research is called to aid managers in decision-making on AI-related activities by providing them well-developed, evidence-based practices for managing AI in organisations.

Against the aforementioned background and motivated by the practical need for 'how' other than 'what', this research aims to empirically explore and work towards theoretical approaches to managing AI in organisations through a sociotechnical lens. It seeks to answer: what does managing AI entail within organisations, and what organisational performance outcomes arise? A qualitative, interpretative study was conducted to investigate and conceptualise the design practices, governance mechanisms, and organisational outcomes of managing AI. The study developed a conceptual model that integrates 1) actionable AI design tactics concerning multifaceted responsibility; 2) governance mechanisms facilitating the management of AI design; and 3) organisational performance outcomes. This study heeds the calls for IS research to adopt a more critical and holistic approach to the current debate about responsible AI (Mikalef et al., 2022) and to study the phenomenon of managing AI through sociotechnical thinking (Berente et al., 2019).

This study provides several contributions to research and practice, paving the way for responsible AI management in organisations. Theoretically, it enhances clarity in the literature concerning AI ethics, AI governance, responsible AI, and managing AI by identifying core elements—design tactics and governance mechanisms—of

managing AI from a sociotechnical perspective. It uncovers relevant responsibility dimensions, associated design tactics, governance mechanisms, and organisational outcomes on AI management practices, contributing to the initial theorising on managing AI. Practically, this study offers some actionable design and governance practices that organisations can implement to improve AI management. The model has a checklist character that may offer some granular ideas for organisational decision-makers on what to do to responsibly manage AI and how to do it. Also, the findings could serve as a means to problematise ongoing research discourse and real-world practices.

Our paper proceeds as follows. First, it starts with an overview of managing AI, which is followed by a discussion of literature on the responsible design and governance of AI. The next section introduces the research method, data collection, and data analysis procedures. For clarity, we iterated between initial findings and literature during the empirical analysis. Third, it presents the findings by providing detailed descriptions of and relationships between theoretical concepts with supporting quotes. Finally, the conclusion spells out the theoretical and practical implications arising from this research, and discusses limitations and future research directions.

2. Conceptual Background

2.1 Sociotechnical approach to AI management

AI has recently captured substantial interest in organisational research (e.g., Asatiani et al., 2021, Keding, 2021, Marabelli et al., 2021), as it rashapes various facets of the business landscape, with its transformative potential touching almost

every sector (Davenport et al., 2020; Kopalle et al., 2022). Indeed, organisations undergo significant changes when using AI to transform their business offerings, processes, models, and the nature of work with unprecedented efficiency, effectiveness, and scale. However, AI also introduces tensions related to business capabilities like decision-making (Benbya et al., 2021). Given the autonomy, learning, and inscrutability facets of AI systems (Asatiani et al., 2021; Berente et al., 2021), foremost among those challenges is the tension surrounding decision accountability in organisational settings that touches technical, ethical, and societal aspects (Marabelli et al., 2021). Consequently, some are calling for organisations to responsibly manage their use of AI and bear greater accountability for the consequences that arise from it in many circumstances (Berente et al., 2021; Martin, 2019).

Despite the rise of public-private initiatives globally to define frameworks for the ethical and responsible implementation of AI, they are mainly from a top-down principled approach that may not effectively help organisations guarantee the ethical and responsible development and deployment of AI (Mittelstadt, 2019). AI is continuously presenting managerial challenges due to the evolution of its performance ('the ever-improving execution of tasks to which AI is applied') and scope ('the ever-expanding range of contexts to which AI is applied') brought by its expanding frontiers in complex decision making (Benbya et al., 2020; Berente et al., 2021, p.1438). The real work of managing AI in organisations, including communication, leadership, coordination, and control of AI (Berente et al., 2021), should be undertaken from a more bottom-up approach in time and over time.

Although there is little direct reference to managing AI per se (for exceptions, see Berente et al., 2021), we found recent research has shifted towards a more comprehensive perspective on responsible AI, moving beyond isolated factors like elimination of bias and explainability of outcomes (Mikalef et al., 2022). A nuanced and integrated understanding of responsible AI allows a more proactive rather than reactive approach in the endeavours to offset AI challenges (Buhmann and Fieseler, 2021) and balance the use of AI with responsibility in organisations (Buhmann and Fieseler, 2023; Senoner et al., 2022). These efforts also stem from a deduction of intentions behind managing AI through responsible approaches.

Key stakeholders in AI design and governance need to reflect thoughtfully and actively shape organisational efforts related to AI activities. They need to be informed and understand how to make decisions with and about AI. Thus, it is suggested to pursue ethics as a process (Mittelstadt, 2019) and propose definitions and explanations for responsible design and governance expertise from sociotechnical thinking (Sarker et al., 2019). IS research, with its sociotechnical tradition, can approach the phenomenon of managing AI by considering both social and technical factors (Berente et al., 2021). IS research has been called to aid managers in their decision-making by providing well-developed, evidence-based practices for managing AI in organisations.

2.2 Responsible AI design

Responsible AI[14] is well-discussed in the literature, covering aspects from its underlying technology and applications to stakeholders (e.g., developers, users, regulators, legislators), and social context (Stahl, 2023). It has been inspired by the seminal work from Wiener (1954) on the *automatic age* as well as the ethical and societal implications of AI artifacts. Following a long discussion of the responsibility concept in social sciences and moral philosophy (Fischer, 1999), there is a growing consensus on the notion of responsible AI, that is, the practice of implementing AI with good intention from ethical, societal, and legal points of view to empower individuals and businesses, and to have a fair impact on societies. However, it is still lacking conceptual weight within academic research.

While responsibility is often viewed as a relational concept (Paul et al., 1999), described as linking a subject (who is responsible) to an object (what they are responsible for), it turns out to be more complex in the context of managing AI. Given the fact that the frontier of AI will shift with time (Berente et al., 2021) and the development of AI lacks common aims (Mittelstadt, 2019), many factors may affect how responsibility is perceived, allocated, and realised to have consequences benefiting individuals, businesses, and societies. For instance, beyond the connection between subject and object, different stakeholders are often involved (Mikalef et al., 2022). There is often an authority defining and enforcing the

[14] *Responsible AI design* is an interdisciplinary approach that incorporates fairness, transparency, accountability, human-centric principles, privacy protection, and sustainability into AI systems. Scholars emphasise ethical frameworks and participatory design strategies to ensure AI aligns with human values and serves the public good while mitigating risks.

practical outcomes of the responsibility relationship as well as the evolving normative basis for these decisions (Stahl, 2023).

Despite realising these complexities, most recent guidelines proposed for the ethical and responsible use of AI (Schiff et al., 2021) seem to lack actionable practices that organisations can implement to ensure their AI systems are designed responsibly. More promisingly, however, recent research has begun to explore the work practices adopted by AI designers to promote responsibility and make the best of AI (e.g., Holstein et al., 2019; Lee et al., 2020; Madaio et al., 2022; Rakova et al., 2021). In working towards responsible AI, while AI designers have noted the importance of the intervention of training datasets (Denton et al., 2020; 2021) and algorithmic mitigations (Agarwal et al., 2018), they may still face challenges stemming from biases influenced by their personal perspectives and experiences. To provide additional guidance and support, many publicly available toolkits have been developed (Morley et al., 2020). However, these resources often fall short of explicitly offering clear instructions for practical implementation (Deng et al., 2022) and are primarily designed to support specific types of technical work (Wang et al., 2023). This may neglect the broader sociotechnical aspects of responsible AI. Thus, there is a need for IS research to work towards a framework that supports designers in their 'responsible work' during the design of AI systems, ensuring that both social and technological elements are adequately addressed.

2.3 Responsible AI governance

The importance of governed AI has been acknowledged in a growing body of research (Mäntymäki et al., 2022; Georgieva et al., 2022). Thus, regulatory efforts at

various levels are evolving to manage AI in and across business sectors, reflecting the understandable tensions between the urgency to implement AI and the need for it to be done so responsibly (Floridi, 2019; Wang et al., 2020). As private corporations, governments, and societies are working to determine where and how AI fits into existing data and technology regulations and policies, and human rights law (Schiff et al., 2021), the updating and creation of corresponding ones that touch upon machine learning (ML) models and AI systems also becomes necessary.

While there has been considerable development of knowledge and growing consensus around AI ethics principles (Fjeld et al., 2020; Jobin et al., 2019), how to translate them into practicable governance processes to manage AI remains unclear for organisations (Hickok, 2021; Mäntymäki et al., 2022; Schiff et al., 2021) and draws scholarly attention (Schneider et al., 2023). Moreover, the advent of AI outpaces the development of new regulations, leading to gaps in how emerging issues associated with data, ML models, and AI systems can be addressed (Berente et al., 2021; Schneider et al., 2023). Concerns have therefore been raised as regulators and supervisory agencies mainly react rather than anticipate AI risks (Butler et al., 2023). Such a reactive approach, combined with business board oversight lags in AI implementation, will largely neglect its negative affordances on organisational stakeholders and the wider society (Mikalef et al., 2022). Hence, the idea of AI governance has emerged (e.g., Butcher & Beridze, 2019; Cihon et al., 2021; Floridi, 2018; Gahnberg, 2021) but not to the extent of explicitly defining practices that support the management of AI within organisations (Schneider et al., 2023).

Additionally, researchers in the IS field generally fail to investigate the dark side of AI (Mikalef et al., 2022), which has also led to the call for the responsible governance

for AI technologies. Organisations struggling to oversee their use of AI would benefit from understanding potential AI regulatory actions at the organisational level and regulatory concepts institutionalised in policymaking (Butler et al., 2023, Mäntymäki et al., 2022) to translate abstract AI ethics into practice. Hence, governance mechanisms play an important role in mitigating challenges and raising potential in organisations (Schneider et al., 2023). They need to be identified to define how to attain the needs dictated by responsible AI principles and therefore generate positive outcomes.

3. Research Method

3.1 Research approach

Recognising the nascent stage of knowledge on managing AI, this study adopts an inductive, interpretive approach using selected features of grounded theory (Corbin & Strauss, 2015; Glaser & Strauss, 1967). Known for its flexibility and rigour in IS research, the grounded theory method (GTM) helps develop rich descriptions and practice-oriented conceptual models (Wiesche et al., 2017) of managing AI, while grounding in empirical data through a series of cumulative coding cycles (Strauss & Corbin, 1990). This method is particularly well-suited for studying sociotechnical behaviours (Birks et al., 2013), making it ideal for exploring the management of AI at the intersection of social and technological perspectives (Berente et al., 2021).

We adopted this approach for two key reasons. First, while responsible approaches to design, develop, and deploy AI are widely recognised within the practitioner community, they lack explicit attention in academic literature (Arrieta et al., 2020). This gap underscores the need for a qualitative study to examine how practitioners

apply and manage responsible AI in real-world settings. By capturing nuanced insights, we can bridge the disconnect between theory and practice, informing both academia and industry. Second, managing AI is inherently complex in a real-world setting, partly due to its evolving computational frontiers (Benbya et al., 2021; Berente et al., 2021), and principle-to-practice gaps (Morley et al., 2020). Thus, an in-depth, interpretative examination of multiple perspectives of elite informants (Solarino & Aguinis, 2021) is particularly useful for defining what managing AI entails in complex scenarios, and for constructing a novel, practice-oriented model (Corbin & Strauss, 2015; Jones & Noble, 2007).

Methodological bricolage (Pratt et al., 2020) offers an alternative to templates (e.g., Gioia methodology; Gioia et al., 2013) for qualitative data analysis, highlighting researchers' agency and creativity in combining various methodologies. It operates iteratively to construct deeper connections between data and literature. Given GTM's flexibility and analytical power, Birks et al. (2013, p.2) proposed key criteria for its use in IS research, including "theory development, constant comparison, iterative coding, theoretical sampling, management of preconceptions, inextricable link between data collection and analysis". We used analytical moves (Pratt et al., 2020) from various methodologies while referring to these criteria to inform the data collection and analysis.

We adopted a partial portfolio of GTM procedures for rich and rigorous descriptions of AI management and conceptual model building (i.e., which defines variables and their relationships without fully justify them or specify their boundaries; Wiesche et al., 2017), contributing theoretically (Corbin & Strauss, 2015). Existing literature, as well as the analytical and theoretical memos in the forms of text narratives and

mind maps are used for constant comparison when analysing and interpreting the data (Charmaz, 2006). Several iteratives of data coding refined interrelated concepts and dimensions, allowing systematic model development from data (Strauss & Corbin, 1990). Theoretical sampling ensured broad data coverage, and data collection ceased upon reaching theoretical saturation (Saunders et al., 2018), meaning data fully represent relevant conceptual categories (Stark et al., 2007). No a prori theory specifically addressed the phenomenon AI management. While we were aware of relevant literature, it did not constrain creativity or idea generation (Myers, 2009). Data collection and analysis remained intrinsically related.

3.2 Data collection

This study relies on triangulation of primary and secondary data from multiple sources to gain a richer understanding of AI management. Iterative cross-checking of theoretical themes based on multiple data enhances research validity through data triangulation (Patton, 2002).

For primary data, we conducted twenty semi-structured interviews virtually via Google Meet (due to pandemic restrictions) from April through July 2021 until theoretical saturation (when no codes/themes emerge) (Glaser & Strauss, 1967). These interviews provided the most core evidence for our findings. Additionally, analytic memo writing documented coding processes and reflections, contributing to theory development and ensuring the qualitative rigour.

Participants[15] included academic and professional elite informants with expertise in responsible AI and AI management, spanning machine learning, data science, ethics, and law (see Table III.A1 in Appendix A). They were chosen following theoretical sampling logics (Patton, 2002) via direct emails, with snowballing technique used in some instances. All participants hold crucial roles and have experience in designing, evaluating, overseeing, or consulting on responsibility in AI projects, offering rich insights into responsible AI management.

To facilitate discussions, an interview guide (see Appendix B) was developed and adjusted throughout data collection. We made tweaks where needed to account for new areas of inquiry related to responsible AI management that had been identified during the interview and coding process. Written consents were obtained from all participants before the interviews, which were digitally recorded, transcribed verbatim, and appropriately formatted, yielding 287 pages of data. Informants reviewed transcripts to ensure accuracy.

The secondary data comprised the analysis of data from company websites, news, reports, videos, and other periodicals on responsible AI management, to the extent possible. Data were selected following a representative case selection logic and served as an important data source to triangulate our findings (Patton, 2002). Specifically, we chose AI giants and start-ups known for responsible AI practices, given the absence of a definitive, universally recognised list. Over 2020-2022, we

[15] Their contacts were obtained through personal connections and LinkedIn messages, as they were part of shared groups focused on responsible AI, AI ethics, AI governance, AI for social good etc.

collected 190 documents. Table III.1 summarises the sources of evidence and their use in analysis.

Table III.1 Summary of data sources and analysis

Data Sources	Type of Data	Use in the Analysis
Semi-structured Interviews	• Perspectives from 20 elite informants from academia and practice across various domains relevant to responsible AI. • On average, interviews lasted between 45 and 75 minutes. • Personal notes, based on interactions with all respondents.	• Capture how informants experience and interpret responsible AI to frame research analysis. • Balance both the width and the depth of perspectives. • Discuss emerging findings. • Triangulate with document analysis and observations.
Archival sources	• Over 190 company web pages, reports, public documents, and other periodicals. • On average, document length ranges from one page to 32 pages.	• Capture how organisations design, develop, and deploy responsible AI in the real-world. • Capture how responsible AI was portrayed in the media, and trace the changing representation of responsible AI in the media. • Familiarise with the context of responsible AI management, and triangulate with informants' interpretations.
Media coverage	• Press articles, videos. • Media coverage was gathered sporadically in the early stage of data collection.	

3.3 Data analysis

The analysis of our qualitative data used analytical moves (Pratt et al., 2020; Grodal et al., 2021) and followed an iterative process (Birks et al., 2013). We adopted a more structured approach (i.e., Straussian GTM) consisting of three steps: open, axial, and selective coding (Wiesche et al., 2017), during and after the collection of data for a crucial exploratory analysis. Constant comparative analysis identified *first-order concepts* (informant-derived) until saturation occurred, linking them to *higher-order categories* (researcher-induced), which were critical for model building (Charmaz, 2006; Glaser, 2016). Connections between categories suggested by

interviewees were examined across data, forming a robust basis for our grounded model (Urquhart, 2013). Divergences were resolved through discussions among the research team to ensure consensus.

Open coding established a contextual understanding of responsible AI design and governance, capturing a broad range of emerging patterns and distinct concepts (Corbin & Strauss, 2015). Both 'lumping' and 'splitting' coding strategies were used to scrutinise AI management practices and organisational outcomes, getting to the essence of categorising while avoiding leading to superficial or overwhelming analysis (Saldaña, 2016). Single occurrences in the data were coded for significance, with concepts grouped into subcategories through consensus among the research team, yielding 29 first-order concepts. *Axial coding*, conducted alongside open coding, connected subcategories and restructured concepts into more focused categories (Corbin & Strauss, 2015). Through constant comparative, we iteratively refine codes, allowing second-order analytical themes to fully emerge, while ensuring prior theoretical concepts and assumptions helped us to structure and make sense of the data, but not to the extent of restricting our interpretations. This stage distilled 29 first-order concepts into 12 second-order themes. *Selective coding* was followed to identify a set of seven aggregate dimensions that explain 'managing AI'. A sensemaking process was involved, which enabled us to understand how responsible AI design leads to organisational consequences via responsible AI governance mechanisms.

3.4 Analysis of research trustworthiness

The trustworthiness of this research was assessed using two sets of criteria (Flint et al., 2002). While *credibility, transferability, dependability, confirmability,* and *integrity* form a set of criteria appropriate to interpretative research (Hirschman, 1986; Lincoln & Guba, 1985), the criteria of *fit, understanding, generality,* and *control* are common to grounded research (Strauss & Corbin,1990). We partially addressed the combined trustworthiness criteria as we adopt a partial portfolio of grounded theory procedures. Table III.2 demonstrates how our data and analyses satisfy these criteria.

Table III.2 Trustworthiness of the study and findings

Trustworthiness criteria	Methods of addressing trustworthiness
Credibility Extent to which the results appear to be an acceptable representation of the data	• Multiple data sources (interviews, archival data etc.) • Research team members gave input during data interpretations • Fellow peer debriefing on initial interpretations • *Result*: Revised data structure and expanded model
Transferability Extent to which findings from one study in one context will apply to other contexts	• Theoretical sampling • Thick description • Model was evaluated against published industry reports • *Result*: Theoretical concepts were represented by data from participants, key constructs were relevant across settings
Dependability Extent to which findings are unique to time and place, and explanations are stable or consistent	• Elite informants were recruited from academia and practice across various domains relevant to AI/responsible AI • Participants reflected on their perceptions and experiences covering most important topics • Facts and interpretations triangulated by other data sources • *Result*: Found consistency across participants' interpretations about managing AI
Confirmability Extent to which interpretations are the results of participants and phenomenon as	• Interpretations and supportive evidence were reviewed by a fellow peer who is not familiar with responsible AI • Summary of preliminary findings to research team members • Data analysis procedures recorded and described

opposed to researcher biases	• *Result*: Coding and interpretations were extended and refined, and conclusions were confirmed flowing from the information collected
Integrity Extent to which interpretations are influenced by misinformation form or evasions by participants	• Interviews were conducted professionally in line with the university approved participant information sheet and consent form • *Result*: Participants have no attempt to mislead or evade the issues discussed, the researcher was sensitive to cultural nuances and ethical biases on sensitive topics around responsible AI
Fit Extent to which findings fit with substantive areas under investigation	• Research methods used to address credibility, dependability, and confirmability • *Result*: Concepts were more deeply described, and the theoretical integration was made more fluid, capturing the complexities of responsible AI discovered in the data
Understanding Extent to which participants accept results as possible representation	• Summary of findings to participants, fellow peers, and research team members • *Result*: They accepted the findings
Generality Extent to which findings reveal multiple aspects of the phenomenon	• Multiple sources of data were collected and analysed • Interviews were of sufficient length and depth • Interviewees include different key actors involved in responsible AI • *Result*: Multiple aspects and perspectives were captured
Control Extent to which participants influence aspects of the model	• Participants have control over some aspects of the conceptual model • *Result*: Participants can influence responsible AI management

Note. We partially address the trustworthiness criteria for grounded theory as we adopt a partial portfolio of its procedures.

4. Findings

Taking a sociotechnical perspective, we found that responsibly managing AI within organisations rests on the interplay between *responsible AI design* (i.e., covering the technical aspect) and *governance* (i.e., covering the social aspect). In this section, we report findings by presenting our data and the theoretical insights they led to. We identified design tactics that AI designers can implement to meet evolving multifaceted responsibilities in AI ethics. We also identify governance mechanisms

that help enact these responsibilities to ensure desired organisational outcomes (*instrumental, humanistic*). Relevant literature is interspersed within the interpretive results to provide theoretical support.

Figure III.1 presents the grounded model, which culminated from our data analysis, depicting key concepts and associations (i.e., pre-theoretical understanding). Illustrative data and example quotations supporting the emergent theoretical model are included in Appendix C, with Table III.A1 listing the interviewees quoted in the study. Figures III.2 and III.3 illustrate the data structure, providing a chain of evidence from raw data to emerging concepts and aggregate dimensions (Gioia et al., 2013).

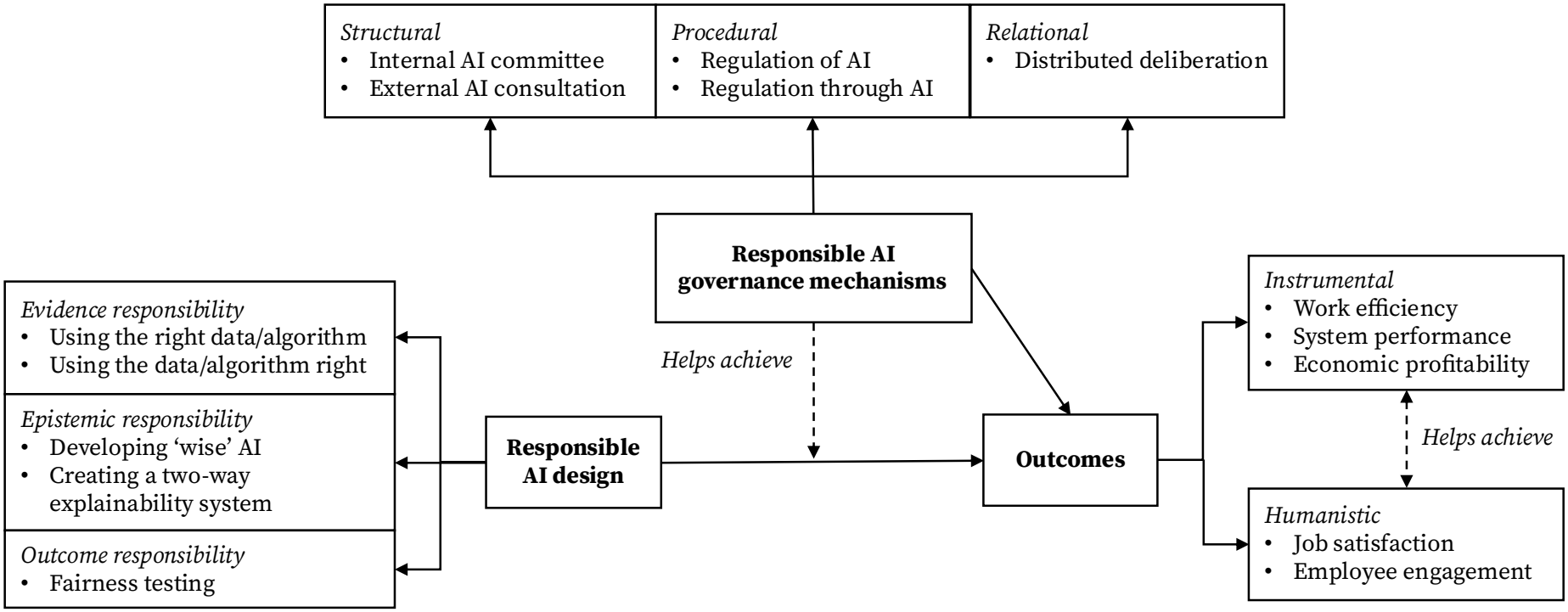


Figure III.1 Conceptual model of managing AI

4.1 Responsibilities in AI design: evidence, epistemic, outcome

Understanding the responsibilities associated with AI design can serve as a stepping-stone to fostering responsible AI management within organisations. By embedding these responsibilities into their work, AI designers can better navigate managerial complexities with heightened awareness and increased accountability. This section explores the dimensions of responsibility and the corresponding AI design tactics that emerged as we theorised our data in consultation with the literature. The ongoing discourse on AI ethics (Buhmann et al., 2020; Mittelstadt et al., 2016; Tsamados et al., 2022) highlights three sets of key challenges concerning AI systems: *evidence*, *epistemic*, and *outcome*. In response, AI design tactics are categorised into three constitutive responsibilities, each addressing one set of challenges to guide ethical and effective AI implementation.

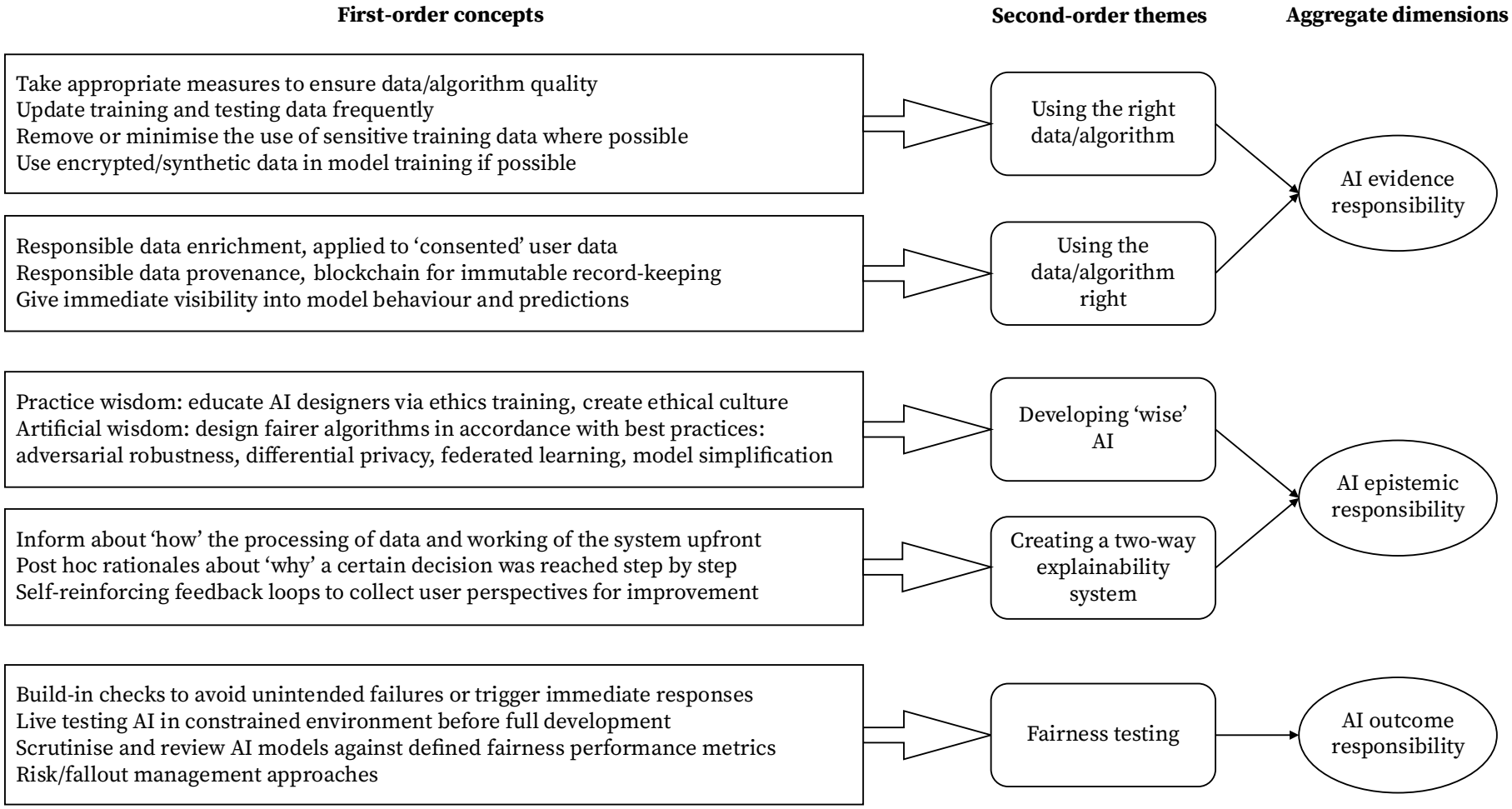


Figure III.2 Data structure for responsible AI design

4.1.1 Evidence responsibility

This dimension of responsibility focuses on ensuring AI design are based on reliable, accurate, and valid *evidence,* including both data and analytical processes. This help address concerns that arise when self-learning AI systems transform large datasets into decision-making insights (Buhmann & Fieseler, 2023). Our study identifies two theoretical themes, guiding designers in avoiding bias-laden systems, ensuring accurate algorithmic predictions, and mitigating serious ethical implications.

An important aspect is ***using the right data/algorithm***, which underscores the need for data and algorithms that are appropriate, accurate, relevant, and aligned with the responsible objectives of AI applications. Poor-quality or unethical data and non-deterministic algorithms can result in flawed decisions and bias since algorithmic decisions rely on data associations and correlations (Tsamados et al., 2022). Interviewee #13 emphasised the need for organisations to critically assess their data sources, warning:

> *"I think it's also important that…organisations that are using that kind of AI, they actually take a step back and look at the information and the data they're using…that leads to the various recommendations that are being made, because there's a big potential for bias there."*

Ensuring data quality is essential for AI designers to create reliable AI models that produce accurate and fair analyses and decisions. Google's *Know Your Data campaign*[16] aims to help AI product teams and researchers understand, assess, and manage datasets to improve data quality, either before training a model or when

[16] https://knowyourdata.withgoogle.com/

debugging it. *Taking appropriate measures*, such as ensuring accuracy, adequacy, representativeness, can help *ensure data quality*, as Interviewee #12 stated:

> *"I think in any stage of the data process from collecting to processing and annotating and so on, questions must be asked as where is the data coming from, who has collected the data, is it annotated, who's done it, how is it done, are there any forms of writing guidelines for people to think about it... make sure you use those data as needed, but as much as necessary..."*

Ensuring the quality of algorithms is also crucial for evidence responsibility, especially as evidence-based decision-making aided by algorithms becomes increasingly common. AI designers need to select *appropriate algorithms* to achieve fairer outcomes, as Interviewee #3 explained:

> *"AI should be at least a fair agent, if you use a model where you use the algorithms...the algorithms should be in a neutral position, I mean without bias..."*

Additionally, consistently improving the quality of data fed into AI algorithms will lead to ethically aligned recommendations to guide decision-making. AI designers need to frequently *update the training and testing data* by identifying redundant, obsolete, or trivial data and actioning it for archival or deletion. As Interviewee #4 mentioned:

> *"...it was for archival, and you have to identify whether a document should be conserved for a long period of time, or like something that we didn't need to keep, and we could like trash in the bin, or something needed to be done..."*

Another identified tactic is to *remove or minimise the use of sensitive training data where possible*. This can prevent inherent biases in data that could lead to unfair

results in AI models, which may perpetuate or exacerbate existing inequalities. As interviewee #18 stated:

> *"AI systems work with the aggregation of data, and the main objective of these systems is to, whether it's with a good intention or not, profile people and classify them into categories... there is a way to eliminate those kinds of biases that can lead to discrimination... we get that the data sets can be debiased at the very beginning..."*

In addition, AI designers should *use encrypted/synthetic data to replace the use of sensitive data in model training*, if possible, as Interviewee #14 remarked:

> *"...there is increasing work on ways to train models that don't require direct access to data or can operate on encrypted data and synthetic data... so the data is not actually ever directly seen by an individual or by the company..."*

While encrypted data is converted from readable data using encryption techniques, synthetic data is often generated through algorithms that simulate real data patterns. These approaches are particularly useful when real data is either not available or its use is restricted due to ethical or legal considerations. They provide feasible solutions for maintaining the robustness of AI models while adhering to strict data protection standards.

The second theoretical theme is ***using the data/algorithm right***, which means ethical and responsible practices in all aspects of managing data and algorithm processing. One identified tactic is to ensure *responsible data enrichment*, referring to the ethical approach to improving the quality and value of an existing dataset while maintaining its integrity. As Interviewee #7 noted:

> *"I think we try to make sure that we control every single step of the data ingestion, so from the first reception to the transformation of data… we do run scenarios, we do take a lot of time under data quality, which is not the case in the digital world… I think we do try to make a lot of thoughts about cleaning the data, making the first banner before we do any conclusion… and that not only we preserve confidential information which I think is the basic, but also that what we are proposing to a user is for the user's own benefit…"*

Moreover, such practice needs to be applied to '*consented' user data*, as Interviewee #8 mentioned: *"… seeking permission, asking for consent… those kinds of things will become procedural, and they'll be translated into processes in large institutions."*

In addition, it is important to ensure *responsible data provenance*, a detailed record describing the origins, processing, and use of data in both forward and backward directions, as well as how it has been altered or handled over time. Interviewee #14 stated that:

> *"… I had a group, this group was responsible for what we call data governance… their job was to make sure that access to data was controlled, and that the provenance of data was recorded… so we knew which data passed from which person to which person…"*

Implementing data provenance in AI involves using tools and techniques that can automatically log data changes, for example, *blockchain solutions* can be used to provide immutable record on a decentralised network (where no single entity has control over the entire dataset), as Interviewee #17 put it: *"…if you use blockchain to…record and automate them in a permanent traceable ledger…"*

Existing literature also highlights the importance of implementing data retention and reproducibility measures to mitigate evidence concerns (e.g., Buhmann et al.,

2019; Werder et al., 2022). These measures are essential for ensuring that data, once collected, is preserved in a manner that allows it to be accessed and analysed consistently over time. Also, Tsamados et al. (2022) discuss algorithmic traceability, which is the ability to trace back and understand the decision-making process of an algorithm for normative purposes.

Another tactic designers can adopt is to *give immediate visibility into model behaviours and predictions,* as Interviewee #7 explained: *"… inside the algorithm itself, I think it's about externalising a lot of parameters, and being able to give access to those parameters as much as we can."* This visibility is essential not only for operational effectiveness but also for supporting broader objectives like regulatory compliance, ethical governance, and continuous improvement in AI technologies.

4.1.2 Epistemic responsibility

This dimension of responsibility is described as providing an account for the reliability and precision of the ethical AI system throughout its lifecycle. It arises from the need to address epistemic concerns that relate to challenges posed by ways of AI opacity, for example, the inscrutability and poor traceability of algorithmic inputs and their processing for laypeople and even for AI experts (Buhmann & Fieseler, 2023; Tsamados et al., 2022). Two theoretical themes emerged in this study.

The first one is ***developing 'wise' AI***, which refers to the process of creating inherently ethical AI systems that embody human wisdom and possess capabilities to assist with ethical decision making. AI designers play a significant role in this, as mentioned by Interviewee #6, *"I think that whoever is designing and creating these*

things obviously needs to be responsible when designing and creating these things." Concerns may arise from designers' deliberate intent to optimise AI performance, avoid accountability, or even evade regulations (Ananny & Crawford, 2018). It is necessary to improve *practical wisdom*, which enables AI designers not only to know what to design but also to understand how to design AI in a way that is ethically responsible.

One approach is to *educate AI designers* via ethics training, as Interviewee #9 described:

> *"I think it should start at even the education of data scientists, people who are going to be building those… when they're learning about how to build models, it should be ingrained in even the methodology, like the data science methodology should have those ethical guidelines, in my mind…"*

These training initiatives can help eliminate bias in AI team, provide practical guidance that enables designers to incorporate ethical considerations into the system development process, thereby fostering an environment where ethical and unbiased decision-making becomes the norm.

In addition, creating *ethical culture* may help motivate AI designers as Interviewee #15 mentioned:

> *"you need to create this culture, data culture and ethical culture, so that the data scientist who is building it has the trigger to raise an ethical flag or red flag if something is wrong and escalated… because that trigger is not there, then of course having an ethical framework doesn't make a lot of sense."*

Responsible AI is a field deeply imbued with values, where the virtues of AI designers play a crucial role in determining how AI machines augment or fully

automate human decisions (Buhmann et al., 2020). While AI designers often face pressures that compel them to prioritise organisations' interests, such as high quality systems and increased profits, when designing AI systems (Mittelstadt, 2019), their personal moral convictions can drive them toward more ethical and responsible behaviours during the design process (Kish-Gephart et al., 2010). As Interviewee #19 explained:

> *"I was able to give the leadership all the projects... remove all these bad practices and make a transparent product recommender, for example in this industry... there are also a lot of good people who are willing to raise better voices and implement these best practices."*

The process of self-actualisation, where AI designers strive to realise their capabilities in designing ethical AI systems, is a fundamental element of developing wisdom. It involves continuously adapting one's understanding and behaviours to align more closely with personal virtues and collective ideals (broader cultural and social norms) on ethical and responsible AI. As Interviewee #15 mentioned, *"there are a lot of people... who are a bit sick of the rat race and are looking for upskilling to challenge their skills but also to have a social impact, to have a purpose in their life, and are open for voluntary work."* Thus, organisations will need to consider creating ethical culture to cultivate virtues among AI designers. This is essential for them making morally sound and well-informed decisions when designing AI systems.

Another important aspect relates to *artificial wisdom*, that is, leveraging computational advances to create AI systems that, rather than merely intelligence, will share many features of wisdom in human (e.g., consciousness, morality) to promote greater individual and societal wellbeing. The first-order concepts for this

theme involve many design tactics for AI designers to create fairer AI systems by incorporating wisdom-like attributes, which we discuss below.

Adversarial robustness refers to the ability of an AI system to withstand adversarial attacks intended to fool or mislead it. These attacks involve slightly altering input data in ways that are often imperceptible to humans but can cause the AI to make errors. This is a way to ensure AI safety and security as Interviewee #1 mentioned:

> *"You might have some people like some individuals that will start developing affection for those systems because of the way the systems treat them… if you can understand these adversarial attacks, if you can achieve some level of confidence in your results, then you can commit your systems safer and more predictable…"*

Differential privacy is a privacy-protecting approach that allows developers to extract valuable insights from datasets to train AI models while withholding personal information. The core technique involves adding controlled, random noise to the data. Such noise is enough to protect the identity and data of individuals, making it difficult to infer specific details about any person from the aggregate data (Zhu et al., 2020). Interviewee #11 discussed *"private preserving machine learning" and suggested that "privacy can actually be built into responsible AI"*. Also, Microsoft proposed *Differential Privacy for Everyone*[17], which aims at reaping the full benefits offered by the data when the individual privacy is protected at the same time.

Another way for enhancing both the performance and privacy of ML systems is *federated learning*. As Interviewee #14 put:

[17] https://download.microsoft.com/download/D/1/F/D1F0DFF5-8BA9-4BDF-8924-7816932F6825/Differential_Privacy_for_Everyone.pdf

> *"you have to also invest in and stay connected to the community that is researching these methods where you can do training of models that operate on encrypted data or operate data… it's called federated learning… in a federated way so that the company never actually sees the details of the data…"*

It is a technique where the model training process is distributed across multiple decentralised devices. The model learns from data that remains on local devices and send only model updates to a central server to improve the model (Wei et al., 2020). This helps reduce the risks of data breaches by minimising data movement. Also, IBM has been exploring *Federated Learning*[18] as a part of its broader initiatives in advancing ethical AI technologies, which is crucial for compliance with data privacy regulations.

Model simplification is a strategy used to make complex ML models (often referred to as 'black box' models due to opaque nature) more understandable and interpretable to the users. Some designer informants mentioned that they would use highly interpretable *gate models* to help explain the inner workings of black box systems. Interviewee #9 stated that:

> *"One way that we have thought about doing that is trying to approximate those low interpretability models with high interpretability models… if we're doing classification using a machine learning model, we might try to interpret the results with like logistic regression type of model, for instance… then we're still doing the results with the machine learning model, but we're trying to provide an explanation or some kind of like reason code maybe, or like seeing these are the features most likely to contribute to that result…"*

[18] https://ibmfl.res.ibm.com/

In addition to gate models, sometimes the best way to achieve simplicity is by using inherently simpler models. Decision trees and logistic regression are examples that will provide good interpretability for classification problems. As Interviewee #3 mentioned, *"I think the decision tree is very basic and it's quite easy to understand how it works..."*. However, if there is a need for achieving higher prediction accuracy, *model distillation*[19] could be another potential tactic. For instance, Google has explored the idea of creating a simpler 'student' model that learns from the outputs of a complex 'teacher' model. This helps retain the performance characteristics while benefiting from the lower computational demands and greater interpretability.

Designing inherently explainable AI systems can help push the boundaries imposed by trade-offs between the interpretability and accuracy of AI outcomes. It allows AI experts to investigate as well as gives end users the abilities to understand the inner workings of AI systems. As Interviewee #13 put:

> *"I would say a starting point would probably be moving as much as possible towards explainable AI... so that you're able to at least understand what the machine is doing... then at least, you know what's going on so that the human may be able to intervene at some point earlier than the earliest stage before losing control."*

The second theoretical theme is ***creating a two-way explainability system***. The concept of two-way explainability extends the traditional notion of AI explainability (the capability of AI systems to be understood as least by some party) to involve not only how AI systems provide explanations on AI processes and decisions to end

[19] https://research.google/pubs/distilling-the-knowledge-in-a-neural-network/

users (AI-to-human explainability) but also how users can provide feedback to improve the AI's decision-making processes (human-to-AI feedback). This can bridge the gap between technical and ethical considerations by enhancing the explainability, as Interviewee #1 put it:

> *"I think the idea of explainability in the future should be a two-way system… the next is when we start thinking about the user providing explanations to the system because this is how humans exchange information… we explain things to each other."*

For *AI-to-human explainability*, there is a need to ensure *prospective and retrospective transparency*, which refers to disclosing insights into the processes used for future decision-making as well as the analysis of decisions already made. As AI legislation evolves, there is a growing push for a 'right to explanation'. Many data protection regulations such as GDPR[20] would require organisations to communicate to users the logic behind AI algorithms. This gains importance as the decision-making processes of AI become less predictable, leading to calls for increased transparency to ensure ethical and responsible AI in the first place (Buhmann & Fieseler, 2023).

The opacity nature of AI systems can make assessing the trustworthiness of algorithm decisions challenging even for technical experts (Dennehy et al., 2023). In order to ensure explainability, the data suggested that *informing about 'how' the processing of data and working of the system in reaching ethical decisions upfront* is useful for both AI developers and end users. Interviewee #7 mentioned this as follows:

> *"I think the thing is… it's about the transparency of algorithms… if you can explain in simple words, what kind of routines you are working with, how you*

[20] https://www.privacy-regulation.eu/en/recital-71-GDPR.htm

treat the data when you ingest the data, and what kind of output data you produce…"

To support end users to understand the considerations when making decisions on the use of AI applications, the data suggest that designers need to ensure retrospective transparency. This involves *providing post hoc rationales, step by step, explaining 'why' a certain decision was reached*. Interviewee #14 described it in this way:

"…there were some very concrete things that are easier, which is to be able to explain the causal factors that led to a decision, that may be possible… so that's not explaining how the inner workings of the system performed, but it may be as simple as saying the reasons that the system chose to provide you with this…"

For *human-to-AI feedback*, Interviewee #4 has suggested to create *self-reinforcing feedback loops* to collect user perspectives for AI system improvement:

"… they're going to rectify or validate, and this is what being used to feedback back into the model… So, this is one case where you can have a continuous feedback loop to evaluate and improve the model…"

4.1.3 Outcome responsibility

This dimension of responsibility emphasises the need to manage the impacts of AI decisions/actions on individuals and societies to ensure positive outcomes. ***Fairness testing*** has been identified as the main theoretical theme from the data, which involves evaluating AI algorithms and systems to ensure that they operate without bias and treat all users equitably. As Interviewee #6 put:

"I think these are the kinds of fundamental building blocks that AI should also include… if we look at AI systems as an extension of people systems, I'm

thinking of AI as another type of, probably a smarter version but another type of supporting technology, then I think things can be managed in a more positive way which benefits people, whether it's me or you or us as a society."

AI designers ought to perform fairness testing on models they develop, as perpetuated biases or the biased design of systems will lead to discriminatory practices and unfair outcomes (Tsamados et al., 2022), particularly in sensitive areas like hiring, credit scoring, and healthcare. Some fairness testing practices that designers could follow is to *create built-in checks to avoid unintended failures*, as Interviewee #19 suggested:

"The best way to do that is trying to make those explainable machine learning models to try to look or to find how they took in the position... and you can detect if there is something wrong there, and even if they are using a fightable tool to get some bias or to discriminate some ethics, to remove them all because you initially don't know that until you see those predictions..."

Also, AI designers should *create built-in checks to trigger immediate responses*, according to Interviewee #2, *"...where there could be errors in recognition, errors in detection, and that might lead to actually even physical harm to people... there have to be checks and balances, emergency switches, or buttons to actually guard against those sorts of issues that might occur."* Similarly, *Microsoft Azure Sentinel*[21] is a solution to help streamline security operations by automating threat responses to alerts based on predefined conditions.

Another approach mentioned by Interviewee #19 is to conduct *live testing in constrained environment before full development*:

[21] https://learn.microsoft.com/en-us/shows/azure-friday/automate-threat-response-with-azure-sentinel

> *" what we do is, instead of designing or implementing this machine learning model in going through to production, usually to make a step at the stage of post analysis of the model, is kind of pre-production… and we analyse with our business, for example, usually with people from the business, analysts, these outcomes… for example, if there are some things that we have biases or overfeed because we are cantering in some features that we shouldn't do, we usually detect that with business"*

This practice, often termed 'sandbox' testing, allows developers to identify potential issues and make necessary adjustments within a secure, controlled setting to minimise risks associated with full-scale deployment.

After deployment, developers will need to *scrutinise and review AI models against defined fairness metrics* to help evaluate the impact of the algorithms processes on fairness and privacy. As Interviewee #20 mentioned:

> *"When you're building the model, you have to iterate it and optimise it for certain situations… and then once you know that there's a situation where it's failing, you need to continue to sort of iterate and optimize it for that situation.…so to consider that metric is how you would make a metric for it, but like to consider that almost as a metric for optimisation, like how do we reduce bias as a metric of success."*

Fairness performance metrics are essential tools designed to assess how fair an AI model is across different groups. Many researchers stated the importance of embedding ethics into AI system design (e.g., Bauer, 2020; Henningsson & Eaton, 2023), as it helps address the challenges of implementing top-down driven changes and avoids getting lost in translation when bringing abstract principles into applicable practices (Schiff et al., 2021). Similarly, Interviewee #4 commented that

AI designers can add ethics and other regulatory contents as model performance metrics when developing AI systems:

> *"when you work on an AI model, you are going to work with an optimising metric along to improve the accuracy or something… and you might have like to invest in one metric you are playing against, you try to get as best as possible, besides that, you're going to have like satisfying metrics… and then you could use exactly that with ethical features, what is the minimum we can tolerate…"*

After deployment, developers need to learn *risk/fallout management approaches* to identify and mitigate the potential adverse effects, as Interviewee #18 stated, *"there are so many ways where things could go wrong, and that's why we need to use like frameworks that can assess all the risks, and that we can assess systematically the risks associated with one precise AI system."*

One specific tactic mentioned by Interviewee #14 is to conduct *sensitivity testing* to check the robustness of the model, not only to gain confidence in the solutions but also to identify who is accountable for specific actions or outcomes within a process or situation:

> *"…to do sensitivity testing on some particular cause… I think that would be very helpful because then you can direct people who have the logistical resources to address the problem, but to have some confidence that if they took that measure, it would actually address the problem…"*

Another tactic suggested by Interviewee #16 is to apply *verification loops* into the operational processes to continuously check and confirm that systems perform as intended:

> *"…that calls for a verification loop for the results, those are the kind of things where you have to plan for this, and you have to invest…I think that once again,*

responsible AI would mean that, it's not about making sure that there is no mistake, it's about making sure that when there is a mistake made by the AI system, like you have some contingency plans that are there, like verification loops…"

4.2 AI governance mechanisms: structural, procedural, relational

A deep understanding of how to sustain the positive organisational results of managing AI in organisations paves a proactive way to attain responsible AI design and ensure it adheres to ethical standards. Scholars have suggested investigating the role of *AI governance* as a key mechanism that influence AI implementation and foster transparency, accountability, and trust (Mäntymäki, 2022; Schneider et al., 2023). This study delineates AI governance at the organisational level and conceptualises it as the structures of rules, practices, and processes designed to develop, direct, and control AI design practices. Such a definition is normative in that the intention is to be action-oriented to guide organisations to manage AI. As Interviewee #8 mentioned that:

"…it has to have the usual accountability structures, like policy, standards, procedures, to go with it, to make sure that it delivers what it's meant to… it covered somewhere along the document, ethical policies, ethical standards, ethical procedures, and everything else along with it becomes more important."

After comparing the data with existing literature, insights were borrowed from studies on IT governance architectures (Chau et al., 2020). The governance aspect of managing AI was approached through three distinct but complementary mechanisms: *structural*, *procedural*, and *relational*. This approach articulates specific and practical organisational AI governance that can be effectively followed by various actors involved in AI management.

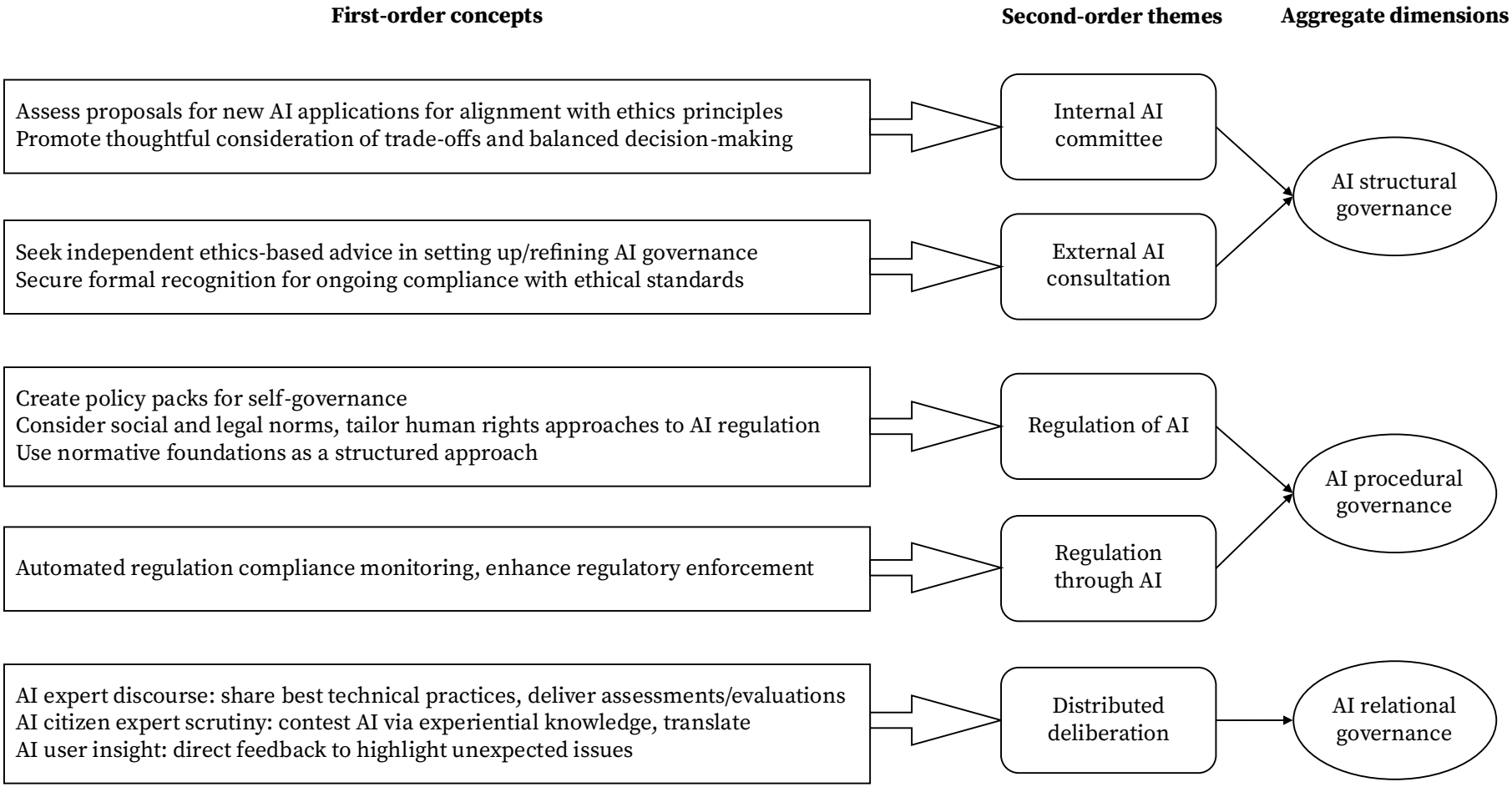


Figure III.3 Data structure for responsible AI governance

4.2.1 Structural AI governance

This dimension of governance mechanism is defined as the allocation of specific roles, responsibilities, and decision-making authorities for supervising, directing, and planning AI governance activities in organisations. Such governance can be initiated in two primary ways, which are discussed below.

The first approach involves establishing ***internal dedicated AI committee***, recognised as a professionally diverse group of executives formed within an organisation, focused on AI-related matters. As Interviewee #14 commented, *"it probably needs to be some sort of level of oversight committee or something that reviews (AI) decisions."* The rationale for establishing a committee to manage AI is discussed by Interviewee #15:

> *"It's not up to a data scientist, it's up to the board or the C-level, that stares our managers of a company… because if an AI model is not responsible, meaning it harms, it produces ethical risk, or it generates a bias in source, it discriminates, for example… it brings with a lot of risk reputationally, legally, legislatively, which entails of course the examples of companies getting massive fines and so on… so, it's not up until the data scientists provide that ownership, it's on a bigger level."*

Governing AI should be a multidisciplinary endeavour (Minkkinen et al., 2023) to ensure that the management of AI technologies is guided by a broad spectrum of ethical perspectives and expertise. As mentioned by Interviewee #12, *"… I think maybe the key could be in the diversity of the team working on it, with regards to looking at potential consequences…"*

Organisations should establish specific roles related to data, AI models, and AI systems to ensure fairness, as humans remain partially responsible for final AI

system decisions—whether made through supervised reliance or proactive oversight (Serban et al., 2020; Teodorescu et al., 2021). In addition to the roles of chief information officers (CIOs), chief data officers (CDOs), and boards of directors in setting strategic directions for implementing AI technologies in organisations, research suggests having executives in the upper echelons with AI experience, or at least possessing the ability to scrutinise AI technologies and their strategic implications (e.g., Li et al., 2021).

To enhance oversight and improve decision-making within organisations, decisions can be brought to an internal review board, to *assess proposals of new AI applications for alignment with ethics principles*, as Interviewee #5 described:

> *"my current company has a set of rules and a group of people who review how we do, and how we use technology... we have a review board... make sure it meets the principles of ethical AI, responsible AI... and we can go talk to someone who's in our legal organisation to say, did we do this the right way..."*

Similarly, Google has deployed a *decision-making process*[22] for their multidisciplinary team to review new AI projects, weighing the nature of the ethical and social benefits involved. In one project using generative adversarial networks (GANs) to create photo-realistic synthetic faces, they chose not to publicly release ML models due to concerns about the potential misuse of these models by malicious actors to produce "deepfakes" aimed at spreading misinformation.

In addition, the AI committee can *promote thoughtful consideration of trade-offs and balanced decision-making*, according to Interviewee #15:

[22] https://ai.google/responsibility/ai-governance-operations/

> *"... something needs to be discussed, and the ethical issue arises, or red flags pop up in the data science team, that can be escalated towards the C-level, to a specific board and ethical board, and then the ethical issues can be tackled... because in the end I had to meet the decision if an AI model or an AI product is responsible enough."*

The second approach includes ***external AI consultation***. It involves seeking advice and expertise from specialised group from outside the organisation on responsible AI projects and strategies. As Interviewee #6, a consultant, explained, *"I go in and out of businesses and give advice, and then sometimes help them implement emerging technologies... I do a lot of work around governance and quality assurance as well..."* Research has also emphasised the importance of conducting independent AI audits that allow for the interrogation of complex processes according to company policy, industry standards, or regulations (Brundage et al., 2020; Mittelstadt et al., 2016).

Organisations, especially smaller ones, need to *seek independent ethics-based advice in setting up or refining AI governance*, as suggested by the consulting experience shared by Interviewee #15,

> *"I consult clients in setting up data and AI governance structures, meaning also some ethical boards or a data governance council... building out AI stewardship, data stewardship, code of conduct, metrics, KPIs, to really measure if the AI model you build complies with those metrics you put forward or with the principles you put forward."*

In addition to seeking advice and expertise, organisations need to *secure formal recognition for ongoing compliance with ethical standards*. This is essential for organisations to manage risks and ensure regulatory compliance in the increasingly complex landscape of AI governance. As Interviewee #7 put:

"I think there's a necessity to have an external eye on it, a third party that can guarantee and levelise the fact that your algorithm is responsible and certified… I think there's a necessity for a third party to intervene in those kinds of topics… we would like to be certified on that transparency…"

4.2.2 Procedural AI governance

This dimension of governance mechanism refers to the policies, standards, processes, and monitoring activities associated with strategic decision-making that ensures AI operates responsibly within organisations. Perceptions of the legitimacy of decisions are likely to improve when fair procedural governance mechanisms are in place, as indicated by ethics (Martin and Waldman, 2023) and legitimacy scholarship (Tyler, 2006). Also, establishing a procedural regularity for validating and verifying the design of AI can help translate abstract principles into technical specifications (Floridi, 2019; Morley et al., 2021). Two theoretical themes were identified.

Regulation of AI. The first theme refers to as the specified responsible AI rules and standard procedures created for organisations to regulate decision-making/monitoring. Scholars (e.g. Mäntymäki et al., 2022; Mittelstadt, 2019) and Interviewee #15 has called for 'bottom-up' approach to organisational AI ethics:

"…to me, it should come bottom-up… if you create something bottom-up, there's a platform that carries the message, and having a bottom-up built standard on ethical usage of AI that industries can use and self-regulate, so themselves have a say in their regulation instead of enforced, I think that's a better way…"

To enhance compliance with AI ethics and achieve strategic objectives, managers need to *create policy packs for self-governance*:

> *"We will tell what the standard is for ethical usage of AI, and we will do that by merging best practices from academia, from legislative sources, and from the industry, and we come up with the best… having it independently, it's will always be better."*

In addition, when implementing and governing AI technologies, managers can *consider social and legal norms* as Interviewee #10 mentioned:

> *"there are some regulations, and we've seen also with such companies as Microsoft and IBM and Google, they have called for it because of their personal regulations… they value their reputation, and they want sometimes to engage in these voluntary regulations… self-regulation is okay, but it's only one component…the other components are structural regulations and political regulations, and society should be also aware of possibilities on the one hand… on the other hand, also stress of AI, so it should be collective work"*

Moreover, managers can *tailor human rights approaches to AI regulation* as Interviewee #13 said:

> *"… the number one thing for me is having human rights due diligence, processes, or impact assessments… and this is something that's also coming back through more and more regulations now, it's really getting kind of centre stage… but for me, that's the most important thing, I mean, overall, it's about businesses respecting human rights as they develop and use AI, but especially the human rights due diligence can help them to detect potential issues with the harmful effects of AI, that they can then try to mitigate, avoid altogether by not developing that AI."*

Koniakou (2023) has also engaged in the discourse of AI governance and argued for the necessity to extend human rights obligations to private entities, such as

companies that develop or utilise AI technologies. This approach is advocated to ensure that the development and application of AI respect and protect human rights, offering a more robust basis for AI governance.

Normative foundations can provide a more structured approach to exploring possible governance practices, as Interviewee #18 stated that:

> *"... ethics is part of a very ancient tradition, and it cannot be done properly without referring to that tradition...and there are many different ways that ethics can contribute to the normative like the governance of AI..."*

In reviewing existing literature, a two-level utilitarian approach (Bauer, 2020; see Appendix D) may provide organisations with a robust framework to effectively address the complex challenges involved in managing AI through actionable governance practices.

The second theoretical theme is ***regulation through AI***, which refers to the use of AI technologies in organisations to automatedly *monitor, enforce, and ensure compliance with regulations*. As with regulatory concerns surrounding the use of AI there is opportunity for it to be involved in the regulatory processes and activities. This approach not only increases the reach and precision of regulatory efforts but also reduces the administrative burden and costs associated with the compliance monitoring and regulatory enforcement. One of our participants commented:

> *"I think that there are several ways of controlling the complexity of AI... you can have a system like...you try to enforce that it's doing some kind of like something you consider good..." (Interviewee #1)*

Evidence can also be found in existing scholarly studies, for example, Akhigbe et al. (2017) mentioned that AI can be used for regulatory compliance to check and

enforce business compliance activities. Thus, organisations could develop tools for the technical realisation of principles guidance.

4.2.3 Relational AI governance

This dimension refers to communication, training, and coordination approaches that facilitate collaborative relationships among stakeholders in the oversight of AI systems. Designers need to grasp the potential of responsible AI design tactics whereas managers must understand, participate in, and support key management activities. Together, they can build a wider dialogue that bridges the gap between 'techies' and 'users'. This study revealed one main theoretical theme.

Distributed deliberation refers to a decision-making process that is spread out across multiple participants, including the general population, citizen experts, and AI experts. While cooperative audits of AI systems have been suggested to address ethical challenges (Mittelstadt, 2019), the literature has mostly focused on expert settings (Buhmann & Fieseler, 2023). Organisations need mechanisms to widen participation, getting citizen insights and social evaluations from outside the industry to govern AI. As Interviewee #17 discussed:

> *"companies should start engaging the public sector, private sector, and do almost like a co-creation of these public policies, only then they will take charge...we need to work with people, and governments, and public, and companies to get a consensus..."*

Different deliberative venues can be explored to elucidate AI governance. First, the functions of *AI expert discourse* are to share best technical practices and deliver assessments or evaluations to wider audience. As suggested by Interviewee #14:

> *"…I know of many volunteer organisations that do all sorts of things, even one of my friends (AI expert) built a mechanism which he put on GitHub, anyone can access it…"*

In addition, the main function of *AI citizen expert* is to contest AI via experiential knowledge. As commented by Interviewee #18, *"I think they should also be able to contest or question that decision, and be given answers about how the machine has come to a conclusion…"*. This group of individuals may not have technical knowledge but have expertise and practical experience in a particular field of responsible AI. They can scrutinise the results, assumptions, and methodology of AI systems to identify potential biases, inaccuracies, or misapplications. As Interviewee #17 described their experience:

> *"…part of my job is to work with the technical geeks, as well as the development professionals or international professionals, to act as a bridge between the people who create new algorithms, new models, new machine learning tools, new AI, and the ones who are in the forefront of communities… I tried to bring them together; that's part of my job."*

Organisations need to set channels to facilitate the discussion according to Interviewee #10:

> *"I think there are quite a lot of such examples in our real world… whistle blow and channel in the company, it's one possible way of regulating this, that people just can discuss there and complain there… if there are some points of discussion at the regulatory level, where people can also discuss them and put them out…open discussion, I think it's the way to go…"*

AI end users can also play a crucial role in providing insight, such as direct feedback on the performance of AI systems. For instance, Interviewee #13 suggested that, *"businesses can also check with individuals, they could conduct surveys to see, okay, did*

the customers or the people affected actually understand this information that they're getting that." They can highlight unexpected issues that might not be apparent in development or testing environments, as Interviewee #6 put:

> *"I think we…as individuals, need to take personal responsibility around those kinds of things… and I think without that, we are always going to be in danger of bias, creeping into systems bias… and designers of these systems potentially having the bias that's unchecked, uncontrolled, and probably unreported."*

4.3 Outcomes of managing AI

While significant attention has been devoted to the responsible design and governance of AI, a small yet expanding body of research is beginning to investigate its impact on organisational outcomes. From a sociotechnical perspective, managing AI effectively involves understanding two possible values of organisational outcomes: instrumental and humanistic (Beath et al., 2013; Sarker et al., 2019). These outcomes emerge from the interaction between the technical artifacts (i.e., AI) and the social contexts in which individuals and collectives develop and use AI. Joint optimisation of both technical and social components is likely to generate better outcomes (Wallace et al., 2004).

4.3.1 Instrumental outcomes

This theoretical theme refers to practical, measurable results that are often linked to achieving broader strategic objectives. In the context of managing AI, instrumental outcomes include *better system performance*, *higher work efficiency*, and *economic profitability*.

Managing AI by responsible design helps improve system performance and therefore meet end users' need as Interviewee #1 commented:

> *"…just by implementing the explainability in that system, it's already helped you find out the technical problems… by acting on making our systems explainable, you get these benefits of the expectations of the users are much better met, because if you have no explanations, then you would get something like the users doesn't know what is, why it's not working…"*

It can also help employee make informed decisions and improve their *work efficiency* as Interviewee #8 explained:

> *"… by compliance, what you're getting is a standardised predictable behaviour… that's the biggest advantage, when you comply, you're becoming predictable… just because of the technological capability, and the gains that you will have with time, like whatever you were doing previously in years, you could now do in minutes and seconds… that's a big gain… with that additional time, there's a lot more you could do, you could invest that in many more capabilities to do more good…."*

Through improving the AI system performance and human work efficiency, *economic profitability* such as higher profits and external investments, can be brought to organisations as Interviewee #13 said:

> *"… I worked for the company that develops AI, and they have AI embedded like throughout the whole value of the business, so their whole point and the purpose of the business is to make people-powered AI, which is basically responsible AI, so that's really what they're about… and I've seen that having positive effects for them in terms of…investments, so investors are now getting really interested in companies that are taking responsible AI seriously…"*

4.3.2 Humanistic outcomes

This theoretical theme pertains to the enhancement of the well-being, satisfaction, and empowerment of individuals. In the context of managing AI within organisations, humanistic outcomes are *job satisfaction* and *employee engagement*. As according to Interviewee #10:

> *"...if a company has responsible ways of working, they will get also good employees who stick to those values, because really good employees are when they have a choice of where to work, then they will look for ethical companies, for responsible companies, because they want to have a clear mind and clear consciousness in order to work in a good company."*

Evidence can also be found in literature, for example, by promoting and ensuring transparency of data collection, processing, and use throughout AI project lifecycles, responsible AI design and governance are argued to improve knowledge flows among organisational units, thus enhancing inter-departmental collaboration (Rantane et al., 2021).

Sarker et al. (2019) suggested that "instrumental and humanistic outcomes can form a virtuous cycle wherein both are synergistically connected" (pp. 710). Organisations pursuing humanistic outcomes would encourage more positive actions among managers and designers, and this can lead to feedback to create more instrumental outcomes.

5. General Discussion

5.1 Theoretical contributions

This study provides several important contributions to the theory. First, it advances the nascent literature on managing AI (Berente et al., 2021) and the body of knowledge on implementing AI ethics in practice (Georgieva et al., 2022; Morley et al., 2020). By introducing three constitutive responsibilities that address distinct AI challenges (Buhmann & Fieseler, 2023; Tsamados et al., 2022), this study conceptualises and integrates responsible AI design (Benjamins et al., 2019) and governance (Schneider et al., 2023) as actionable strategies for organisations to manage AI responsibly. Additionally, it builds bridges to the AI governance literature by highlighting the role of governance mechanisms in enforcing these responsibilities. These perspectives extend current AI ethics discourse and offer novel insights into managing AI in organisations. The study also provides empirical support for the proposition that responsible AI management can drive positive organisational outcomes.

Second, this study advances the sociotechnical perspective on AI management by emphasising ethical and inclusive AI use. Responding to calls to recommit to the sociotechnical perspective in the IS field (Sarker et al., 2019), it explores responsible AI design and governance as both distinctive and interconnected elements. This perspective is relevant not only to philosophy and social sciences scholars exploring AI ethics and governance, but also to technical experts responsible for designing and deploying AI systems responsibly. The findings are grounded in diverse perspectives from different AI stakeholders, enriching the understanding of responsible AI management and offering theoretical implications about the

positioning of these actors. Engaging private stakeholders in AI governance draws on participatory democracy, promoting inclusive decision-making in AI development and deployment.

Third, the study highlights the need to go beyond analysing responsible AI design and governance mechanisms to uncover the normative foundations underlying AI management. This approach informs future scholarly efforts to define AI management concepts and articulate practical principles of responsible AI design and governance. It supports a 'bottom-up' approach to AI ethics in the private sector and transforms professional AI ethics into organisational ethics (Mittelstadt, 2019). A pluralistic and theoretically informed approach, drawing from ethical perspectives, offers a nuanced framework for responsible AI management in organisations. Aligning on some crucial and fundamental moral principles will help organisations adapt to new challenges brought by technological advancements and identify best practices, such as a two-level utilitarian approach, in the future.

5.2 Practical implications

This research has implications for practice. Drawing on findings regarding responsible AI design and governance, the study provides strategic considerations for formulating and implementing strategies for managing AI. While governments, non-profits, and tech giants develop ethical AI frameworks, organisations that use or develop AI systems should also adopt responsible actions to manage AI frontiers, gaining benefits beyond mere legal compliance, such as enhanced reputation.

Organisations seeking to leverage AI often find limited guidance in existing IS research on effective and responsible AI management. It is increasingly recognised that AI

designers and managers must be accountable for the intended and unintended consequences of AI. However, organisational AI governance remains inconsistent across industries, complicating efforts to address the ethical and societal issues posed by AI. The proposed model identifies two key facets for effective AI management: design responsibilities and governance mechanisms. These provide a structured approach for organisations to navigate AI-related ethical and societal challenges. The model offers actionable recommendations for AI designers and managers to implement responsible practices and predict organisational performance outcomes. The findings can help organisations assess their current AI-related practices and provide insights for policymakers to enhance AI governance. Additionally, organisations are advised to develop complementary capabilities to effectively leverage AI, aligning it with business needs while reinforcing their ethical and operational responsibilities.

5.3 Limitations and future research

This study has several limitations that could be addressed by further research. First, this study adopts a broad, cross-sectoral perspective, which, while beneficial for identifying overarching themes, may lack the depth required to fully understand industry-specific nuances. By not restricting the investigation to particular sectors, this study captures shared best practices and implementation strategies across disciplines. However, this approach limits the exploration of sector-specific complexities, regulatory environments, and domain-specific responsible AI applications. Future research should conduct industry-specific case studies, particularly in sectors such as healthcare and finance, to provide deeper insights into domain-specific responsible AI challenges, regulatory adaptations, and ethical dilemmas unique to each industry.

Second, qualitative research inherently relies on subjective experiences, which, while rich in context, may not always be generalisable across industries or geographical regions. The study's findings are context-dependent, shaped by the perspectives of the selected participants, and may not fully capture the diversity of AI design and governance practices. Additionally, this study relies heavily on participant interviews over archival data, as the focus is on capturing micro-level perspectives of AI management. However, the small sample size may not fully represent the broader AI community. Furthermore, this study primarily examines managerial viewpoints, excluding end-user perspectives, which are critical for understanding trust-building in responsible AI, addressing information asymmetry, and ensuring data privacy. Future research should explore different stakeholder perspectives through extensive interviews and observations to assess whether the grounded model developed here can comprehensively explain AI management practices.

Third, the study predominantly involved informants from Western countries, specifically the United Kingdom (UK), the United States (US), and the European Union (EU), due to the convenience of participant recruitment in these regions. This geographical limitation underscores the need for comparative cross-cultural studies, as responsible AI practices vary based on cultural, legal, and economic factors. A study comparing responsible AI implementation across different regions would provide insights into how AI ethics and governance are shaped by local regulatory landscapes, cultural values, and societal expectations, helping to develop globally adaptable AI management frameworks.

Additionally, the lack of quantitative validation limits the generalisability of the results across different sectors and regions. Future research could adopt a mixed-methods approach, integrating large-scale surveys to measure the prevalence and impact of responsible AI design and governance practices. As AI regulations, design tactics, and governance frameworks continue to evolve, this study reflects only a specific point in time. Future developments may introduce ethical considerations and best practices not accounted for in this research. A longitudinal study tracking how organisations adapt to evolving AI management frameworks over time would offer valuable insights into policy effectiveness, corporate adaptation, and trends in ethical AI adoption, providing a dynamic perspective on the implementation of responsible AI principles.

Chapter III References

Ågerfalk, P. J. (2020). Artificial intelligence as digital agency. *European Journal of Information Systems*, 29(1), 1–8. https://doi.org/10.1080/0960085X.2020.1721947

Agarwal, A., et al. (2018). A reductions approach to fair classification. *Proceedings of the 35th International Conference on Machine Learning, 80,* 60–69.

Agrawal, A., Gans, J., & Goldfarb, A. (2019). *The economics of artificial intelligence: An agenda.* Chicago: University of Chicago Press.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: imitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society, 20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Arnold, T. H., & Scheutz, M. (2018). The "big red button" is too late: An alternative model for the ethical evaluation of AI systems. *Ethics and Information Technology, 20*(4), 59–69. https://doi.org/10.1016/10.1007/s10676-018-9447-7

Arrieta, A. B., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58,* 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Asatiani, A., et al. (2021). Sociotechnical envelopment of artificial intelligence: An approach to organizational deployment of inscrutable artificial intelligence systems. *Journal of the Association for Information Systems, 22*(2), 325–352. https://doi.org/10.17705/1jais.00664

Baird, A., & Maruping, L. M. (2021). The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. *MIS Quarterly, 45*(1), 315–341. https://doi.org/10.25300/MISQ/2021/15882

Bauer, W. A. (2020). Virtuous vs. utilitarian artificial moral agent. *AI & Society, 35,* 263–271. https://doi.org/10.1007/s00146-018-0871-3

Beath, C., et al. (2013). Expanding the frontiers of information systems research: Introduction to the special issue. *Journal of the Association for Information Systems, 14*(4), i–xvi. https://doi.org/10.17705/1jais.00330

Benbya, H., et al. (2020). Complexity and information systems research in the emerging digital world. *MIS Quarterly, 44*(1), 1–17. https://doi.org/10.25300/MISQ/2020/13304

Benbya, H., Pachidi, S., & Jarvenpaa, S. (2021). Special issue editorial: Artificial intelligence in organizations: Implications for information systems research. *Journal of the Association for Information Systems, 22*(2), 10. https://doi.org/10.17705/1jais.00662

Benjamins, R., Barbado, A., & Sierra, D. (2019). Responsible AI by design in practice. *arXiv preprint.* arXiv: 1909.12838. https://arxiv.org/abs/1909.12838

Berente, N., et al. (2021). Managing artificial intelligence. *MIS Quarterly, 45*(3), 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Birks, D. F., et al. (2013). Grounded theory method in information systems research: Its nature, diversity, and opportunities. *European Journal of Information Systems, 22*, 1–8. https://doi.org/10.1057/ejis.2012.48

Bolton, W. J., et al. (2022). Developing moral AI to support decision-making about antimicrobial use. *Nature Machine Intelligence, 4*, 912–915. https://doi.org/10.1038/s42256-022-00558-5

Buhmann, A., & Fieseler, C. (2021). Towards a deliberative framework for responsible innovation in artificial intelligence. *Technology in Society, 64*, 101475. https://doi.org/10.1016/j.techsoc.2020.101475

Buhmann, A., & Fieseler, C. (2023). Deep learning meets deep democracy: Deliberative governance and responsible innovation in artificial intelligence. *Business Ethics Quarterly, 33*(1), 146–179. https://doi.org/10.1017/beq.2021.42

Buhmann, A., Paßmann, J., & Fieseler, C. (2020). Managing algorithmic accountability: Balancing reputational concerns, engagement strategies, and the potential of rational discourse. *Journal of Business Ethics, 163*(2), 265–280. https://doi.org/10.1007/s10551-019-04226-4

Butcher, J., & Beridze, I. (2019). What is the state of artificial intelligence governance globally? *The RUSI Journal, 164*, 88–96. https://doi.org/10.1080/03071847.2019.1694260

Butler, T., Gozman, D., & Lyytinen, K. (2023). The regulation of and through information technology: Towards a conceptual ontology for IS research. *Journal of Information Technology, 38*(2), 86–107. https://doi.org/10.1177/02683962231181147

Brundage, M., et al. (2020). Toward trustworthy AI development: Mechanisms for supporting verifiable claims. *arXiv preprint,* arXiv:2004.07213. https://doi.org/10.48550/arXiv.2004.07213

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis.* Thousand Oaks, CA: Sage.

Chau, D. C., et al. (2020). The effects of business-IT strategic alignment and IT governance on firm performance: A moderated polynomial regression analysis. *MIS Quarterly, 44*(4), 1679–1703. https://doi.org/10.25300/MISQ/2020/12165

Cihon, P., et al. (2021). AI certification: Advancing ethical practice by reducing information asymmetries. *IEEE Transactions on Technology and Society, 2*(4), 200–209. https://doi.org/10.1109/TTS.2021.3077595

Cihon, P., Schuett, J., & Baum, S. D. (2021). Corporate governance of artificial intelligence in the public interest. *Information, 12*(275). https://doi.org/10.3390/info12070275

Clarke, R. (2019). Principles and business processes for responsible AI. *Computer Law & Security Review, 35,* 410–422. https://doi.org/10.1016/j.clsr.2019.04.007

Corbin, J. M., & Strauss, A. (2015). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (4th ed.). Thousand Oaks, CA: Sage.

Davenport, T., et al. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science, 48,* 24–42. https://doi.org/10.1007/s11747-019-00696-0

Deng, W. H., et al. (2022). Exploring how machine learning practitioners (try to) use fairness toolkits. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency,* 473–484.

Dennehy, D., et al. (2023). Artificial intelligence (AI) and information systems: Perspectives to responsible AI. *Information Systems Frontiers, 25*(1), 1–7. https://doi.org/10.1007/s10796-022-10365-3

Denton, E., et al. (2020). Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv preprint,* arXiv:2112.04554. https://doi.org/10.48550/arXiv.2007.07399

Denton, E., et al. (2021). Whose ground truth? Accounting for individual and collective identities underlying dataset annotation. *arXiv preprint,* arXiv:2112.04554. https://doi.org/10.48550/arXiv.2112.04554

Díaz-Rodríguez, N., et al. (2023). Connecting the dots in trustworthy artificial intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion, 99,* 101896. https://doi.org/10.1016/j.inffus.2023.101896

Ding, F., Li, D., & George, J. F. (2014). Investigating the effects of IS strategic leadership on organizational benefits from the perspective of CIO strategic role. *Information & Management, 51*(7), 865–879. https://doi.org/10.1016/j.im.2014.08.004

Drucker, P. E. (2008). *Management* (Rev. Ed.). New York: HarperCollins.

Dwivedi, Y. K., et al. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice, and policy. *International Journal of Information Management, 57,* 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Eitel-Porter, R. (2021). Beyond the promise: Implementing ethical AI. *AI Ethics, 1*, 73–80. https://doi.org/10.1007/s43681-020-00011-6

Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization, 28*(1), 62–70. https://doi.org/10.1016/j.infoandorg.2018.02.005

Fischer, J. M. (1999). Recent work on moral responsibility. *Ethics, 110*, 93–139.

Fjeld, J., et al. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center for Internet & Society. 1. http://dx.doi.org/10.2139/ssrn.3518482*

Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence, 1*(6), 261–262. https://doi.org/10.1038/s42256-019-0055-y

Floridi, L. (2018). Soft ethics, the governance of the digital, and the general data protection regulation. *Philosophical Transactions of the Royal Society A, 376*(2133). https://doi.org/10.1098/rsta.2018.0081

Gahnberg, C. (2021). What rules? Framing the governance of artificial agency. *Policy and Society, 40*(2), 194–210. https://doi.org/10.1080/14494035.2021.1929729

Gal, U., Hansen, S., & Lee, A. S. (2021). Towards theoretical rigor in ethical analysis: The case of algorithmic decision-making systems. *Journal of the Association for Information Systems, 21*(1), 1634–1661. https://doi.org/10.17705/1jais.00784

Georgieva, I., et al. (2022). From AI ethics principles to data science practice: A reflection and a gap analysis based on recent frameworks and practical experience. *AI Ethics, 2,* 697–711. https://doi.org/10.1007/s43681-021-00127-3

Gioia, D. A. (2021). A systematic methodology for doing qualitative research. *The Journal of Applied Behavioral Science, 57*(1), 20–29. https://doi.org/10.1177/0021886320982715

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. *Organizational Research Methods, 16*(1), 15–31. https://doi.org/10.1177/1094428112452151

Glaser, B. (2016). Open coding descriptions. *Grounded Theory Review: An International Journal, 15*(2).

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Mill Valley, CA: Sociology Press.

Gomes, C., et al. (2019). Computational sustainability: Computing for a better world and a sustainable future. *Communications of the ACM, 62*(9), 56–65. https://doi.org/10.1145/3339399

Grodal, S., Anteby, M., & Holm, A. L. (2021). Achieving rigor in qualitative analysis: The role of active categorization in theory building. *Academy of Management Review, 46*(3), 591–612. https://doi.org/10.5465/amr.2018.0482

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Henningsson, S., & Eaton, B. D. (2023). Governmental regulation and digital infrastructure innovation: The mediating role of modular architecture. *Journal of Information Technology, 38*(2), 126–143. https://doi.org/10.1177/0268396222111442

Hevner, A. R., et al. (2004). Design science in information system research. *MIS Quarterly, 28*(2), 75–105. https://doi.org/10.2307/25148625

Hickok, M. (2021). Lessons learned from AI ethics principles for future actions. *AI and Ethics,* 41–47. https://doi.org/10.1007/s43681-020-00008-1

Hirschman, E. C. (1986). Humanistic inquiry in marketing research: Philosophy, method, and criteria. *Journal of Marketing Research, 23*(3), 237–249. https://doi.org/10.2307/3151482

Ho, C., et al. (2019). Governance of automated image analysis and artificial intelligence analytics in healthcare. *Clinical Radiology, 74*(5), 329–337. https://doi.org/10.1016/j.crad.2019.02.005

Hoffman, M., Kahn, L., & Li, D. (2018). Discretion in hiring. *The Quarterly Journal of Economics, 133*(2), 765–800. https://doi.org/10.1093/qje/qjx042

Holstein, K., et al. (2019). Improving fairness in machine learning systems: What do industry practitioners need? *Conference on Human Factors in Computing Systems Proceedings.* Association for Computing Machinery.

Huang, M. H., & Rust, R. T. (2021). Engaged to a robot? The role of AI in service. *Journal of Service Research, 24*(1), 30–41. https://doi.org/10.1177/1094670520902266

Huang, M. H., & Rust, R. T. (2023). EXPRESS: The caring machine: Feeling AI for customer care. *Journal of Marketing.* https://doi.org/10.1177/00222429231224748

Jain, H., et al. (2021). Editorial for the special section on humans, algorithms, and augmented intelligence: The future of work, organizations, and society. *Information Systems Research, 32*(3), 675–687. https://doi.org/10.1287/isre.2021.1046

Jarvenpaa, S. L., & Ives, B. (1991). Executive involvement and participation in the management of information technology. *MIS Quarterly, 15*(2), 205–227. https://doi.org/10.2307/249382

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

Jones, R., & Noble, G. (2007). Grounded theory and management research: A lack of integrity? *Qualitative Research in Organizations and Management, 2*(2), 84–103. https://doi.org/10.1108/17465640710778502

Jussupow, E., et al. (2021). Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research, 32*(3), 713–735. https://doi.org/10.1287/isre.2020.0980

Kant, I. (2002). *Groundwork for the Metaphysics of Morals.* New Haven: Yale University Press.

Keding, C. (2021). Understanding the interplay of artificial intelligence and strategic management: Four decades of research in review. *Management Review Quarterly, 71*(1), 91–134. https://doi.org/10.1007/s11301-020-00181-x

Kish-Gephart, J. J., Harrison, D. A., & Treviño, L. K. (2010). Bad apples, bad cases, and bad barrels: Meta-analytic evidence about sources of unethical decisions at work. *Journal of Applied Psychology, 95*(1), 1–31. https://doi.org/10.1037/a0017103

Koniakou, V. (2023). From the "rush to ethics" to the "race for governance" in artificial intelligence. *Information Systems Frontiers, 25*(1). https://doi.org/10.1007/s10796-022-10300-6

Kopalle, P. K., et al. (2022). Examining artificial intelligence (AI) technologies in marketing via a global lens: Current trends and future research opportunities. *International Journal of Research in Marketing, 39*(2), 522–540. https://doi.org/10.1016/j.ijresmar.2021.11.002

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

Lebovitz, S., Levina, N., & Lifshitz-Assaf, H. (2021). Is AI grounded truth really true? The dangers of training and evaluating AI tools based on experts' know-what. *MIS Quarterly, 45*(3), 1501–1525. https://doi.org/10.25300/MISQ/2021/16564

Lee, M. K., et al. (2020). Human-centered approaches to fair and responsible AI. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems,* 1–8.

Li, J., et al. (2021). Strategic directions for AI: The role of CIOs and boards of directors. *MIS Quarterly, 45*(3), 1603–1643. https://doi.org/10.25300/MISQ/2021/16523

Lincoln, Y., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage.

MacIntyre, A. (2013). *After virtue: A study in moral theory*. London, UK: Bloomsbury.

Madaio, M., et al. (2022). Assessing the fairness of AI systems: AI practitioners' processes, challenges, and needs for support. *Proceedings of the ACM on Human-Computer Interaction, 6*, CSCW1, 1–26.

Mäntymäki, M., et al. (2022). Defining organizational AI governance. *AI and Ethics, 2*(4), 603–609. http://dx.doi.org/10.2139/ssrn.4553185

Marabelli, M., Newell, S., & Handunge, V. (2021). The lifecycle of algorithmic decision-making systems: Organizational choices and ethical challenges. *The Journal of Strategic Information Systems, 30*(3), 101683. https://doi.org/10.1016/j.jsis.2021.101683

Martin, K. (2019). Ethical implications and accountability of algorithms. *Journal of Business Ethics, 160*(4), 835–850. http://dx.doi.org/10.2139/ssrn.3056716

Martin, K., & Waldman, A. (2023). Are algorithmic decisions legitimate? The effect of process and outcomes on perceptions of legitimacy of AI decisions. *Journal of Business Ethics, 183*, 653–670. https://doi.org/10.1007/s10551-021-05032-7

McCorduck, P. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence* (2nd ed.). Natick, MA: A. K. Peters, Ltd.

Mikalef, P., et al. (2022). Thinking responsibly about responsible AI and "the dark side" of AI. *European Journal of Information Systems, 31*(3), 257–268. https://doi.org/10.1080/0960085X.2022.2026621

Mingers, J., & Walsham, G. (2010). Toward ethical information systems: The contribution of discourse ethics. *MIS Quarterly, 34*(4), 833–854. https://doi.org/10.2307/25750707

Minkkinen, M., Zimmer, M. P., & Mäntymäki, M. (2023). Co-shaping an ecosystem for responsible AI: Five types of expectation work in response to a technological frame. *Information Systems Frontiers, 25*(1). https://doi.org/10.1007/s10796-022-10269-2

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*, 501–507. https://doi.org/10.1038/s42256-019-0114-4

Mittelstadt, B. D., et al. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2). https://doi.org/10.1177/2053951716679679

Morley, J., et al. (2021). Ethics as a service: A pragmatic operationalisation of AI ethics. *Minds and Machines, 31*(2), 239–256. https://doi.org/10.1007/s11023-021-09563-w

Morley, J., et al. (2020). From what to how: An initial review of publicly available AI ethics tools, methods, and research to translate principles into practices. *Science and Engineering Ethics, 26*(4), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Myers, M. D. (2009). *Qualitative research in business & management* (3rd ed.). Thousand Oaks, CA: Sage.

Paradice, D., et al. (2018). A review of ethical issue considerations in the information systems research literature. *Foundations and Trends in Information Systems, 2*(2), 117–236. https://doi.org/10.1561/2900000012

Patton, M. Q. (2002). *Qualitative research & evaluation methods.* Thousand Oaks, CA: Sage.

Paul, E. F., Miller, F. D. M., & Paul, J. (1999). *Responsibility.* Cambridge University Press.

Perrault, R., et al. (2019). *Artificial Intelligence Index 2019 Annual Report.* Stanford, CA: AI Index Steering Committee, Human-Centered AI Institute, Stanford University.

Peters, D., et al. (2020). Responsible AI—two frameworks for ethical design practice. *IEEE Transactions on Technology and Society, 1*(1), 34–47. https://doi.org/10.17863/CAM.49394

Pratt, M. G., Sonenshein, S., & Feldman, M. S. (2020). Moving beyond templates: A bricolage approach to conducting trustworthy qualitative research. *Organizational Research Methods,* 1–28. https://doi.org/10.1177/1094428120927466

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science, 48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Rakova, B., et al. (2021). Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction, 5*, CSCW1, 1–23.

Ricoeur, P. (1992). *Oneself as another* (K. Blamey, Trans.). Chicago: University of Chicago Press.

Saldaña, J. (2021). *The coding manual for qualitative researchers* (4th ed.). Thousand Oaks, CA: Sage.

Sarker, S., et al. (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

Saunders, B., et al. (2018). Saturation in qualitative research: Exploring its conceptualization and operationalization. *Quality & Quantity, 52*(4), 1893–1907. https://doi.org/10.1007/s11135-017-0574-8

Serban, A., et al. (2020). Adoption and effects of software engineering best practices in machine learning. *Proceedings of the 14th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM),* 1–22. https://doi.org/10.1145/3382494.3410681

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

Schiff, D., et al. (2021). AI ethics in the public, private, and NGO sectors: A review of a global document collection. *IEEE Transactions on Technology and Society, 2*(1), 31–42. https://doi.org/10.1109/TTS.2021.3052127

Schneider, J., et al. (2023). Artificial intelligence governance for businesses. *Information Systems Management, 40*(3), 229–249. https://doi.org/10.1080/10580530.2022.2085825

Senoner, J., Netland, T., & Feuerriegel, S. (2022). Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing. *Management Science, 68*(8), 5704–5723. https://doi.org/10.1287/mnsc.2021.4190

Solarino, A. M., & Aguinis, H. (2021). Challenges and best-practice recommendations for designing and conducting interviews with elite informants. *Journal of Management Studies, 58*(3), 649–672. https://doi.org/10.1111/joms.12620

Stahl, B. C. (2023). Embedding responsibility in intelligent systems: From AI ethics to responsible AI ecosystems. *Scientific Reports, 13*(1), 7586. https://doi.org/10.1038/s41598-023-34622-w

Starks, H., & Trinidad, S. B. (2007). Choose your method: A comparison of phenomenology, discourse analysis, and grounded theory. *Qualitative Health Research, 17*(10), 1372–1380. https://doi.org/10.1177/1049732307307031

Strauss, A., & Corbin, J. M. (1990). *Basics of qualitative research: Grounded theory procedures and techniques.* Thousand Oaks, CA: Sage.

Teodorescu, M. H. M., et al. (2021). Failures of fairness in automation require a deeper understanding of human-ML augmentation. *MIS Quarterly, 45*(3). https://doi.org/10.25300/MISQ/2021/16535

Tsamados, A., et al. (2022). The ethics of algorithms: Key problems and solutions. *AI & Society, 37,* 215–230. https://doi.org/10.1007/s00146-021-01154-8

Tyler, T. R. (2006). Restorative justice and procedural justice: Dealing with rule breaking. *Journal of Social Issues, 62*(2), 307–326. https://doi.org/10.1111/j.1540-4560.2006.00452.x

Urquhart, C. (2013). *Grounded theory for qualitative research.* Thousand Oaks, CA: Sage.

Wang, Q., et al. (2023). Designing responsible AI: Adaptations of UX practice to meet responsible AI challenges. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23).* Hamburg, Germany: ACM. https://doi.org/10.1145/3544548.3581278

Wang, Y., Xiong, M., & Olya, H. (2020). Toward an understanding of responsible artificial intelligence practices. *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS),* Maui, Hawaii, 4962–4971.

Wei, K., et al. (2020). Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security, 15,* 3454–3469. https://doi.org/10.1109/TIFS.2020.2988575

Werder, K., Ramesh, B., & Zhang, R. (2022). Establishing data provenance for responsible artificial intelligence systems. *ACM Transactions on Management Information Systems, 13*(2), Article 22. https://doi.org/10.1145/3503488

Wiener, N. (1954). *The human use of human beings* (2nd ed.). New York, NY: Doubleday Anchor Books.

Wiesche, M., et al. (2017). Grounded theory methodology in information systems research. *MIS Quarterly, 41*(3), 685–702. https://doi.org/10.25300/MISQ/2017/41.3.02

Zhu, T., et al. (2020). More than privacy: Applying differential privacy in key areas of artificial intelligence. *IEEE Transactions on Knowledge and Data Engineering, 34*(6), 2824–2843. https://doi.org/10.1109/TKDE.2020.3014246

Chapter III Appendix A: List of Interviews

Table III.A1 The profiles of interviewees

Interviewees	Based in	Age	Gender	Education	Job title	Duration of the Interview
Interviewee #1	United Kingdom	31	Male	PhD	Data Scientist, Head of AI	59:50
Interviewee #2	United Kingdom	31	Male	PhD	Researcher in Affective Intelligence and Robotics	30:33
Interviewee #3	United Kingdom	27	Male	PhD	Researcher in Machine Learning and Deep Learning	34:09
Interviewee #4	France	32	Male	Master	Data Scientist	36:45
Interviewee #5	United States	57	Male	Master, MBA	Principal AI Architect	36:29
Interviewee #6	United Kingdom	50	Male	PhD	CEO & Founder	47:47
Interviewee #7	France	49	Male	Master	President & Co-Founder	33:39
Interviewee #8	United States	47	Male	Master	Co-Founder	37:28
Interviewee #9	United States	33	Male	Master	Co-Founder & Chief Product Officer	27:31
Interviewee #10	Germany	30	Male	Master	Consultant	30:19
Interviewee #11	United Kingdom	28	Male	Master	Chief Technology Officer	51:13
Interviewee #12	Germany	42	Female	Master	Business Strategy Manager	35:06
Interviewee #13	Netherlands	31	Female	PhD	Professor in Law, Human Rights, and AI	33:04
Interviewee #14	United States	63	Male	Master	Principal Investigator, Technologist	43:45
Interviewee #15	Belgium	34	Male	Master	Senior Data Governance Expert	42:58
Interviewee #16	Belgium	27	Male	Master	AI Governance Director	28:15
Interviewee #17	United Kingdom	42	Male	Master	Chief of Party, Senior Technical Expert	37:17
Interviewee #18	Canada	37	Female	PhD	Professor in AI Ethics and AI Governance	40:28
Interviewee #19	Spain	41	Male	Master	Lead Data Scientist	57:18
Interviewee #20	United States	43	Male	Bachelor	Director of Data Science & Engineering	38:59

Chapter III Appendix B Interview Guide

Table III.B1 An example interview guide for AI designers

Interview stages/topics	Main and prompting questions
Introduction	1) Greetings and talk through the research idea and design. 2) Emphasise the confidentiality. a) No identifiable information that either I have already known about you, or you might be mentioning in our conversations will be accessible form the public. b) This identifiable information will be deleted as early as possible. 3) Ask for consent to be recorded. 4) Ask if they have read the explanations/examples about responsible AI that I send them via email a) If yes, start the interview b) If not, briefly shared some news on responsible/irresponsible use of AI 5) Ask if there are any questions to be addressed before interview. 6) Thank them for their time and support for this research.
Part 1: Opening questions to build rapports	1) Could you tell me a few things about your role in your company? a) What does this role entail? b) What are your main responsibilities?
Part 2: Understanding of ethical/responsible AI in practice	*Goal*: Understand the concepts, scope, foundational principles of responsible AI in real-world setting.
	1) What does responsible AI mean to you? 2) Can you explain the core principles of responsible AI in practice according to your understanding/experiences? 3) Why is responsible AI critical to AI adoption, why is it important for you?
Part 3: Experience with ethical/responsible AI design	*Goal*: Understand practical experience in embedding responsible AI principles into system design.
	Screening question: 1) Have you considered incorporating responsible AI principles into AI system design?

	For interviewees with experience: a) What frameworks or tools have you used to ensure your AI models are responsible? (e.g., transparency) b) How did you ensure they are effective? What processes do you follow to monitor AI models after deployment? c) Can you describe a time when you identified and mitigated an ethical issues in an AI project? *Possible questions following conversations:* a) How do you identify and address biases in datasets during development? Have you ever faced a situation where an AI model exhibited bias? b) How do you balance accuracy with fairness in AI models? c) How do you ensure AI models are interpretable and explainable? What challenges have you faced? d) How would you handle sensitive or personally identifiable information in an AI project? e) How do you involve end-users and stakeholders in AI design to ensure inclusivity and accessibility? f) What measures would you implement to prevent unintended consequences of AI models? g) How do you deal with the AI mistakes? *For interviewees without direct experience:* a) Have your company considered incorporating the use of AI into CSR strategies? i) How does your company manage to do that? ii) What do you think about it? Where do you think could be improved? iii) How does your company communicate these initiatives with stakeholders? (*whether they are explainable, reliable, transparent*) b) What is your suggestion for organisations in the future?
Part 3: Emerging trends	1) What recent developments in responsible AI are you most excited about? 2) How do you stay updated with trends and best practices?
End	1) Do you have something to add that we did not cover in our conversations? 2) Do you have anyone you would recommend me to interview for getting more insights on responsible AI, if possible? a) Can I get their contacts? 3) Ask interviewee to share some demographic information. Age, cultural background, education level etc.

Table III.B2 An example interview guide for AI managers

Interview stages/topics	Main and prompting questions
Introduction	1) Greetings and talk through the research idea and design. 2) Emphasise the confidentiality. a) No identifiable information that either I have already known about you, or you might be mentioning in our conversations will be accessible form the public. *b)* This identifiable information will be deleted as early as possible. 3) Ask for consent to be recorded. 4) Ask if they have read the explanations/examples about responsible AI that I send them via email a) If yes, start the interview b) If not, briefly shared some news on responsible/irresponsible use of AI 5) Ask if there are any questions to be addressed before interview. 6) Thank them for their time and support for this research.
Part 1: Opening questions to build rapports	1) Could you tell me a few things about your role in your company? a) What does this role entail? b) What are your main responsibilities? 2) What are the roles of AI in your company? a) How does AI perform? i) How do you ensure that AI is performed as intended? ii) Where do you think could be improved? iii) How do you manage the potential performance risk?
Part 2: Understanding of ethical/responsible AI in practice	*Goal*: Understand the AI management, governance, and policy implementation in real-world setting.
	1) What does responsible AI mean to you? 2) Can you explain what is an AI governance framework according to your understanding/experiences?
Part 3: Experience with responsible AI management/ governance	*Goal*: Understand practical experience in managing/governing AI.
	Screening question: 1) Have you considered incorporating some responsible AI principles into system design?
	For interviewees with experience (if yes): a) How do you ensure compliance with AI regulations such as GDPR, CCPA, or AI Act? What types of ethical rules or principles? b) How do you establish accountability for AI decisions? c) Describe how you've implemented governance processes for an AI project. d) Describe a time when you collaborated with legal, compliance, or other teams on an AI-related project. (if have) e) What are the key risks associated with deploying AI systems, and how do you address them? *Possible questions following conversations:* a) How do you stay in control of a complex AI system?

	b) How do you monitor models for issues like drift, fairness degradation, or performance loss after deployment? c) How do you conduct a risk assessment for an AI project? d) How do you ensure AI systems are regularly reevaluated, audited for compliance and ethical standards? e) How do you communicate AI risks and decisions to non-technical stakeholders? *For interviewees without direct experience (if no):* a) Can you describe a governance structure for responsible AI in an enterprise setting? b) What are the key risks associated with deploying AI systems, and how can we address them? c) What regulatory frameworks or standards are shaping the responsible AI globally? d) Are you seeing consensus on AI principles? Or do you notice any disagreements?
Part 4: Understanding about responsible AI from economic, societal, ethical perspective	*Potential questions if they have not been covered:* 1) Economic perspective: a) How has AI impacted your company's productivity, efficiency, and profitability? b) What are the benefits and risks of AI for your company? c) Do you see AI contributing economic inequality? 2) Societal perspective: a) Has your company used AI to promote social sustainability or human well-being? b) What kinds of AI-driven social initiatives? c) How does AI affect consumer trust and public perceptions? d) What kinds of benefits could be created? 3) Ethical perspective: a) What are the biggest ethical concerns in AI, and how do they impact your company, organisations, and society? b) How did you address those unintended consequences? c) By addressing them, what kinds of benefits could be created?
Part 3: Emerging trends	1) How do you think about releasing laws rather than voluntary ethical rules? 2) What recent innovations or tools in responsible AI governance excite you, and why?
End	1) Do you have something to add that we did not cover in our conversations? 2) Do you have anyone you would recommend me to interview for getting more insights on responsible AI, if possible? a) Can I get their contacts? 3) Ask interviewee to share some demographic information. Age, cultural background, education level etc.

Note. These questions were designed to facilitate discussion, meaning that not all of them will be covered in a single interview but will be addressed across multiple conversations as much as possible. However, as responsible AI is a relatively new concept, it may not be fully explored. While interviewees may not have had detailed experience, they shared some of their opinions on the topics.

Chapter III Appendix C: Selected Indicative Codes and Quotes

Table III.C1 Selected indicative codes and quotes for AI design

Aggregate dimension: **Evidence responsibility**		
Second-order themes	*First-order concepts*	*Exemplar quotes*
Using the right data/algorithm	Take appropriate measures to ensure data/algorithm quality	"Assuring that your data is well distributed across like the populations and the stuff, certainly that the process is unbiased..." *(Interviewee #1)* "I think in any stage of the data process from collecting to processing and annotating and so on, questions must be asked as where is the data coming from, who has collected the data, is it annotated, who's done it, how is it done, are there any forms of writing guidelines for people to think about it... make sure you use those data as needed, but as much as necessary..." *(Interviewee #12)*
	Update training and testing data frequently	"...it was for archival, and you have to identify whether a document should be conserved for a long period of time, or like something that we didn't need to keep, and we could like trash in the bin, or something needed to be done..." *(Interviewee #4)*
	Remove or minimise the use of sensitive training data where possible	"what we do with FairLearn is we say, we're going to obscure those sensitivities labels and give us an estimate of what our biases based on those sensitive topics... we have to decide when we're using algorithms, what are the things that could result in a problem..." *(Interviewee #5)* "AI systems work with the aggregation of data, and the main objective of these systems is to, whether it's with a good intention or not, profile people and classify them into categories... there is a way to eliminate those kinds of biases that can lead to discrimination... we get that the data sets can be debiased at the very beginning..." *(Interviewee #18)*
	Use encrypted/synthetic data in model training if possible	"...there is increasing work on ways to train models that don't require direct access to data or can operate on encrypted data and synthetic data... so the data is not actually ever directly seen by an individual or by the company, and so these methods can be employed, and there's active research going on and how to do this better..." *(Interviewee #14)* "... you may even have synthetics features to see how to understand better, how the decision the algorithm is making some decisions, especially when it's going to involve customers and human beings in production." *(Interviewee #19)*
Using the data/algorithm right	Responsible data enrichment, applied to 'consented' user data	"I think we try to make sure that we control every single step of the data ingestion, so from the first reception to the transformation of data... we do run scenarios, we do take a lot of time under data quality, which is not the case in the digital world... I think we do try to make a lot of thoughts about cleaning the

		data, making the first banner before we do any conclusion… and that not only we preserve confidential information which I think is the basic, but also that what we are proposing to a user is for the user's own benefit…" *(Interviewee #7)* "… seeking permission, asking for consent, and making things very transparent, those kinds of things will become procedural, and there'll be translated into processes in large institutions." *(Interviewee #8)*
	Responsible data provenance, blockchain for immutable record-keeping	"… I had a group, this group was responsible for what we call data governance… their job was to make sure that access to data was controlled, and that the provenance of data was recorded… so we knew which data passed from which person to which person." *(Interviewee #14)* "… one way probably can deal with this kind of problem (privacy issues) is to use blockchain techniques, because blockchain is a system you can store lots of information…" *(Interviewee #3)* "… for instance, if you use blockchain to…record and automate them in a permanent traceable ledger…" *(Interviewee #17)*
	Give immediate visibility into model behaviour and predictions	"… inside the algorithm itself, I think it's about externalising a lot of parameters, and being able to give access to those parameters as much as we can." *(Interviewee #7)*
Aggregate dimension: **Epistemic responsibility**		
Second-order themes	*First-order concepts*	*Exemplar quotes*
Developing 'wise' AI	Practice wisdom: educate AI designers via ethics training, create ethical culture	"I think they (AI designers) need to be adequately educated and trained to understand that technology is meant to do good for society, and that's the primary purpose of it… and we should have the right checks and balances in making sure that happens the ways it is intended to…" (Interviewee #8) "I think it should start at even the education of data scientists, people who are going to be building those… when they're learning about how to build models, it should be ingrained in even the methodology, like the data science methodology should have those ethical guidelines, in my mind…" *(Interviewee #9)* "you need to create this culture, data culture and ethical culture, so that the data scientist who is building it has the trigger to raise an ethical flag or red flag if something is wrong and escalated… because that trigger is not there, then of course having an ethical framework doesn't make a lot of sense." *(Interviewee #15)*
	Artificial wisdom: design fairer algorithms in	"You might have some people like some individuals that will start developing affection for those systems because of the way the systems treat them… if you can understand these adversarial attacks, if you can achieve some level of confidence in your results, then you can commit your systems safer and more predictable…" *(Interviewee #1)*

	accordance with best practices: adversarial robustness, differential privacy, federated learning, model simplification	"There is an area of research called adversarial robustness… of how to make AI not sensitive to small disturbances with the kind of like observing a big difference in that output… so exactly that, how to make sure that you're never ever gonna make a bad decision." *(Interviewee #11)* "you do have to address them at the trust level… you have to also invest in and stay connected to the community that is researching these methods where you can do training of models that operate on encrypted data or operate data… it's called federated learning… in a federated way so that the company never actually sees the details of the data… *(Interviewee #14)* "One way that we have thought about doing that is trying to approximate those low interpretability models with high interpretability models… if we're doing classification using a machine learning model, we might try to interpret the results with like logistic regression type of model, for instance… then we're still doing the results with the machine learning model, but we're trying to provide an explanation or some kind of like reason code maybe, or like seeing these are the features most likely to contribute to that result…" *(Interviewee #9)*
Creating a two-way explainability system	Inform about 'how' the processing of data and working of the system upfront	"I think the idea of explainability in the future should be a two-way system… the next is when we start thinking about the user providing explanations to the system because this is how humans exchange information… we explain things to each other." *(Interviewee #1)* "…they're [AI designers] going to rectify or validate, and this is what being used to feedback back into the model… this is one case where you can have a continuous feedback loop to evaluate and improve the model." *(Interviewee #4)* I think the thing is… it's about the transparency of algorithms… if you can explain in simple words, what kind of routines you are working with, how you treat the data when you ingest the data, and what kind of output data you produce… *(Interviewee #7)*
	Post hoc rationales about 'why' a certain decision was reached step by step	"…there were some very concrete things that are easier, which is to be able to explain the causal factors that led to a decision, that may be possible… so that's not explaining how the inner workings of the system performed, but it may be as simple as saying the reasons that the system chose to provide you with this…" *(Interviewee #14)* "… disclose enough information that people actually are using these systems, in a meaningful way, in a way that is useful but also like cautious enough, understanding the limitation and all that while being relatively easy for them to use it…" *(Interviewee #16)*
	Self-reinforcing feedback loops to collect user	"… they're going to rectify or validate, and this is what being used to feedback back into the model… So, this is one case where you can have a continuous feedback loop to evaluate and improve the model…" *(Interviewee #4)*

	perspectives for improvement	
Aggregate dimension: **Outcome responsibility**		
Second-order themes	*First-order concepts*	*Exemplar quotes*
Fairness testing	Build-in checks to avoid unintended failures or trigger immediate responses	"I think these are the kinds of fundamental building blocks that AI should also include… if we look at AI systems as an extension of people systems, I'm thinking of AI as another type of, probably a smarter version but another type of supporting technology, then I think things can be managed in a more positive way which benefits people, whether it's me or you or us as a society." *(Interviewee #6)* "The best way to do that is trying to make those explainable machine learning models to try to look or to find how they took in the position… and you can detect if there is something wrong there, and even if they are using a fightable tool to get some bias or to discriminate some ethics, to remove them all because you initially don't know that until you see those predictions…" *(Interviewee #19)* "I think it's gonna be recognising the mistakes, and putting into place systems where companies can assess and care enough to assess the results for sort of the impact of AI that they're using…" *(Interviewee #20)* "…where there could be errors in recognition, errors in detection, and that might lead to actually even physical harm to people… there have to be checks and balances, emergency switches, or buttons to actually guard against those sorts of issues that might occur." *(Interviewee #2)*
	Live testing AI in constrained environment before full development	" what we do is, instead of designing or implementing this machine learning model in going through to production, usually to make a step at the stage of post analysis of the model, is kind of pre-production… and we analyse with our business, for example, usually with people from the business, analysts, these outcomes… for example, if there are some things that we have biases or overfeed because we are cantering in some features that we shouldn't do, we usually detect that with business" *(Interviewee #19)*
	Scrutinise and review AI models against defined fairness performance metrics	"You have to develop test mechanisms, to prove that the biases are not present in the system… it's a very mechanistic thing… there are many different methods, each method depends on what kind of algorithm you're using, but they exist… and so, what you need to do is to include them as part of your development process, just like you would include a test programme in your development process, and you do it from the beginning…" *(Interviewee #14)* "When you're building the model, you have to iterate it and optimise it for certain situations… and then once you know that there's a situation where it's failing, you need to continue to sort of iterate and optimize it for that situation….so to consider that metric is how you would make a metric for it, but like to

		consider that almost as a metric for optimisation, like how do we reduce bias as a metric of success." *(Interviewee #20)* "when you work on an AI model, you are going to work with an optimising metric along to improve the accuracy or something… and you might have like to invest in one metric you are playing against, you try to get as best as possible, besides that, you're going to have like satisfying metrics… and then you could use exactly that with ethical features, what is the minimum we can tolerate…" *(Interviewee #4)*
	Risk/fallout management approaches	"there are so many ways where things could go wrong, and that's why we need to use like frameworks that can assess all the risks, and that we can assess systematically the risks associated with one precise AI system." *(Interviewee #18)* "…to do sensitivity testing on some particular cause… I think that would be very helpful because then you can direct people who have the logistical resources to address the problem, but to have some confidence that if they took that measure, it would actually address the problem…" *(Interviewee #14)* "…that calls for a verification loop for the results, those are the kind of things where you have to plan for this, and you have to invest…I think that once again, responsible AI would mean that, it's not about making sure that there is no mistake, it's about making sure that when there is a mistake made by the AI system, like you have some contingency plans that are there, like verification loops…" *(Interviewee #16)*

Table III.C2 Selected indicative codes and quotes for AI governance

Aggregate dimension: **AI structural governance**		
Second-order themes	*First-order concepts*	*Exemplar quotes*
Internal AI committee	Assess proposals for new AI applications for alignment with ethics principles	"my current company has a set of rules and a group of people who review how we do, and how we use technology… we have a review board… make sure it meets the principles of ethical AI, responsible AI… and we can go talk to someone who's in our legal organisation to say, did we do this the right way…" *(Interviewee #5)*
	Promote thoughtful consideration of trade-offs and balanced decision-making	"it should be discussed with them in terms of what is the real lead, who should be constantly involved in the decision-making process of the company itself… in terms of whether what they are trying to achieve is actually happening or not, or whether it's happening the opposite effect, are they having the opposite effect on people… I think that's one way forward for it to make the decision-making bit more inclusive…" *(Interviewee #2)* "… something needs to be discussed, and the ethical issue arises, or red flags pop up in the data science team, that can be escalated towards the C-level, to a specific board and ethical board, and then the ethical issues can be tackled… because in the end I had to meet the decision if an AI model or an AI product is responsible enough." *(Interviewee #15)*
External AI consultation	Seek independent ethics-based advice in setting up/refining AI governance	"I consult clients in setting up data and AI governance structures, meaning also some ethical boards or a data governance council… building out AI stewardship, data stewardship, code of conduct, metrics, KPIs, to really measure if the AI model you build complies with those metrics you put forward or with the principles you put forward." *(Interviewee #15)*
	Secure formal recognition for ongoing compliance with ethical standards	"I think there's a necessity to have an external eye on it, a third party that can guarantee and levelise the fact that your algorithm is responsible and certified… I think there's a necessity for a third party to intervene in those kinds of topics… we would like to be certified on that transparency…" (Interviewee #7) "I think the best place to start would be to, if a company has the resources, would be to have sort of an external review process of the AI that they're using… non-internal, and then once you have the external review, and the sort of the biases that are gonna be there, come to light and you know about them, then you have to look at them and see why…" *(Interviewee #20)*
Aggregate dimension: **AI procedural governance**		
Second-order themes	*First-order concepts*	*Exemplar quotes*

Regulation of AI	Create policy packs for self-governance	"…I already heard a lot of executives talk about ethics, and that starts with making adjustments to the policies, enterprise-level policies, or institutional-level policies, to make sure that ethics is an important element while you're defining your policies…" *(Interviewee #8)* "We will tell what the standard is for ethical usage of AI, and we will do that by merging best practices from academia, from legislative sources, and from the industry, and we come up with the best… having it independently, it's will always be better" *(Interviewee #15)* "I'm trying in my company to make those initiatives, maybe even instead of working with those people like me who would spend some hours every week to develop some ethics guidelines for our company…" *(Interviewee #19)*
	Consider social and legal norms, tailor human rights approaches to AI regulation	"there are some regulations, and we've seen also with such companies as Microsoft and IBM and Google, they have called for it because of their personal regulations… they value their reputation, and they want sometimes to engage in these voluntary regulations… self-regulation is okay, but it's only one component…the other components are structural regulations and political regulations, and society should be also aware of possibilities on the one hand… on the other hand, also stress of AI, so it should be collective work" *(Interviewee #10)* "I think there have to be government-level frameworks in place for this. Because I think that certain companies will do it automatically because their business relies on being trusted. So, they will do the right thing, but that's not a universal thing by any means, so there have to be frameworks." *(Interviewee #14)* "… the number one thing for me is having human rights due diligence, processes, or impact assessments… and this is something that's also coming back through more and more regulations now, it's really getting kind of centre stage… but for me, that's the most important thing, I mean, overall, it's about businesses respecting human rights as they develop and use AI, but especially the human rights due diligence can help them to detect potential issues with the harmful effects of AI, that they can then try to mitigate, avoid altogether by not developing that AI." *(Interviewee #13)*
	Use normative foundations as a structured approach	"… ethics is part of a very ancient tradition, and it cannot be done properly without referring to that tradition…and there are many different ways that ethics can contribute to the normative like the governance of AI…" *(Interviewee #18)*
Regulation through AI	Automated regulation compliance monitoring, enhance regulatory enforcement	"I think that there are several ways of controlling the complexity of AI… you can have a system like…you try to enforce that it's doing some kind of like something you consider good…" *(Interviewee #1)*

Aggregate dimension: **AI relational governance**		
Second-order themes	*First-order concepts*	*Exemplar quotes*
Distributed deliberation	AI expert discourse: share best technical practices, deliver assessments/evaluations	"…I know of many volunteer organisations that do all sorts of things, even one of my friends (AI expert) built a mechanism which he put on GitHub, anyone can access it…" *(Interviewee #14)* "there will be people (AI experts) challenging it, there will be people looking into it… and, I think even though it's a very complex statistical model from outer space, I think there will be people who will look into it and challenge it and understand it." *(Interviewee #15)*
	AI citizen expert scrutiny: contest AI via experiential knowledge, translate	"…part of my job is to work with the technical geeks, as well as the development professionals or international professionals, to act as a bridge between the people who create new algorithms, new models, new machine learning tools, new AI, and the ones who are in the forefront of communities… I tried to bring them together, that's part of my job" *(Interviewee #17)*
	AI user insight: direct feedback to highlight unexpected issues	"I think we…as individuals, need to take personal responsibility around those kinds of things… and I think without that, we are always going to be in danger of bias, creeping into systems bias… and designers of these systems potentially having the bias that's unchecked, uncontrolled, and probably unreported." *(Interviewee #6)* "I think in the end, the consumer will tell whether the company is right or wrong" *(Interviewee #15)*

Table III.C3 Selected indicative codes and quotes for outcomes

Aggregate dimension: **Managing AI outcomes**		
Second-order themes	*First-order concepts*	*Exemplar quotes*
Instrumental outcomes	Work efficiency, system performance, economic profitability	"…just by implementing the explainability in that system, it's already helped you find out the technical problems… by acting on making our systems explainable, you get these benefits of the expectations of the users are much better met…" "… AI helps a lot in organisations…and particularly helping people do stuff quickly in less amount of time and less effort…" *(Interviewee #1)* "…they probably helped them achieve lean organisations and lean production lines… for example, in factories…you increase the robustness of robots, you increase the predictability of your product line, and you also increase your performance and efficiency…" *(Interviewee #2)* "…and we were doing cutting-edge data science, a lot of AI system development, particularly to drive efficiencies, autonomy, and all those other good things…" *(Interviewee #6)* "… growth, efficiency, safety, and compliance could be multiplied, just because of the technological capability, and the gains that you will have with time, like whatever you were doing previously in years, you could now do in minutes and seconds… that's a big gain… with that additional time, there's a lot more you could do, you could invest that in many more capabilities to do more good…." *(Interviewee #8)* "in the end, you reduce costs, you mitigate risks, you also generate revenue because you have the brand awareness of being a responsible company. The products you develop will be developed much sooner, if you have proper data sets to use, it will potentially generate other products as well." *(Interviewee #15)*
Humanistic outcomes	Job satisfaction, employee engagement	"I think also from the employees, because people work for a company where they can wake up and look in the mirror in the mornings and feel good, they want to support something good. So, I think it helps to get the buy-in from the employees and all their energy, if there is a good cause to work for." *(Interviewee #12)* "…if a company has responsible ways of working, they will get also good employees who stick to those values, because really good employees are when they have a choice of where to work, then they will look for ethical companies, for responsible companies, because they want to have a clear mind and clear consciousness in order to work in a good company." *(Interviewee #10)*

Chapter III Appendix D: Two-level Utilitarian Approach

Responsible AI has been largely related to a broad discourse, ethics, in academic and policy circles in recent years (Jobin et al., 2019; Mittelstadt, 2019). Ethics is a normative philosophical discipline that concerns what is good, right, or acceptable and how an agent ought to act and be towards others (Ricoeur, 1992). Research in the IS field has long been concerned with the ethical implications of various IS elements (Gal et al., 2021; Paradice et al., 2018), but much of the current ethical exploration of AI systems draws upon the atheoretical applications of ethics (Mingers & Walsham, 2010). This includes broad classes of ethical concerns like privacy, biases, and fairness (e.g., Hagendorff, 2020, Jobin et al., 2019). Given the profound multi-faceted implications of the growing prevalence of AI and the challenges in embedding ethics into practices (Morley et al., 2020), there is a need to understand the ethical significance of AI in a more structured and theoretically informed way (Gal et al., 2021).

Various ethical principles have been proposed by philosophers to conciliate morals, desires, and capacities of moral agents. Among them, three moral paradigms: deontology, utilitarian, and virtue ethics are most dominant and applicable in the ethical analysis of AI use (Bolton et al., 2022; Gal et al., 2021). Specifically, deontological ethics believes an act is morally *right* if it is in accordance with one's duty and adheres to moral rules (Kant, 2002). Using a deontological approach, organisations managing AI should conform to AI principles and guidelines, which can be challenging to implement, as previously noted. Virtue ethics focuses on the moral character of an agent carrying out actions, and the 'virtue' is assessed based on comparisons with the moral acts of a virtuous person (MacIntyre, 2013). A virtue

ethics viewpoint would emphasise virtuous intentions and behaviours at both individual and collective levels in the design, development, and deployment of AI systems. However, one would argue that having dispositions for virtuous behaviours is similar to possessing and following rules.

Utilitarian ethics evaluates the ethicality of an act based on its consequences with the guiding value of *good* (frequently measured as pleasures/happiness). However, it is possible that the best possible results would conflict with moral intuitions, since acts are not adjudicated based on their inherent moral value. Thus, in exploring the phenomenon of managing AI, a two-level utilitarian approach: rule utilitarianism and act utilitarianism proves superior and can approximate the dispositions of the deontological and virtuous approaches (Bauer, 2020). Specifically, while rule utilitarianism emphasises intuitive thinking, requiring managers to diligently follow a set of rules in managing AI, act utilitarianism consists of critical thinking, urging managers to take actions that would tend to maximise good, especially when this is rule conflict or no rules. Thus, two-level utilitarianism combines duties and virtue and determines the appropriateness of AI management based on the overall balance of positive and negative implications.

Chapter IV

Goal Versus Duty: Marketing Communications for Generative Artificial Intelligence (GenAI) Healthcare Service

Abstract

Effective marketing communications, as the *voice* of the organisations, can build strong brands that shift consumer perceptions and behaviours. Through real-world observations of generative artificial intelligence (GenAI) powered healthcare services, we propose that digital healthcare providers might emphasise *goal-* or *duty-oriented* marketing messages. However, current knowledge lacks insights into which approach is more effective and why. Drawing on the perspectives of construal level theory (CLT), accessibility-diagnosticity model (ADM), and cue utilisation theory (CUT), this study uses an experimental design to examine the factors and processes that influence how consumers evaluate GenAI healthcare marketing communications in different healthcare domains: prevention, diagnosis, and treatment. Across a series of experiments involving measured (i.e., high vs. low) and primed (i.e., why vs. how) construal levels, the findings collectively indicate that consumers react more favourably to duty-oriented messages for preventive healthcare services; however, they prefer goal-oriented messages when paired with *why* explanations for prevention and *how* explanations for diagnosis. These combinations, influenced by perceived information diagnosticity (the relevance and usefulness of the information), increase consumers' intention to use the services. Together, these experiments shed light on how construal level and accessibility-diagnosticity effects offer novel and valuable insights into marketing communications for GenAI healthcare services. With these insights, marketers can effectively enhance marketing strategies by tailoring optimal messages and message combinations to influence consumer perceptions and choices in their practice.

Keywords: artificial intelligence (AI), generative AI (GenAI), healthcare service, construal level theory, accessibility-diagnosticity model

1. Introduction

The rapid advancements in technology (e.g., machine learning, ML) have led to the recent development of interactive generative artificial intelligence (GenAI)[23], such as OpenAI's GPT models and Google's Gemini, notable for their algorithmic ability to interact and communicate directly with consumers in natural language (Huang & Rust, 2023; Jo, 2023). The underlying algorithms of GenAI are capable of interpreting, learning from, and adapting in response to external data inputs and generating novel data outputs (often in text format) aimed at catering to specific tasks (Brown, 2020). Due to its prompt-response design, GenAI shows great potential for addressing many healthcare issues previously handled by human professionals, such as providing diagnostic and therapeutic support (Meskó & Topol, 2023). It can act as a health advisor, offering novel, personalised medical insights to support decision-making while leaving the final decision to users. However, an important issue common to this context (and possibly others) is that users often exhibit distrust toward algorithm-generated advice, even when artificial agents can sometimes outperform humans in specific tasks (Burton et al., 2020) such as diagnosing potential diseases. This phenomenon, often referred to as *algorithm aversion* (Dietvorst et al., 2015; Turel & Kalhan, 2023), can prevent individuals, companies, and societies from fully harvesting the benefits of GenAI's advising role in healthcare.

Initially, empirical studies such as the one by Longoni et al. (2019) revealed that consumers were reluctant to utilise medical services provided by AI, primarily due

[23] Our focus in this study is on the transformer-based generative models, which are considered widely as the foundation model of GenAI, rather than other classes of generative models (Bommasani et al., 2021).

to concerns about *uniqueness neglect*—the fear that personal nuances of conditions might be overlooked. However, the narrative is shifting, as subsequent research has explored the potential of more sophisticated AI in consumption contexts where consumers share sensitive information (e.g., health data) (Wichmann et al., 2022). This shift suggests that, while a personal touch remains important, the efficiency and capabilities of AI in delivering healthcare cannot be overlooked, particularly given the wide variation in the abilities of human professionals. Recognising this balance, researchers, such as Huang and Rust (2023), have begun exploring how advanced AI systems can address these concerns. They highlighted the significant potential of GenAI to enhance emotionally charged interactions, improving customer care as it evolves with advanced thinking, emotional intelligence, and communication capabilities. It is likely to transform digital healthcare by marrying technology efficiency with the empathetic and personalised interactions consumers seek, thereby fostering greater acceptance of GenAI in healthcare.

However, GenAI does not add value if consumers neither understand its potential nor leverage its advice (Dietvorst et al., 2018) for health maintenance. This may nullify companies' efforts to add accuracy or safety to their GenAI services as a point of differentiation. One of the means identified to motivate consumers is through effective marketing communications[24], as studies have shown that people can change their perceptions of AI after gaining more knowledge about it (e.g., Allen & Choudhury, 2021; Filiz et al., 2021). It is thus imperative to investigate how marketers can effectively communicate their GenAI-powered healthcare services

[24] While marketing communications refers to different promotional messages, channels, and media, this paper focuses on digital promotions such as information on the brand's official website commonly used by app providers to reach potential consumers.

with customers (maybe in the future) to facilitate positive consumer responses so as to reap the benefits of enhanced service design. Additionally, given GenAI's potential for emotional connection (Huang & Rust, 2023), emotionally connected customers are expected to be more valuable, loyal, and capable of bringing steady profit streams to firms (Magids et al., 2015; Rust et al., 2004). Thus, effective marketing communications for GenAI healthcare services can ultimately achieve a win-win situation.

The present research examines how marketing communication shapes consumers' intention to use GenAI healthcare services. Adopting an empirics-first approach (Golder et al., 2023; see Appendix A), we base our exploration of *message orientation* on real-world marketing observations (see Appendix B). Empirics-first research "reveals novel research questions untethered to theory and lends itself well with today's data-rich environment" (p. 319). With this empirics-first perspective, we observe that some real-world AI/GenAI healthcare app providers use marketing communications with specific message orientations that reflect their selling point (see Table IV.B1 in Appendix B). We refer to these as *goal- and duty-oriented messages*[25], designed to inform and engage their target consumers.

In our study, goal-oriented messages focus on the effectiveness of GenAI-powered services in achieving specific health outcomes for users, emphasising functional benefits (e.g., *"detects potential risks with great accuracy to improve your health outcomes"*). In contrast, duty-oriented messages highlight the ethical and societal

[25] In the current study, message orientation is a theoretically grounded persuasive marketing communication strategy, designed to motivate consumer behaviour by presenting appeals that are framed in terms of either goal or duty.

responsibilities of the company, emphasising user safety, inclusiveness, and broader values (e.g., *"ensures safe and inclusive healthcare with ethical use of GenAI"*). Additionally, we observed frequent mentions of *how* and *why* explanations in companies' marketing communications about their use of AI/GenAI-powered healthcare services. How explanations focus on the specific processes or features of GenAI technology that enable its functionality (e.g., *"how the system analyses data to generate insights"*), while why explanations emphasise the broader purpose or goals behind using GenAI in healthcare (e.g., *"why it helps improve patient outcomes or enhances inclusivity"*). An important question arising from this real-world observation is to determine which message features, or combinations thereof, are most effective in communicating GenAI healthcare services. Moreover, the process through which these effects happen has yet to be discovered. To address this, we aim to answer the following research question: *How can marketers effectively communicate GenAI healthcare services to persuade consumers to use them?* Specifically, we seek to understand when a goal- or duty-oriented message is more effective in generating consumers' intention to use the GenAI healthcare services and the underlying mechanisms.

Our primary emphasis of outcome variable across all studies is on *intention to use*, supported by meta-analysis findings (Keller & Lehmann, 2008) that the nature of marketing communication can significantly affect consumer intentions. Taking the perspectives of the construal level theory (CLT, Trope & Liberman, 2003; 2010) the accessibility-diagnosticity model (ADM, Feldman & Lynch, 1988; Herr et al., 1991), and the cue utilisation theory (CUT, Cox, 1962), we propose that marketing communications for GenAI healthcare services are influenced by construal levels

and evaluated based on information attributes (e.g., accessibility, relevance, usefulness). These evaluations may vary depending on the types of healthcare services delivered by GenAI, such as prevention, diagnosis, and treatment. In order to fully test the effectiveness of varying levels of construal in influencing consumer evaluations of GenAI healthcare marketing communications, this research first included a measured chronic construal (high- vs low-level mindsets) to explore individual differences. It then implemented a more explicit and proven manipulation of message construal (how vs. why) for comparison, reflecting our observations from real-world settings. This approach also addresses a crucial methodological gap[26] in applying CLT to marketing communication concepts (Lee, 2019) by offering empirical validations on whether exposure to specific message features (goal vs. duty) indeed elicits different levels of construal mindsets before manipulating them. Additionally, employing both measured and manipulated construal level[27] allowed us to investigate whether the accessibility of additional information influences consumer perceptions of information diagnosticity and their intention to use in the context of GenAI healthcare.

In addition to intention, we tested the perceived trustworthiness as an outcome variable in Study 2 due to its critical role in consumer decision-making, particularly for advanced technologies like GenAI in sensitive areas (e.g., healthcare), where uncertainty and perceived risks are high. Consumers often exhibit algorithm aversion due to concerns about transparency and other ethical considerations

[26] Much of the existing research manipulate construal levels directly without first confirming if the messages inherently align with or activate high- or low-level mindsets.
[27] In the manipulated conditions, participants were exposed to more information, enabling us to explore how varying levels of detail affect their evaluations and decision-making processes.

(Castelo et al., 2019). Testing trust alongside intention provides a holistic view of consumer responses, as trust influences confidence in GenAI's capabilities. This dual focus provides a more comprehensive understanding of the factors that shape consumer responses to GenAI healthcare services.

This study advances theoretical understanding and practical applications of GenAI healthcare marketing by integrating CLT, ADM and CUT to explain how strategic message framing influences consumer responses. Theoretically, we demonstrate that psychological distance and message construal shape engagement with AI healthcare services, highlighting the context-dependent nature of persuasion. Our findings identify perceived information diagnosticity as a key psychological mechanism driving AI adoption, extending CUT beyond traditional product evaluations to digital healthcare contexts where consumers rely on extrinsic cues in the absence of direct experience. Additionally, we contribute to AI ethics and responsible AI literature by showing that well-calibrated marketing messages can reduce algorithm aversion and foster greater trust and use intention in GenAI healthcare services.

Practically, we provide a strategic framework for designing persuasive and responsible AI marketing communications. Our findings highlight the importance of tailoring message framing to specific healthcare services—using duty-oriented messages for prevention and goal-oriented messages with procedural clarity for diagnosis and treatment. These insights offer actionable guidance for marketers, policymakers, and AI developers seeking to enhance consumer trust and adoption of AI-driven healthcare solutions while ensuring responsible and effective communication in high-stakes industries.

In the remainder of the paper, we first summarise the literature pertaining to our variables, and theorise their interactions to form hypotheses based on CLT and ADM perspectives. We then present empirical studies testing the premise that the matching effect of message orientation and construal level (both measured and manipulated) in GenAI healthcare marketing communications leads to enhanced consumer responses (i.e., intention to use, trust), and investigating the process that underlies these effects. Followed we present the findings, coupled with detailed discussion. Finally, we include the theoretical and practical implications of our work, acknowledge its limitations and suggest directions for future investigation.

2. Theoretical Background and Hypothesis Development

2.1 Message orientation, healthcare service, construal mindset: high, low

In the digital age, firms increasingly adopt AI-driven competitive strategies[28], yet they face a persistent tension between financial motivations and ethical responsibilities. This trade-off is particularly pronounced in the healthcare sector, which ranks among the top industries for AI investment (Sun & Medaglia, 2019) and is subject to significant regulatory and public scrutiny (Morley, 2020). Given the transformative potential of GenAI in healthcare, where it can assist in medical tasks (e.g., conversational chatbot) such as diagnosis, treatment planning, and patient engagement (Meskó & Topol, 2023), firms must strategically navigate their *AI orientation*—a firm's strategic aspiration and direction in introducing and

[28] Some are technology firms developing AI and mostly use traditional strategies that exploit AI capabilities. Others are technologically enabled firms using AI to deliver products and/or services, and might develop new business models to create and capture value. Also, those firms may operate in one or across multiple markets.

applying AI to guide its business practices (Ding et al., 2014; Li et al., 2021). Drawing from literature on corporate purpose[29] in for-profit firms (George et al., 2023) and our observations from real-world examples (see Appendix B), we distinguish between goal- and duty-based AI orientations. Goal-based AI orientation emphasises *value capture*, where firms leverage ethical practices primarily as a strategic asset to enhance competitiveness and profitability with improved algorithmic system performance. In contrast, duty-based AI orientation emphasises *value creation*, where firms integrate AI into their businesses based on an ethical commitment to societal welfare.

Beyond internal corporate strategies, AI orientation can be strategically framed in marketing communication messages to persuade consumers and differentiate themselves in the market, particularly in sensitive industries like healthcare. In this study, we propose a framework to predict consumer reactions to GenAI healthcare services based on the framing of marketing communications: goal- (e.g., that highlight AI's effectiveness in achieving specific user health outcomes) versus duty-oriented messages (e.g., that highlight AI's role in ensuring user safety and inclusivity). The key question for GenAI healthcare marketers is whether goal- or duty-oriented marketing communications are more effective, and whether this effect varies across different healthcare service segments. GenAI healthcare services—which leverage GenAI tools to enhance medical outcomes—can be classified into three distinct service types: prevention, diagnosis, and treatment

[29] Corporate purpose is used to signal intent and define what value a firm seek to create for its stakeholders. While goal-based purpose is organisation-specific and basic, duty-based purpose links to ethical obligations and emerges from broader social values (George et al., 2021).

(Achar et al., 2020; Mathur et al., 2013; Longoni et al., 2019), with each targets a different stage of the healthcare process.

In our case, *GenAI prevention* focuses on the use of GenAI to identify potential health risks and prevent illness before it occurs (e.g., skin cancer prevention). *GenAI diagnosis* means that GenAI can analyse observed symptoms to provide diagnostic suggestions, helping identify conditions that require further human investigation or immediate medical intervention (e.g., initial diagnostic triage for headaches). GenAI treatment involves the application of GenAI in the procedures of treating an existing medical condition (e.g., digital therapy for eating problems). Since a wide range of decisions related to health maintenance are associated with and influenced by the operation of psychological factors (Keller & Lehmann, 2008; Salovey et al., 2000) such as thought levels (Achar et al., 2020; how, why), it is critical to examine how message framing affects consumer acceptance of GenAI healthcare services. Consumer psychology differs significantly across prevention, diagnosis, and treatment contexts, particularly in how individuals process health-related risks, emotions, and decision-making considerations (Achar et al., 2020; Mathur et al., 2013).

To further examine these psychological differences, we integrate CLT (Trope & Liberman, 2003; 2010). CLT posits that psychological distance (e.g., temporal, spatial, social, or hypothetical) shapes how individuals process information and make decisions (Adler & Sarstedt, 2021; Aggarwal et al., 2015). When individuals perceive events as distant, they rely primarily on high-level construals—abstract, desirability-driven features, to form their evaluations and actions as psychological distance increases (Liberman & Trope, 1998; Todorov et al., 2007; Trope &

Liberman, 2010). In contrast, proximal events can activate low-level construals, leading individuals to focus more on pragmatic, feasibility-driven considerations.

Applying CLT to GenAI healthcare marketing, we suggest that consumer decisions could be based on different manifestations of high and low construal aspects of message focus when considering different healthcare service contexts (prevention, diagnosis, treatment), for example, idealistic values versus pragmatic concerns. Values and ideals are perceived superordinate to pragmatic concerns, influencing consumer evaluations and choices more in distant than immediate situation (Fujita et al., 2006; Kivetz & Tyler, 2007; Trope et al., 2007). For GenAI preventive and therapeutic services (which involve a future-oriented, psychologically distant event), consumers will engage in high-level construal, focusing on value-based attributes such as privacy protection and inclusivity. Therefore, duty-oriented messages (highlighting ethical responsibility) should be more persuasive for GenAI prevention and treatment services. For GenAI diagnosis services (which involve immediate medical concerns, representing a psychologically proximal event), consumers will engage in low-level construals, prioritising feasibility and instrumental benefits (Liu & Chang, 2020) such as diagnostic accuracy, efficiency. Thus, goal-oriented messages (highlighting GenAI's effectiveness) should be more persuasive for diagnosis services.

Trust is a critical determinant of consumer adoption of GenAI-driven healthcare services, as it influences not only usage intentions but also long-term engagement. While intention to use captures immediate behavioural inclinations, trust reflects a deeper psychological acceptance of the technology. Including trust as an outcome variable allows us to assess whether duty-oriented messages not only encourage

initial adoption but also foster sustained confidence in GenAI healthcare solutions. If consumers perceive GenAI as ethical and responsible, they may be more likely to integrate it into their healthcare decisions over time, making trust a key factor in the widespread acceptance of AI-driven healthcare services. Taken together, we hypothesise:

H1: Duty-oriented messages will be more persuasive for GenAI prevention (H1a) and treatment (H1b) services, while goal-oriented messages will be more persuasive for GenAI diagnosis (H1c) services.

Before further exploring induced message features (i.e., how/why explanations in the current research), it has been determined that individual cognitive differences may explain some variability in AI appreciation/aversion (Oliver, 2020). This perspective suggests that assessing individual chronic levels of construal may provide critical and nuanced insights (Freitas et al., 2004) into how to better predict consumer attitudes or behaviours and therefore design effective marketing communications in the context of GenAI healthcare. Notably, a high-construal level individual may not constantly have abstract thinking but should find high level construal easier to elicit and react more favourably to a stimulus that is more psychologically distant than would a low-construal level individual (Trope & Liberman, 2010). Given the varying nature of GenAI healthcare services, it is possible that participants' responses to GenAI marketing communications may differ based on their triggered construal mindsets.

H2: Chronic construal level moderates the effect of message orientation on GenAI usage intention, with high-level construal individuals favouring duty-

oriented messages for prevention (H2a) and treatment (H2b) services, and low-level construal individuals favouring goal-oriented messages for diagnosis (H2c) services.

2.2 Message orientation, healthcare service, message construal: how, why

Research suggests that people are more likely to engage in systematic-reflective thinking when information is readily accessible and sufficient to motivate them to override prejudice to emerging technologies (Herr et al., 1991; Turel & Kalhan, 2023) like GenAI. As discussed in the preceding section, one key factor that may moderate the effectiveness of message orientation (goal vs. duty) in shaping consumer responses is construal level, which influences how individuals perceive the outcomes of GenAI healthcare services. One notable feature that represents CLT is feasibility versus desirability (Trope & Liberman, 2010), which has been extensively studied in consumer research (e.g., Han et al., 2016; Lin & Chang, 2021; Yan et al., 2006; White et al., 2011). Feasibility captures the how aspect of an object or action, while desirability focuses on the why aspect—why one should use an object or perform an action. Our research conceptualises *why* and *how* explanations as representations of high versus low construal, respectively—an approach commonly adopted in literature and observed in marketing communications by real-world AI healthcare providers (see Appendix B).

Since predictions about future (vs. present) health experiences tend to be more abstract and schematic, we expect that framing GenAI prevention services (e.g., preventing future skin problems) and GenAI treatment services (e.g., improving future eating behaviours) using higher-level construal (why) will increase the

salience of values and ethical responsibility (duty), making such messages more persuasive. Conversely, because GenAI diagnosis services (e.g., determining current medical status) focus on immediate, concrete decision-making, consumers may respond more favourably to feasibility-focused messages (how), which highlight instrumental benefits (goal). A meta-analysis (Soderberg et al., 2015) has shown that the features deemed as high or low construal would have implications for judgement and decision-making in the face of psychological distance. Thus, in the case of GenAI-powered healthcare, consumers may weigh the trade-offs of delegating health-related decision-making to GenAI, depending on the construal level activated by the message.

In summary, we propose that duty-oriented messages and why construal, both operating at higher levels of abstraction, would facilitate a congruent information processing style, making them more effective for prevention and treatment services. Similarly, goal-oriented messages and how construal, which encourage a more concrete and practical mindset, should enhance persuasion in diagnosis services. Consumers will evaluate GenAI healthcare services more favourably when message framing aligns with their activated mode of thinking. We thus hypothesise:

> H3: Duty-oriented messages with why explanations will be more persuasive for GenAI prevention (H3a) and treatment (H3b) services, while goal-oriented messages with how explanations will be more persuasive for GenAI diagnosis (H3c) services.

2.3 Mechanism underlying the matching effects: information diagnosticity

Cue Utilisation Theory (CUT), rooted in cognitive psychology (Ullah et al., 2022), explains how individuals rely on informational cues to make judgments and decisions, particularly in uncertain and complex conditions (Cox, 1962). The theory posits that a cue's influence on decision-making depends on its *perceived diagnosticity*—how relevant and informative it appears (Dick et al., 1990; Feldman & Lynch, 1988; Ahluwalia, 2002; Pham & Avnet, 2004). A widely applied research model incorporating information diagnosticity is the ADM (Feldman & Lynch, 1988; Herr et al., 1991), which suggests that individuals' judgments depend on both the accessibility and diagnosticity of information, determining its usefulness in shaping impressions and decision goals (Lynch et al., 1988).

When making behavioural decisions, consumers rely on information retrieved from memory, external contexts, or a combination of both (Menon et al., 1995). This study focuses on the diagnosticity of external, context-based information, specifically the effects of message orientation (goal vs. duty) and construal level (how vs. why) on GenAI healthcare adoption. While various factors may influence information diagnosticity (Lynch et al., 1988), two key attributes under marketers' control are information quantity and information strength. Greater information quantity (i.e., multiple message combinations) allows consumers to accumulate evidence for decision-making, while greater information strength (i.e., perceived reliability, validity, and relevance) enhances the message's persuasive impact. When both are high, consumers should perceive the information as more diagnostic, leading to stronger intentions to use GenAI healthcare services.

Duty-oriented messages, which emphasise ethical responsibility and societal values, align with high-level construal (why) explanations that promote abstract reasoning (Trope & Liberman, 2003). Since GenAI prevention and treatment services involve future-oriented, schematic medical outcomes, this alignment should enhance perceived diagnosticity, as consumers see why explanations (e.g., reasons for using GenAI to promote safety and inclusiveness) as more relevant and persuasive. Conversely, goal-oriented messages, which highlight instrumental benefits, align with low-level construal (how) explanations that focus on concrete, procedural details. Since GenAI diagnosis services require immediate, data-driven decision-making, consumers should perceive how explanations (e.g., procedures for improving accuracy and efficiency) as more diagnostic and persuasive in this context.

To summarise, marketing communications in GenAI healthcare that align message orientation (goal vs. duty) with message construal (how vs. why) can enhance perceived diagnosticity, which in turn influences consumer intention to use and trust in GenAI healthcare services. Since perceived diagnosticity determines how informational cues shape judgment, consumers are more likely to accept and trust AI recommendations when they perceive the information as useful and relevant. Prior research supports this claim (e.g., Turel & Kalhan, 2023), showing that informative AI messages can mitigate AI aversion, leading consumers to rely more on AI-driven insights. This reinforces the crucial role of diagnosticity in shaping consumer attitudes and decision-making. Building on this, we propose:

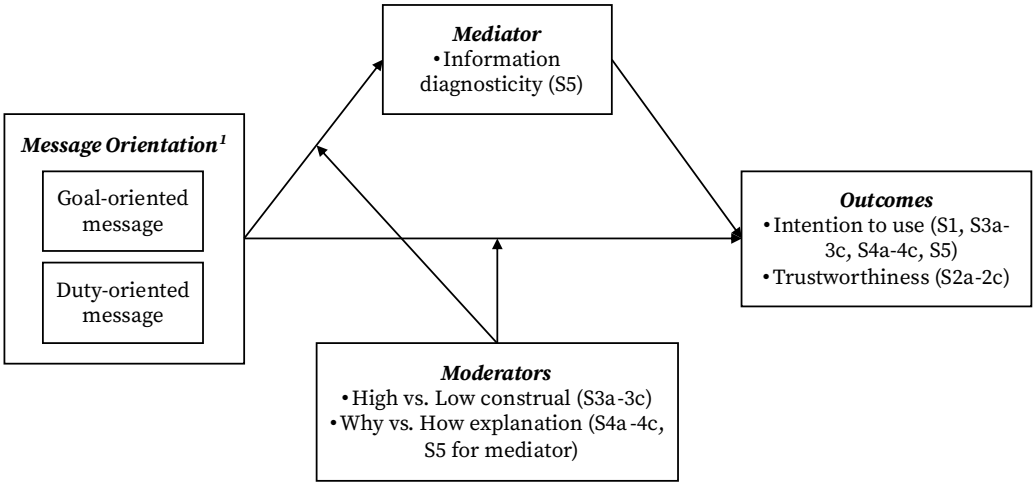
H4: Perceived information diagnosticity mediates the effectiveness of matched marketing communications on consumer evaluations of GenAI healthcare services.

3. Overview of Studies

These experiments examine how marketing communications influence consumer responses to GenAI healthcare services, with a focus on the interplay between message orientation, construal levels (measured and manipulated), and consumer evaluations across prevention, diagnosis, and treatment contexts. Figure IV.1 provides the empirical framework, while tables IV.1 and IV.2 summarise study details and research variables.

Study 1 provides evidence that consumers have higher intention to use a GenAI prevention service when its marketing communications feature a duty-oriented message (vs. goal-oriented message). Studies 2a-2c extend this investigation to three healthcare domains (prevention, diagnosis, treatment), testing consumers' perceived trustworthiness of GenAI healthcare services. Trust-enhancing effects of duty-oriented messages appear only in prevention (Study 2a) but not in diagnosis or treatment, likely due to the psychological nature of these contexts. Studies 3a-3c replicate Studies 2a-2c with intention to use as the outcome variable, and further operationalise construal levels through measured construal mindsets (high vs. low) to examine individual psychological differences in message evaluation. Study 3a reveals that goal-oriented messages are more effective for GenAI prevention services among high-construal consumers, suggesting that additional explanations

(why, how) are needed to encourage adoption of GenAI diagnosis and treatment services. Studies 4a-4c test marketing communications in a controlled laboratory setting, manipulating message construal (how vs. why). Results show that goal-oriented messages paired with why explanations increase intention for GenAI prevention services (Study 4a), while how explanations are more effective for GenAI diagnosis services (Study 4b). Study 5 further confirms that perceived information diagnosticity mediates these effects, highlighting the mechanism driving consumer responses to GenAI healthcare marketing.



Notes: [1] Goal- and duty-oriented messages were observed in real-world marketing communication examples for AI-powered healthcare apps, see Appendix B for detailed information; Healthcare delivered: (a) prevention, (b) diagnosis, and (c) treatment.

Figure IV.1 Overview of the studies

Table IV.1 Summary of study details

Study	Rationales	Objectives	Findings	Study design with Variables	Context†			Sample size††
					a	b	c	
Study 1	*Goal* and *duty* marketing communications observed from real-world examples. Duty-oriented (vs goal-oriented) messages are associated with higher CL.	*Testing the main effect*: the impact of message orientation (goal vs. duty) on consumers' intention to use GenAI prevention services.	Duty-oriented messages ->> Intention to use GenAI prevention services H1a is supported.	Message orientation (IV)—Intention to use (DV_1)	✓			147
Study 2a – 2c	A nuanced, segment-specific approach is crucial for effectively navigating considerations for GenAI healthcare service marketing communications.	*Testing the boundary conditions:* the impact of message orientation on consumers' perceived trustworthiness of GenAI healthcare services across three contexts.	Duty-oriented messages ->> Perceived trustworthiness; H1a is supported. H1b and H1c are not.	Message orientation (IV)—Perceived trustworthiness (DV_2)	✓	✓	✓	436
Study 3a – 3c	Individuals' preferences of message orientation may vary depending on their construal mindsets (high vs. low). Addressing a methodological gap in applying CLT to marketing communication concepts.	*Testing the moderating effect:* whether individual construal mindsets (psychological differences) moderate consumers' intention to use GenAI healthcare services across three contexts.	Goal-oriented messages × High-construal mindset ->> Intention to use GenAI for GenAI prevention services H2a is supported while H2b and H2c are not. Higher-construal respondents report greater intention across studies, indicating that additional explanations (*why* and *how*) may be necessary to persuade consumers.	Message orientation (IV) × Construal mindset (Mod_1)—Intention to use (DV_1)	✓	✓	✓	300

Study 4a – 4c	*How* and *why* marketing communications, observed from real-world examples, have been extensively explored in CLT research as explicit and proven manipulations of message construal.	*Testing the moderating effect:* whether message construal (how vs. why explanations) moderate consumers' intention to use GenAI healthcare services across three contexts.	Goal-oriented messages × why explanation ->> Intention to use GenAI for GenAI prevention services. Goal-oriented messages × how explanation ->> Intention to use GenAI for GenAI diagnosis services. H3c is supported while H3a and H3b are not.	Message orientation (IV) × Message construal (Mod_2) —Intention to use (DV_1)	✓	✓	✓	724
Study 5	CUT posits that the use of a cue in decision-making depends on its perceived diagnosticity. Individuals are likely to perceive higher diagnosticity in combined marketing communications.	*Testing the mediating effect:* the role of information diagnosticity as a driver of consumers' intention to use GenAI healthcare services.	Goal-oriented messages × why explanation × GenAI prevention services/Goal-oriented messages × how explanation × GenAI diagnosis services ->> perceived information diagnosticity ->> Intention to use H4 is partially supported.	Message orientation (IV) × Message construal (Mod_2) —Information diagnosticity (Med) — Intention to use (DV_1)	✓	✓		479

† GenAI healthcare contexts: (a) prevention, (b) diagnosis, and (c) treatment.

†† The sample size was calculated using G*Power to ensure 80% power ($\beta = 0.2$) to detect a medium effect size ($d = 0.5$; ensure sufficient sensitivity to detect a moderate, noticeable difference despite practical constraints) at a two-tailed significance level ($\alpha = 0.05$). At least 75 participants per group were required for the one-factorial design and 50 for the two-factorial design, accounting for dropout and void data rates.

Table IV.2 Summary of variables

Variables	Definitions
Independent variable:	
IV: **Message orientation** (goal-oriented vs. duty-oriented)	
Goal-oriented message	Focuses on achieving specific healthcare outcomes or objectives, emphasising the functional and instrumental benefits of the GenAI powered healthcare service.
Duty-oriented message	Highlights ethical and societal responsibilities, emphasising values such as inclusiveness, safety, and moral obligations.
Moderating variable:	
Mod_1 : **Construal mindset** (high vs. low); Mod_2 :**Message construal** (how vs. why)	
High-construal mindset	Represents abstract thinking, focusing on broader meanings, overarching purposes, and distant implications of actions or events. Individuals with a high-construal mindset prioritise desirability, considering the overall value or significance of an action.
Low-construal mindset	Represents concrete thinking, focusing on immediate details, specific actions, and the feasibility of actions or events. Individuals with a low-construal mindset prioritise feasibility, considering the practical steps or immediate outcomes of an action.
How explanation	Emphasises the specific processes, features, or mechanisms through which GenAI delivers healthcare services, focusing on feasibility and practical implementation.
Why explanation	Highlight the overarching purpose, benefits, or broader value of using GenAI in healthcare services, focusing on desirability and long-term implications.
Mediating variable:	
Med: **Information diagnosticity**	
Information diagnosticity *Adapted from Baker & Lutz (2000), Baker (2001)*	The degree to which consumers perceive the provided information as relevant, credible, and sufficient to evaluate the utility and effectiveness of the service. It reflects the extent to which the information facilitates informed decision-making, particularly in complex and high stakes such as healthcare contexts, where uncertainty and perceived risks are significant.
Dependent variable:	
DV_1: **Intention to use**; DV_2: **Perceived trustworthiness**	
Intention to use *Adapted from Venkatesh et al.(2012)*	The extent to which a consumer intends to adopt or engage with a GenAI healthcare service reflects a deliberate and conscious decision-making process regarding future use.
Trustworthiness *Adapted from McKnight et al.(2002)*	The extent to which consumers believe that GenAI and its powered services are reliable, ethical, and capable of delivering accurate/safe, and effective outcomes. It reflects consumer confidence in the system's technical abilities/adherence to ethical standards, and alignment with user interests and values.

3.1 Study 1: Intention to use GenAI prevention service by message orientation

In Study 1, we investigated whether consumers exhibit a higher intention to use a GenAI prevention healthcare app when its marketing communications adopt a duty-oriented (vs. goal-oriented) message. Respondents' intention to use preventive services was assessed as a measured factor.

We first focused on GenAI prevention services as they are more prevalent and widely promoted, emphasising proactive health management. Unlike diagnosis or treatment, prevention involves lower perceived risk, providing a clearer test of the baseline model for message orientation effects on consumer responses to GenAI healthcare applications without the confounding influence of heightened risk perceptions. Additionally, its long-term commitment aspect makes it a meaningful context for examining the persuasive power of duty-oriented versus goal-oriented messages.

3.1.1 Participants, design, and procedures

A total of 147 participants (60.5% female, M_{age} = 43.7 years) were recruited from Prolific for monetary compensation to participate in a single-factorial design study (message orientation: goal vs. duty). They were asked to evaluate adverts for a digital app that is powered by GenAI technologies. To enhance the generalisability of results, we adopted an actual derived cases approach, presenting lifelike scenarios as recommended by Aguinis and Bradley (2014). Skin cancer prevention (i.e., SkinVision, see Table IV.B1 in Appendix B) was selected as the healthcare context due to its real-world application. To minimise the effects of brand or product familiarity, we used the fictitious brand name *CareAI* for the app.

After participants agreed to take part in the study, they were instructed to imagine needing digital healthcare services for illness prevention and came across an advert introducing the app. Participants were informed that CareAI uses a GPT-4 model to generate skin cancer risk predictions in natural language based on user-provided images (see Appendix C for details). Following a briefing on the medical context, app functions, and AI/GenAI technologies, participants viewed adverts with varying text in the advertising appeals.

To manipulate message orientation, we used pretested adverts ($n = 52$, see Appendix D). The experimental stimuli were carefully designed following realistic scenarios to enhance experimental realism (listed in Appendix C). Goal-oriented messages highlight CareAI's aims to improve health outcomes, focusing on functional benefits like accuracy and effectiveness in illness prevention. In contrast, duty-oriented messages emphasise on higher-order values, addressing safety and inclusiveness in communications with consumers.

Participants were then asked to evaluate the advert messages and report their intention to use CareAI for skin cancer prevention, measured as the dependent variable (three, seven-point items; adapted from Venkatesh et al., 2012) that is held to be predictive of behaviours (consistent with the theory of planned behaviour [TPB]; Ajzen, 1991). It was presented on a separate page of the online questionnaire to reduce the potential for common method bias in self-reported data (Hulland et al., 2018, similar strategies have been adopted in all following studies). They were also asked to provide demographic information.

3.1.2 Results and discussion

Intention to use. Our main premise was that a duty-oriented message would result in increased consumer intention to use a GenAI prevention service. To test this hypothesis, we conducted a one-way analysis of variance (ANOVA) comparing participant preferences for message orientation (goal vs. duty) in the advert. As expected, a duty-oriented message (M_{duty} = 3.86, SD = 1.69) was found to yield significantly higher consumer intention to use the app than a goal-oriented message (M_{goal} = 3.17, SD = 1.82; $F(1, 145) = 5.736$, $p = .018$). H1a was supported.

Discussion. The results provide initial evidence that consumers exhibit a preference for duty-oriented messages, as they showed significantly higher intention to use the app for prevention services when the advert emphasised duty. Research suggests that people would hold (on average) an implicit bias against AI, often perceiving it as less trustworthy when the reference point is a human professional (Turel & Kalhan, 2023), particularly in sensitive contexts like healthcare. When consumers consider using GenAI for preventive purposes without observed symptoms, their concerns often focus on potential harms and risks, such as privacy issues, rather than the accuracy of AI-driven predictions. A duty-oriented message helps alleviate these broader concerns by emphasising the ethical and responsible use of technology in GenAI prevention, thereby increasing consumer confidence in GenAI prevention services.

Building on these findings, the next study will examine how different healthcare service domains (i.e., prevention, diagnosis, treatment) influence the perceived trustworthiness of GenAI healthcare services. Specifically, we will explore the

impact of message orientation (goal vs. duty) on consumer perceptions across these domains.

3.2 Study 2a-2c: Perceived trustworthiness by message orientation

In Study 2, we sought to test consumers' trust in GenAI healthcare services across three contexts: prevention (2a), diagnosis (2b), and treatment (2c). In each of these separate studies, respondents evaluated adverts for CareAI and reported their level of trust, defined as 'the willingness to depend on or be vulnerable to a technology for whatever one needs from it' (McKnight et al., 2002, p.1018). Unlike trusting beliefs, technology-related trusting intention is more action-oriented and can predict risk-related behaviours, such as the willingness to rely on GenAI results.

We test trust across prevention, diagnosis, and treatment services to capture variations in trust formation—prevention involves long-term reliance, diagnosis requires accuracy, and treatment entails direct intervention. However, intention to use is tested selectively in Study 1, given that adoption likelihood varies. Prevention services are discretionary and highly influenced by persuasion, whereas diagnosis and treatment are need-driven, making intention to use more dependent on medical necessity. Thus, trust applies broadly, while persuasion effects on usage are most relevant in prevention, without considering moderators in Study 3 and 4.

3.2.1 Participants, design, and procedures

A total of 436 participants were recruited from Prolific for single factorial (message orientation: goal, duty) experiments, with monetary compensation provided (Study 2a: n = 147, 60.5% female, M_{age} = 44.2 years; Study 2b: n = 146, 51.4% female, M_{age} =

42.3 years; Study 2c: n = 143, 57.3% female, M_{age} = 47.3 years). The experimental procedure was identical in all three studies. We chose skin cancer prevention, headache diagnosis, and digital therapy for healthcare contexts (i.e., SkinVision, Ada, and Youper, see Table IV.B1 in Appendix B), and created stimuli following realistic scenarios (listed in Appendix C). Participants were first introduced to the medical contexts, app functions, and AI/GenAI technologies before examining adverts for GenAI healthcare services. They rated their trust in these services (three, eleven-point items; adapted from McKnight et al., 2002) as the dependent variable and provided demographic information.

3.2.2 Results and discussion

Manipulations. For each study, we included a single-item question to confirm the effectiveness of our manipulation check for message orientation (goal vs. duty), ensuring it remained robust despite slight wording changes. After applying the exclusion criterion, ANOVA results showed that respondents exposed to goal-oriented messages rated them significantly lower than those exposed to duty-oriented messages (Study 2a: $F(1, 145) = 7.719$, $p = .006$; Study 2b: $F(1, 144) = 12.205$, $p < .001$; Study 2c: $F(1, 141) = 4.226$, $p = .042$). These results confirm that the manipulations were successful.

Trustworthiness. We conducted one-way ANOVAs with message orientation (goal vs. duty) as the independent variable and perceived trust as the dependent variable across three healthcare contexts: prevention, diagnosis, treatment. For GenAI prevention services (Study 2a), participants reported significantly higher trust after viewing a duty-oriented message (M_{duty} = 5.95, SD = 2.05) compared to a goal-

oriented message (M_{goal} = 5.15, SD = 2.49; F(1, 145) = 4.466, p = .036). However, the effects were insignificant for either GenAI diagnosis (Study 2b: p = .396) or GenAI treatment (Study 2c: p = .977) services. Thus, H1a was supported, while H1b and H1c were not.

Discussion. The results suggest that duty-oriented messages enhance perceived trust in the CareAI app for GenAI prevention services. This may be because preventive measures often involve long-term health management, where ethical responsibility and inclusivity resonate more strongly with consumers. In contrast, for diagnosis and treatment services, participants did not report significant differences in levels of perceived trustworthiness. This indicates that in these contexts, consumers may rely less on external message cues and instead use intuitive reasoning or medical necessity to assess credibility, especially when the information provided is insufficient to support confident judgements (Akdeniz et al., 2013; Tsai & McGill, 2011).

Notably, while trust is often a key determinant of technology adoption (McKnight et al., 2002), it does not appear to be a major barrier in this study. Instead, practical benefits and service relevance may drive consumer engagement more than perceived credibility. This underscores the need to focus on persuasive messaging that directly influences user engagement, rather than trust-building interventions, in subsequent studies. Consequently, we drop perceived trustworthiness as an outcome variable.

When including moderators, we exclude trust because moderation effects typically assess how or when persuasive messaging influences consumer responses, which

aligns more with intention to use. Trust formation is gradual and multifaceted, shaped by factors beyond message framing, such as brand reputation and prior experience. Unlike intention to use, which can be immediately influenced by persuasive messaging, trust develops over time and is less likely to fluctuate based on message framing alone. Moreover, moderation effects highlight situational influences (e.g., construal level) that alter persuasion strength, making trust less responsive to these transient message variations. Instead, intention to use is a more dynamic, context-sensitive outcome variable, making it more suitable for testing moderation mechanisms.

In our next study, we aim to extend our findings by exploring psychological states and proposing construal level as a potential moderator, as individuals may differ in their habitual abstract vs. concrete thinking styles. Understanding these variations is crucial for accurately interpreting consumer responses to message orientation (goal vs. duty) in experiments applying CLT.

3.3 Study 3a-3c: Construal mindsets as measured moderator

In Study 3, we examined whether individuals' construal mindsets moderate their intention to use GenAI healthcare services across the same healthcare contexts as Study 2: prevention (3a), diagnosis (3b), and treatment (3c). The design for each study was a single factorial design (message orientation: goal vs. duty) with construal level (high vs. low) used as a measured variable. Participants first rated their intention to use the app, then completed a purportedly unrelated questionnaire measuring their construal mindsets, followed by demographic information.

3.3.1 Participants, design, and procedure

A total of 436 participants were recruited from Prolific for monetary compensation, with 300 participants retained for further analysis (Study 3a: n = 105, 58.1% female, M_{age} = 45.5 years; Study 3b: n = 101, 50.5% female, M_{age} = 42.2 years; Study 3c: n = 94, 57.4% female, M_{age} = 46.2 years). All participants were first presented with introductory messages detailing the medical contexts, app functions, and clarifications about AI/GenAI technologies (see Appendix C for details). After confirming they had read the messages, participants completed two separate tasks.

In the first task we had participants evaluate the message orientation (goal vs. duty) of the adverts using the same stimuli and experimental procedures as in Study 2 and report their intention to use the app in each of the three separate studies. The second task involved a purportedly unrelated activity examining what specific behaviours imply to individuals. Their construal mindsets (tendency to construe information at a high vs. low level) were measured using Vallacher and Wegner's (1989) multi-item Behaviour Identification Form (BIF), based on action-identification theory.

Participants reviewed a list of 25 actions (e.g., making a list, as shown in Table IV.C4) and chose between two options for each action (e.g., getting things organised [high-level construal] vs. writing something down [low-level construal]) that best represented their interpretation. Responses were scored as 0 for low-level construal and 1 for high-level construal, with the total summed to form a construal level index ranging from 0 to 25. Participants also completed demographic questions, attention checks, and a suspicion probe, with none aware of the research hypotheses.

3.3.2 Results and discussion

Manipulations. Similar to Study 2, the ANOVA results for the single-item manipulation check in each study confirmed the success of our manipulations (all *ps* < .05).

Intention to use. For individual construal mindsets, we performed a split by retaining the upper third (high-level construal participants) and lower third (low-level construal participants) of the data for further analysis, while excluding the middle third. This resulted in datasets of 105 participants in Study 3a, 101 participants in Study 3b, and 94 participants in Study 3c. This approach allowed us to focus on how message orientation and construal mindsets interact to influence participants' intention to use the app. Additionally, without sacrificing the interpretability—which is the motivation for categorising a continuous variable in the first place—this approach is perceived as more effective than a binary split (Gelman & Park, 2007).

We ran two-way between-subjects ANOVAs with message orientation (goal vs. duty) and construal mindset (high vs. low) as independent variables, and consumer intention to use the app as the dependent variable. In the prevention context (Study 3a), the model was significant ($F(3, 101) = 5.311$, $p = .002$, $\eta^2 = .136$). Significant main effects were found for message orientation ($F(1, 101) = 5.624$, $p = .020$, $\eta^2 = .053$) and construal mindset ($F(1, 101) = 5.502$, $p = .021$, $\eta^2 = .052$), as well as a significant interaction effect ($F(1, 101) = 4.795$, $p = .031$, $\eta^2 = .045$).

These results suggested that participants, particularly those with a high-construal mindset ($M_{high} = 4.08$, SD = 1.64; $M_{low} = 3.32$, SD = 1.81), showed higher intention to

use the app for illness prevention when exposed to a duty-oriented message (M_{duty} = 4.09, *SD* = 1.57) compared to a goal-oriented message (M_{goal} = 3.70, *SD* = 1.76). However, in Study 3b (diagnosis) and 3c (treatment), no significant main effects of message orientation or interaction effects were found (3b: *ps* > .491; 3c: *ps* > .105), but high-construal participants consistently reported higher intention to use the app overall (3b: $F(1, 97) = 8.067$, $p = .005$, $\eta^2 = .077$; 3c: $F(1, 90) = 4.625$, $p = .034$, $\eta^2 = .049$). Thus, H2a is supported, while H2b and H2c are not.

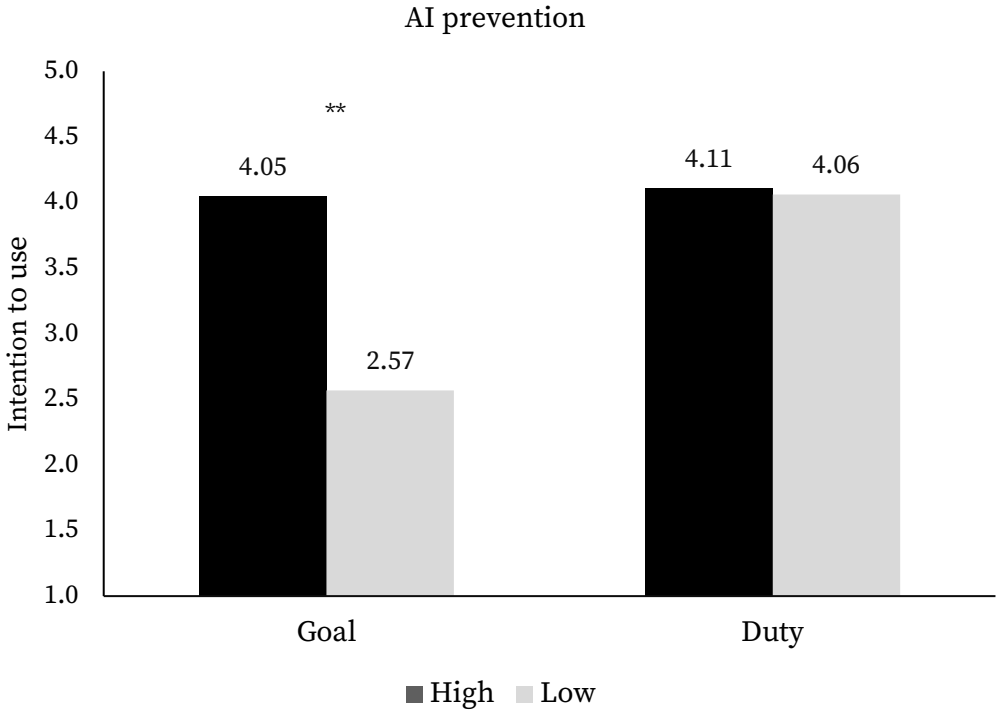
Probing of the interaction in Study 3a showed that, high-construal participants reported higher intention to use AI prevention services in the goal condition ($M_{goal\text{-}high}$ = 4.05, *SD* = 1.87) compared to low-construal participants ($M_{goal\text{-}low}$ = 2.57, *SD* = 1.61; $F(1, 48) = 8.942$, $p = .004$, $\eta^2 = .157$). In the duty condition, participants' intention to use the app was similarly high across construal mindsets, with no significant difference ($M_{duty\text{-}high}$ = 4.11, *SD* = 1.46; $M_{duty\text{-}low}$ = 4.06, *SD* = 1.72; $p = .906$).

Additionally, consistent with findings from Study 3b and 3c, high-construal participants demonstrated a higher overall intention to use the app for prevention services across conditions (M_{high} = 4.08, *SD* = 1.64; M_{low} = 3.32, *SD* = 1.81, $p = .021$, $\eta^2 = .052$). A possible explanation is that high-construal individuals tend to engage in abstract thinking, making them more receptive to the broadly framed language used in both goal- and duty-oriented adverts.

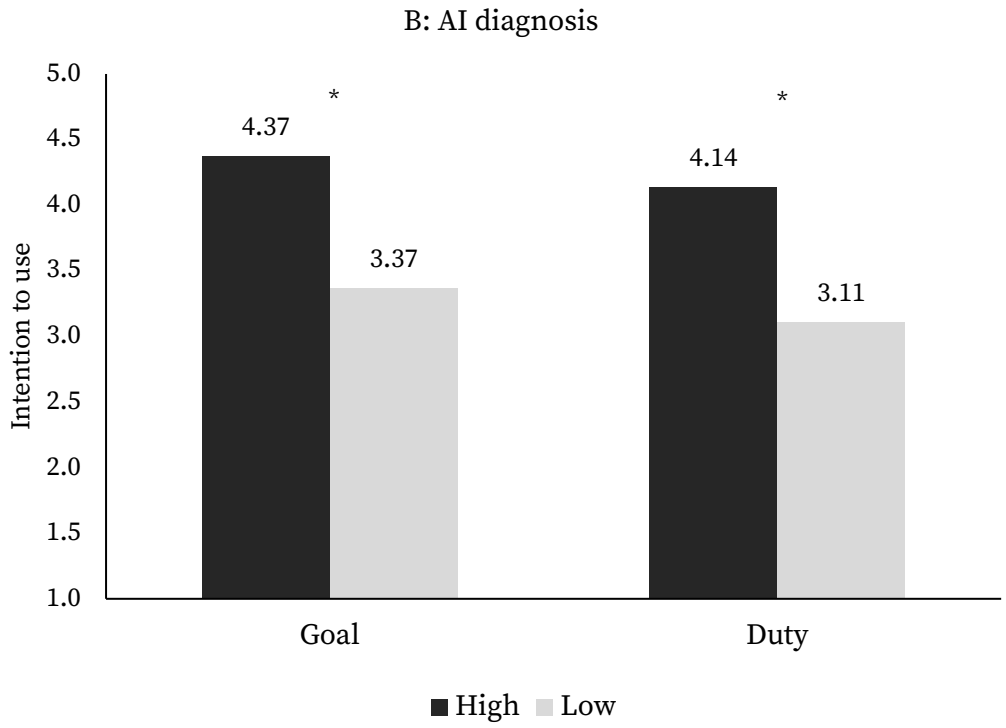
Table IV.3 Intention to use by message orientation, construal mindset (measured by BIF score) and healthcare services (S3a-3c)

	Goal Condition		Duty Condition	
	High	Low	High	Low
Intention to use				
GenAI prevention	4.05 (1.87)[b]	2.57 (1.61)[a]	4.11 (1.46)[b]	4.06 (1.72)[b]
GenAI diagnosis	4.37 (1.75)[b]	3.37 (2.00)[a]	4.14 (1.78)[b]	3.11 (1.44)[a]
GenAI treatment	4.08 (1.38)[a]	3.59 (1.58)[a]	4.00 (1.66)[a]	3.11 (1.51)[a]

Note. Participants were randomly assigned to each condition. Cells contain means with standard deviations in parentheses. Higher scores indicate higher behavioural intentions. Different superscripts within a row indicate the means are significantly different at $p < .05$.



Note. ** p < 0.01.



Note. * p < 0.05.

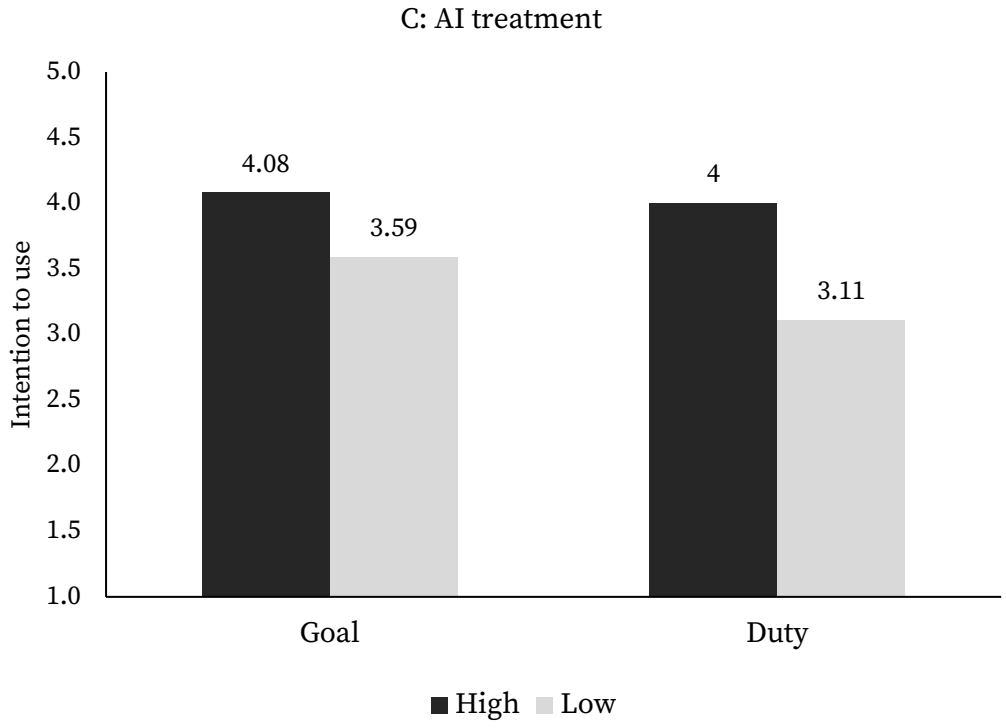


Figure IV.2 Intention to use GenAI healthcare service varies by message orientation, moderated by construal mindset

Discussion. For GenAI prevention services, evidence shows that duty-oriented messages lead to higher consumer intention to use the app compared to goal-oriented messages. This may be because clear objectives and outcomes, typically emphasized in goal-oriented messages, are less critical when individuals do not have observed symptoms.

For goal-oriented messages, participants' intention to use the app varied significantly based on their construal mindsets. This prompted the consideration of goal systems in the context of CLT, specifically superordinate goals and focal goals (Bagozzi & Dholakia, 1999; Kruglanski et al., 2018). Superordinate goals emphasise the higher-order purpose of an action, while focal goals describe the concrete means of performing the action. A potential explanation is that participants perceived goal-oriented messages for GenAI prevention services as encompassing both a superordinate aspect (e.g., maintaining health) and a focal aspect (e.g., performing checks). These aspects may be weighted differently depending on individuals' construal mindsets. When preventive health outcomes are perceived as psychologically distant, individuals with high construal mindsets may find goal-oriented messages more compelling, as they align with their preference for abstract, higher-order thinking.

For GenAI diagnosis and treatment services, the results suggest a need for further marketing communication strategies, as only high-construal participants reported higher intention to use the app. From the view of ADM (Herr et al., 1991), individuals may rely on surface cues (e.g., implicit biases) when they lack relevant information about GenAI's performance (Turel & Kalhan, 2023). In the contexts of diagnosis and treatment, some consumers may exhibit negative prejudice toward

AI technologies due to easily retrievable memories of AI ethics issues. When symptoms are present, they may become risk-averse in decision-making and undervalue GenAI advice compared to that of trained experts, especially when domain knowledge is required for judgment (Logg et al., 2019).

This aversion could also be explained by the somatic marker hypothesis (Bechara & Damasio, 2005), which posits that in the absence of reliable and relevant information, consumers—particularly those with low construal mindsets—may rely on gut feelings and make random decisions under uncertainty. Unlike GenAI preventive services, diagnosis and treatment involve more critical, high-stakes health decisions where inaccurate information could result in delays in proper care. Consumers, especially those with observed symptoms, might distrust GenAI's ability to diagnose diseases or recommend treatments due to the complexity of medical conditions. To address this, further marketing communications emphasising the feasibility and desirability of GenAI healthcare services may help consumers evaluate and make more confident decisions about using such technologies.

The above inferences are supported by a pilot study, where many participants, particularly those in the goal-oriented message condition, expressed a need for additional explanations before deciding to use GenAI healthcare services. The data also indicated that some participants preferred consulting a human doctor if they observed symptoms. This may explain why individuals with low-construal mindsets showed lower intention to use the app, as they tend to focus more on feasibility.

These findings attest to the need of accounting for individual differences in construal levels. When both goal- and duty-oriented messages provide general information, exposure to new or more detailed information may influence individuals' motivation to reflect and override their biases (Lai et al., 2013; Payne & Correll, 2020; Serenko & Turel, 2020), even though implicit associations are often stable. In such cases, individuals may rely more on explicit information for decision-making (FitzGerald et al., 2019), reinforcing the need for tailored, informative marketing communications to address potential biases and enhance trust.

Therefore, in Study 4, we manipulate construal levels (rather than measure them) and include multiple messages per condition to test our premise: whether the matching effect (message orientation × message construal) on consumer intention to use GenAI healthcare services emerges when incorporating "how" and "why" explanations in the marketing communications. We anticipate that across the three contexts (prevention, diagnosis, treatment), participants exposed to a specific message focus will be more persuaded by a corresponding explanatory message compared to other message combinations.

3.4 Study 4a-4c: message construal as moderator

Study 4 was conducted to investigate how message orientation (goal vs. duty) and message construal (how vs. why) influence consumers' intentions to use GenAI healthcare services across the same contexts as the previous study: prevention (4a), diagnosis (4b), and treatment (4c). A two-factor factorial design was employed, with

participants randomly assigned to either a goal- or duty-oriented message condition, combined with either a why or how explanation. In each of the three studies, the primary outcome variable of interest was consumers' intention to use the CareAI app.

3.4.1 Participants, design, and procedures

We conducted Studies 4a-c using the Prolific subject pool with 724 participants (Study 4a: n = 240, M_{age} = 42.2 years, 59.2% female; Study 4b: n = 239, M_{age} = 44.9 years, 61.1% female; Study 4c: n = 245, M_{age} = 44.3 years, 56.3% female). The studies had a 2 (message orientation: goal vs. duty) × 2 (message construal: how vs. why) between-subjects factorial design and were carried out in three healthcare service contexts. As such, 12 different advert stimuli were created to reflect our message combinations (see Appendix C). After participants confirmed they had read the introductory messages about CareAI, they were presented with adverts for message orientation from previous studies, along with new explanation stimuli about GenAI usage in healthcare services, to manipulate both message orientation (goal vs. duty) and message construal (how vs. why). Specifically, *why* explanations (high-level construal) focused on abstract and broader objectives, highlighting the overall purpose of using GenAI for healthcare services. In contrast, *how* explanations (low-level construal) focused on concrete details, emphasising the specific features and processes involved in using GenAI for healthcare services. Messages were slightly varied for different healthcare services (see Appendix C).

In each study, participants were randomly assigned to one of four marketing message combinations. Following the manipulated messages, participants were

asked to rate their agreement with the statement: "This advert focuses on explaining how over why CareAI uses GPT-4 for achieving accurate health outcomes/ensuring user safety and inclusiveness" where 1 = *why* and 7 = *how*, for a manipulation check of message construal. Participants then completed the use intention measures and provided demographic information.

3.4.2 Results and discussion

Manipulation. For each study, we included a manipulation check and an attention check question to assess the effectiveness of the message construal manipulations. The results from ANOVAs indicated significant differences between the why and how conditions across all studies (Study 4a: $F(1, 238) = 5.099$, $p = .025$; Study 4b: $F(1, 237) = 11.136$, $p < .001$; Study 4c: $F(1, 243) = 6.606$, $p = .011$). Additionally, 85.4%, 90.8%, and 86.5% of participants passed the attention checks in Study 4a, 4b, and 4c, respectively. Thus, our manipulations were successful.

Intention to use. We ran between-subjects ANOVAs with message orientation (goal vs. duty) and message construal (how vs. why) as independent variables and intention to use the app as the dependent variable. In the prevention context (study 4a), the model was significant ($F(3, 236) = 3.216$, $p = .024$, $\eta^2 = .039$). The analysis revealed a significant interaction effect ($F(1, 236) = 4.596$, $p = .033$, $\eta^2 = .019$), and a main effect of message construal ($F(1, 236) = 5.403$, $p = .021$, $\eta^2 = .022$), but no main effect of message orientation ($F(1, 236) = 0.040$, $p = .841$, $\eta^2 = .000$). Probing of the interaction showed that participants in the goal (vs. duty) condition reported higher intention to use GenAI prevention services when exposed to a why (vs. how) explanation ($M_{\text{goal-why}} = 3.91$, $SD = 1.61$; $M_{\text{goal-how}} = 2.91$, $SD = 1.66$; $p = .001$; $\eta^2 = .087$).

However, in the duty condition, participants' intention to use the app did not differ across message construal (p =.901). Thus, H3a was not supported.

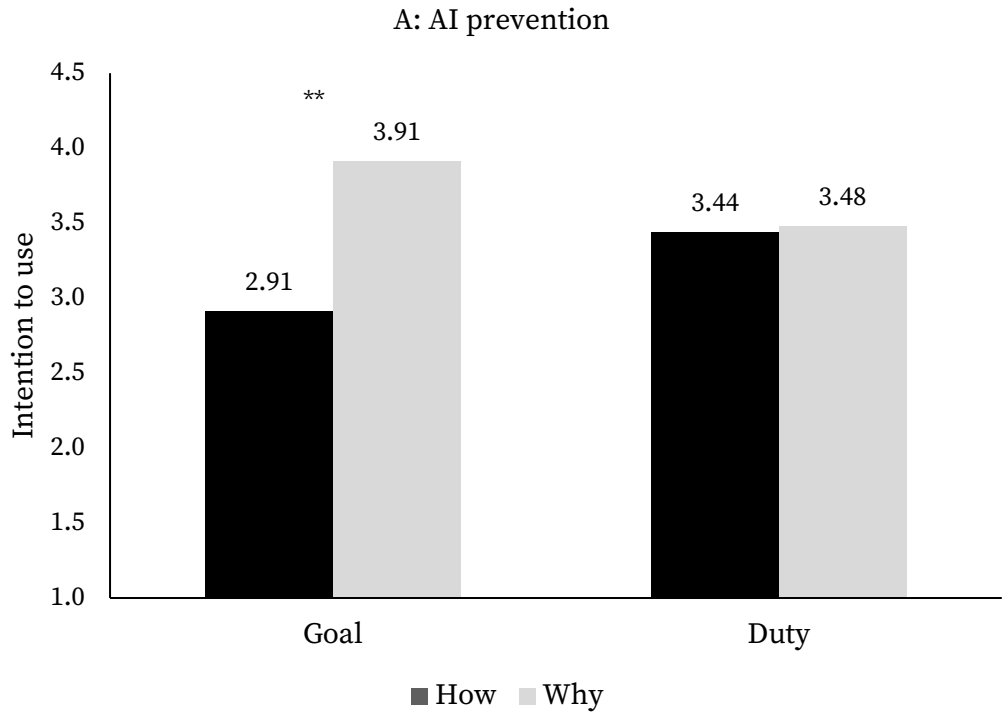
In the diagnosis context (Study 4b), the model was found to be significant (F(3, 235) = 3.420, p =.018, η^2 = .042). The analysis revealed a significant interaction effect (F(1, 235) = 4.335, p =.038, η^2 = .018), and a significant main effect of message construal (F(1, 235) = 5.859, p =.016, η^2 = .024), but no significant main effect of message orientation (F(1, 235) = .250, p =.617, η^2= .001). Probing the interaction showed that participants in the goal condition reported significantly higher intention to use the app when exposed to a how (vs. why) explanation ($M_{\text{goal-how}}$ = 3.86, SD = 1.82; $M_{\text{goal-why}}$ = 2.83, SD = 1.75; p = .002). However, participants in the duty condition did not show differences in intention across message construal (p = .808). Thus, H3c was supported.

In the treatment context (Study 4c), the model was significant (F(3, 241) = 2.745, p =.044, η^2 = .033); however, the analysis did not reveal any significant main or interaction effects (all ps > .066). Despite this, planned comparisons showed that participants reported higher intention to use the app when presented with goal × why communications compared to duty × why communications ($M_{\text{goal-why}}$ = 4.01, SD = 1.57; $M_{\text{duty-why}}$ = 3.35, SD = 1.65; F(1, 118) = 4.977, p = .028) or goal × how communications ($M_{\text{goal-why}}$ = 4.01, SD = 1.57; $M_{\text{goal-how}}$ = 3.35, SD = 1.70; F(1, 119) = 4.865, p = .029). These findings suggest that construal level had a significant effect within the goal condition. Thus, H3b was not supported.

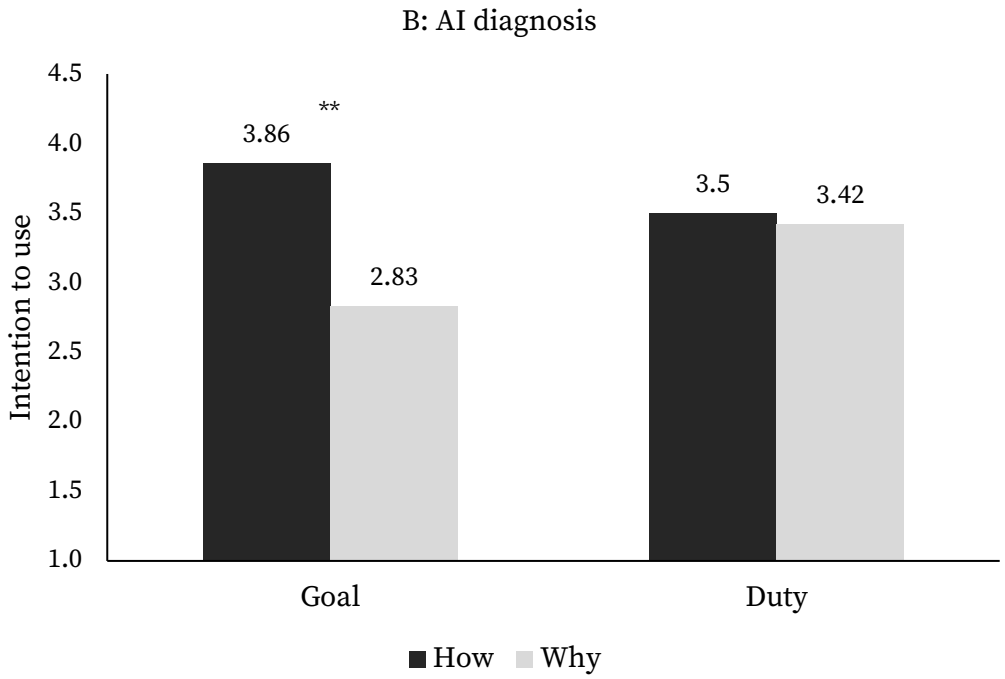
Table IV.4 Intention to use by message orientation, message construal, and healthcare services (S4a-4c)

	Goal Condition		Duty Condition	
	How	Why	How	Why
Intention to use				
GenAI prevention	2.91 (1.66)[a]	3.91 (1.61)[b]	3.44 (1.77)[a]	3.48 (1.88)[a]
GenAI diagnosis	3.86 (1.61)[b]	2.83 (1.75)[a]	3.50 (1.72)[a]	3.42 (1.78)[a]
GenAI treatment	3.35 (1.70)[a]	4.01 (1.57)[b]	3.21 (1.81)[a]	3.35 (1.65)[a]

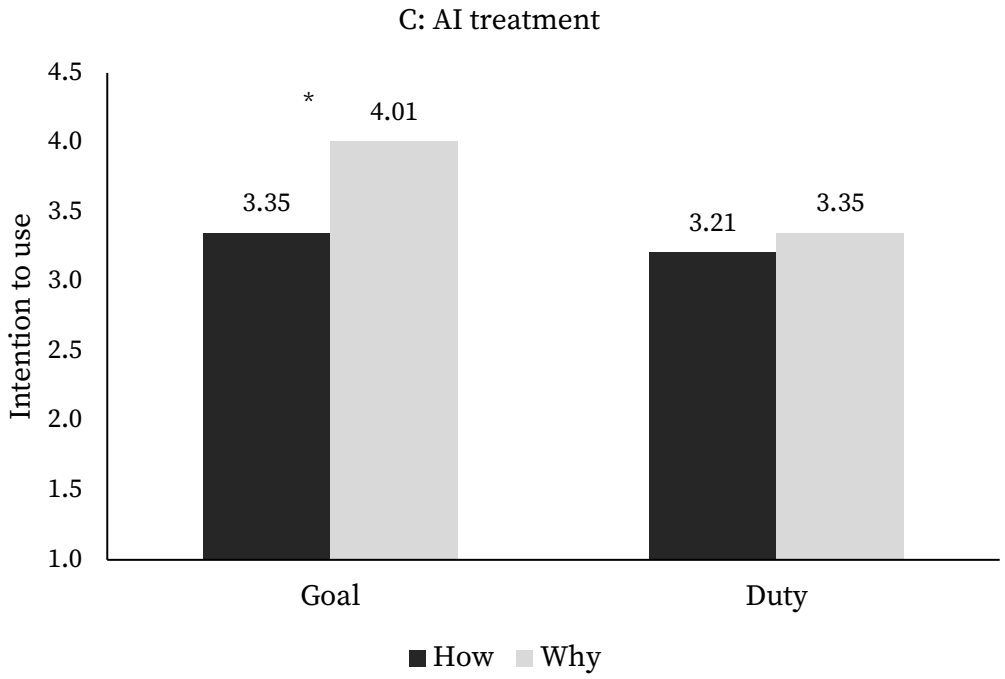
Note. Participants were randomly assigned to each condition. Cells contain means with standard deviations in parentheses. Higher scores indicate higher behavioural intentions. Different superscripts within a row indicate the means are significantly different at $p < .05$.



Note. ** p < 0.01.



Note. ** $p < 0.01$.



Note. * $p < 0.05$.

Figure IV.3 Intention to use GenAI healthcare service varies by message orientation, moderated by message construal

Discussion. Study 4 explored the impact of message orientation and message construal on consumer intention to use GenAI healthcare services. As predicted, the *goal* × *how* message significantly outperformed other message combinations in fostering consumer intentions to use GenAI for diagnosis purposes. Conversely, the *goal* × *why* message was more effective than alternative combinations in enhancing consumer use intention for GenAI prevention services. For treatment services, while the *goal* × *why* message directionally resulted in higher intentions to use than other combinations, the difference was not statistically significant. This result was unexpected, but a noteworthy finding was that participants in the *goal* × *why* condition exhibited significantly higher intentions compared to those in the *duty* × *why* condition. Taken together, these findings suggest that with multiple messages, evoking an abstract mindset should be able to draw consumer attention toward the desirability of goal-oriented messages over duty-oriented messages, without compromising interpretability, one of the key motivations for categorising a continuous variable—thereby leading to more favourable evaluations.

When preventive and therapeutic health outcomes are perceived as psychologically distant events (not directly experienced), participants tend to place greater emphasis on the superordinate aspects of goal-oriented messages (e.g., maintaining/achieving health). These messages, paired with why explanations (evoking higher construal levels), are more likely to activate higher construal mindsets. Consequently, the combination of a goal-oriented message and why explanation serves as a distal driver of construal level, increasing the likelihood that participants will delegate illness prevention and treatment actions to GenAI agents.

According to the ADM, accessible information is not used for judgment and decision-making if more probative, relevant, or useful information is available (Feldman & Lynch 1988; Lynch et al., 1988; Herr et al., 1991). When little or no performance information is known, consumers may hold implicit prejudices against GenAI and exhibit aversion to its associated risks (Logg et al., 2019). However, when provided with additional accessible information that indicates GenAI's performance (diagnostic), which is often linked to explanations in goal-oriented messages, consumers may develop a clearer sense of the likelihood of success regarding GenAI's preventive actions. This allows them to make trade-off decisions under risk rather than ambiguity. With this diagnostic information (why explanations), consumers engage in systematic-reflective thinking, aligning their evaluations with the new information while relying less on surface cues (e.g., automatically retrieved prejudices towards GenAI technologies) (Bechara & Damasio, 2005). As a result, they may shift their focus away from higher-order aspirations like inclusiveness and instead prioritise the desirability of instrumental and functional GenAI services. Triggering an abstract mindset in this context should not add to the high-level processing for duty-oriented messages, thus producing a ceiling effect. From this lens, the transition is driven by access to relevant GenAI performance information, which helps reshape consumer perceptions and decision-making.

For GenAI prevention and diagnosis services, goal-oriented messages align with CLT, wherein the effectiveness of why (high-level construal) or how (low-level construal) explanations depends on the psychological distance of the goals. In contrast, duty-oriented messages appear less influenced by these effects.

Specifically, duty-oriented messages, regardless of the type of explanation paired with them, tend to yield more consistent and moderate consumer intentions across experimental conditions. One possible explanation is that extrinsic motivations tied to duty-oriented messages are less dependent on the specifics of an action (how) or its purpose (why). Also, the appeal to duty often carries moral weight that triggers emotional responses, making individuals less sensitive to detailed information. This may explain why pairing duty-oriented messages with either type of explanation led to similarly moderate evaluations. From the ADM perspective, diagnostic information is the crucial factor influencing judgments rather than merely accessible information (Herr et al., 1991). Respondents may prioritise the quality and relevance of duty-oriented messages over additional explanations. This interpretation, while reasonable to some extent, merits further research.

Given the reversal findings in the matching effects within the contexts of prevention and diagnosis, Study 5 was designed to investigate the underlying mechanism of the marketing communications for these two healthcare services. Specifically, we examined the nature of the interaction among three variables and tested whether the effects on consumer intention to use GenAI services were mediated by perceived information diagnosticity, providing further evidence of the persuasion process.

3.5 Study 5: The mediating role of information diagnosticity

Study 5 tested the role of perceived information diagnosticity as a driver of consumers' intention to use GenAI healthcare services when exposed to combined marketing communications (message orientation × message construal) in the

contexts of prevention and diagnosis. We hypothesise that participants will perceive the *goal* × *why* message as more diagnostic, leading to higher intention to use the app in the prevention context. Conversely, we expect participants to find the *goal* × *how* message more diagnostic, resulting in higher intention to use the app in the diagnosis context.

3.5.1 Participants, design, and procedures

A total of 479 Prolific workers (M_{age} = 41.8 years, 60.1% female) were each randomly assigned to one of the conditions in a 2 (message orientation: goal vs. duty) × 2 (message construal: how vs. why) × 2 (healthcare: prevention vs. diagnosis) factorial design. The healthcare manipulation involved describing the services as using GenAI to either prevent skin cancer (prevention) or diagnose headaches (diagnosis). After viewing the message stimuli (same as in previous studies), participants completed perceived information diagnosticity questions (three items, seven-point scales, α = .89), adapted from Baker and Lutz (2000) and Baker (2001), to serve as the mediator for the predicted effect in the marketing communications. The primary outcome variable was consumers' intention to use the CareAI app. Finally, participants provided demographic information.

3.5.2 Results and discussion

Manipulations. Consistent with Study 4, ANOVA results for the manipulation checks on message orientation (goal vs. duty) and message construal (how vs. why) confirm that the manipulations were successful across all studies (all *ps* < .05).

Intention to use. A three-way ANOVA was conducted with message orientation (goal vs. duty), message construal (how vs. why), and healthcare service (prevention vs. diagnosis) as independent variables and intention to use as the dependent variable. The analysis yielded a significant three-way interaction effect ($F(1, 475) = 8.926$, $p = .003$, $\eta_p^2 = .019$), indicating that the effectiveness of combined marketing communications (message orientation × message construal) on consumer intention to use the app differs between prevention and diagnosis contexts. Additionally, a significant two-way interaction between message construal and healthcare service ($F(1, 475) = 11.262$, $p < .001$, $\eta_p^2 = .023$) was found, suggesting that participants exposed to a higher (vs. lower) construal mindset (why message) would have higher intention to use GenAI prevention (vs. AI diagnosis) services. No significant main effects or other interaction effects were detected (*ps.* > .619).

Mediation. To further investigate these interactions, we tested whether perceived information diagnosticity mediates their effect on consumers' intention to use the app. We first conducted a three-way ANOVA to examine the effects of message orientation (goal vs. duty), message construal (how vs. why), and healthcare service (prevention vs. diagnosis) on perceived information diagnosticity. The analysis revealed a significant interaction effect ($F(1, 475) = 5.655$, $p = .018$, $\eta_p^2 = .012$). Pairwise comparisons confirmed that the pattern of results was similar to the observed pattern of behavioural intention. Next, we conducted a mediation analysis with information diagnosticity as a process variable to examine whether it mediates the effect of the interaction (i.e., paired marketing communications for different health services) on intention to use the app. In the procedures, message orientation, message construal, and healthcare service were contrast coded (GenAI prevention:

goal, why, prevention = 1, others = 0; GenAI diagnosis: goal, how, diagnosis = 1, others = 0). To test the hypothesised mediation, we performed mediated moderation analyses (Muller et al., 2005) using PROCESS Model 4 (10,000 bootstrap resamples; Hayes, 2013).

In this analysis, message orientation × message construal × healthcare service served as the independent variable, information diagnosticity (mean centered) as the mediator, and intention to use as the dependent variable. All independent variables (message orientation, message construal) and their two-way interaction terms (message orientation × message construal, message orientation × healthcare service, message construal × healthcare service) were included as covariates. As predicted, bootstrapping procedures revealed that the indirect effect (i.e., the path through the mediator) was significant (GenAI prevention: β = 1.34, SE = .38; 95% CI = [.584, 2.121]; GenAI diagnosis: β = 1.04, SE = .43; 95% CI = [.198, 1.904]). The direction of the effects confirmed that *goal* × *why* × *prevention* and *goal* × *how* × *diagnosis* marketing communications can lead to higher perceived information diagnosticity, which in turn increased intention to use GenAI healthcare services. Thus, H4 was supported.

Table IV.5 Intention to use and information diagnosticity by message orientation, message construal, and healthcare services (S5)

	Goal Condition		Duty Condition	
	How	Why	How	Why
Intention to use				
GenAI prevention	2.91 (1.66)[a]	3.91 (1.61)[b]	3.44 (1.77)[a]	3.48 (1.88)[a]
GenAI diagnosis	3.86 (1.61)[b]	2.83 (1.75)[a]	3.50 (1.72)[a]	3.42 (1.78)[a]
GenAI treatment	3.35 (1.70)[a]	4.01 (1.57)[b]	3.21 (1.81)[a]	3.35 (1.65)[a]
Information diagnosticity				
GenAI prevention	4.02 (1.17)[a]	4.42 (1.45)[b]	3.94 (1.33)[a]	3.80 (1.42)[a]
GenAI diagnosis	4.60 (1.25)[b]	3.92 (1.66)[a]	4.30 (1.35)[a]	4.30 (1.52)[a]
GenAI treatment	4.21 (1.36)[a]	4.73 (1.23)[b]	4.11 (1.56)[a]	4.00 (1.45)[a]

Notes. Participants were randomly assigned to each condition. Cells contain means with standard deviations in parentheses. Higher scores indicate higher behavioural intentions. Different superscripts within a row indicate the means are significantly different at $p < .05$.

Discussion. The results of Study 5 provide additional empirical evidence that supports the proposed matching effects and the underlying processes—perceived information diagnosticity—through which a match among message orientation, message construal, and healthcare service affects intention to use. As predicted, a match between a goal-oriented message and high-level construal (why explanations) in the prevention context, as well as a match between a goal-oriented message and low-level construal (how explanations) in the diagnosis context, both led to higher intention to use GenAI healthcare services via perceived information diagnosticity. This suggests that when GenAI healthcare service marketing communications align with consumers' construal mindsets and the nature of the healthcare service, they perceive the message as more informative and relevant, ultimately increasing adoption intent.

4. General Discussion

Across five experiments in different health domains, this research identifies conditions that enhance consumer responses (intention to use, trust) toward GenAI-powered healthcare services. Studies 1 and 2 show duty-oriented messages are more effective for GenAI prevention services, while Study 3 finds high-construal consumers prefer goal-oriented messages for illness prevention, aligning with their abstract thinking and long-term health goals. No message orientation advantage emerges for diagnosis and treatment services. Studies 4 and 5 confirm that aligning message orientation with message construal enhances perceived information diagnosticity, boosting consumer intention to use GenAI healthcare services. Goal-oriented messages paired with why explanations work best for prevention and treatment, while how explanations are more effective for diagnosis. Duty-oriented messages remain equally effective across conditions. These findings provide a strategic framework for optimising GenAI healthcare marketing, ensuring greater consumer engagement and adoption.

4.1 Theoretical contributions

Theoretically, our study contributes to the literature on three main fronts. First, it adds to the knowledge of CLT in the context of GenAI healthcare marketing by demonstrating how psychological distance influences consumer responses. While prior studies have primarily linked CLT to consumer decision-making and persuasive messaging (Adler & Sarstedt, 2021), relatively little attention has been given to societal focus (e.g., responsible AI communication) in CLT-related research (Davis & Ozanne, 2019; Viglia, 2020). We address this gap by empirically showing

that strategic message framing influences consumer engagement with GenAI healthcare services. We highlight the role of psychological distance and reveal a matching effect between message orientation and construal level across different service types. This context-dependent effect underscores the need for tailored marketing communication strategies that align with consumer mindsets.

Beyond its traditional role of defining psychological distance as a determinant of construal levels (Trope & Liberman, 2010), our study extends CLT to GenAI communication strategies, demonstrating that message design itself serves as a cognitive cue that activates construal mindsets, influencing how consumers process and evaluate services. By examining message orientation and message construal, we provide novel insights into how psychological distance interacts with marketing communications in high-stakes domains like healthcare. Our findings suggest that consumers prefer different construal levels depending on when the health outcome is expected, and that the effectiveness of message orientation depends on both the activated mindset and perceived information diagnosticity.

Second, this research advances CUT and ADM by identifying perceived information diagnosticity as a key psychological mechanism in GenAI healthcare adoption. Our study bridges these theories with marketing communication, demonstrating how consumers evaluate informational cues—assessing their relevance, usefulness, and weight—when processing matched marketing messages. Our findings extend prior research on trust in AI and algorithmic decision-making (Dietvorst et al., 2015; Turel & Kalhan, 2023) by showing that alignment between message orientation and construal level enhances perceived information diagnosticity, making GenAI healthcare communications more persuasive across different healthcare contexts.

This alignment reduces algorithm aversion and increases consumer acceptance of GenAI healthcare services.

Moreover, we expand CUT's application beyond traditional product evaluations to digital GenAI healthcare services, where intrinsic and extrinsic cues are indicators of quality (Richardson et al., 1994). However, in the context of GenAI healthcare, intrinsic cues (e.g., firsthand experience) are often unavailable. Consumers rely on extrinsic informational cues, such as message orientation and message construal, to assess service credibility, safety, and effectiveness to guide decision-making. Additionally, our findings support and extend research on the situational variability of CUT (e.g., Wang et al., 2016), demonstrating that the diagnosticity of message framing varies across healthcare services. We further contribute to consumer psychology research by demonstrating how individual construal mindsets shape cue reliance, reinforcing the influence of cognitive processing styles on cue weighting and perceived information diagnosticity in consumer responses to GenAI healthcare.

Third, this study contributes to the IS literature on AI ethics and responsible AI adoption, as well as literature on AI healthcare marketing, by highlighting the critical role of marketing communication in shaping consumer responses to GenAI healthcare. Prior research has focused on technical transparency and regulatory compliance (Mittelstadt et al., 2016; Tsamados et al., 2022), but our findings show that strategic message framing is equally vital in fostering AI trust and adoption in high-stakes healthcare contexts. We introduce a message design framework that operationalises GenAI healthcare communication, demonstrating that message orientation and message construal interact to shape consumer responses. Our

findings reveal that well-calibrated, aligned marketing messages reduce algorithm aversion and foster greater acceptance of AI-powered healthcare services, offering actionable insights for tailoring persuasive messaging in GenAI healthcare.

4.2 Practical implications

Our findings offer clear practical implications for marketing managers of GenAI healthcare brands. Recognising the strategic importance of marketing communications for different GenAI healthcare services, we encourage marketers to consider the cognitive mindsets of their target consumers. As GenAI technologies evolve and become increasingly integrated into healthcare, our study explores communication strategies that have received limited attention, particularly in conveying AI's role in delivering human-centric and empathetic care. Understanding how to effectively communicate GenAI healthcare services provides valuable insights for marketers to remain forward-thinking and socially responsible in their messaging. These insights also hold relevance for the digital healthcare industry, technology developers, policymakers, and consumers alike.

We advise marketers to capitalise on strategic message framing by deliberately employing message orientation (goal vs. duty) and message construal (how vs. why explanations) to enhance consumer engagement with GenAI healthcare services. Specifically, for GenAI prevention services, we recommend using a duty-oriented message that highlights ethical and societal responsibilities associated with AI-driven illness prevention. Alternatively, a goal-oriented message combined with why explanations can effectively communicate the benefits and purpose of adopting GenAI for proactive health management. For GenAI diagnosis and

treatment services, the messaging strategy shifts from a dual approach to a more targeted, outcome-focused approach. Communicating the operational aspects and accuracy of GenAI in diagnosing health conditions can enhance transparency, increasing consumer understanding and appreciation of its technological advantages. For treatment services, emphasising personalised care plans and improved health outcomes can address consumer needs for effective and tailored healthcare solutions, reinforcing GenAI's role in enhancing treatment effectiveness.

By identifying the contexts in which marketing communications are most effective, we advocate for a nuanced approach to GenAI healthcare marketing. Demonstrating that message framing influences abstract or concrete thinking, our research extends CLT's applicability and highlights its impact on perceived diagnosticity, trust, and adoption. These insights underscore the strategic importance of message design in GenAI marketing, providing a framework for crafting persuasive yet ethical communications that balance technological innovation with societal responsibility, ultimately fostering consumer confidence in GenAI healthcare solutions.

4.3 Limitation and future research

This research is subject to several limitations that offer opportunities for future research to expand upon its findings. First, while this study focuses on GenAI's advisory role in healthcare, its generalisability to other, particularly more sensitive, domains remains uncertain, requiring further investigation across diverse applications. As GenAI adoption expands into high-stakes decision-making—such

as automated financial advising, GenAI-driven legal analysis, or even GenAI-assisted policymaking—the ethical implications of its widespread use must be carefully considered when marketing these products/services.

For instance, in the finance sector, GenAI-powered investment and risk assessment services may require a balance between goal-oriented efficiency messaging and duty-oriented assurances emphasising security and ethical responsibility. In legal services, where fairness and compliance are critical, consumers may expect greater transparency and explainability in GenAI-generated recommendations. Future research could explore industry-specific factors influencing consumer responses, as well as the effectiveness of combining goal- and duty-oriented messages and determining the optimal sequencing to maximise consumer engagement and trust.

Second, in the healthcare domain, while message orientation, construal level, and message construal are relevant, other conditions such as behavioural knowledge (Snefjella & Kuperman, 2015), message-related features (Lee, 2019), and contextual factors (e.g., severity of medical conditions; Longoni et al., 2019) may shape consumer responses to GenAI healthcare service marketing communications. Future research should explore how these factors interact with different communication strategies, as this could provide a more nuanced understanding of the complex dynamics involved in promoting GenAI healthcare services.

Moreover, individual differences may also play a role. For example, additional analyses (Appendix D) reveal that higher-income participants prefer duty-oriented messages for GenAI diagnosis services, potentially indicating that affordability concerns (Longoni et al., 2019) influence consumer receptivity to marketing

communications in GenAI-powered healthcare. Future studies could further investigate socio-demographic factors, such as age, health literacy, and prior experience with AI-driven services, to understand heterogeneous consumer responses.

Furthermore, while this study highlights perceived information diagnosticity as a key driver of consumer intention to use GenAI healthcare services in response to well-matched marketing communications, consumer preferences for these services are inherently complex and likely shaped by multiple psychological mechanisms. For instance, Longoni et al. (2019) suggest that dehumanisation concerns and morality considerations may lead consumers to rely more on human clinical judgments, even when AI-based solutions offer comparable or superior accuracy. Future research should explore additional psychological factors to better understand the boundaries of consumer acceptance of GenAI-driven healthcare in response to goal- or duty-oriented marketing communications.

Third, as for methodological limitations, the experimental manipulations relied on hypothetical scenarios, which may limit ecological validity. Future research could strengthen manipulations by employing quasi-experimental designs with actual users of GenAI healthcare services, ensuring that participants' engagement and decision-making more closely reflect real-world behaviour. Also, it is important to control for unintended influences in study stimuli, such as perceptions of service quality, which may inadvertently shape consumer reactions.

Moreover, this study relied on a Western sample (UK residents), yet prior research (e.g., Aaker, 2000) highlights cultural differences in persuasion processes and

marketing communication effects. Consumers in different cultures may hold distinct values and preferences, influencing how they respond to message focus in marketing communications. Future research could explore cultural variations, as studies suggest that Asian consumers may have lower sensitivity to AI use compared to their Western counterparts.

Furthermore, to mitigate brand influence, this study used fictitious brands in GenAI healthcare scenarios. However, real-world consumers typically prefer familiar brands, especially in high-stakes decision-making contexts (Saini & Lynch, 2016; Vaidyanathan, 2000). Future studies should investigate whether brand familiarity influences consumer responses to GenAI-powered healthcare services and whether consumers react differently to marketing communications when the GenAI service is offered by a well-known versus unfamiliar brand.

While this research provides valuable insights into marketing communications for GenAI healthcare services, further studies are needed to refine these findings by incorporating contextual, psychological, cultural, and brand-related factors. Addressing these limitations will offer a more comprehensive understanding of how to optimise marketing strategies for GenAI adoption in healthcare and beyond.

Chapter IV References

Aaker, J. L. (2000). Accessibility or diagnosticity? Disentangling the Influence of Culture on Persuasion Processes and Attitudes. *Journal of Consumer Research*, 26(4), 340–357. https://doi.org/10.1086/209567

Achar, C., Agrawal, N., & Hsieh, M. H. (2020). Fear of detection and efficacy of prevention: Using construal level to encourage health behaviours. *Journal of Marketing Research, 57*(3), 582–598. https://doi.org/10.1177/0022243720912443

Adler, S., & Sarstedt, M. (2021). Mapping the jungle: A bibliometric analysis of research into construal level theory. *Psychology & Marketing, 38*(9), 1367–1383. https://doi.org/10.1002/mar.21537

Aggarwal, P., & Zhang, M. (2015). Seeing the big picture: The effect of height on the level of construal. *Journal of Marketing Research, 52*(1), 120–133. https://doi.org/10.1509/jmr.12.0067

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods, 17*(4), 351–371. https://doi.org/10.1177/1094428114547952

Ahluwalia, R. (2002). How prevalent is the negativity effect in consumer environments? *Journal of Consumer Research, 29*(2), 270–279. https://doi.org/10.1086/341576

Akdeniz, B., Calantone, R. J., & Voorhees, C. M. (2013). Effectiveness of marketing cues on consumer perceptions of quality: The moderating roles of brand reputation and third-party information. *Psychology & Marketing, 30*(1), 76–89. https://doi.org/10.1002/mar.20590

Allen, R., & Choudhury, P. (2022). Algorithm-augmented work and domain experience: The countervailing forces of ability and aversion. *Organization Science, 33*(1), 149–169. https://doi.org/10.1287/orsc.2021.1554

Andreasen, A. R. (2006). *Social marketing in the 21st century*. Thousand Oaks, CA: Sage Publications.

Bagozzi, R. P., & Dholakia, U. (1999). Goal setting and goal striving in consumer behavior. *Journal of Marketing, 63*(4), 19–32. https://doi.org/10.2307/1252098

Baird, A., & Maruping, L. M. (2021). The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. *MIS Quarterly, 45*(1), 315. https://doi.org/10.25300/MISQ/2021/15882

Baker, W. E. (2001). The diagnosticity of advertising generated brand attitudes in brand choice contexts. *Journal of Consumer Psychology, 11*(2), 129–139. https://doi.org/10.1207/S15327663JCP1102_05

Baker, W. E., & Lutz, R. J. (2000). An empirical test of an updated relevance-accessibility model of advertising effectiveness. *Journal of Advertising, 29*(1), 1–14. https://doi.org/10.1080/00913367.2000.10673599

Bechara, A., & Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and Economic Behavior, 52*(2), 336–372. https://doi.org/10.1016/j.geb.2004.06.010

Brown, T., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33,* 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Burton, J. W., Stein, M. K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making, 33*(2), 220–239. https://doi.org/10.1002/bdm.2155

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research, 56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Cox, D. F. (1962). The measurement of information value: A study in consumer decision-making. In *Emerging concepts in marketing* (pp. 413–421). Chicago, IL: American Marketing Association.

Davis, B., & Ozanne, J. L. (2019). Measuring the impact of transformative consumer research: The relational engagement approach as a promising avenue. *Journal of Business Research, 100,* 311-318. https://doi.org/10.1016/j.jbusres.2018.12.047

Dick, A., Chakravarti, D., & Biehal, G. (1990). Memory-based inferences during consumer choice. *Journal of Consumer Research, 17*(1), 82–93. https://doi.org/10.1086/208539

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2016). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Ding, F., Li, D., & George, J. F. (2014). Investigating the effects of IS strategic leadership on organisational benefits from the perspective of CIO strategic roles. *Information & Management, 51*(7), 865–879. https://doi.org/10.1016/j.im.2014.08.004

Feldman, J. M., & Lynch, J. G. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology, 73*(3), 421–435. https://doi.org/10.1037/0021-9010.73.3.421

Filiz, I., et al. (2021). Reducing algorithm aversion through experience. *Journal of Behavioral and Experimental Finance, 31,* 100524. https://doi.org/10.1016/j.jbef.2021.100524

FitzGerald, C., et al. (2019). Interventions designed to reduce implicit prejudices and implicit stereotypes in real-world contexts: A systematic review. *BMC Psychology, 7*(Article 29). https://doi.org/10.1186/s40359-019-0299-7

Freitas, A. L., Gollwitzer, P., & Trope, Y. (2004). The influence of abstract and concrete mindsets on anticipating and guiding others' self-regulatory efforts. *Journal of Experimental Social Psychology, 40*(6), 739–752. https://doi.org/10.1016/j.jesp.2004.04.003

Fujita, K., et al. (2006). Construal levels and self-control. *Journal of Personality and Social Psychology, 90*(3), 351–367. https://doi.org/10.1037/0022-3514.90.3.351

Gelman, A., & Park, D. K. (2007). Splitting a predictor at the upper quarter or third and the lower quarter or third. *The American Statistician, 63*(1), 1–8. https://doi.org/10.1198/tast.2009.0001

George, G., et al. (2021). Purpose in the for-profit firm: A review and framework for management research. *Journal of Management, 49*(6), 1841–1869. https://doi.org/10.1177/01492063211006450

Golder, P. N., et al. (2023). Learning from data: An empirics-first approach to relevant knowledge generation. *Journal of Marketing, 87*(3), 319–336. https://doi.org/10.1177/00222429221129200

Han, D., Duhachek, A., & Agrawal, N. (2016). Coping and construal level matching drives health message effectiveness via response efficacy or self-efficacy enhancement. *Journal of Consumer Research, 43,* 429–447. https://doi.org/10.1093/jcr/ucw036

Herr, P. M., Kardes, F. R., & Kim, J. (1991). Effects of word-of-mouth and product-attribute information on persuasion: An accessibility-diagnosticity perspective. *Journal of Consumer Research, 17*(4), 454–462. https://doi.org/10.1086/208570

Huang, M. H., & Rust, R. T. (2021). Engaged to a robot? The role of AI in service. *Journal of Service Research, 24*(1), 30–41. https://doi.org/10.1177/1094670520902266

Huang, M. H., & Rust, R. T. (2023). EXPRESS: The caring machine: Feeling AI for customer care. *Journal of Marketing.* https://doi.org/10.1177/00222429231224748

Hulland, J., Baumgartner, H., & Smith, K. M. (2018). Marketing survey research best practices: Evidence and recommendations from a review of JAMS articles. *Journal of the Academy of Marketing Science, 46,* 92–108. https://doi.org/10.1007/s11747-017-0532-y

Jo, A. (2023). The promise and peril of generative AI. *Nature, 614*(1), 214–216.

Kahn, B. E., & Luce, M. F. (2003). Understanding high-stakes consumer decisions: Mammography adherence following false-alarm test results. *Marketing Science, 22*(2), 393–410. https://doi.org/10.1287/mksc.22.3.393.17737

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist, 39*(4), 341–350. https://doi.org/10.1037/0003-066X.39.4.341

Keller, P. A. (2006). Regulatory focus and efficacy of health messages. *Journal of Consumer Research, 33*(1), 109–114. https://doi.org/10.1086/504141

Keller, P. A., & Lehmann, D. R. (2008). Designing effective health communications: A meta-analysis. *Journal of Public Policy & Marketing, 27*(2), 117–130. https://doi.org/10.1509/jppm.27.2.117

Kivetz, Y., & Tyler, T. R. (2007). Tomorrow I'll be me: The effect of time perspective on the activation of idealistic versus pragmatic selves. *Organizational Behavior and Human Decision Processes, 102*(2), 193–211. https://doi.org/10.1016/j.obhdp.2006.07.002

Kruglanski, A. W., et al. (2018). A structural model of intrinsic motivation: On the psychology of means-ends fusion. *Psychological Review, 125*(2), 165–182. https://doi.org/10.1037/rev0000095

Lai, C. K., Hoffman, K. M., & Nosek, B. A. (2013). Reducing implicit prejudice. *Social and Personality Psychology Compass, 7*(5), 315–330. https://doi.org/10.1111/spc3.12023

Lee, S. J. (2019). The role of construal level in message effects research: A review and future directions. *Communication Theory, 29*(3), 319–338. https://doi.org/10.1093/ct/qty030

Li, J., et al. (2021). Strategic directions for AI: The role of CIOs and boards of directors. *MIS Quarterly, 45*(3), 1603–1644. https://doi.org/10.25300/MISQ/2021/16523

Lin, Y. C., & Chang, C. C. A. (2021). Influencing consumer responses to highly aesthetic products: The role of mindsets. *Journal of Retailing, 97*(3), 459–476. https://doi.org/10.1016/j.jretai.2020.10.004

Liberman, N., & Trope, Y. (1998). The role of feasibility and desirability considerations in near and distant future decisions: A test of temporal construal theory. *Journal of Personality and Social Psychology, 75*(1), 5–18. https://doi.org/10.1037/0022-3514.75.1.5

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151,* 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Luo, X., et al. (2019). Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science, 38*(6), 937–947. https://doi.org/10.1287/mksc.2019.1192

Lynch, J. G., Marmorstein, H., & Weigold, M. F. (1988). Choices from sets including remembered brands: Use of recalled attributes and prior overall evaluations. *Journal of Consumer Research, 15*(2), 169–184. https://doi.org/10.1086/209155

Magids, S., Zorfas, A., & Leemon, D. (2015). The new science of customer emotions. *Harvard Business Review,* 66–74

Mathur, P., et al. (2013). The influence of implicit theories and message frame on the persuasiveness of disease prevention and detection advocacies.

Organizational Behavior and Human Decision Processes, 122(2), 141–151. https://doi.org/10.1016/j.obhdp.2013.05.002

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research, 13*(3), 334–359. https://doi.org/10.1287/isre.13.3.334.81

Menon, G., Raghubir, P., & Schwarz, N. (1995). Behavioral frequency judgments: An accessibility-diagnosticity framework. *Journal of Consumer Research, 22*(2), 212–228.

Meskó, B., & Topol, E. J. (2023). The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digital Medicine, 6*(120), 1–6. https://doi.org/10.1038/s41746-023-00873-0

Morley, J., et al. (2020). The ethics of AI in healthcare: A mapping review. *Social Science & Medicine, 260*(113181), 1–14. https://doi.org/10.1016/j.socscimed.2020.113172

Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology, 89*(6), 852–863. https://doi.org/10.1037/0022-3514.89.6.852

Oliver, K. O. L. L. (2020). *Reactions on algorithms: A systematic literature review of algorithm aversion and algorithm appreciation* (Doctoral dissertation). The University of Innsbruck.

Payne, B. K., & Correll, J. (2020). Race, weapons, and the perception of threat. In *Advances in Experimental Social Psychology* (Elsevier). https://doi.org/10.1016/bs.aesp.2020.04.001

Pham, M. T., & Avnet, T. (2004). Ideals and oughts and the reliance on affect versus substance in persuasion. *Journal of Consumer Research, 30*(4), 503–518. https://doi.org/10.1086/380285

Richardson, P. S., Dick, A. S., & Jain, A. K. (1994). Extrinsic and intrinsic cue effects on perceptions of store brand quality. *Journal of Marketing, 58*(4), 28-36. https://doi.org/10.1177/002224299405800403

Rothman, A. J., Bartels, R. D., & Wlaschin, J. (2006). The strategic use of gain- and loss-framed messages to promote healthy behavior: How theory can inform practice. *Journal of Communication, 56*(1), S202–S220. https://doi.org/10.1111/j.1460-2466.2006.00290.x

Rust, R. T., Katherine, N. L., & Valarie, A. Z. (2004). Return on marketing: Using customer equity to focus marketing strategy. *Journal of Marketing, 68*(1), 109–127. https://doi.org/10.1509/jmkg.68.1.109.24030

Saini, Y. K., & Lynch, J. G. (2016). The effects of the online and offline purchase environment on consumer choice of familiar and unfamiliar brands. *International Journal of Research in Marketing, 33*(3), 702–705. https://doi.org/10.1016/j.ijresmar.2016.02.003

Salovey, P., et al. (2000). Emotional states and physical health. *American Psychologist, 55*(1), 110–121. https://doi.org/10.1037/0003-066X.55.1.110

Serenko, A., & Turel, O. (2020). Measuring implicit attitude in information systems research with the implicit association test. *Communications of the Association for Information Systems, 47,* 397–431. https://doi.org/10.17705/1CAIS.04719

Soderberg, C. K., et al. (2015). The effects of psychological distance on abstraction: Two meta-analyses. *Psychological Bulletin, 141,* 525–548. https://doi.org/10.1037/bul0000005

Snefjella, B., & Kuperman, V. (2015). Concreteness and psychological distance in natural language use. *Psychological Science, 26,* 1449–1460. https://doi.org/10.1177/0956797615591771

Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly, 36*(2), 368–383. https://doi.org/10.1016/j.giq.2018.09.008

Todorov, A., Goren, A., & Trope, Y. (2007). Probability as a psychological distance: Construal and preferences. *Journal of Experimental Social Psychology, 43*(3), 473–482. https://doi.org/10.1016/j.jesp.2006.04.002

Trope, Y., & Liberman, N. (2003). Temporal construal. *Psychological Review, 110*(3), 403–421. https://doi.org/10.1037/0033-295X.110.3.403

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117*(2), 440–463. https://doi.org/10.1037/a0018963

Trope, Y., Liberman, N., & Wakslak, C. (2007). Construal levels and psychological distance: Effects on representation, prediction, evaluation, and behavior. *Journal of Consumer Psychology, 12*(7), 83–95. https://doi.org/10.1016/S1057-7408(07)70013-X

Tsai, C., & McGill, A. L. (2011). No pain, no gain? How fluency and construal level affect consumer confidence. *Journal of Consumer Research, 37*(5), 807–821. https://doi.org/10.1086/655855

Turel, O., & Kalhan, S. (2023). Prejudiced against the machine? Implicit associations and the transience of algorithm aversion. *MIS Quarterly, 47*(4). https://doi.org/10.25300/MISQ/2022/17961

Ullah, S., et al. (2022). Assessing the influence of celebrity and government endorsements on bitcoin's price volatility. *Journal of Business Research, 145,* 228–239. https://doi.org/10.1016/j.jbusres.2022.01.055

Vaidyanathan, R. (2000). The role of brand familiarity in internal reference price formation: An accessibility-diagnosticity perspective. *Journal of Business and Psychology, 14*(4), 605–624.

Vallacher, R. R., & Wegner, D. M. (1989). Levels of personal agency: Individual variation in action identification. *Journal of Personality and Social Psychology, 57*(4), 660–671. https://doi.org/10.1037/0022-3514.57.4.660

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly, 36*(1), 157–178. https://doi.org/10.2307/41410412

Viglia, G. (2020). The sharing economy: Psychological mechanisms that affect collaborative consumption. Psychology & Marketing, 37(5), 627-629. https://doi.org/10.1002/mar.21358

Wang, Y., Minor, M. S., & Wei, J. (2016). Aesthetics and the online shopping environment: Understanding consumer responses. *Journal of Retailing, 92*(4), 494–509. https://doi.org/10.1016/j.jretai.2016.08.005

White, K., MacDonnell, R., & Dahl, D. W. (2011). It's the mind-set that matters: The role of construal level and message framing in influencing consumer efficacy and conservation behaviors. *Journal of Marketing Research, 48,* 472–485. https://doi.org/10.1509/jmkr.48.3.472

Wichmann, J. R., et al. (2022). A global perspective on the marketing mix across time and space. *International Journal of Research in Marketing, 39*(2), 502–521. https://doi.org/10.1016/j.ijresmar.2021.09.001

Yan, D., Sengupta, J., & Hong, J. (2016). Why does psychological distance influence construal level? The role of processing mode. *Journal of Consumer Research, 46,* 598–613. https://doi.org/10.1093/jcr/ucw045

Chapter IV Appendix A: Empirics-first Research Approach

Table IV. A1 Empirics-first research approach

Steps in research process	Empirics-first approach
Testing or developing theory	One of several potential outcomes
Determining research focus	Research is exploration-minded
Recognising the nature of the research process	Iterative
Tolerating research messiness	Messiness is common, can be an asset, and should be fully exploited and reported
Reviewing the literature	Conformable with an absence of prior research but grateful for any literature that provide insights or inspiration
Searching for explanations	Multiple angles encouraged
Formulating priors	More diffuse
Developing conceptual framework	Loosely tied constructs and relationships; conceptual framework may develop along the way
Collecting data	Collect empirical observations to explore and understand the focal phenomenon
Analysing data	Document the empirical outcomes, including null results
Checking robustness	Tolerance of (or desire for) multiple explanations and nuanced results
Dealing with failed robustness checks	Viewed as learning opportunities
Writing the article	Suggested template: Motivate the phenomenon, describe the various analyses, and end with insights gained; theoretical implications may (but need not) emerge

Note. Adopted from Golder et al. (2023)

Chapter IV Appendix B: Examples from Real-world AI Healthcare App Providers

Table IV.B1 Selected real-world examples of marketing communications

Marketing message orientation		**Marketing message explanation**	
Goal-oriented perspective	Duty-oriented perspective	How explanation	Why explanation
***SkinVision**: skin cancer melanoma detection*			
AI prevention; Goal-oriented slogan: Check & track every spot with SkinVision			
Accuracy and effectiveness • smart about skin health • a reliable skin assessment in 30 seconds • accurate and timely skin cancer detection • the most reliable personalised skin health advice and health path recommendation • clinically validated • can detect signs of most common skin cancers with up to 95% sensitivity • proven accuracy that been clinically tested through studies • clinical evidence of our algorithm's accuracy • have a large, varied, and well-controlled databased	**Safety and security** • regulated medical service • regulated medical device • regulated & CE marked • privacy • risk assessments • fully committed to protecting and safeguarding the personal data you share with us • rest assured that before we assist you with our service, we will always ask for your explicit consent regarding the use of personal data	**Improve accuracy and effectiveness** • merges AI technology with the expertise of skin health professionals and dermatologists • highly accurate AI-related algorithms • supported by an advanced quality team • combines cutting-edge technology with best-in-class dermatologists expertise and support • offers a possibility to take a self-exam anywhere, anytime and receive an immediate risk assessment along with an indication of whether to visit a doctor, with what urgency • invests in research and collects clinical evidence, striving to provide the most reliable, accurate, personalised experience • algorithm looks for patterns by analysing an enormous amount of information, or data, patterns are used to predict outcome • data scientists train algorithms with images, so it can find patterns between skin spot photos and dermatologist-generated risk labels • test algorithms against the golden standards of skin cancer—measure the sensitivity of algorithms **Improve safety and inclusiveness** • CE mark, TGA approval and ISO certification • provide a secure analysis, and a personal profile and archive	**AI perspective** • algorithms are programs (math and logic) that perform better over time when trained with more information **Healthcare perspective** • expands your ability to self-examine your skin, elevates your knowledge of when to act, how and why • goal: make sure you visit the right doctor at the first sign of risk for skin cancer • helps you identify your skin type to learn the most effective ways to protect yourself from skin cancer • a service of choice whether you want to address your most immediate concerns, learn what steps you should take next, understand your skin profile, and introduce the most intelligent skin health regime to your seasonal rhythm • a supportive tool to check the spots you worry about and receive an instant risk indication • it is an app advised by doctors, pharmaceutical companies, and health insurers • it is endorsed by melanoma patients' support and advocacy groups as a vital part of skin cancer prevention self-exam toolbox

***Ada**: symptom checks and tracks*

AI diagnosis; Goal-oriented slogan: Health. Supporting better health outcomes and clinical excellence with intelligent technology.

Accuracy and effectiveness

- better health outcomes and clinical excellence
- inform health decisions
- enhance triage to appropriate care
- reduce avoidable costs
- research looks at the performance and efficacy
- check symptoms online 24/7
- personalised assessment report, health information
- smart results
- share relevant information with your doctor
- track symptoms and severity in the app
- 97% of patients found Ada easy to use
- provides condition suggestions closest to human doctors
- 33% more accurate, on average, than other symptom assessment apps
- recommendations comparable to triage nurses
- clinical studies show Ada is the most accurate symptom checker app

Safety and security

- medical guidance in 7 languages

Inclusiveness in design

- optimised with human doctors
- most comprehensive condition coverage
- 99% condition coverage including common, rare, high-risk obstetric, paediatric, and mental health conditions

Social benefit

- supporting pandemic responses

Safety and security

- European medical device regulation (MDR)
- class II classification
- clinical governance
- 'security by design'
- Audits from external authorities and internal teams

Improve accuracy and effectiveness

- the AI of the Ada app assesses your answers against its medical dictionary of thousands of disorders and medical conditions
- core system connects medical knowledge with intelligent technology
- guidance is personal to unique health profile
- understand, manage, and get care for symptoms with trusted medical expertise in minutes
- convert medical knowledge and clinical excellence into better outcomes
- doctors created Ada to think like a doctor
- clinical precision and medical oversight from human doctors

Improve safety and inclusiveness

- develops an AI that raises the bar for user accessibility and industry regulation
- apply the strictest data regulations to protect and keep your information private
- complies with relevant medical device regulations and security standards and follows the highest data protection standards
- user safety and quality feedback addressed
- personal health information shared with Ada is confidential and encrypted, it's never shared without explicit consent
- committed to increasing transparency and understanding about how we manage security
- regularly attempts to breach our own security to spot and fix any weak points
- protect data in accordance with the highest data protection standards

AI perspective

- medical AI simplifies healthcare journeys and helps people take care of themselves
- support better health outcomes and clinical excellence with intelligent technology
- people provide the right input to get the right guidance from technology
- develops technology to be as private and secure as possible
- designed to be secure from the beginning and throughout the product lifecycle

Healthcare perspective

- take good care of yourself with Ada
- health management app
- help you if you have common or less common
- healthcare services around the world face urgent challenges and uncertain future.
- millions can't access the care they need
- medical accuracy can help decrease disease detection timeframes to accelerate treatment
- advice safety can help make sure healthcare professionals see the patients who need them most in the right setting
- comprehensive condition coverage means high-risk and uncommon diseases are not missed or misdiagnosed

***Youper**: self-guided therapy for anxiety and depression*

AI treatment; Duty-oriented slogan: Empathetic, safe, and clinically validated chatbot for mental healthcare

Accuracy and effectiveness

- 83% experience better moods
- studied by researchers: showed significant improvement in symptoms
- has been proven clinically effective at reducing symptoms of anxiety and depression
- quick checks, big effects
- the most effective way to improve mental health
- over 60,000 5-star reviews
- proprietary technology

Safety and security

- makes mental healthcare accessible to more people/for everyone
- has supported the mental health/wellbeing of over two million people
- building the future of mental health care

Inclusiveness in design

- safe and clinically validated AI
- diversity, equity, inclusion, belonging
- diverse workforce
- culture of belonging

Social benefit

- AI to solve the mental health
- supports healthcare providers: continuing an initiative that started during the COVID-19 pandemic
- supports students in underserved communities

Safety and security

- safe AI for mental health
- safety principles: safety first, human-centred, protect privacy, guided by science, monitor, learn, and iterate

Improve accuracy and effectiveness

- combines psychology and artificial intelligence to understand users' emotional needs and engage in natural conversations
- powered by the state-of-the-art large language model
- mental health chatbot understands users and guides them with personalised interventions
- based on decades of research in psychology and evidence-based therapies (e.g., Cognitive Behavioural Therapy [CBT])
- provides instant support whenever and wherever needed

Improve safety and inclusiveness

- led by clinicians and backed by science
- rigorous testing and safety assessment
- private, encrypted, and secure
- crisis and harmful language detection
- apply AI to disrupt mental healthcare with safe and effective solutions to support patients
- never sell or share users' data for advertising or marketing purposes
- designs products based on the best research available
- monitors the usage of chatbot and uses anonymised data to learn and continuously improve the safety of systems over time
- bring great people to our team who are always striving to improve and be their best selves

AI perspective

- AI can expand how we offer support to patients and extend the capabilities of the mental health workforce
- the first responder for someone experiencing emotional health issues
- suitable as a companion application to a drug or potentially as standalone prescription digital therapeutics
- studies show that patient real-world engagement is a challenge to the effectiveness of digital solutions, Youper is the most engaging digital health solution for anxiety and depression

Healthcare perspective

- demand for mental health support is increasing, but the supply of clinicians is not
- it takes patients years of suffering in silence before they could get the help they needed
- everyone faces emotional challenges in life, but treatment can help—the real problem is the series of barriers we need to overcome in order to access that treatment
- seeing a mental health professionals can be intimidating and expensive, and those two issues hold many people back

Chapter IV Appendix C: Full Stimuli

Table IV.C1 Stimuli used in studies for GenAI prevention

Medical Care	Prevention: Skin cancer screening
Description of medical context *(all conditions)*	Imagine you have decided to have a skin cancer screening. A skin cancer screening is a visual inspection of your skin. The objective of this screening is the early detection of skin cancer. It does not involve any physical examinations or laboratory tests. Skin cancer is the most common cancer, with two main categories: melanoma and non-melanoma. While melanoma is the most aggressive but rarest type of skin cancer, non-melanoma skin cancers such as carcinoma are most common. Skin cancer is slow-growing and often free of glaring warning signs. However, there are a few signs of skin cancer symptoms you can keep an eye on when performing a skin self-exam. These indicators let you know if a visit to your doctor might be a good idea. With early detection and proper treatment, skin cancer has a high cure rate.
Description of app function *(all conditions)*	CareAI is a paid-for medical service that takes control over your skin health to a new level. The CareAI app employs GPT-4, a generative AI model, to check the spots you worry about, perform risk assessment, generate instant health advice, and give you access to health professionals.
Description of AI/GenAI technology	Another thing you need to know before beginning the survey is that GPT-4 is more advanced than previous rule-based AI. The GPT-4 model accepts text and image input and generates text output. You can ask any questions you have about skin spots through natural conversations. Given the images and descriptions of skin spots you provide, GPT-4 can generate a list of possible skin conditions and health advice. This is based on its training through pattern recognition and probabilistic inference rather than fixed diagnostic rules.
Comprehension check	To ensure that you have understood the information, please select the correct statement. *goal* × *why*: CareAI offers accurate skin cancer screening with up to 95% (90%) sensitivity *goal* × *how*: CareAI can hold conversations and generate text (image) output *duty* × *why*: CareAI is a safe and inclusive (accurate and effective) medical service *duty* × *how*: CareAI is designed carefully to generate safe and inclusive (accurate and effective) content
Mediator	Information diagnosticity (7 Likert scale; Baker & Lutz, 2000, Baker, 2001) Did the advert make you believe that the screening provided can help prevent skin cancer Did the advert create credible impression Did the advert make you feel better about your decision
Dependent variables	Trustworthiness (11 Likert scale; McKnight et al., 2002) For getting the skin cancer screening, I feel I can depend on CareAI I can always rely on CareAI to provide me the skin cancer screening I need I feel that I could count on CareAI to give the skin cancer screening I require Intention to use (7 Likert scale; Venkatesh et al., 2012) I intend to use the app for skin cancer screenings I will try to use the app for skin cancer screenings I plan to use the app for skin cancer screenings

Table IV.C2 Stimuli used in studies for GenAI diagnosis

Medical Care	Diagnosis: Triaging of a potential emergency
Description of medical context *(all conditions)*	Imagine you are experiencing a mild headache and decide to get an initial diagnosis. The initial headache diagnosis is based on your reported symptoms and medical history. The objective of this diagnosis is to determine whether your symptoms require immediate medical attention. It does not involve any physical examinations or laboratory tests. Headaches are a very common condition that most people will experience many times during their lives. While most headaches are not dangerous, certain types can be a sign of a more serious condition. It is important to diagnose headaches correctly to determine the urgency of getting immediate medical attention and appropriate medical care.
Description of app function *(all conditions)*	CareAI is a paid-for medical service that takes control over your headaches to a new level. The CareAI app employs GPT-4, a generative AI model, to diagnose your symptoms, determine the urgency, generate instant health advice, and give you access to nervous system health expertise.
Description of AI/GenAI technology	Another thing you need to know before beginning the survey is that GPT-4 is more advanced than previous rule-based AI. The GPT-4 model accepts text and voice input and generative text output. You can ask any questions you have about headaches through natural conversations. Given the descriptions of headache symptoms and medical history you provide, GPT-4 can determine your medical status and generate health advice. This is based on its training through pattern recognition and probabilistic inference rather than fixed diagnostic rules.
Comprehension check	To ensure that you have understood the information, please select the correct statement. *goal* × *why*: CareAI offers effective initial headache diagnosis with up to 95% (90%) accuracy *goal* × *how*: CareAI can hold conversations and generate text (image) output *duty* × *why*: CareAI is a safe and inclusive (accurate and effective) medical service *duty* × *how*: CareAI is designed carefully to generate safe and inclusive (accurate and effective) content
Mediator	Information diagnosticity (7 Likert scale; Baker & Lutz, 2000, Baker, 2001) Did the advert make you believe that the initial diagnosis provided can help determine medical status Did the advert create credible impression Did the advert make you feel better about your decision
Dependent variables	Trustworthiness (11 Likert scale; McKnight et al., 2002) For getting the headache diagnosis, I feel I can depend on CareAI I can always rely on CareAI to provide me the headache diagnosis I need I feel that I could count on CareAI to give the headache diagnosis I require Intention to use (7 Likert scale; Venkatesh et al., 2012) I intend to use the app for headache diagnoses I will try to use the app for headache diagnoses I plan to use the app for headache diagnoses

Table IV.C3 Stimuli used in studies for GenAI treatment

Medical Care	Treatment: Digital therapies
Description of medical context *(all conditions)*	Imagine you are considering digital therapeutics based on clinically proven behavioural therapies like cognitive behavioural therapy (CBT). A CBT treatment is an early intervention for eating problems. Unhealthy eating behaviours may include eating too much or too little or worrying about your weight or body shape. The objective of this treatment is to re-establish regular healthy eating, maintain a healthy weight, and address any emotions and behaviours that are keeping the eating problems stuck. It does not involve any physical examinations or laboratory tests.
Description of app function *(all conditions)*	CareAI is a paid-for medical service that takes control over your eating problems to a new level. The CareAI app employs GPT-4, a generative AI model, to assess eating problems, provide instant support, monitor symptoms, and give you access to diet and mental health expertise.
Description of AI/GenAI technology	Another thing you need to know before beginning the survey is that GPT-4 is more advanced than previous rule-based AI. The GPT-4 model accepts text and voice inputs and generative text outputs. You can ask any questions you have about eating problems through natural conversations. Given the descriptions of eating problems you provide and your chats with the CareAI chatbot, GPT-4 can generate health advice and offer instant support. This is based on its training through pattern recognition and probabilistic inference rather than fixed diagnostic rules.
Comprehension check	To ensure that you have understood the information, please select the correct statement. *goal* × *why*: CareAI offers accurate and reliable treatments with up to 95% (90%) effectiveness *goal* × *how*: CareAI can hold conversations and generate text (image) output *duty* × *why*: CareAI is a safe and inclusive (accurate and effective) medical service *duty* × *how*: CareAI is designed carefully to generate safe and inclusive (accurate and effective) content
Mediator	Information diagnosticity (7 Likert scale; Baker & Lutz, 2000, Baker, 2001) Did the advert make you believe that the treatment provided can help recover from eating problems Did the advert create a credible impression Did the advert make you feel better about your decision
Dependent variables	Trustworthiness (11 Likert scale; McKnight et al., 2002) For getting the eating problem treatment, I feel I can depend on CareAI I can always rely on CareAI to provide me eating problem treatments I need I feel that I could count on CareAI to give eating problem treatments I require Intention to use (7 Likert scale; Venkatesh et al., 2012) I intend to use the app for eating problem treatments I will try to use the app for eating problem treatments I plan to use the app for eating problem treatments

Table IV.C4 Behaviour Identification Form (BIF) for study 3a-3c

Action	Description
1. Making a list	Getting organised* Writing things down
2. Reading	Following lines of print Gaining knowledge*
3. Joining the army	Helping the Nation's defence* Signing up
4. Washing clothes	Removing odours from clothes* Putting clothes into the machine
5. Picking an apple	Getting something to eat* Pulling an apple off a branch
6. Chopping down a tree	Wielding an axe Getting firewood*
7. Measuring a room for carpeting	Getting ready to remodel* Using a yardstick
8. Cleaning the house	Showing one's cleanliness* Vacuuming the floor
9. Painting a room	Applying brush strokes Making the room look fresh*
10. Paying the rent	Maintaining a place to live* Writing a check
11. Caring for houseplants	Watering plants Making the room look nice*
12. Locking a door	Putting a key in the lock Securing the house*
13. Voting	Influencing the election* Marking a ballot
14. Climbing a tree	Getting a good view* Holding on to branches
15. Filling out a personality test	Answering questions Revealing what you're like*
16. Toothbrushing	Preventing tooth decay* Moving a brush around in one's mouth
17. Taking a test	Answering questions Showing one's knowledge*
18. Greeting someone	Saying hello Showing friendliness*
19. Resisting temptation	Saying "no" Showing moral courage*
20. Eating	Getting nutrition* Chewing and swallowing
21. Growing a garden	Planting seeds Getting fresh vegetables*
22. Traveling by car	Following a map Seeing countryside*
23. Having a cavity filled	Protecting your teeth* Going to the dentist
24. Talking to a child	Teaching a child something* Using simple words
25. Pushing a doorbell	Moving a finger Seeing if someone's home*

Note. High-construal descriptions are marked with asterisks; Vallacher & Wegner (1989).

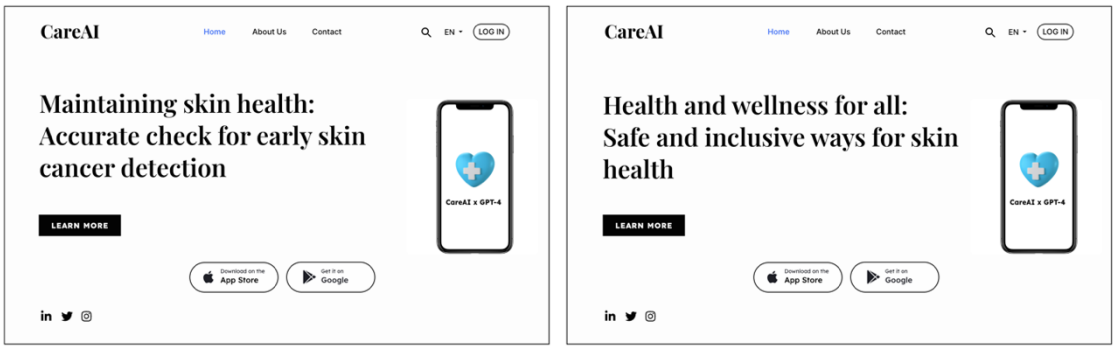


Figure IV.C1 Message focus stimuli for GenAI prevention

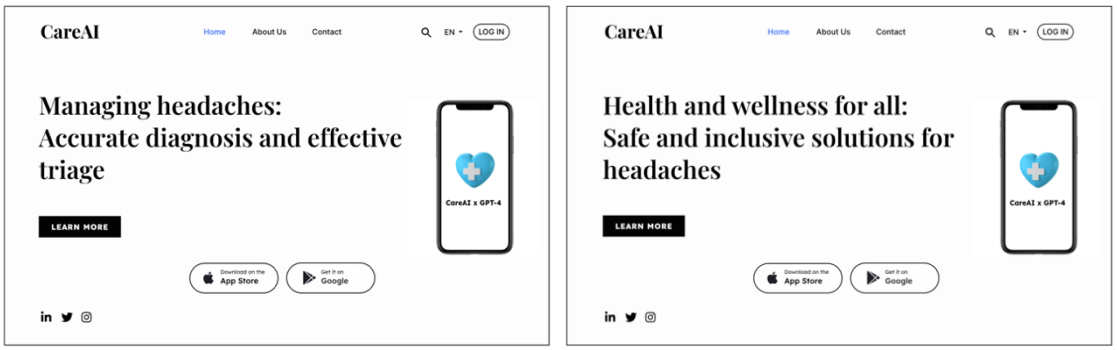


Figure IV.C2 Message focus stimuli for GenAI diagnosis

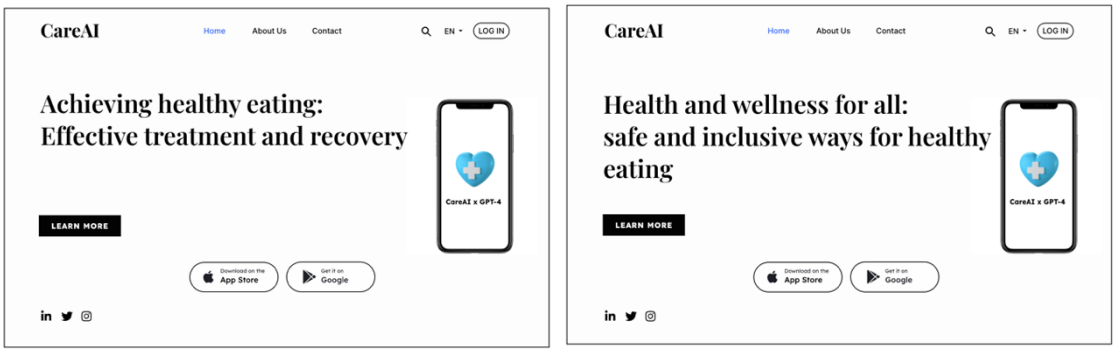


Figure IV.C3 Message focus stimuli for GenAI treatment

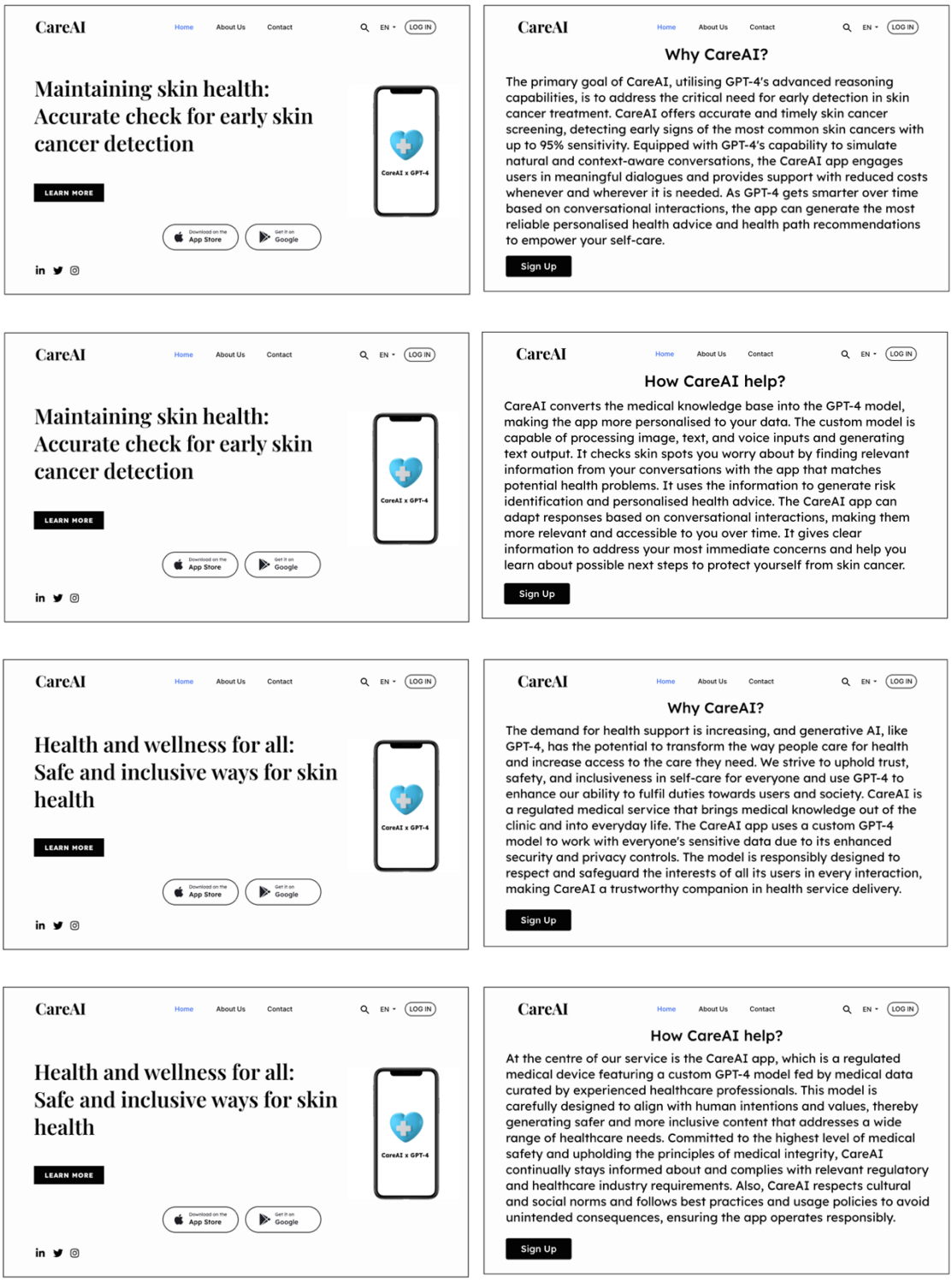


Figure IV.C4 Message focus and message construal stimuli for GenAI prevention

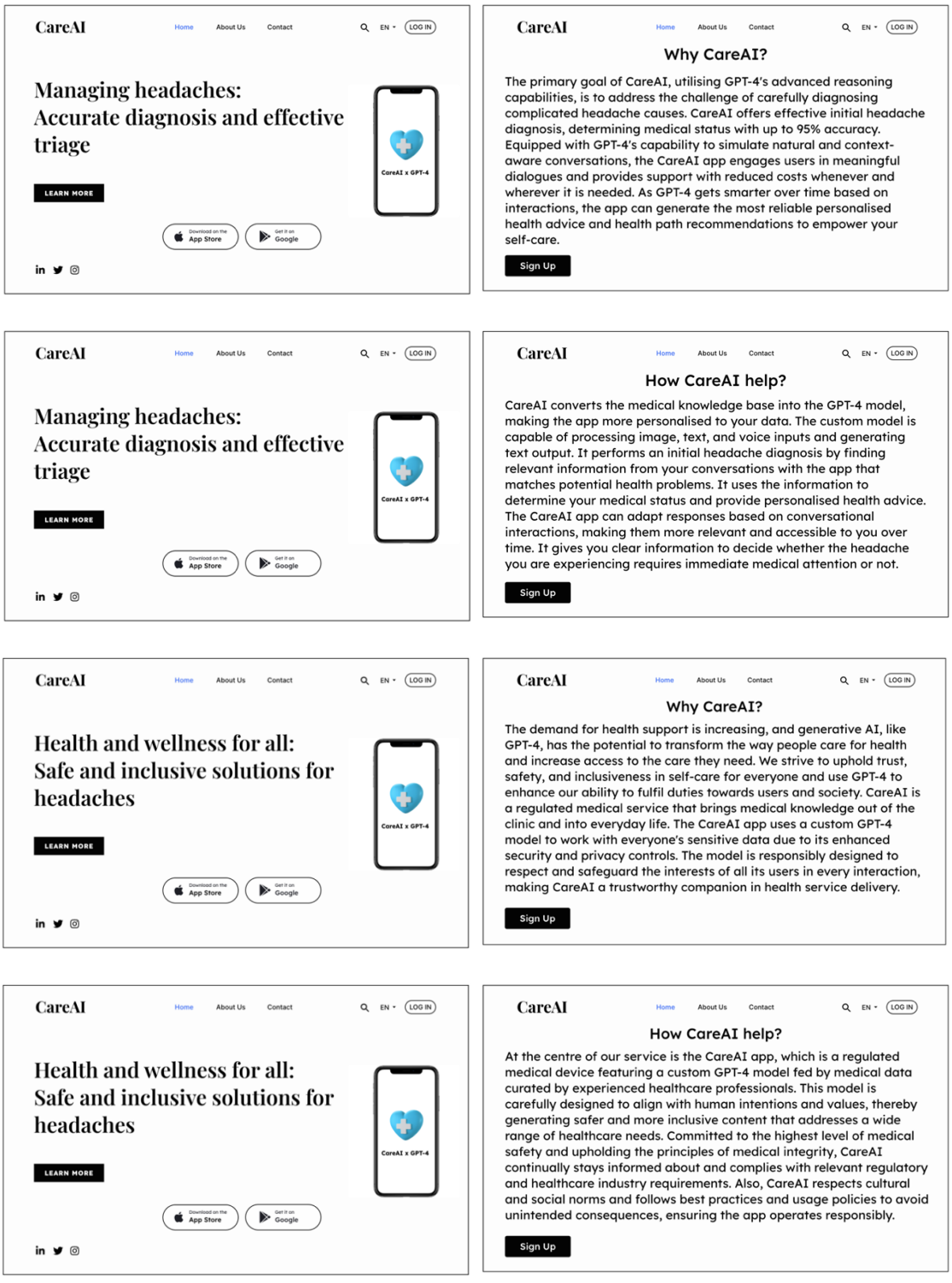


Figure IV.C5 Message focus and message construal stimuli for GenAI diagnosis

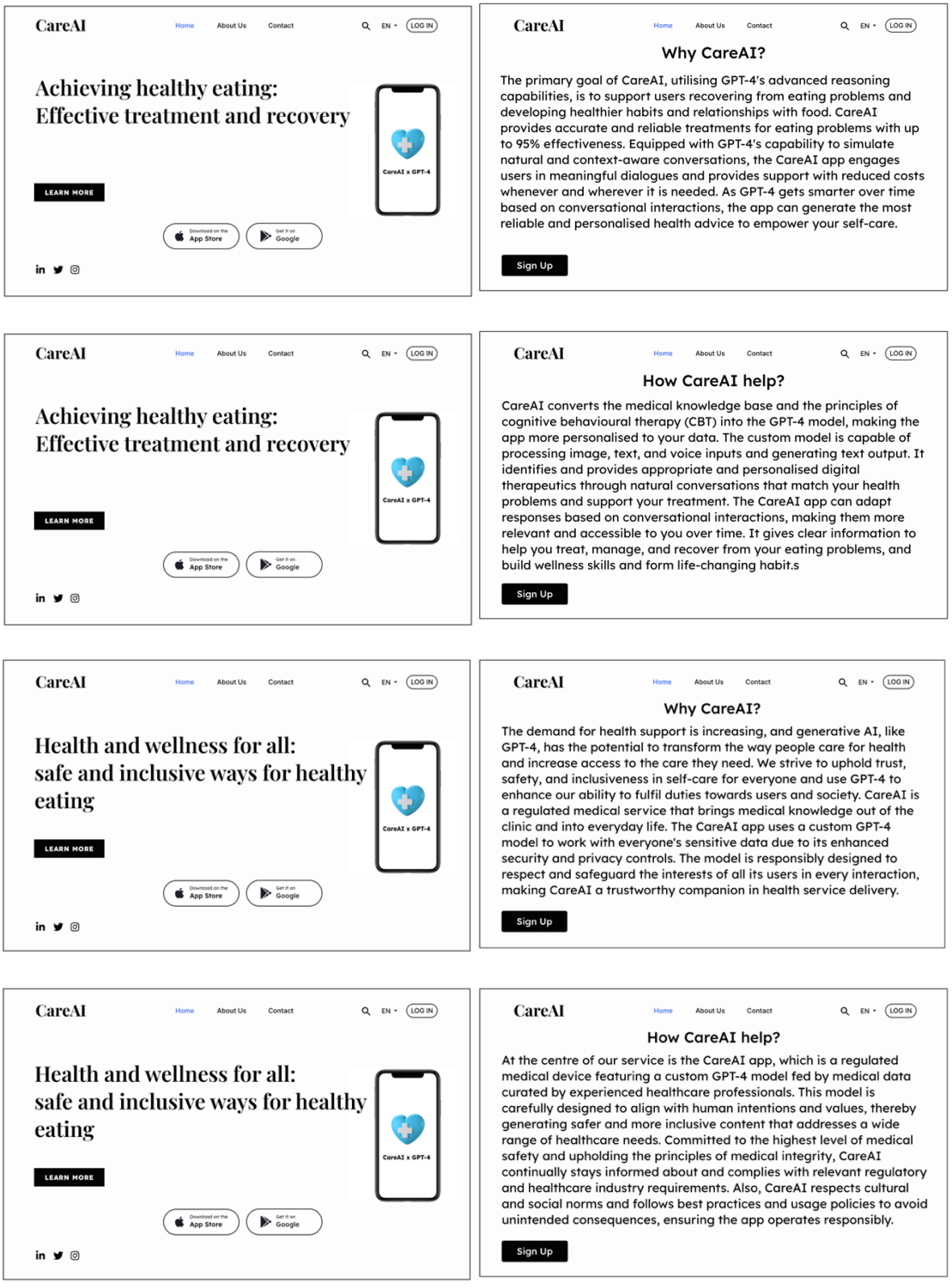


Figure IV.C6 Message focus and message construal stimuli for GenAI treatment

Chapter IV Appendix D: Pretest, Pilot Test, and Additional Analysis

Pretest Study. To pretest the adverts used in this study, 52 participants (M_{age} = 41.57 years, 57.7% female) who were excluded from the main study were randomly assigned to one of two conditions to evaluate either a goal-oriented or duty-oriented advert message. The primary objectives were to identify adverts that vary according to their message orientation on goal or duty, and demonstrate equivalency on all related variables. Participants answered a series of questions evaluating the adverts. First, they responded to a three-item measure (self-developed) of whether the message emphasised goal or duty, using 7-point Likert scales. The first item asked, " In your opinion, does this advert focus more on CareAI's goal of achieving users' health outcomes or on CareAI's duty to ensure users' safety and inclusiveness?" where 1 = *Duty* and 7 = *Goal*. The second item asked participants to agree or disagree with the following statement: "The advert message focuses on goal over duty"; whereas the third item was reverse coded and asked participants to agree or disagree with the following statement: "The advert emphasises duty over goal" (1 = *Strongly disagree*; 7 = *Strongly agree*). All three items were standardised and averaged to form a goal index (α = .90). The ANOVA results showed the goal-oriented message (vs. duty-oriented message) had a significantly higher score (M_{goal} = .43, *SD* = .87 vs. M_{duty} = -.50, *SD* =. 70; $F(1, 50) = 16.381$; $p < .001$), and both means were significantly different from the neutral point of the scale (*ps* < .05). This indicates that the manipulation of message orientation was successful.

Pilot Study. We conducted a pilot study with 139 Prolific workers (M_{age} = 39.51 years, 61.2% female) to test whether our experimental design, procedures, and instruments were practical and feasible (we used the general term *healthcare* rather than focusing on specific contexts of healthcare services). In our analysis, 87.5% of respondents indicated that the clarifications on the differences between AI and GenAI were clear. Additionally, 93% of participants reported that they would use CareAI, but only for minor to moderate illnesses, with most stating they would prefer consulting a human doctor if they observed symptoms. Based on these findings, we decided to focus on preventive healthcare as the initial context to test the main effect of message orientation on consumers' intention to use the app. A one-way ANOVA with message orientation (goal vs. duty) as the independent variable and intention to use CareAI as the dependent variable revealed that the framing effect was not significant ($F(1, 137) = 0.452$, $p = 0.503$). This suggested the need to explore different healthcare service contexts in our main studies.

Additional Analysis. Chi-square tests of independence were performed to examine the randomisation across experimental conditions. In Study 5, the analysis showed no significant differences in respondents' gender ratio (χ^2 (14, n = 479) = 16.508, p = .283) or age group (χ^2 (35, n = 479) = 43.409, p = .156) across eight combinations of marketing communication. However, a significant association was found between participants' income and their responses to different communications (χ^2 (42, n = 479) = 61.022, p = .029, Cramer's V = .128), indicating that consumer evaluations of GenAI healthcare marketing communications varied significantly by income. An examination of the standardised residuals revealed that lower-income participants tended to favour goal-oriented messages when paired with *why* explanations for GenAI diagnostic services, while higher-income participants preferred duty-oriented messages. From the perspective of abstraction in thought processes, why explanations provide higher-order rationales for using GenAI in diagnostic actions. For lower-income participants, these explanations may bridge the gap between abstract and concrete thinking by contextualising immediate and tangible benefits, making goal-oriented messages more appealing. In contrast, for higher-income participants, why explanations may reinforce the broader, duty-oriented implications of using GenAI for diagnosis in healthcare, aligning with their tendency toward abstract thinking and consideration of wider societal impacts.

Chapter IV Appendix F: ANOVA Analysis Results

Table IV.F1 ANOVA results for Study 3a-3c

Dependent Variable: ***Intention to use***						
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Study 3a						
Corrected Model	44.264	3	14.755	5.311	.002	.136
Intercept	1433.212	1	1433.212	515.941	<.001	.836
Message Orientation	15.623	1	15.623	5.624	.020*	
Construal Mindset	15.285	1	15.285	5.502	.021*	
Orientation * Construal	13.320	1	13.320	4.795	.031*	
Error	280.564	101	2.778			
Total	1768.476	105				
Corrected Total	324.828	104				
Study 3b						
Corrected Model	27.689	3	9.230	2.907	.039*	.082
Intercept	1414.484	1	1414.484	445.469	<.001	.821
Message Orientation	1.518	1	1.518	.478	.491	.005
Construal Mindset	25.615	1	25.615	8.067	.005**	.077
Orientation * Construal	.004	1	.004	.001	.973	.000
Error	308.001	97	3.175			
Total	1747.889	101				
Corrected Total	335.690	100				
Study 3c						
Corrected Model	12.685	3	4.228	1.784	.156	.056
Intercept	1270.898	1	1270.898	536.208	<.001	.856
Message Orientation	1.815	1	1.815	.766	.384	.008
Construal Mindset	10.963	1	10.963	4.625	.034*	.049
Orientation * Construal	.894	1	.894	.377	.541	.004
Error	213.314	90	2.370			
Total	1524.256	94				
Corrected Total	225.999	93				

Note. *p < 0.05, **p < 0.01.

Table IV.F2 ANOVA results for Study 4a-4c

Dependent Variable: ***Intention to use***						
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Study 4a						
Corrected Model	29.197	3	9.732	3.216	.024	.039
Intercept	2832.571	1	2832.571	936.142	<.001	.799
Message Orientation	.122	1	.841	.000	.122	.040
Message Construal	16.349	1	16.349	5.403	.021*	.022
Orientation * Construal	13.906	1	13.906	4.596	.033*	.019
Error	714.087	236	3.026			
Total	3579.222	240				
Corrected Total	743.285	239				
Study 4b						
Corrected Model	32.167	3	10.722	3.420	.018	.042
Intercept	2776.261	1	2776.261	885.601	<.001	.790
Message Orientation	.784	1	.784	.250	.617	.001
Message Construal	18.369	1	18.369	5.859	.016*	.024
Orientation * Construal	13.588	1	13.588	4.335	.038*	.018
Error	736.699	235	3.135			
Total	3545.778	239				
Corrected Total	768.867	238				
Study 4c						
Corrected Model	23.611	3	7.870	2.745	.044	.033
Intercept	2971.797	1	2971.797	1036.639	<.001	.811
Message Orientation	9.771	1	9.771	3.408	.066	.014
Message Construal	9.788	1	9.788	3.414	.066	.014
Orientation * Construal	4.010	1	4.010	1.399	.238	.006
Error	690.890	241	2.867			
Total	3684.333	245				
Corrected Total	714.501	244				

Note. *p< 0.05, **p < 0.01.

Table IV.F3 ANOVA results for Study 5

Dependent Variable: ***Intention to use***						
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	61.465	7	8.781	2.851	.006	.041
Intercept	5608.646	1	5608.646	1820.855	<.001	.794
Message Orientation	.764	1	.764	.248	.619	.001
Message Construal	.031	1	.031	.010	.920	.000
Healthcare Service	.103	1	.103	.033	.855	.000
Orientation * Construal	.001	1	.001	.000	.988	.000
Orientation * Health	.144	1	.144	.047	.829	.000
Construal * Health	34.690	1	34.690	11.262	< .001***	.023
Orientation * Construal *Health	27.493	1	27.493	8.926	.003**	.019
Error	1450.786	471	3.080			
Total	7125.000	479				
Corrected Total	1512.251	478				

Note. *p< 0.05, **p < 0.01, ***p < 0.001

Chapter V

Thesis Conclusion

1. Research Contributions and Implications

This thesis examines responsible AI from organisational, managerial, and consumer perspectives. Specifically, it 1) integrates theoretical insights (i.e., CSR, IS business value, PBV) to position responsible AI as a strategic asset rather than merely a compliance obligation; 2) highlights the role of institutional leadership in responsible AI management from a sociotechnical perspective, expanding AI governance frameworks beyond technical and regulatory oversight to incorporate managerial, stakeholder-driven approaches; 3) bridges ethical AI communication strategies with consumer psychology, demonstrating how responsible message framing influences AI acceptance and adoption through CLT, ADM, and CUT, which explain how psychological distance, information relevance and usefulness, and extrinsic cues shape consumer perceptions of AI. While each essay presents distinct theoretical and empirical insights, collectively, they provide key research contributions and practical implications for responsible AI in the following meaningful ways.

1.1 Theoretical contributions

First, at the organisational level, this thesis advances the theoretical foundations of responsible AI and positions it as a strategic business imperative. It integrates CSR theory (Windsor, 2006) and IS business value framework (i.e., ES benefits; Shang & Seddon, 2002) within a PBV lens (Bromiley & Rau, 2014) to reframe responsible AI as more than a normative ethical discourse. It empirically establishes responsible AI as both a regulatory necessity and a business enabler, demonstrating its link to IT resilience, operational efficiency, and stakeholder trust. By moving beyond prior

conceptualisations that position responsible AI solely as a moral duty (Jobin et al., 2019; Floridi et al., 2018), this thesis provides an empirically validated framework that connects responsible AI practice with business performance, reinforcing its role in long-term business value creation. Moreover, applying the ES benefits framework (Shang & Seddon, 2002), this thesis extends IS business value research by positioning responsible AI as a competitive advantage rather than mere compliance. By integrating ethics with performance metrics, this thesis demonstrates that responsible AI fosters both technological resilience and strategic differentiation.

Furthermore, this thesis reconceptualises responsible AI through a PBV lens (Bromiley & Rau, 2014) as an evolving, transferable industry-wide practice, rather than a firm-specific resource as framed by RBV (Barney, 1991), which views AI governance as a proprietary asset for competitive differentiation. PBV highlights responsible AI's diffusion across industries, shaped by regulatory mandates, stakeholder expectations, and shared governance frameworks. By situating responsible AI within PBV, this thesis advances the understanding of how responsible AI evolves through shared organisational practices rather than exclusive firm-level capabilities.

Second, at the managerial level, this thesis highlights the sociotechnical challenges of AI implementation, advocating for multi-stakeholder responsible AI management frameworks that balance AI autonomy with human oversight to scale responsible AI adoption. It contributes to the sociotechnical perspective (Berente et al., 2021) by integrating responsible AI design and governance mechanisms, moving beyond technocentric AI ethics that focus narrowly on technical

transparency and algorithmic fairness, as well as abstract regulatory discussions. Instead, it emphasises broader organisational governance structures essential for ethical AI design and deployment, reinforcing the need for multi-stakeholder accountability across policymakers, industry leaders, and end-users, shifting from top-down regulatory enforcement toward inclusive, multi-level governance. Additionally, this thesis advances the human-AI collaboration debate, addressing the shift toward AI-driven decision-making. As AI autonomy increases, AI management frameworks must evolve to maintain human oversight and accountability, ensuring AI remains aligned with organisational and societal values. It also highlights the role of institutional leadership in shaping and disseminating responsible AI practices.

Third, at the consumer level, this thesis integrates responsible AI with consumer psychology and marketing communication research. It extends the literature by applying CLT, CUT, and ADM to GenAI healthcare services, demonstrating that strategic, responsible AI messaging (message orientation: goal, duty × message construal: how, why) significantly impacts consumer trust and adoption. While prior AI ethics research has primarily emphasised technical transparency and fairness, this study reveals that message framing and construal alignment collectively shape perceived GenAI reliability and effectiveness. Findings show that psychological distance influences AI acceptance, and that well-calibrated messaging reduces algorithm aversion by enhancing perceived information diagnosticity—a key driver of GenAI adoption.

Unlike traditional products, GenAI healthcare lacks direct experiential cues, requiring consumers to rely on extrinsic informational cues (e.g., message framing)

for credibility and risk assessment. This study bridges AI ethics with marketing communication, offering a novel framework for persuasive yet responsible AI messaging. It also expands CLT's application beyond traditional marketing contexts, demonstrating that message design itself functions as a cognitive cue, shaping consumer perceptions. By positioning responsible AI communication as a trust-building tool, this research provides theoretical insights for GenAI adoption in healthcare and public policy.

1.2 Practical implications

First, a critical practical implication of this research is the need for comprehensive AI ethics training to support responsible AI implementation. Organisations must equip managers, AI developers, and employees with the necessary knowledge and skills to navigate ethical AI challenges through structured training programs, cross-functional workshops, and self-directed learning. Implementing mentorship programs and certification frameworks (e.g., IEEE's EAD) fosters an ethical AI culture, reinforcing responsible decision-making and strengthening organisational AI governance. The importance of AI ethics training is reinforced across all three essays. Essay 1 highlights its role in aligning AI development with corporate governance and regulatory compliance, ensuring ethical standards are met at an organisational level. Essay 2 demonstrates how training empowers AI stakeholders to engage in ethical decision-making at a managerial level, particularly as AI systems become more autonomous, requiring stronger oversight. Essay 3 underscores the significance of AI literacy for both consumers and businesses, showing that consumer trust and AI adoption are directly influenced by how well

AI's ethical safeguards are communicated. By embedding AI ethics training within corporate structures, organisations can mitigate risks, foster a responsible AI culture, and build stakeholder trust, ensuring AI implementation aligns with both ethical principles and business objectives.

Second, responsible AI management necessitates robust risk management frameworks that cover the entire AI lifecycle—from design to deployment and post-implementation evaluation. Organisations must proactively identify and mitigate risks such as algorithmic bias, data security breaches, and ethical dilemmas. Establishing clear governance structures, data quality protocols, and AI oversight mechanisms ensures AI reliability, transparency, and regulatory compliance. Additionally, prioritising human-AI collaboration safeguards against excessive automation, ensuring AI augments rather than replaces human decision-making. The significance of AI risk management is reinforced in Essays 1 and 2. Essay 1 underscores the role of data governance and algorithmic fairness in ensuring that responsible AI enhances business value while minimising risks. Essay 2 introduces governance mechanisms that facilitate ethical AI management, demonstrating how AI risk mitigation aligns with socio-technical AI governance frameworks. By adopting a human-centric AI risk management approach, organisations can ensure ethical AI deployment, strengthen compliance, and enhance long-term business sustainability.

Third, as AI adoption accelerates, companies must establish strong leadership and accountability mechanisms to embed responsible AI into business functions. Appointing a CRAiO ensures that AI governance aligns with ethical principles, legal compliance, and CSR commitments. Alternatively, organisations can establish AI

ethics advisory boards to provide independent oversight, risk assessment, and strategic direction, ensuring cross-functional alignment in responsible AI strategies. The importance of AI ethics leadership is reinforced in Essays 1 and 2. Essay 1 discusses how firms with strong AI governance structures (e.g., CRAiO roles, advisory boards) can gain strategic advantages by fostering consumer trust and mitigating reputational risks. Essay 2 highlights the role of institutional leadership in driving responsible AI adoption, advocating for multi-stakeholder accountability in AI decision-making. By establishing AI ethics leadership, organisations can institutionalise AI accountability, ensure responsible AI implementation, and embed ethical AI governance into business strategy.

Fourth, AI adoption must balance business growth with social responsibility. Companies should integrate ethical AI governance into corporate sustainability strategies, addressing privacy concerns, employment impacts, and human rights risks. Aligning AI-driven innovation with CSR and environmental, social, and governance (ESG) initiatives strengthens public trust and long-term AI acceptance. Businesses that proactively address AI's societal impact can enhance brand credibility and maintain stakeholder confidence. This connection between AI governance and social responsibility is evident in Essays 1 and 2. Essay 1 highlights how responsible AI governance aligns with CSR objectives, demonstrating that balancing economic performance with ethical considerations enhances firm reputation. Essay 2 presents a socio-technical framework for managing AI's broader impact on society, workers, and consumers, advocating for sustainable AI implementation strategies. By integrating responsible AI governance with CSR and

ESG commitments, firms can drive ethical AI adoption while ensuring long-term stakeholder trust.

Fifth, transparent and responsible AI communication is essential for widespread AI adoption. This research highlights how strategic AI messaging—specifically goal vs. duty framing and how vs. why explanations—significantly influences consumer trust and adoption. Companies must craft persuasive yet ethical AI communications to mitigate algorithm aversion and enhance consumer confidence in AI-driven services, particularly in healthcare, finance, and public policy. Essay 3 applies consumer psychology theories (CLT, CUT, and ADM) to AI communication, demonstrating how message framing and information diagnosticity shape consumer attitudes toward AI adoption. This research underscores the importance of responsible AI branding, showing that well-calibrated messaging improves AI acceptance while reducing perceived ethical risks. By designing transparent, ethical AI communications, organisations can enhance consumer engagement, reduce resistance to AI adoption, and establish long-term trust in AI-enabled services.

1.3 Future research directions

First, future research should adopt mixed- and multiple-methods research designs (Bromiley & Rau, 2014, 2016) to enhance empirical rigour and provide a holistic, multi-stakeholder understanding of responsible AI adoption and management. Given that much of the existing research relies on secondary data or self-reported consumer, organisational perspectives, future studies should incorporate primary data from diverse stakeholders—including consumers, managers, policymakers, AI

developers, and regulators—to triangulate insights and mitigate biases in assessing responsible AI design and governance. This multi-perspective approach would yield a more balanced, contextually rich analysis of how responsible AI is operationalised across sectors, industries, institutions, and geographical regions. In addition, longitudinal studies tracking the evolution of responsible AI governance, business value creation, and organisational adaptation over time would provide critical insights into how firms navigate regulatory shifts, technological disruptions, and evolving societal expectations. Given AI's dynamic nature, it is crucial to examine how responsible AI practices, strategic AI positioning, performance outcomes, and consumer evaluations develop over the long term and how firms institutionalise responsible AI as a sustainable competitive advantage.

Second, this thesis conceptualises responsible AI as an evolving, transferable industry-wide practice shaped by regulatory mandates, stakeholder expectations, and shared governance frameworks, aligning with PBV. However, from an RBV perspective, responsible AI could also be viewed as a firm-specific capability that provides sustained competitive advantage through proprietary governance mechanisms, talent acquisition, and firm-specific technological expertise. Future research should bridge PBV and RBV perspectives, investigating whether responsible AI primarily fosters standardised ethical norms across industries (PBV) or serves as a strategic differentiator for firms investing in exclusive AI governance capabilities (RBV). Empirical studies should assess whether responsible AI adoption follows an industry-wide convergence model (PBV) or remains firm-specific and path-dependent (RBV). Furthermore, longitudinal studies could examine the institutionalisation, diffusion, and adaptation of responsible AI across different

industries and regulatory environments, identifying best practices, barriers, and firm-specific strategies that shape AI governance. This research would contribute to a deeper theoretical understanding of whether responsible AI is primarily driven by institutional pressures (PBV) or firm-level resource accumulation and differentiation (RBV).

Third, this study primarily examines responsible AI in Western contexts (UK, US, EU), where AI governance is heavily influenced by GDPR, AI Acts, and strong institutional accountability mechanisms. However, AI ethics, adoption, and governance are deeply shaped by cultural, legal, and economic factors, leading to varied regulatory approaches and ethical priorities across different regions. Future research should conduct comparative cross-cultural studies to explore how responsible AI governance is implemented in different regulatory landscapes, particularly in emerging economies, non-Western regulatory environments, and global AI markets. These studies would shed light on contextual variations, regulatory divergence, and best practices that could inform globally adaptable AI governance models. Additionally, further industry-specific research is needed to examine responsible AI in high-stakes domains such as finance and legal services, where ethical risks and governance challenges are particularly pronounced. Sectoral studies could provide a granular understanding of domain-specific regulatory constraints, ethical dilemmas, and AI governance best practices, offering actionable insights for policymakers and industry leaders. Moreover, longitudinal research tracking the regional diffusion and industry-specific adaptation of responsible AI would enhance understanding of how AI governance evolves across different cultural and regulatory environments.

Fourth, future research should explore consumer psychology models (e.g., CLT, CUT, ADM) in AI adoption, particularly how responsible AI communication strategies influence trust, ethical perceptions, and algorithm aversion. While prior research focuses on technical transparency, less attention has been given to how message framing, psychological distance, and information diagnosticity shape consumer acceptance across industries and cultures. Additionally, individual differences—such as health literacy, AI familiarity, regulatory awareness, and cultural background—moderate perceptions of AI credibility, fairness, and decision-making authority. Dehumanisation concerns remain a barrier to AI adoption (Longoni et al., 2019), requiring further study on moral considerations, consumer autonomy, and trust in AI vs. human decisions. Future research should also examine how responsible AI communication builds long-term trust, brand loyalty, and engagement in AI-driven services. Longitudinal studies can assess how transparent messaging, fairness framing, and crisis communication shape consumer confidence over time. These insights would inform ethical AI branding, ensuring AI messaging aligns with consumer expectations, ethical concerns, and industry-specific risks to foster trust and responsible AI adoption at scale.

2. Thesis Philosophical Stance: Pragmatist Approach

This thesis adopts a pragmatist philosophical stance, which shapes both conceptual and methodological considerations across the three essays, although each essay maintains its own distinct philosophical stance. Pragmatism, as a philosophical approach, prioritises practical relevance, problem-solving, and real-world application, bridging the gap between theory and practice (Morgan, 2014). Unlike

positivism, which seeks objective truths, or constructivism, which focuses on subjective interpretations, pragmatism is concerned with generating actionable knowledge (Saunders et al., 2019) that is adaptable to evolving social, ethical, and technological challenges. Given that responsible AI is a rapidly emerging field, pragmatism provides the flexibility to combine theoretical perspectives and empirical insights to explore responsible AI from organisational, managerial, and consumer perspectives.

Although each essay in this thesis has its own philosophical underpinnings, the overarching pragmatist stance ensures that research contributions remain relevant, transferable, and actionable across different industries and stakeholders. This is particularly important for responsible AI, where regulatory landscapes, business imperatives, and consumer expectations are continuously evolving. Pragmatism also informs methodological pluralism, integrating qualitative and quantitative approaches to capture the complexities of AI governance, adoption, and communication.

2.1 Conceptual considerations

Essay 1. Pragmatism informs the conceptualisation of responsible AI by bridging CSR theory (Windsor, 2006), IS business value frameworks (Shang & Seddon, 2002), and the PBV (Bromiley & Rau, 2014). Rather than treating responsible AI as a normative ethical discourse, this research positions it as both a regulatory necessity and a business enabler. It moves beyond theoretical debates on AI ethics (Jobin et al., 2019; Floridi et al., 2018) and empirically examines how responsible AI enhances IT resilience, operational efficiency, and stakeholder trust. By adopting a

pragmatist stance, we take a problem-oriented approach, focusing on how organisations can integrate responsible AI into business strategy to generate business value rather than merely discussing why AI ethics matters. This aligns with pragmatist research, which emphasises actionable knowledge—designed to offer practical frameworks that firms can implement rather than abstract theoretical models. Moreover, it supports PBV's emphasis on industry-wide best practices, demonstrating that responsible AI is not a proprietary firm resource (as in RBV; Barney, 1991) but a transferable practice shaped by regulatory, technological, and societal dynamics.

Essay 2. The pragmatist approach also shapes the socio-technical perspective on responsible AI management by integrating both AI design and governance mechanisms (Berente et al., 2021). While many AI ethics frameworks adopt a technocentric view or a governance view, focusing on algorithmic fairness, transparency, and bias mitigation, this research takes a broader, multi-stakeholder governance-oriented approach, examining how organisational structures, leadership, and regulatory environments influence responsible AI adoption and governance. A key contribution is its emphasis on multi-stakeholder governance—highlighting the participatory democracy model, where AI governance is distributed across policymakers, industry leaders, and end-users rather than imposed top-down by regulatory bodies. This pragmatic, problem-solving approach aligns with the view that AI governance should be adaptable and context-sensitive, responding to rapid changes in AI regulation and deployment (Mikalef et al., 2022). Pragmatism also informs the focus on human-AI collaboration—rather than assuming a binary relationship where AI either automates or replaces human

decision-making, this research explores adaptive governance models that balance AI autonomy with human oversight. This reflects pragmatism's rejection of rigid dichotomies (Morgan, 2014) and its focus on dynamic, evolving solutions.

Essay 3. Pragmatism is particularly evident in Essay 3, which integrates consumer psychology theories (CLT, CUT, ADM) to investigate responsible AI communication strategies. This research moves beyond philosophical discussions of AI ethics to ask how AI can be effectively and responsibly communicated to enhance consumer trust and adoption. Unlike traditional research that focuses on technical transparency as the primary trust-building mechanism, this research examines how AI message framing (goal vs. duty message orientation, how vs. why message construal) influences consumer perception of AI credibility. This aligns with pragmatism's emphasis on real-world applicability, ensuring that findings inform practical AI branding and marketing strategies. The pragmatist stance also underscores the need for context-sensitive AI marketing communication—rather than assuming a one-size-fits-all approach, this paper highlights how consumer reactions vary across different psychological factors. It calls for adaptive AI messaging strategies that align with diverse consumer expectations, reinforcing pragmatism's commitment to flexibility and responsiveness in research design (Morgan, 2014).

2.2 Methodological considerations

Pragmatism advocates methodological pluralism (Tashakkori & Teddlie, 2010), integrating diverse research methods to capture the complexities of responsible AI adoption and governance. This thesis employs a multiple-method approach,

combining qualitative multiple case studies, expert semi-structured interviews, and consumer psychology experiments to ensure both theoretical robustness and practical relevance. Essay 1 adopts a cross-industry case study approach to identify best practices in responsible AI implementation, providing empirical insights into how responsible AI practices generate business value. Essay 2 incorporates expert interviews and archival data to theorise a responsible AI management model, encompassing AI design and governance mechanisms using a grounded theory method. This essay aligns AI management with real-world challenges. Essay 3 applies experimental research to examine how consumers interpret responsible AI communication strategies, particularly in healthcare, a high-stakes domain. This multi-stakeholder perspective ensures that responsible AI is studied not just from an organisational standpoint but also through managerial and consumer lenses, enhancing the study's empirical rigor and external validity.

Aligned with pragmatism's emphasis on real-world problem-solving (Creswell, 2013; Morgan, 2014), this thesis prioritises empirical validation over conceptual discussions. Essays 1 and 2 rely on qualitative research and expert validation to ensure that responsible AI business value and responsible AI management models are firmly rooted in organisational realities, addressing how firms navigate regulatory requirements, stakeholder expectations, and operational constraints. Essay 3 extends this empirical focus by using experimental research to assess consumer interactions with AI-driven healthcare services, providing insights into how responsible AI communication strategies shape trust and adoption. By integrating theoretical constructs with practical evidence, this thesis advances

responsible AI research as both an academic and applied field, reinforcing pragmatism's commitment to actionable knowledge.

Pragmatism supports an iterative research design where findings from one study inform and refine subsequent inquiries (Saunders et al., 2019). This thesis follows an iterative structure, where Essay 1's insights on responsible AI business value inform Essay 2's development of responsible AI management frameworks. Essay 2's exploration of AI oversight and stakeholder accountability then informs Essay 3's investigation into consumer reaction and responsible AI communication strategies. This iterative approach ensures that theoretical developments remain empirically grounded and continuously refined through real-world applications. By adopting a pragmatist lens, this thesis demonstrates how responsible AI can be operationalised across different levels—organisational, managerial, and consumer—while addressing both academic and practical concerns.

By adopting a pragmatist approach, this thesis bridges theory and practice, positioning responsible AI as both conceptually and theoretically rigorous while ensuring practical applicability. Pragmatism's emphasis on contextual problem-solving, empirical validation, and interdisciplinary integration enhances the thesis's contributions across responsible AI, AI ethics, AI governance, AI business value, and AI marketing communications. This research underscores that responsible AI is not merely a theoretical construct but a practical imperative, necessitating adaptive governance, strategic communication, and multi-stakeholder collaboration to foster ethical, sustainable, and value-driven AI adoption.

Chapter V References

Barney, J. B. (1991). Firm resources and sustained competitive advantage. *Journal of Management, 17*(1), 99–120. https://doi.org/10.1177/014920639101700108

Berente, N., et al. (2021). Managing artificial intelligence. *MIS Quarterly, 45*(3), 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Bromiley, P., & Rau, D. (2014). Towards a practice-based view of strategy. *Strategic Management Journal, 35*(8), 1249–1256. https://doi.org/10.1002/smj.2238

Bromiley, P., & Rau, D. (2016). Operations management and the resource-based view: Another view. *Journal of Operations Management, 41*, 95–106. https://doi.org/10.1016/j.jom.2015.11.003

Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Sage publications.

Floridi, L., et al. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*, 689–707. https://doi.org/10.1007/s11023-018-9482-5

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

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Mikalef, P., et al. (2022). Thinking responsibly about responsible AI and "the dark side" of AI. *European Journal of Information Systems, 31*(3), 257–268. https://doi.org/10.1080/0960085X.2022.2026621

Morgan, D. L. (2014). Pragmatism as a paradigm for social research. *Qualitative Inquiry, 20*(8), 1045-1053. https://doi.org/10.1177/1077800413513733

Saunders, M., Lewis, P., & Thornhill, A. (2019). *Research methods for business students*. Pearson.

Shang, S., & Seddon, P. B. (2002). Assessing and managing the benefits of enterprise systems: The business manager's perspective. *Information Systems Journal, 12*(4), 271–299. https://doi.org/10.1046/j.1365-2575.2002.00132.x

Tashakkori, A., & Teddlie, C. (2010). *Sage handbook of mixed methods in social and behavioral research*. Sage.

Windsor, D. (2006). Corporate social responsibility: Three key approaches. *Journal of Management Studies, 43*(1), 93–114. https://doi.org/10.1111/j.1467-6486.2006.00584.x

Thesis Appendix A: Ethics Approval

This PhD project (Application Reference Number 037918) was approved on ethics grounds by ethics reviewers from the University of Sheffield on 06/04/2021, on the basis that we adhere to the approved documentations including ethics application and amendment forms, participant information sheets, consent forms, and the following responsibilities when delivering the research project:

- The project must abide by the University's Research Ethics Policy.
- The project must abide by the University's Good Research & Innovation Practices Policy.
- The researcher must inform their supervisor of any significant changes to the project or the approved documentation.
- The researcher must comply with the requirements of the law and relevant guidelines relating to security and confidentiality of personal data.
- The researcher is responsible for effectively managing the data collected both during and after the end of the project in line with best practice, and any relevant legislative, regulatory, or contractual requirements.