Lisbon School of Economics & Management
Universidade de Lisboa

MASTERS IN MANAGEMENT (MIM)

MASTERS FINAL WORK

DISSERTATION

MYSTERY SHOPPING AND ARTIFICIAL INTELLIGENCE: THE IMPLICATIONS ON SERVICE LEVEL OPTIMIZATION

GUILHERME GOUVEIA PIRES

JANUARY 2026

Lisbon School of Economics & Management
Universidade de Lisboa

MASTERS IN MANAGEMENT (MIM)

MASTERS FINAL WORK

DISSERTATION

MYSTERY SHOPPING AND ARTIFICIAL INTELLIGENCE: THE IMPLICATIONS ON SERVICE LEVEL OPTIMIZATION

GUILHERME GOUVEIA PIRES

Jury:

President: Prof. MARIA EDUARDA MARIANO AGOSTINHO SOARES
Rapporteur: Prof. SUSANA CATARINA DE JESUS FERNANDES DOS SANTOS
Supervisor: Prof. RICARDO MARINO FRANCISCO RODRIGUES

April - 2026

ACKNOWLEDGMENTS

This dissertation is not only the realization of another personal milestone, but also the closing of one of the most beautiful cycles of my life, my five-year academic journey at ISEG Lisbon School of Economics & Management. I would like to thank this institution for all these years of growing learning, always surrounded by good professionals and academic excellence.

I would also like to express my sincere gratitude to Professor Ricardo Marino Francisco Rodrigues for all his guidance and knowledge sharing during these months. Despite his enormous scientific knowledge, he also transmitted me values of perseverance, and his message was always one of optimism and confidence in my abilities.

I would also like to thank the participants in the interviews, who were always willing to collaborate in whatever was needed. Without them, it would have been impossible to enrich this research.

Finally, I would like to express my deepest gratitude to my family and friends for all their love and friendship. From day one, they have supported me and made me believe that this would be possible. I would like to particularly give a special thanks to my grandfather, who is no longer with us but who, having never had the opportunity to study, was one of the people who most supported my pursuit of education.

I would like to acknowledge the use of Grammarly (AI-based writing assistant) which was used solely for the purpose of proofreading and improving the linguistic accuracy of this dissertation.

The Google Gemini was also used to assist in organizing and formatting quotes extracted from interviews.

RESUMO

Nos últimos anos, a prática do Cliente Mistério tem testemunhado uma mudança de paradigma na sua essência. Esta transição advém da transformação digital verificada, que, por consequência, fomentou a procura por dados em tempo real. Este estudo consiste na investigação das implicações do uso da Inteligência Artificial nas vertentes operacionais e estratégicas do setor, focando-se no confronto entre a automatização do processo e a simultânea necessidade de uma avaliação humana empática.

A dissertação adota um método qualitativo, onde se triangularam entrevistas feitas a profissionais de topo do setor com uma análise documental da Indústria. A pesquisa revela que a Inteligência Artificial não está a substituir totalmente o avaliador humano, mas sim a redefinir e adaptar o seu papel. Os resultados evidenciam três conclusões principais: Em primeiro lugar, estrategicamente, o serviço tem transitado de uma mera ferramenta de diagnóstico para um importante indicador de performance financeira. A nível operacional, a Inteligência Artificial tem contribuído para a automatização dos processos logísticos e de validação técnica, assumindo um papel de "Controlador de Dados", enquanto a força de trabalho humana permanece insubstituível na avaliação da "Inteligência Intuitiva". Por último, regulações e legislação como o RGPD surgem como a principal barreira à implementação da monitorização contínua, o "Efeito Panóptico".

A investigação conclui que o futuro da Indústria se debruça num "Modelo Híbrido" de Inteligência Aumentada, onde a tecnologia surge como agente de agilização e confirmação dos dados, enquanto os humanos protagonizam a interpretação comportamental, rejeitando assim a teoria da substituição total.

Palavras-Chave: Cliente Mistério, Inteligência Artificial, Automação, Modelo Híbrido, Ética

ABSTRACT

In recent years, the practice of Mystery Shopping has witnessed a paradigm shift in its essence. This transition comes from the digital transformation that has taken place, which has consequently fuelled the demand for real-time data. This study investigates the implications of using Artificial Intelligence in the operational and strategic aspects of the sector, focusing on the conflict between process automation and the simultaneous need for empathetic human evaluation.

The dissertation adopts a qualitative method, triangulating interviews with top professionals in the sector with documentary evidence of the industry. The research reveals that Artificial Intelligence is not completely replacing the human evaluator, but rather redefining and adapting their role. The results highlight three main conclusions: First, strategically, the service has transitioned from a mere diagnostic tool to an important financial driver. At the operational level, Artificial Intelligence has contributed to the automation of logistics and technical validation processes, assuming the role of "Data Controller," while the human workforce remains irreplaceable in the assessment of "Intuitive Intelligence." Finally, regulations and legislation such as the GDPR emerge as the main barrier to the implementation of continuous monitoring, the "Panopticon Effect."

The research concludes that the future of the industry lies in a "Hybrid Model" of Augmented Intelligence, where technology acts as an agent for streamlining and confirming data, while humans play a leading role in behavioural interpretation, thus rejecting the theory of total replacement.

Keywords: Mystery Shopping, Artificial Intelligence, Service Quality, Automation, Hybrid Model, Ethics.

ABBREVIATIONS

AI: Artificial Intelligence

CX: Costumer Experience

MS: Mystery Shopping

MFW: Master Final Work

RQ: Research Question

TABLE OF CONTENTS

Contents

ACKNOWLEDGMENTS ..... **I**
ABSTRACT ..... **IV**
ABBREVIATIONS ..... **V**
TABLE OF CONTENTS ..... **VI**
LIST OF FIGURES ..... **VIII**
CHAPTER 1 - INTRODUCTION **1**
CHAPTER 2 – LITERATURE REVIEW **3**
2.1 Mystery Shopping definition 3
2.2 Service Quality 3
2.2.1. The contemporary relevance of SERVQUAL 3
2.2.2. Quality of Process 4
2.3 Purpose of Mystery Shopping 5
2.4 Limitations of Mystery Shopping 6
2.5 Mystery Shopping on predicting loyalty 7
2.6 AI as the tool for Data Collection and Analysis 7
2.6.1. Validity of the Human Approach 8
2.7 Critical Challenges 8
2.7.1. Ethical Surveillance and Employee Stress 8
2.7.2. The "Black Box" effect of AI 9
2.7.3. The Risk of "Mechanical" Standardization 9
2.7.4. Algorithmic Bias in Service Evaluation 9
CHAPTER 3 - METHODOLOGY **10**
3.1. Introduction 10
3.2. Research Philosophy 10
3.3. Research Approach 11
3.4. Research Strategy 11
3.4.1 Research Strategy: Multiple Case Study 11
3.5 Data Collection Method 12
3.5.1 Sampling Strategy (Idiographic Approach) 12
3.5.2. Secondary Data: Documentary Analysis 13
3.5.3. Primary Data: The Interview Protocol 13
3.6. Data Analysis: Thematic Analysis 14
3.6.1. The Coding Process 14
3.6.2. The Six Phases of Analysis 15
CHAPTER 4 – FINDINGS AND DISCUSSION **16**
4.1. Software Analysis 16
4.2. From Diagnostic Tool to Strategic Driver: The ROI Shift 16
4.2.1. Standardization as a Control Mechanism: The "Whip" Effect 17
4.2.2. The Financial Imperative: Efficiency 18

4.3. Automating "Technical Quality": AI as the Task Controller 19
4.3.1. The Transformation of Logistics and Management 19
4.3.2. Data Validation and Logic Checks 20
4.3.3. From Validation to Advanced Analytics 21
4.4. The Human Element: Cognitive Limitations and the Credibility Challenge 21
4.4.1. The “Superhuman” Burden on Data Integrity 22
4.4.2. The Sampling Dilemma 22
4.4.3. The Paradox of Subjectivity vs. Objectivity 23
4.4.4. Credibility 24
4.5. Future Perspectives: Risks and Ethical Challenges 25
4.5.1. The Inevitability of Job Displacement 25
4.5.2. The Threat of Market Monopolies ("The Google Risk") 25
4.5.3. The Response: Ethical Audits 26
4.6. Real-Time Evaluation 27
4.6.1. The UK Case Study: "Live Streaming" 27
4.6.2. The Blocking Consent 27
4.6.3. Signs of Openness and Future Technology 28
4.7. The Main Barrier: Privacy and Ethical Compliance 29
4.7.1. The European Constraint: GDPR as a Brake 29
4.7.2. Technology vs. Privacy 30
4.7.3. The Mandate for Consent 30
4.8. The Future Model: Hybrid Approach? 31
4.8.1. The Separation of Collection and Analysis 31
4.8.2. The "Second Eye": Augmented Intelligence and the human as tool of final control 32
4.8.3. The "Human Premium" and Virtual Frontiers 32
4.9. Chapter Summary 33
CHAPTER 5- CONCLUSIONS AND RECOMMENDATIONS 35
5.1. Main Conclusions 35
5.2. Theoretical Contributions 35
5.3. Managerial Implications 36
5.4. Limitations 37
5.5. Suggestions for Future Research 37
REFERENCES 38
APPENDICES 43
Appendix 1: Interview Protocol 43
Appendix 2: MAX QDA codes list 44
Appendix 3: Software Desktop 45
Appendix 4: Word trends per document 46

LIST OF FIGURES

Figure 1: Participant Statistics……………………………………………………………14
Figure 2: Code Matrix Browser…………………………………………………………..16

CHAPTER 1 - INTRODUCTION

Customer experience (CX) management has established itself as one of the main areas of competition among contemporary companies. The economic importance of this sector is evident, in 2009, the Mystery Shopping industry was valued at over $1.5 billion annually (Zikmund et al., 2009). In a globalized market, where products can be easily replicated, service quality emerges as the main differentiating factor for financial sustainability. In this scenario, Mystery Shopping has historically been recognized as the benchmark tool for compliance auditing and service experience evaluation (Wilson, 2001).

However, the market research industry is facing a paradigm shift driven by Digital Transformation. Advances in Artificial Intelligence (AI) and the growing demand for real-time data are challenging traditional models, which were previously based on retrospective assessments and manual processes. According to Block et al. (2022), Mystery Shopping is evolving from a passive diagnostic tool to a strategic instrument for generating financial value, in which agility and Return on Investment (ROI) become essential factors.

The quest for greater efficiency presents companies in the sector with a strategic decision: automate processes to reduce costs and increase speed or maintain the human element in order to ensure emotional depth and empathy in assessments.

The integration of Artificial Intelligence into Mystery Shopping presents significant challenges. Although technologies such as Machine Learning and computer vision can increase efficiency in data collection, they also raise important questions regarding the validity of assessments and the ethics of surveillance.

Academic literature highlights a central tension: the dichotomy between “Mechanical Intelligence,” dominated by AI, and “Intuitive/Empathic Intelligence,” unique to humans (Huang & Rust, 2018). Furthermore, the continuous monitoring provided by technology, known as the “Panopticon” effect, conflicts with strict regulations, such as the GDPR in Europe, and with employee acceptance, which constitutes substantial barriers to full adoption (Dwivedi et al., 2021).

The central issue addressed in this dissertation is the uncertainty surrounding how these new technologies are reshaping the Mystery Shopping value chain. Specifically, it questions whether AI will replace human evaluators or whether a hybrid model of collaboration will prevail.

This study analyses the impact of Artificial Intelligence on the Mystery Shopping industry, structured around three research questions (RQ). RQ1 examines how is AI transforming the operational processes and strategic value of Mystery Shopping. RQ2 assesses the impact of automation on the human role, particularly regarding empathy and service quality evaluation. RQ3 identifies how ethical and regulatory challenges shape the adoption of AI-driven monitoring in Mystery Shopping.

The relevance of this study is ambivalent. In the academic context, the study fills a gap in the literature on service auditing by applying frameworks such as TAM-DEF and the Theory of Intelligence Substitution to a new context. In the management sphere, it presents a strategic roadmap for the implementation of a Hybrid Model, clearly distinguishing between tasks that benefit from automation and those that require human intervention, to avoid service obsolescence.

The dissertation is structured in five chapters. Chapter 1 presents the introduction, followed by the literature review of the fundamental theoretical models in Chapter 2. Chapter 3 details the qualitative methodology. In Chapter 4, the analysis and discussion of results triangulate empirical evidence with the literature to map changes in the sector. Chapter 5 presents the final conclusions, answering the research questions and presenting recommendations for the future.

CHAPTER 2 – LITERATURE REVIEW

2.1 Mystery Shopping definition

"Mystery shopping, a form of participant observation, uses researchers to deceive customer service personnel into believing that they are serving real customers or potential customers." Wilson, A. M. (2001, p.721)

Wilson's quote is a very good prelude to what can be the Mystery Shopping in the Customer Experience scenario, in a very deep and meaningful way. This old practice still works and has been used since the 1940s (Xu and He, 2014) because of one basic principle: According to (Wilson, 2001), the service personnel's perception of serving a "real customer" enhance the reliability of the measurement. If we measure the process rather than the outcome, we can optimize the process and, in that way, increase the outcome.

Mystery shopping has been used as a hidden observation tool to evaluate service delivery. Unlike satisfaction surveys, which gather general and vague impressions, mystery shopping focuses on the service process, allowing companies to assess whether the planned processes and behavioural expectations are being met (Wilson, 2001).

Meticulous selection and training of shoppers, fully structured scripts, and objective evaluation parameters have been used to increase the reliability of the process (Algi & Lukman, 2018). However, the limitations of this process remain in subjectivity, memory lapses, and low frequency of visits per store. (Finn & Kayandé,1999).

2.2 Service Quality

In order to assess the impact of Artificial Intelligence on Mystery Shopping (MS), it is important to address the essence of service quality. Starting with some theoretical foundations, the literature suggests that service quality is an abstract construct, contrasting with objective quality, which arises from the difference between a consumer's expectations and their perceptions of the service performance.

2.2.1. The contemporary relevance of SERVQUAL

Although the emergence of new ways of measuring services, the SERVQUAL scale still is the main one for categorizing service quality. According to Parasuraman et al.(1988), SERVQUAL evaluates service quality through five different dimensions that are still relevant for any Mystery

Shopping checklist these days, as Devi and Reddy (2016) often argue that the usual categorization of those observational items into these dimensions is to pinpoint training needs and service gaps. The first one is Tangibles (the appearance of physical facilities, equipment and personnel), followed by Reliability (The ability to perform the promised service dependably and accurately. The third one is Responsiveness (the willingness to help customers and provide prompt service) and the fourth is Assurance (the knowledge and courtesy of employees and their ability to inspire trust). Finally, and most importantly, Empathy (the caring, individualized attention the firm provides its customers).

Despite its importance, SERVQUAL is a generic measurement tool with broad applicability (Ladhari,2009), and it has faced some criticism regarding the accuracy of its “difference scores” (Quality= Perception–Expectation). Cronin & Taylor (1992) discuss that perception-only scores (SERVPERF) explain more variance in overall service quality than the difference scores used in SERVQUAL. The authors argue that the performance-based approach is theoretically superior because it avoids the statistical flaws in measuring expectations. Building on the challenges identified by Ladhari (2009), this is where AI-enabled Mystery Shopping becomes essential, by shifting the paradigm to objective data collection, utilizing what Davenport et al. (2020) describes as “Analytical Intelligence”. In that way, this transition moves the methodology away from subjective "difference scores" toward objective, data-driven "perception-only" analysis, stabilizing the validity issues raised.

2.2.2. Quality of Process

The literature identifies a critical distinction within the topic of Quality. Harvey (1998) distinguishes between “Quality of Results” (what the customer receives) and “Quality of Process” (how the service is delivered). Additionally, Harvey (1998) introduces two major concepts: Technical Process Quality, considered the “objective” component (for example, whether the server performed the steps correctly), and Perceptual Process Quality, which is described as “in the eye of the beholder” (for example, whether the customer felt cared for during the shopping experience).

According to Harvey’s distinction, Mystery Shopping settled on human workforce to assess the perceptual dimensions. Despite that, modern technological applications challenge this strict separation by attempting to shift “perceptual quality” into objective technical data. To illustrate, tools like Sentiment Analysis focus to measure the subjective experience, such as whether a customer felt cared for into binary positive/negative metrics, effectively treating a perceptual attribute as a technical process step. As a result, data points that Harvey identified as subjective 'perceptual quality' are analysed by algorithms employing the same quantitative logic used for 'technical' processes.

2.3 Purpose of Mystery Shopping

Although Mystery Shopping has its own limitations, the literature defends its continuity as an "ancillary"/complementary tool to customer satisfaction survey (CSS) (Devi & Reddy, 2016). The authors argue that, while the ordinary surveys tend to be subjective and depend on customer memory, the Mystery Shopping method captures the real experience at a "snapshot in time".

The main goal of using the Mystery Shopping process is an assessment of the level of customer service, which is done by people who take on the role of "average" buyers of the products (Gębarowski & Siemieniako,2015). Van der Wiele et al. (2005) consider that service excellence requires additional effort beyond conventional satisfaction surveys. In the authors' view, Mystery Shopping is a complementary tool that provides in-depth insights into how the service is provided. Through a case study of a Dutch temporary employment agency, they found that repeated Mystery Shopping visits revealed significant improvements in the visible component of the service, such as branch appearance and friendliness, but not in deeper elements such as empathy, trust, and emotional connection.

The practice of Mystery Shopping emerges as an element that monitors and links the various links in the "chain of excellence" (Van der Wiele et al.,2005). This concept illustrates several steps that, once completed, lead the company to strong financial performance. The chain begins by identifying that exemplary leadership promotes employee excellence, which in turn creates good flex workers and satisfied customers, who may then become loyal and generate greater financial performance. The authors also argue that the practice of Mystery Shopping fosters short-term behavioural improvements in service delivery when integrated into feedback and reward systems.

Block et al. (2022) redefine the practice of mystery shopping as a strategic driver rather than a simple diagnostic tool. Using data from three retail chains, they prove that Mystery Shopping is useful for solving agency problems in multi-site firms. The two levels of decision-making (headquarters/central management and site managers/local staff) clash with each other as the first values the company's good reputation and brand uniformity, while the second prioritizes the performance results of its own store, translated into sales and costs. Store managers deliberately focus on sales tasks because they translate into visible, measurable output that is closely related to their performance. On the other hand, the focus on the consumer experience derived from service (greeting customers warmly, keeping the store clean, consistent brand experience) is neglected due to the high allocation of time and resources and to the fact that the benefits (good reputation and customer loyalty) extend to all stores and not just the one that made that investment. This Performance vs. Service dichotomy leads to misunderstandings between local managers and headquarters. Thus, Block et al. (2022) present the practice of Mystery Shopping as a tool for promoting homogeneity and control.

When used, Mystery Shopping scores show low variation between different sites, low correlation with sales but high correlation with local managers' bonuses, confirming their strategic role as an incentive and control tool. However, these studies were conducted by human shoppers who reported standardized ratings, which, with the emerging practice of Mystery Shopping with AI (Lobo & Shetty, 2025), through NLP and computer vision, raises the question of whether AI can replicate or even enhance this strategic role identified by Block et al (2022).

Also related to sales, Blessing & Natter (2019) present some reasons why Mystery Shopping fails to predict them. The authors note that Mystery Shoppers do not become emotionally involved in purchases, a behaviour that is inherent to real customers (Ladhari, 2009). In addition, other factors are also relevant, such as the small number of samples, which leads to unreliable averages, and surveys that are complex and therefore lead to memory lapses and halo effects. Furthermore, shoppers may not truly represent the real customer base, either demographically or motivationally, while owners' checklists lack psychometric validation.

Blessing & Natter (2019) challenge the viability of human mystery shopping as a parameter for evaluating service quality, which, in contrast, reinforces the viability of AI-driven Mystery Shopping. When accompanied by AI, through sentiment analysis, facial recognition, or even voice analytics, the practice can analyse interactions continuously and objectively, without small samples or bias. The authors thus question managerial reliance based on data from traditional Mystery Shopping for subsequent strategic decisions and incentives. This issue leads us to the importance of AI-driven Mystery Shopping systems, which can overcome human limitations on a large scale.

2.4 Limitations of Mystery Shopping

Wilson (2001) highlights three persistent problems with human-based mystery shopping: time lag between observation and strategic management actions, limited sample of visits, and subjectivity in evaluation (behavioural judgments). According to Finn (2001), the main MS method fragility lies on its psychometric reliability when applied to small samples. The author, with the Generalization Theory, points out that the outcomes are highly sensitive to the variability introduced by both evaluators and the specific contexts of service interaction. Unlike the common practice, the author argues that two or even three visits are not enough to capture the heterogeneity of service delivery, especially regarding more subjective contexts, like the Personal Selling. Although the author recognizes the Mystery Shopping as a valid alternative to customer satisfaction inquires, concludes by pointing out that data validity can only be ensured with a considerably higher sample size (up to about 20 visits per store), that normally surpasses what is budgeted for. If not, the results tend to reflect more statistical noise than a true benchmark of the customer experience.

The main goal of Mystery Shopping is acquiring facts, not perceptions, and therefore, these limitations are the bridge to the adoption of automated and AI-driven alternatives. According to Lobo & Shetty (2025), modern technologies such as AI, through Natural Language Processing, sentiment analysis, Machine learning, Computer Vision, and automation promise to closely monitor, provide real-time feedback, and reduce the bias inherent in subjectivity. Despite this, technology brings new ethical and privacy challenges that Wilson's initial approach does not anticipate. However, comparing human-driven and AI-driven mystery shopping practices in terms of reliability, timing, and costs is a pressing issue that needs to be addressed.

2.5 Mystery Shopping on predicting loyalty

Gremler et al.(2020) conducted a meta-analytic compilation of 224 studies on customer relational benefits, identifying three dimensions that foster customer loyalty: confidence, social, and special treatment. The authors note that confidence and social benefits are the biggest drivers of customer loyalty, mainly through perceived value and relationship quality, with switching costs playing a secondary role. The results show that trust and emotional connection are the basis of continued loyalty, rather than transactional incentives. For research purposes in AI-driven Mystery Shopping, this paper leads us to believe that Sentiment Analysis can measure relational benefits in real time, for example, translated into trust or empathy. In this way, AI-driven sentiment analysis may offer a more objective and scalable means of exploring the relationship between service quality and customer loyalty outcomes, complementing and potentially surpassing the limitations of traditional human evaluations.

2.6 AI as the tool for Data Collection and Analysis

AI-enabled Mystery Shopping represents a shift in “task type” and “intelligence level”. Davenport et al.(2020) highlights that AI enhances “Task Automation”, which involves standardized and rule-based applications. In this scenario, AI plays a role of a “Controller of Data”, being capable of analysing non-numeric inputs such as text, voice and facial expressions, in order to gain insights.

This tool automates the observing and listening functions of a mystery shopper. Liu (2012) defines this as the problem of Sentiment Analysis, in which AI identifies an opinion as a tuple of a target (for example a staff member) and a sentiment (positive or negative). Thus, by processing large volumes of unstructured data, such as reviews and recordings, AI enables continuous mystery shopping that would not be possible by humans alone.

2.6.1. Validity of the Human Approach

AI enhances data velocity, yet the literature suggests a hybrid model. Huang and Rust (2018) categorically divide service tasks into Mechanical, Analytical, Intuitive, and Empathetic intelligences. They make it clear: AI process Analytical tasks but cannot solve Intuitive and Empathetic tasks.

AI must add value, not replace, the mystery shopping process. Chuah and Yu (2021) demonstrate that while facial recognition API's detect emotions such as happiness or surprise, only humans have inherently the judgment to interpret those emotions accurately and counteract algorithmic bias or opacity. Dwivedi et al. (2021) suggests that AI systems should generate an audit trail of their decisions. In Mystery Shopping, AI may provide a score, but only a human manager ensures a fair, equitable, and contextually appropriate evaluation.

2.7 Critical Challenges

While the literature establishes AI's potential to optimize the validity and speed of Mystery Shopping, it also highlights significant societal and managerial challenges that threaten the viability of this hybrid model. The integration of AI introduces new risks regarding employee ethics, data transparency, and the potential erosion of the human element in service delivery.

2.7.1. Ethical Surveillance and Employee Stress

A fundamental shift identified in the literature is the move from "intermittent" to "continuous" monitoring. Wilson (2001) emphasizes that for Mystery Shopping to be effective, "employee acceptance" is critical, that is, staff must view the process as developmental rather than punitive. Traditional mystery shopping, being sporadic, allowed for periods of relaxed performance.

However, AI-enabled monitoring, like video analysis, creates a general sense of “panopticon". Dwivedi et al. (2021) warn that AI adoption often outpaces regulation, leading to "Social Challenges" regarding privacy and workers' rights. The "TAM-DEF" framework proposed by Misra (2021) (in Dwivedi et al., 2021) argues that any AI deployment must be assessed for "Ethics" and "Fairness & Equity" to ensure it does not exploit vulnerable populations, in this case, frontline service workers. The acronym stands for “Transparency”, “Accountability”, “Misuse”, “Digital Divide”, “Ethics” and “Fairness” and it was a mnemonic to present the six challenges on AI adoption. If AI monitoring is perceived as surveillance, it may lead to stress and service sabotage rather than optimization, thereby directly counteracting the program's goals.

2.7.2. The "Black Box" effect of AI

As described above by Misra (2021) (in Dwivedi et al., 2021), “Transparency” is one of the main challenges regarding AI adoption. The author argues that “AI based systems tend to function as a black box where the lack of transparency acts as a barrier to adoption of the technology” (p.6).
For Mystery Shopping to drive optimization, the feedback must be actionable. A human shopper can explain *why* they gave a low rating. In contrast, deep learning algorithms often function as "Black Boxes"(we don´t know what’s behind the process) (Rai, 2020**)**. Edwards (2021) (in Dwivedi et al., 2021) highlights the "Explainability crisis” in AI: neural networks may produce an accurate classification, for example, bad service, but they cannot explain the reasoning behind it.
This lack of transparency is a managerial risk. Lebovitz et al. (2022) argue that a "fog of algorithmic opacity" makes it hard for decision-makers to validate data. The insight becomes useless for strategy. If a manager gets a low AI "Sentiment Score" with no explanation, recovery plans are impossible.
Furthermore, Wilson (2001) notes that communicating results is essential for motivation**.** Therefore, ensure that human intervention actively translates algorithmic scores into effective coaching moments (Jarrahi, 2018).

2.7.3. The Risk of "Mechanical" Standardization

Despite the issue of transparency, excessive reliance on AI can also be a theoretical risk that will distort service standards toward what is easily measurable, rather than what is important (Huang and Rust, 2018). The author alert that while AI is proficient in "Mechanical" and "Analytical" intelligence, it still faces challenges with "Intuitive" and "Empathetic" tasks.
Harvey (1998) highlights that Technical Quality, that is the service delivered, serves merely as a baseline while “Perceptual Process Quality” (the subjective experience of the service) is what distinguishes exceptional service. Although AI may rate a transaction as 100% compliant, if an employee smiles and uses prescribed language, it may fail to detect the “Uncanny Valley”, which refers to discomfort caused by artificial emotion (Chuah and Yu,2021). Excessive reliance on AI optimization can therefore produce service that is technically flawless but emotionally unfulfilling.

2.7.4. Algorithmic Bias in Service Evaluation

The objectivity of AI is not guaranteed. Dwivedi et al. (2021) discuss the serious problem of “Algorithmic Bias,” stating that AI models often replicate biases found in their training data. In the mystery shopping ecosystem, a trained AI model may penalize employees with specific accents or facial expressions if the training set associates these characteristics with low service scores. Unlike a human shopper, who can be trained to recognize and correct their own "method bias" (Ladhari, 2009), an AI system requires rigorous and continuous "transparency and audit" to ensure that service is evaluated fairly across different employee demographics.

CHAPTER 3 - METHODOLOGY

3.1. Introduction

This research aims to explore the role of Artificial Intelligence in optimizing the service level within Mystery Shopping practices. To approach the main topic of investigation, the study addresses a paradigm shift from traditional observation to automated evaluation by questioning how strategic decision-makers are experiencing this transition. This chapter presents research philosophy, data collection, and analysis methods applied to answer the research question.

3.2. Research Philosophy

This investigation assumes an Interpretivist research philosophy. This type of research is conducted under the assumption that reality is socially constructed (Saunders et al.,2023), fully based on experiences and decisions of individuals. To justify this type of research it is imperative that we approach the Positivist paradigm. Bell et al. (2022) argue that a Positivist approach characterizes an objective reality as independent of social actors, in which a statistical measurement produces law-like generalizations. Such an approach would be insufficient for the objectives of this research, despite having been applied in Mystery Shopping research to test psychometric reliability, as proved by Finn's (2001) use of Generalizability Theory.

The incorporation of AI into service is not only a measurement of a variable, but a complex renovation of the human-machine relationship (Huang & Rust, 2018). Therefore, and according to Creswell and Creswell (2018), when the objective is to understand the meaning that individuals attribute to a social problem, rather than assessing cause-and-effect relationships, Interpretivism is the necessary path to follow.

A Positivist approach captures the extent of AI use but does not address the strategic logic or managers' subjective perceptions of risk and value (Block et al., 2022). Therefore, an Interpretivist philosophy is essential for understanding these deeper strategic nuances.

3.3. Research Approach

This research is based on an Abductive research approach, which involves moving back and forth between the data collected (interviews) and theory (literature). In this research, we are gathering data to investigate a phenomenon by identifying and explaining patterns to test an existing theory through additional data collection, a typical approach in an abductive framework (Saunders et al.,2023).

Given that this study is guided by three Research Questions and builds upon an established theoretical framework, neither a purely Inductive nor a purely Deductive approach is adequate. An Abductive approach allows the research to move iteratively between theory and data, remaining open to emerging insights while still being anchored in existing literature (Saunders et al., 2023).

We will conduct interviews based on a protocol (see Appendix 1) that is grounded on Block et al. (2022), Wilson (2001), Lobo & Shetty (2025) , Blessing & Natter (2019) and van der Wiele et al. (2005). The presented literature will guide the conversations, but we will always remain open to surprising insights from the CEOs, that may confirm, contradict or ignore what is established in the theory, allowing for the change of existing frameworks.

3.4. Research Strategy

Qualitative inquiry is fundamental when a topic still is “immature” or the relevant components are not fully developed yet (Creswell & Creswell, 2018). Therefore, semi-structured interviews with top executives of Market Research firms provide the necessary knowledge to understand the mechanisms behind the optimization process that surveys cannot reveal. Saunders et al. (2023) proclaims that semi-structured interviews are the most valid method when the study lacks complex or privileged data (in this case, C-level strategic decision making), that cannot be collected through standardized surveys. This method strikes an important balance: it ensures that "deductive" themes from the literature, such as algorithmic bias, are included. It also enables "abductive" discovery of new and unexpected industry trends (Creswell & Creswell, 2018).

3.4.1 Research Strategy: Multiple Case Study

Following the Interpretivist philosophy, the strategy that most applies to the investigation is a Multiple Case Study. According to Yin (2018), when investigating a current phenomenon in its real-life context, especially when the boundaries between the phenomenon (AI adoption) and the context (organizational strategy) are unclear, this strategy is empirically essential.

Even tough Yin (2018) presents the structural framework for using "Replication Logic" across cases, this investigation also relies on the fact that the primary goal of case study research in not to quantify a phenomenon, but to assess its "intrinsic particularity" (Stake,1995).

This method uses each firm as a unique unit of analysis. The choice behind the multiple case study approach is sustained by the following key methodological arguments: Firstly, as Eisenhardt (1989) concluded, a "cross-case analysis" is permitted when enables the researcher to identify patterns that a single observation might miss. This method is important for assessing the CEO's choices towards the "Hybrid model" presented by Huang and Rust (2018).

Furthermore, the investigation looks for analytical generalization rather than statistical generalization, representing a population, as Yin (2018) said. The main point of the study is not to expand and cover frequencies, but theories (AI viability in Mystery Shopping practices). It's not about trying to prove that everyone does this (statistics), but to prove how and why this happens (analytics), using these cases as proof of concept. The "force of example", presented by Flyvbjerg (2006, p.228), illustrates the misunderstanding that "one cannot generalize from a single case". The author debates that in strategic management, a few in-depth examples ("the black swans") can provide more "information power" than a simplistic survey of thousands. By analysing two market-leading firms, this study catches the "strategic nuance" that the author argues is the core value of social science.

3.5 Data Collection Method

To attenuate the limitation of a small sample size and confirm the validity of the results, this research employs Methodological Triangulation (Denzin,1978). Data were collected from two different sources: primary strategic interviews and secondary in-depth documentation.

3.5.1 Sampling Strategy (Idiographic Approach)

The research uses an idiographic sampling strategy, categorizing "information power" relative to sample volume. The required sample size relies on the information density of the interviews (Malterud et al.,2016). The sample used consists of a kind of "elite" information providers, in this MFW constituted by CEOs of Market Research firms.

Moving back, Block et al. (2022) addresses Mystery Shopping practices as a strategic management practice, which is used to solve high-level agency problems. Although operational staff knows how processes work, only C-level executives can debate the viability of the AI integration.

In such elite inquires, a small sample size (n<20) is often advantageous (Crouch and Mckenzie,2006), since it allows a detailed analysis of the participants strategic logic that would be dispersed in a large amount of data. For that reason, this investigation is focused on two in-depth strategic cases (n=2), triangulated with credited documentary evidence.

3.5.2. Secondary Data: Documentary Analysis

In order to confirm the validity of the Primary Data, a Documentary Analysis was made based on industry standards and strategic reports. This step was crucial and necessary prevent the respondent bias phenomenon, where guests may exaggerate their level of expertise. Data triangulation was based on an intentional sampling strategy (Bowen, 2009), prioritizing documents based on relevance, authority, and contemporaneity. Systematic keyword searches (e.g., “Artificial Intelligence,” “Mystery Shopping”) were used to efficiently extract corroborative evidence from these sources.

The documents that were chosen for the triangulation were: The MSPA Europe/Africa “The Future of Consumer Experience: Evolving Trends, the Impact on Mystery Shopping Companies and how we all need to evolve” (Narsaria, 2024); The MSPA Europe/Africa “Guide to Mystery Shopping” (2024); The MSPA “Common Codes of Professional Standards and Ethical Conduct” (2018); International Chamber of Commerce & ESOMAR “International Code on Market and Social Research” (2008) and “The employee´s guide to Mystery Shopping A tool for developing the whole company” (Boxberg Karlsson & Thomasdotter Schölin, 2013).

These documents represent the industry's normative standard in terms of ethics and strategy, as well as codes of conduct, best practice guidelines, future perspectives, and results. The topics covered in the documents overlap with two key points addressed in the interviews: “Strategic Value” and “Ethical Standards”.

3.5.3. Primary Data: The Interview Protocol

Primary data were obtained through semi-structured interviews conducted via video conferencing, each lasting 45 to 60 minutes. The interview protocol comprised three thematic sections intended to systematically examine the transition from traditional auditing methods to AI-enabled auditing.

The interview protocol (Appendix 1) was carefully developed to address gaps and theories identified in the Literature Review.

Interviews were conducted via Microsoft Teams to accommodate the geographic distribution of experts. The method was always aligned with ethical compliance, since before each interview, participants received a Participation Agreement form to consent and authorize the recording. This form also ensured the anonymity of participants and the right to withdraw, according to ethical research guidelines (Saunders et al.,2019)

Figure 1: Participant Statistics

Interviews made in November/December 2025
Average duration: 45 m

Interviews were conducted via Microsoft Teams to accommodate the geographic distribution of participants

Primary Data: Interviews	Gender	Age Group	Years of Experience in the Industry	Sectors they work with
Participant A	Male	40-45	25 years	Restaurant, Banking, Retail
Participant B	Male	55-60	28 years	Restaurant, Banking, Hospitality, Automotive, Retail

Source: Author's elaboration using Canva

3.6. Data Analysis: Thematic Analysis

The qualitative data from the semi-structured interviews and documents were analysed using Thematic Analysis. This study follows the six-phase framework by Braun and Clarke (2006), which is a widely accepted standard for finding and reporting patterns in data.
To keep the dataset organized and create a clear record of the coding process, the study used MAXQDA 2024 (VERBI Software), a computer-assisted qualitative data analysis tool. MAXQDA helped retrieve codes and show how the CEOs' strategic claims related to the industry documents.

3.6.1. The Coding Process

To address the research questions while remaining receptive to emergent insights, a dual coding approach was implemented using MAXQDA: Deductive Coding (Theory-Driven): Based on the theoretical framework, an initial codebook was developed (see Appendix 2). Some of the examples codes are: “Back-office tasks automation”, “Risks & Job Displacement”, “Empathy and emotions”. The main purpose for that is to confirm or not if the Industry practice aligns with the academic predictions (see Appendix 4).

3.6.2. The Six Phases of Analysis

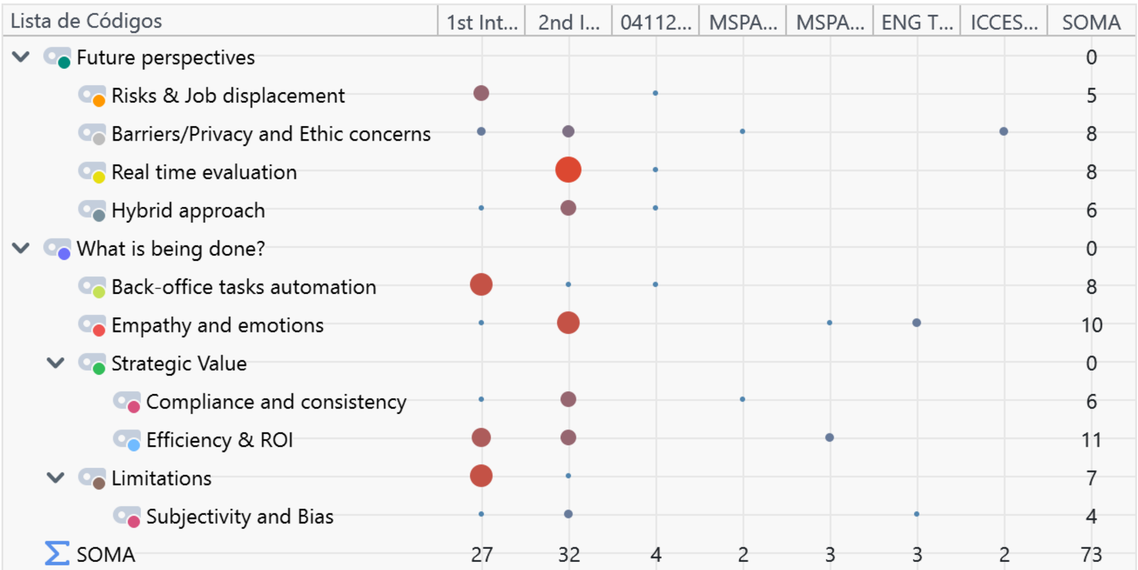
According to Braun and Clarke (2006), the analysis proceeded through six steps. In first place, the process began with the transcription of the interviews, assisted by the MAXQDA software. Then, the whole dataset (interviews and documents) was imported into the software (MAXQDA) and the most interesting features of the data were selected and highlighted using the software's coding function, for example, segments mentioning "human interactions" were associated with the code "Empathy & Emotions". In third place, the long list of codes was reviewed and grouped by themes. For instance, codes related to "task automation", "less cost" and "Efficiency/ROI" were grouped under the theme "Strategic Value". The next step was to discard themes with insufficient data support and soon after, those who remained, were aligned with the research terminology. Finally, the analysis was combined with the literature, selecting explanatory examples form the CEOs and the documented evidence.

CHAPTER 4 – FINDINGS AND DISCUSSION

4.1. Software Analysis

Figure 2 displays the Code Matrix Browser, organized by column similarity to emphasize patterns across data sources. The clustering of interview sources in contrast to documentary evidence illustrates the triangulation process. The row totals (Sum) reflect the relative prominence of each thematic category. The Code Matrix Browser uses what Kuckartz and Rädiker (2019) define as a case-oriented visualization to contrast data sources. In accordance with the cross-case analysis principles outlined by Miles, Huberman, and Saldaña (2014), the visualization reveals distinct clusters that indicate pattern regularity. The proximity among interview sources demonstrates what Yin (2018) describes as "converging lines of inquiry", thereby confirming that the identified themes reflect a shared consensus among experts. In contrast, the clustering of documentary evidence serves as corroboration, indicating that these themes are embedded within the industry's institutional context.

Figure 2: Code Matrix Browser



Source: Author's elaboration using MAXQDA 24 software

4.2. From Diagnostic Tool to Strategic Driver: The ROI Shift

The analysis of the collected information confirms that Mystery Shopping has been through a process of evolution and even for a role change in the industry overview. The practice has evolved from a diagnostic asset to a strategic tool in order to accomplish two critical points: the demand for total Standardization (compliance) and the urge for Operational Efficiency (cost reduction and

ROI). The results suggest that while the goal is to protect the brand, the process is mainly focused on cost reduction and speed.

4.2.1. Standardization as a Control Mechanism: The "Whip" Effect

The first addressed point of strategic value is the power to allow a certain level of uniformity across large-scale operations. The concept of "Compliance and Consistency" is often referred in the dataset and suggests that the main value of Mystery Shopping is the ability to stamp a unified brand standard across diverse geographies.

Participant A offered a valuable metaphor for this function. He characterized the service as a tool for centralized command and control:

> *"They want to standardize. It is as if they had a whip... being that this whip is just for the employees to know that this exists, so that there is uniformity across the whole country, whether it is marketing materials exposed or the commercial conversation they have."* (Participant A)

The findings show that standardization is still a core concept within Mystery Shopping. Participant A used a prominent metaphor, explaining that their clients use their services aiming a certain type of uniformity, characterizing the evaluation process as a "whip".

The use of this metaphor corroborates Wilson's (2001) view of Mystery Shopping as a management control mechanism. This confirms the author's warning that the method frequently acts as a disciplinary tool to enforce compliance, applying intensified surveillance. The participant's description shows the risk that staff may see the process as punitive rather than "developmental", creating friction that management must address.

Participant B confirms this perspective by describing it as **"**Brand Protection**"**. He stated that in a globalized market, heavy deviations from the standard are a threat to the brand's identity:

> *"Excellence is consistency... It is fundamental to see if there are deviations and to be able to act and correct. It is 'brand protection', we are defending the brand. The brand says it works like this, and then it expanded, and then we have oscillations, and that is not going to be good."* (Participant B)

This conceptualization of "Brand Protection" agrees with Van der Wiele et al. (2005), who describe Mystery Shopping as a monitor of the "chain of excellence", essential for financial performance. In addition, it reflects the specific "headquarters perspective" discussed by Block et al. (2022), where the priority is to maintain brand uniformity and homogeneity across diverse locations to solve agency problems.

For Participant B, the strategic value settles in the clarity of expectations. He argued that *"it is important for people to know what is expected,"* because only then can management *"measure... present results and correct to always raise the level."* In this way, the service (Mystery Shopping) acts as the only tool capable of *"measuring the moment of truth"*(Participant B), ensuring that the defined standards are implemented in practice.

The documentary evidence presents this as a fundamental rule: Mystery Shopping Professionals Association [MSPA] (2018) explicitly state that the mission is *"to ensure the maintenance of consistent professional standards and ethical behaviour, in line with its mission and value statements."*

4.2.2. The Financial Imperative: Efficiency

Despite the Mystery Shopping role as a control mechanism to achieve standardization and, consequently, brand image, the analysis of the code "Efficiency & ROI" in MAXQDA suggests that the principal cause for adopting Artificial Intelligence is cost reduction. The primary data (interviews) shows a paradigm shift from effectiveness (doing the things right) to doing the things faster/cheaper.

Participant A was clear about the objective of integrating AI in his business: *"I have every interest in having artificial intelligence, which is to reduce costs."* He gave a specific operational transformation where Artificial Intelligence replaces human workforce to increase throughput. By automating one of the Mystery Shopping stages, the validation process, he predicts a threefold reduction in costs:

> *"Instead of validating 5 visits per hour, they start validating 15, ok? And for us, it is three times less cost. Therefore, there is this component of cost reduction, which is what will happen."*
> (Participant A)

This quest extends to the restructuring of the workforce. Participant A stated that in the future, with the AI fully integrated, his goal is to *"finish with those... external people who are on their computer doing [validation],"* confirming that the innovative way of doing is being used to *"remove people from these processes"* (Participant A).

Participant B corroborated this thought but argued it from the standpoint of time and productivity. He argued that scaling up the operation makes manual work obsolete, once the impact of small efficiency gains has been verified.

> *"If we save five minutes with increased productivity, that is 10,000 minutes. Therefore, 10,000 minutes is many hours... It is very, very significant."* (Participant B)

Participant B concludes by saying that the adoption of these tools is not optional but a survival necessity in the market: *"Those who do not follow this path... stay working more manually, and get left behind."* He further noted that this efficiency translates into *"efficiencies of price and speed,"* citing the United Kingdom market as a benchmark where clients already have *"video recordings immediately available."*

These findings support the main argument of Block et al. (2022), who say that Mystery Shopping is now seen as a strategic driver rather than just a diagnostic tool. The shift from focusing on quality checks to "efficiencies of price and speed" demonstrates that companies now use it mainly to address "agency problems" in multi-site firms by producing measurable, scalable results, not just vague impressions of satisfaction.

The triangulation of these sources demonstrates that the strategic value of Mystery Shopping has been shifted.

4.3. Automating "Technical Quality": AI as the Task Controller

The main objective of the investigation was to identify how AI is currently being integrated into the Mystery Shopping process. The analysis of the MAXQDA code "Back-office tasks automation" ended up indicating that the principal application of Artificial Intelligence is not yet in the field (the collection), but in the huge automation of management, logistics, and data validation.

4.3.1. The Transformation of Logistics and Management

The data reveals a notable change in how projects are structured. Participant A addressed the logistical challenge with the usual questions that arise: *"where are we going," "how are we going to organize this,"* and *"how many people do we need"*. He noted that traditionally, this process was labour-intensive:

> *"Today it is all done manually, with emails, with phone calls, but artificial intelligence will be able to, in the automation process... receive information from the client side."* (Participant A)

The analysis shows that the “management work” has been taken over by AI. Participant A imagined a future where the system is autonomous, relying on the project parameters. Instead of being a human calling manually the analysts/shoppers, the AI executes that:

"Artificial intelligence will be able to receive the file... go to the database and start sending emails to the people it identified based on geographical location, etc. In other words, management work." (Participant A)

Participant B confirmed the dependence on software to enhance efficiency, particularly emphasizing its application in recruitment. He explained that Artificial Intelligence is used to screen the workforce to maintain quality and productivity:

"Then the machine helps us to recruit in regions and exclude scores. We have whole software that helps us be increasingly productive." (Participant B)

Automating these administrative workflows aligns with Davenport et al.'s (2020) typology of "Task Automation". Activities, such as autonomous email routing and database management, fit Davenport's definition of standardized, rule-based applications that increase speed. In this initial phase, the data shows that AI is used to manage administrative tasks within the logistical pipeline, rather than replace high-level strategy. This approach frees humans from what Huang and Rust (2018) describe as "Mechanical Intelligence" tasks.

4.3.2. Data Validation and Logic Checks

In addition to logistics, the most significant finding in this section concerns Data Integrity. Participant A asserted that the influence of AI in this domain is dominant: *"Intelligence will help a lot in data validation, of course, it is very clear, and in the speed of validation."*
The analysis presents a specific example illustrating how AI conducts "Logic Checks" to prevent human error. According to Participant A, manual validation frequently results in mistakes due to human fatigue, while AI can rapidly identify contradictions within a report:

"In the validation part, see the logic of the questions... for example, they had to go to a restaurant and had to order fish but asked for meat. If that happened, the visit itself has to be repeated. In the answers, there must be logic between them, if the analyst put in the description: 'The food was not good', and then put in the option 'Was the food good?' a 'yes'. This is done humanly today, therefore, with errors and takes time." (Participant A)

This statement validates that AI is employed to enforce "Technical Quality" by ensuring alignment between quantitative data, such as scores, and qualitative data, such as descriptions, without the need for human intervention.

This shift from manual error correction to automated logic checks aligns with Harvey's (1998) concept of "Technical Process Quality", which represents the objective side of service delivery and defines the standard for accuracy. By using algorithms to check facts, like making sure the ordered item matches the report, the agency makes its data more reliable and reduces concerns about human bias that are often mentioned in SERVQUAL critiques.

4.3.3. From Validation to Advanced Analytics

The findings indicate that "Back-office automation" encompasses not only data cleaning but also the processing of insights. After logistics are completed and data is validated, Artificial Intelligence is responsible for analysing the results.

Narsaria (2024) discusses that Mystery Shopping companies must shift from basic checklists to advanced analytics that happen automatically in the background:

> *"Instead, mystery shopping companies must integrate AI-driven analytics into their evaluations. Tools that track how well a retailer customizes its services for different shopper personas will be critical in measuring the effectiveness of personalized experiences."* (Narsaria, 2024)

To provide customized service, it is also necessary to have a well-prepared and trained analyst. That person should be selected based on many variables, and this heavy process can be optimized with AI.

The triangulation of these sources from fixing logic errors (Participant A) to managing recruitment (Participant B) and generating strategic analytics (Narsaria,2024), indicates that Artificial Intelligence has fully taken on the role of "Controller of Data" as defined by Davenport et al. (2020). By automating the entire back-office chain, the industry is applying "Task Automation" to handle standardized, rule-based applications. This shows that AI's prioritized value right now is in improving data speed and accuracy, serving as the backbone that aids human analysis.

4.4. The Human Element: Cognitive Limitations and the Credibility Challenge

Previous sections showed AI's potential to automate logical processes. However, the analysis of the code "Limitations" indicates that relying solely on human evaluators in the field poses significant challenges in cognitive capacity, statistical representativeness, and managing subjectivity.

4.4.1. The "Superhuman" Burden on Data Integrity

One of the key findings is that the Mystery Shopping firms' clients demands and requirements often exceed the processing capacity of a human observer. Participant A identified human fallibility as the primary limitation of the traditional method:

> *"I think the biggest failure is because the mystery shopper is a person, is a human being who has limitations, cannot control everything."* (Participant A)

This limitation threatens study quality when evaluation checklists are overly complex. Participant A noted that data integrity is compromised by cognitive overload rather than bad faith.

> *"Data integrity is often called into question because the amount of information the client wants is often superhuman... I think that is a big limitation, the large amount of information the client wants to extract from the Mystery Shopper."* (Participant A)

These findings validate Wilson's (2001) point on the topic, who identifies "memory lapses" as a persistent issue in human-based Mystery Shopping. The cognitive overload reported by participants illustrates the difficulty in maintaining accuracy within what Harvey (1998) defines as "Technical Process Quality," which refers to the objective and sequential delivery of service. Contrary to human limitations with extensive checklists, this area is well connected with the strengths of "Mechanical Intelligence," where AI presents superior performance, as noted by Huang and Rust (2018).

4.4.2. The Sampling Dilemma

In addition to individual limitations, the analysis highlights a structural constraint: the cost of scaling. Participant A described the Mystery Shopping report as "a verdict of a single person," which is statistically insufficient. While the ideal is to sample each store, economic constraints make this approach prohibitively expensive.
This dependence on individual evaluators is intensified by the nature of the labour force. The industry depends on *"freelancers"*, introducing variability in delivery quality. Participant A addressed this challenge within the "gig economy":

> *"One of the things that happens is that normally the people working are freelancers... The degree of delivery... besides there being many people who are good, the majority are not."* (Participant A)

Participant B illustrated the logistical challenges of recruiting qualified freelancers with an example from the automotive sector. The specificity required (owning a specific car brand but lacking evaluation bias) requires ongoing recruitment and training:
"Limitations are, for example, are we going to work with [Major automotive firm A] or [Major automotive firm B]? (...) So, we need a person to go to the workshop who has a [Car brand A]... The person has no experience in these systems but has a [Car brand A], so we have to train them with this mindset." (Participant B)

The Finn's (2001) "Sampling Dilemma" is supported by the participant's concern about the "verdict of a single person". Finn argues that Mystery Shopping's reliance on small samples makes outcomes highly sensitive to variability and inclined to "statistical noise" rather than providing a reliable benchmark. These assessments are described by Van der Wiele et al. (2005) as a "snapshot in time". The use of freelancers further improves the variability argued by Finn (2001) in "Generalizability Theory," since evaluator context introduces bias that is more difficult to control than in automated systems.

4.4.3. The Paradox of Subjectivity vs. Objectivity

Analysis of "Subjectivity and Bias" code reveals a fundamental tension between theoretical guidelines dictates and clients' demand. The documentary evidence, via "The Book for Employees", establishes the primary rule, objectivity:

> *"The purpose of mystery shopping... is not to find out the mystery shopper's personal opinion. Therefore the questions in the checklist must be defined so that they reduce the possibility of subjective opinions."* (Boxberg Karlsson & Thomasdotter Schölin, 2013).

Participant A confirmed this view, stating that *"the Mystery Shopper theoretically cannot give opinions."* However, the data shows that commercial reality often contradicts this theory. Clients request subjective feedback, which places agencies in a challenging position

> *"It won't say if they liked the meal or not, that is not the Mystery Shopper's role, ok? But clients want that too. Therefore, there is a strong component here that sometimes the client wants and then we... are very irrelevant in the decision process."* (Participant A)

This challenge between strict internal guidelines and client requests demonstrates the conflict described by Harvey (1998) between "Technical Quality" (objective steps) and "Perceptual Process Quality" (subjective experience).
The agency works to maintain the objectivity described by Ladhari (2009) to prevent statistical flaws and bias. At the same time, clients often ask for "Perceptual" insight, or what is "in the eye of the beholder," which is inherently subjective.

4.4.4. Credibility

In order to manage this subjectivity and ensure the evaluation is valid, Participant B selected Credibility as the most valuable, and fragile, asset of the human process. Unlike a machine, a human must "act" a convincing role:

"The number one rule of the Mystery Shopper is credibility. If we appear before the salesperson... we have to have credibility. If we don't go with credibility, the person on the other side suspects it is an evaluation." (Participant B)

Achieving this naturalness requires significant training, as the workforce is made up of non-specialists. Participant B noted that success relies on turning "people like you or me" into capable evaluators:

"We have to train our Mystery shoppers very well so they can evaluate well... We are working with people who do not work with us every day, who are people like you or like me, who do this part-time, that is the truth." (Participant B)

This need for some sort of “acting normal” corroborates the critique by Blessing and Natter (2019), who argues that Mystery Shoppers frequently don’t get emotionally involved like real customers, which creates a gap in evaluating authentic experiences. Participant B’s main challenge is about overcoming the "observer effect" described by Wilson (2001), so that staff act naturally, rather than performing for someone they know is an auditor.
These quotes also point to Ladhari (2009) , which argues that there is no emotional involvement in the purchase. This limitation is overcome when AI turns “any customer into an auditor”. By analysing large amounts of spontaneous opinions, companies no longer need shoppers to simulate emotions and can now assess NPS based on the true feelings of customers. Therefore, technology allows careful compliance with processes to be linked to real customer loyalty.
In summary, the data shows that the human factor has inherent limitations. It is cognitively fallible when facing "superhuman" tasks, costly to sample due to reliance on a "verdict of a single person,"

and requires ongoing training to maintain credibility and balance technical objectivity with consumer opinion.

4.5. Future Perspectives: Risks and Ethical Challenges

The final thematic code addresses the future consequences of AI adoption. Although efficiency drives this transition, the analysis reveals significant fears about its socio-economic impact. The findings highlight two threats: job displacement at the micro level and market concentration at the macro level.

4.5.1. The Inevitability of Job Displacement

The immediate risk addressed by the participants is the reduction of the human workforce. Participant A was direct in his evaluation of the short-term consequences of automation, articulating that increased efficiency often incurs costs to labour:

> *"Therefore, the risk is unemployment. There will be fewer people working, there will be less work."* (Participant A)

He explained that this is a result of the technological changes discussed in Section 4.3. As machines handle more back-office and validation tasks, fewer human operators are needed.

> *"It is the automation of processes that will lead to companies having fewer workers, so there is more unemployment in the short term."* (Participant A)

This finding supports the "Societal Challenges" identified by Dwivedi et al. (2021) in AI integration. It demonstrates the shift described by Huang and Rust (2018): as AI becomes proficient in "Mechanical Intelligence" tasks, human labour displacement in standardized areas becomes inevitable.

4.5.2. The Threat of Market Monopolies ("The Google Risk")

Participant A highlighted an existential threat to the Mystery Shopping industry: the potential entry of major tech companies. He expressed concern that firms with extensive data resources, such as Google, could enter the market and make traditional agencies obsolete.

"The big risk is these companies ending. Because... Google can start doing this." (Participant A)
The concern extends beyond competition to the imbalance of power. Participant A argued that the scale of these technological monopolies distorts the market, making it difficult for traditional firms to compete:

"They can destroy any company. And that is a danger which is the concentration of wealth in a few companies." (Participant A)

This issue is a broader version of the "Digital Divide" described in the TAM-DEF framework (Misra, 2021). Most research focuses on individual access or "Fairness", but Participant A points out a digital divide between companies. This means that the "Societal Challenges" mentioned by Dwivedi et al. (2021) may also include whether specialized firms can survive against data monopolies, rather than just only workers' rights.
This finding indicates that while AI improves efficiency, it also lowers barriers for tech giants to enter the sector. This could result in the concentration of wealth and the loss of specialized service providers.

4.5.3. The Response: Ethical Audits

Documentary evidence indicates that the industry should establish new regulatory mechanisms to address unemployment and unchecked corporate power. Narsaria (2024) argues that future evaluation must incorporate a moral dimension, rather than relying solely on technical criteria:

"Evaluations will need to include ethical audits." (Narsaria,2024)

The industry's internal appeal for regulation supports Dwivedi et al. (2021), who argue that AI systems need rigorous and continuous transparent auditing to combat algorithmic bias. This aligns with the TAM-DEF framework's requirement for ethics and responsibility, confirming that, without human intervention to audit the “black box”, automated decisions risk losing the necessary validity for a fair evaluation of the service.

As the human element diminishes and corporate power becomes more concentrated, the industry must increasingly prioritize ethical audits of automated systems to prevent their use as solely extractive instruments.
The findings in this section indicate a pessimistic outlook for future developments. While the efficiency gains outlined in Section 4.2 offer benefits, they also introduce significant risks, including short-term unemployment (Participant A) and the emergence of data monopolies

(Google). To address these challenges, the implementation of ethical audits is proposed, implying that the subsequent evolution of Mystery Shopping will necessitate rigorous oversight of automated systems.

4.6. Real-Time Evaluation

The analysis of the code "Real time evaluation" indicates a clear dichotomy: the technology for real time evaluation already exists and is operational, but its implementation depends on the legislative framework. The data suggests that the future of the industry involves reducing the time between observation and reporting.

4.6.1. The UK Case Study: "Live Streaming"

Participant B provided compelling evidence that real-time evaluation is already a market reality in other regions, rather than a speculative concept. The interviewee compared the situation in Portugal to that of the United Kingdom, where legislation is comparatively more permissive:

> *"We, for example, in the UK, in the automotive sector, which is a sector that uses Mystery Shopping a lot... they are already doing video and live streaming. They are in a room watching, that is already possible. Therefore, if there is legislation, technology already exists."*
> (Participant B)

This example shows that the technical ability to transmit data *"straight to the report"* (Participant B) is already installed. Therefore, the obstacle is not technical, but legal.

This operational reality demonstrates the change described by Wilson (2001) in the evolution of control mechanisms. Wilson addressed a shift from "intermittent observation," such as traditional mystery shopping, to "continuous monitoring." The "Live Streaming" capability described by Participant B exemplifies the "Panopticon" effect, where the "period of relaxed performance" allowed by sporadic visits is replaced by continuous digital surveillance.

4.6.2. The Blocking Consent

Unlike in Britain, applying this technology in Portugal is subject to strict consent requirements. Participant B noted that real-time recording of human interactions is only allowed "if there is authorization for that," and acknowledged, "in some projects we have had it, but it is difficult."

The barrier settles in the employee's right to refusal. The interviewee detailed that implementation relies on explicit agreements, often framed as training:

> *"We have a project or two where all teams signed, accepting for training and continuous improvement purposes. They accept to be recorded, but if they don't authorize, no, it cannot be done."* (Participant B)

Participant B concluded that widespread adoption of this practice is conditional: *"Real time will be increasingly used, if legislation helps, it will be increasingly a fact."*

The participant's emphasis on framing the recording as a tool for "training and continuous improvement" directly activates the critical success factor identified by Wilson (2001). The author warns that, for continuous monitoring to be accepted, the team must see the process as "developmental rather than punitive." By ensuring consent through the promise of training instead of policing, the agency navigates the friction of "surveillance" described in the literature and mitigates the risks of stress and "service sabotage" that Wilson associates with forced monitoring.

4.6.3. Signs of Openness and Future Technology

Despite the restrictions, the data suggests a slow, yet possible evolution. Participant B referenced discussions with regulatory authorities, suggesting a modest change in the traditionally conservative banking sector:

"We had our session with the EBA-European Banking Association… and in banks it is not yet possible, but some openness was seen for it to be possible in the future." (Participant B)

Documented evidence indicates that "real time" will expand beyond video to include data traceability technologies. Narsaria (2024) identifies Blockchain as the next frontier for instant auditing:

> *"These audits could be based on real-time blockchain data, enabling shoppers to trace the entire product lifecycle, from manufacturing to the point of sale."* (Narsaria, 2024)

The analysis demonstrates that the industry is technologically prepared for "Real-Time Evaluation", with ongoing functional examples of Live Streaming in the UK. However, the global implementation of this approach is constrained by the requirement for Explicit Consent, resulting in a situation where progress is determined more by legislative developments than by technological advancements.

The current body of research has not directly suggested that Blockchain would be used for this particular industry. However, the findings of this study suggest a technological solution to the "transparency" concerns presented by Dwivedi et al. (2021). More specifically, the use of blockchain's immutable ledger technology has two key outcomes: first, it addresses the "black box" opacity which is a concern mentioned by Rai (2020) as well as creates an audit trail for traceability, second, it eliminates the "method bias" identified by Ladhari (2009), by allowing the use of objective verifiable data rather than subjective human observation. As such, the adoption of Blockchain technology will assist in achieving the "fairness & equity" defined by the TAM-DEF framework.

4.7. The Main Barrier: Privacy and Ethical Compliance

Although real-time AI monitoring is technologically feasible, privacy remains the main barrier to its implementation. The findings highlight a clear conflict between technological capabilities and legal restrictions, especially in Europe.

4.7.1. The European Constraint: GDPR as a Brake

The interview data indicates a consensus that the regulatory environment in Europe slows down in a significant way innovation compared to other markets. Participant B also noted that the General Data Protection Regulation (GDPR) is the primary constraint:

> *"But there is always the GDPR. We in Europe are much more limited… And no, it does not allow us to move so fast."* (Participant B)

Participant A corroborates this limitation specifically regarding the Portuguese market, stating categorically that the main mechanisms of Artificial Intelligence monitoring (video and audio recording) are currently illegal: *"In Portugal, we cannot record visits."* He emphasized that the current legal blockade restricts discussions regarding AI implementation to a theoretical level until the regulatory framework is updated:

> *"The person cannot be filmed and artificial intelligence would be recording. Until the law changes, nothing of this will be permitted."* (Participant A)

The warning given by Dwivedi et al. (2021) that AI adoption often outpaces regulation deadlocks this legislative. Consequently, this leads to important "Social Challenges" about workers' rights. The "legal blockade" mentioned by Participant A serves as an important safeguard against the unchecked spread of the "Panopticon" effect described by Wilson (2001). It helps make sure that moving toward continuous monitoring does not break the core privacy protections set out in European law.

4.7.2. Technology vs. Privacy

A key finding is that privacy has emerged as the predominant theme due to the increasing power of technology. Participant B stated that without these barriers, technology would be limitless:

> *"Regarding data privacy, it is always a theme. We are talking about all this speed and use of data, then privacy I think is the number one theme, because technology will... will be able to do everything. And that is why it is privacy really."* (Participant B)

However, he acknowledged that this dynamic could change depending on the global context yet emphasized that "having ethics always present... is fundamental" regardless of the pace of change.

The TAM-DEF framework proposed by Misra (2021) is straightly corroborated by Participant B's assertion that technology is "limitless", but must be constrained by ethics
Specifically, it highlights the "Ethics" and "Misuse" pillars of the framework, emphasizing that deployment should be assessed not only for capability but also for moral suitability. This finding confirms that without the "Ethics" component of TAM-DEF, rapid data use may exploit vulnerable frontline workers.

4.7.3. The Mandate for Consent

Triangulating the interview data with industry standards, the documented evidence reinforce that these barriers are not just legal suggestions but mandatory ethical requirements. According to the MSPA (2018), agencies are required to obtain permission prior to conducting any recordings:

> *"Obtain release statements from clients if their employees are to be recorded (audio or video)."* (MSPA, 2018)

Furthermore, the International Chamber of Commerce & ESOMAR (2008) instructs transparency, requiring that *"researchers shall have a privacy policy which is readily accessible to respondents"* and make sure that *"respondents are aware of the purpose of the collection."*

The analysis finds that privacy is the main theme shaping the industry's limits. Although AI technology "can do everything" (Participant B), the GDPR (Participant B) and the rule against "record[ing] visits" (Participant A) help make sure that efficiency does not come before the need for consent and transparency.

The connection between industry codes (MSPA,2018 and ICC & ESOMAR,2008) and the primary data highlights the significance of “Transparency”, which is the first pillar of the TAM-DEF framework.

When the industry requires respondents to be "aware of the purpose," it aims to remove the opacity of the "Black Box" that Rai (2020) criticized. Furthermore, obtaining "release statements" puts into practice Wilson's (2001) idea of "employee acceptance," ensuring that the evaluation process is seen as fair rather than secretive.

4.8. The Future Model: Hybrid Approach?

The peak of the analysis indicates that the industry is not heading towards total automation, but rather towards a Hybrid Model. The findings suggest a clear division of labour where Artificial Intelligence manages data collection, while humans keep control over analysis, validation, and empathy.

4.8.1. The Separation of Collection and Analysis

Participant B outlined a structural framework for the future, emphasizing a clear separation between data collection and its interpretation. He asserted that the labour-intensive aspects of data entry will no longer require human involvement:

> *"In market studies, there is a complete distinction here between collection and analysis. We are talking about collecting, and that will be possible increasingly automatically... collecting, collecting."* (Participant B)

He confirmed that the collection phase will be much more automated, and report drafting will be supported by software. However, he emphasized that influencing behaviour remains a human responsibility:

> *"The analysing, the doing afterwards, is also necessary... it requires more of the human element, because we are talking about training people, transforming people, and this involves more, let's say, a bidirectional conversation."* (Participant B)

This clear division of labour closely aligns with the theoretical framework established by Huang and Rust (2018). Participant B’s difference between "collecting" (automation) and "transforming people" (human) confirms the authors' classification of intelligence. The results indicate that while AI has achieved expertise in "Mechanical" and "Analytical" intelligence, mainly in data collection, the industry still depends exclusively on human workforce for "Intuitive" and "Empathetic" intelligence tasks, including coaching and behavioural change, which remain beyond the current capabilities of machines.

4.8.2. The "Second Eye": Augmented Intelligence and the human as tool of final control

Both participants addressed AI as a support mechanism rather than a replacement for decision-making. Participant A explained a collaborative workflow in which people either assess machines or work with them.

> *"Even humans with machines, even if it is me evaluating machines, I think there will be space for evaluation in the sense that the human will have to intervene and will have to create a report. This report can be helped by artificial intelligence."* (Participant A)

Participant B confirmed this concept of "Augmented Intelligence". Beyond admitting that decision-making will have *"increasing support"*, he designed a boundary for quality control. The human expert serves as the final gatekeeper to verify accuracy before data is shared with the client.

> *"Even in that final part of the analysis and decision making… we still want to have the second eye and validate before it goes out."* (Participant B)

The idea of the human role as a "second eye" or gatekeeper aligns with Davenport et al. (2020), who see Artificial Intelligence as a tool for "Task Automation" that still needs human oversight. This "validation" procedure directly validates the "Algorithmic Bias" concern pointed out by Dwivedi et al. (2021). When a person is moved to "verify accuracy," agencies help reduce the risk of unclear automated decisions, known as the "Black Box" problem (Rai, 2020). This process helps make sure the final report keeps the credibility clients expect.

4.8.3. The "Human Premium" and Virtual Frontiers

A paradoxical finding emerged from the analysis: as technology advances, the distinct value of human qualities increases. Participant B summarized this trend, observing that in an automated world, empathy becomes a luxury good:

"Measuring the customer experience. And it is increasingly a human component, a component of empathy... The more technology, the more the human factor counts." (Participant B)

Finally, the documentary evidence suggests a completely new hybrid frontier: the Metaverse. According to Narsaria (2024), the hybrid model will expand from physical to virtual stores, allowing shoppers to explore digital environments.

"Mystery shopping firms will need to develop virtual mystery shopping methodologies. This will involve sending mystery shoppers into the metaverse to interact with virtual sales staff, evaluate virtual environments, and assess the functionality of digital customer support." (Narsaria, 2018)

The findings indicate that the future of Mystery Shopping will be characterized by a symbiotic model. In this system, artificial intelligence ensures the efficiency and scale of data collection, while human evaluators contribute validation and the empathy necessary for staff training and adaptation to emerging virtual environments such as the Metaverse.
The paradox that "technology makes the human factor count more" strongly supports the findings of Chuah and Yu (2021). As connection with artificial intelligence become increasingly common, the risk of the "Uncanny Valley" (discomfort caused by artificial emotion) also rises. Therefore, authentic human empathy arises as a key differentiator, aligning with Harvey's (1998) conceptualization of high "Perceptual Process Quality."

Although the current academic literature cited in this study (e.g., Block et al., Wilson) does not specifically address virtual reality auditing, the present findings are a significant empirical extension. These results show that the "Strategic Driver" role of Mystery Shopping, as identified by Block et al. (2022), is expanding beyond physical retail into digital ecosystems. This evolution needs an updated hybrid model in which human judgment is applied to the assessment of virtual environments.

4.9. Chapter Summary

This long chapter of Findings and Discussion approached the implications of Artificial Intelligence on the Mystery Shopping Industry, triangulating interviews with experts with certified documentary evidence and literature.

The analysis demonstrates a profound paradigm shift, occurring in three different dimensions: Firstly, there has been a notable strategic shift in the role of the method, which has moved from being a control and diagnostic tool to an important financial driver. In agreement with Block et al. (2022), the results show that mystery shopping service providers now prioritize ROI and speed, using AI as a real-time decision engine rather than a static compliance check. Empirical data shows that customers demand “price and speed efficiencies” (Participant B), validating the transition to a high-frequency-oriented model.

Second, we have witnessed an operational change, where AI has successfully taken on the role of “Data Controller,” as defined by Davenport et al. (2020), automating “Technical Quality” tasks such as logistics, logical checks, and validation.

However, a critical “Empathy Gap” remains. The data validates Huang and Rust's (2018) theory that while “Mechanical Intelligence” is easily automated, “Intuitive Intelligence” requires a human presence to maintain credibility and read emotional nuances. Finally, there has been an ethical shift, grounded in the Opacity Barrier. The transition carries significant risks, from the “Panoptic” effect of continuous surveillance (Wilson, 2001) to the “fog of algorithmic opacity” described by Lebovitz et al. (2022). The results indicate that privacy laws (GDPR) and the need for “Ethical Audits” (TAM-DEF) currently act as the main friction against total automation, prioritizing consent over technological capability.

In short, the research is not based on a “replacement” hypothesis, but rather a Hybrid Model. The future of Mystery Shopping lies in Augmented Intelligence, where blockchain and algorithms ensure the verification of truth (Technical Quality), while humans focus on interpreting meaning (Perceptual Quality).

CHAPTER 5- CONCLUSIONS AND RECOMMENDATIONS

5.1. Main Conclusions

This dissertation aimed to analyse the impact of Artificial Intelligence on the Mystery Shopping industry.

Based on the results presented in Chapter 4, it can be concluded that AI is structurally transforming the sector, both operationally and strategically. The data show AI's transition to a central role as a "Data Controller," taking over the automation of tasks associated with Technical Process Quality, such as logistics routing and consistency checking. This result corroborates the theoretical premise that so-called "Mechanical Intelligence," based on deterministic rules and processes, is the first domain to be fully automated, replacing manual verification with automatic validation systems.

The results also demonstrate a significant temporal change in service delivery, marked by the transition from one-off and retrospective assessments to real-time monitoring models. This evolution validates the conceptual change proposed by Wilson (2001), from intermittent observation to continuous monitoring, and complies with the strategic requirement for speed identified by Block et al. (2022). Analysis of more advanced markets shows that technologies such as live streaming are transforming Mystery Shopping into an immediate management tool, whose value increasingly lies in its ability to provide actionable information in a timely manner.

Contradicting the hypothesis of a possible total replacement of the human factor, the research predicts the consolidation of a hybrid model. Although AI is crucial for data efficiency and consistency, human intervention emerges as an essential factor in interpreting behaviours and evaluating empathic and contextual dimensions. This result confirms the existence of an "Empathy Gap" in current technology and supports the sector's evolution towards an Augmented Intelligence paradigm, in which algorithms ensure data reliability, while human evaluators contribute analytical depth and credibility.

Finally, it is concluded that privacy and ethical compliance issues are the main obstacle to the full automation of Mystery Shopping. Despite the existence of technology that allows high levels of surveillance, its application is conditioned by legal and ethical frameworks, by the GDPR. These results are in line with the TAM-DEF model, demonstrating that factors such as ethics and transparency, rather than technical limitations, are decisive in the adoption and extension of Artificial Intelligence in field observation.

5.2. Theoretical Contributions

This research provides empirical evidence in three main areas, in addition to contributing to academic literature. First, the results achieved allow us to deepen the Strategic Driver theory,

empirically confirming the view of Block et al. (2022), who stated the transition from a diagnostic tool to a true driver of financial value.
The main contribution of this work lies in identifying AI as the mechanism that provides and facilitates this transition by converting static reports into real-time decision support systems.

Second, the study enables the applicability of the theoretical framework of Huang and Rust (2018) to the service sector. The operational distinction is well defined, with AI taking on logistical and procedural functions related to mechanical intelligence, while the human workforce is responsible for coaching and empathetic assessment activities. This difference in roles presents itself as a practical validation of the theory on the order of replacement of service tasks.
Finally, this study helps to rethink the problem of the "black box" in the adoption of AI-based technologies. Previous research has pointed to algorithmic opacity and bias as separate limitations (Rai, 2020; Ladhari, 2009). The results show that technologies such as blockchain can offer a joint solution to these two challenges. The use of immutable and auditable records can reduce the opacity of automated systems and meet the need for transparency highlighted by the TAM-DEF model.

5.3. Managerial Implications

This study provides strategic recommendations to help industry professionals and agency managers adapt to the new technological landscape of Mystery Shopping. With Artificial Intelligence making data collection and validation more accessible, the competitive advantage of agencies now depends on the quality of analysis performed by people. Therefore, it is important to invest in the training of evaluators, developing empathic and interpretive skills, while reducing the focus on mechanical tasks, which tend to be automated. The goal is to strengthen skills related to emotional reading, understanding context, and consulting.
In addition, the adoption of AI monitoring technologies should have a clear focus on development. To avoid resistance within organizations or the idea that there is punitive surveillance, it is important to present these tools as support for learning and continuous improvement, rather than as forms of disciplinary control. As Wilson (2001) points out, implementation will only be effective if employees perceive the technology as performance support, rather than constant panoptic surveillance.
Finally, the study highlights the importance of conducting regular ethical audits on automated systems, mainly because of the risks of algorithmic bias. To maintain trust in data and credibility with customers, agencies should adopt a human-in-the-loop approach, ensuring that final validation and oversight of automated processes remain the responsibility of people.

5.4. Limitations

This study has some limitations that should be considered when interpreting the results. First, although the number of qualitative interviews conducted allowed for an in-depth understanding of the phenomenon analysed, the sample size does not allow for statistical generalization of the conclusions to the entire Mystery Shopping industry. Second, the analysis is strongly embedded in the European regulatory context, especially within the scope of the General Data Protection Regulation (GDPR), which may limit the extrapolation of results to less regulated markets, such as the United States or certain regions of Asia, where the pace of adoption of Artificial Intelligence may be different. Finally, as this is a cross-sectional study, the results reflect only a specific moment in time, capturing a "snapshot" of a rapidly evolving technology. Recent and future advances in Generative Artificial Intelligence are likely to influence or even reshape some of the conclusions presented.

5.5. Suggestions for Future Research

Based on the results obtained, several relevant directions for future research can be outlined. One line of research refers to the adaptation of Mystery Shopping methodologies to virtual and immersive environments, especially in the context of the metaverse, following the evolution of digital spaces for interaction between brands and consumers. In addition, empirical studies on the application of blockchain in service validation can help mitigate the method bias associated with qualitative reports, promoting greater transparency and data reliability. Finally, it is important to develop longitudinal studies that examine the long-term psychological effects of AI-based real-time monitoring on frontline employees, deepening the understanding of the social challenges related to these technologies, as highlighted by Dwivedi et al. (2021).

REFERENCES

Algi, A., & Lukman, E. (2018). Analysis of Mystery Shoppers Program as a Benchmark of Service Culture at Shopping Center "X". https://doi.org/10.2991/BCM-17.2018.10

Bell, E., Bryman, A., & Harley, B. (2022). *Business research methods* (6th ed.). Oxford University Press. https://doi.org/10.1093/hebz/9780198869443.001.0001

Blessing, G., & Natter, M. (2019). Do mystery shoppers really predict customer satisfaction and sales performance? *Journal of Retailing, 95*(3), 47–62. https://doi.org/10.1016/j.jretai.2019.04.001

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, 9(2), 27-40. https://doi.org/10.3316/QRJ0902027

Boxberg Karlsson, V., & Thomasdotter Schölin, L. (2013). *The employee's guide to mystery shopping: A tool for developing the whole company* (2nd ed.). Better Business Books.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101 https://doi.org/10.1191/1478088706qp063oa

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

Chuah, S. H.-W., & Yu, J. (2021). The future of service: The power of emotion in human-robot interaction. *Journal of Retailing and Consumer Services, 61*, 102551. https://doi.org/10.1016/j.jretconser.2021.102551

Cronin, J. J., & Taylor, S. A. (1992). Measuring Service Quality: A Reexamination and Extension. *Journal of Marketing*, *56*(3), 55–68. https://doi.org/10.2307/1252296

Crouch, M., & McKenzie, H. (2006). The logic of small samples in interview-based qualitative research. *Social Science Information*, *45*(4), 483-499.https://doi.org/10.1177/0539018406069584

Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, *48*(1), 24–42. https://doi.org/10.1007/s11747-019-00696-0

Denzin, N. K. (1978). *Sociological methods: A sourcebook* (2nd ed.). McGraw-Hill.

Douglas, J. (2015). Mystery shoppers: An evaluation of their use in monitoring performance. *The TQM Journal*, *27*(6), 705–715. https://doi.org/10.1108/TQM-04-2015-0052

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilangakoon, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., ... Williams, M. D. (2021). Artificial intelligence in business: State of the art and future research agenda. *International Journal of Information Management*, *57*, Article 102190. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review*, *14*(4), 532–550. https://doi.org/10.2307/258557

Finn, A. (2001). Mystery Shopper Benchmarking of Durable-Goods Chains and Stores. *Journal of Service Research*, *3*(4), 310-320. https://doi.org/10.1177/109467050134004

Finn, A., & Kayandé, U. (1999). *Unmasking a phantom: A psychometric assessment of mystery shopping*. Journal of Retailing, 75(2), 195–217. https://doi.org/10.1016/S0022-4359(99)00004-4

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry*, *12*(2), 219–245. https://doi.org/10.1177/1077800405284363

Gębarowski, M., & Siemieniako, D. (2015). Mystery Visitor as Research Method of Trade Show Performance. Economics and Management, 7(3). https://ssrn.com/abstract=2715135

Gremler, D. D., Van Vaerenbergh, Y., Brüggen, E. C., & Gwinner, K. P. (2020). Understanding and managing customer relational benefits in services: A meta-analysis. *Journal of the Academy of Marketing Science*, *48*(3), 565–583. https://doi.org/10.1007/s11747-019-00701-6

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods*, *18*(1), 59–82. https://doi.org/10.1177/1525822X05279903

Harvey, J. (1998). Service quality: A tutorial. *Journal of Operations Management*, *16*(5), 583–597. https://doi.org/10.1016/S0272-6963(97)00026-0

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

Heinonen, J., & Murto, M. (2024). Enhancing event quality through mystery shopping: Integrating theoretical and practical perspectives. In *Proceedings of the 23rd European Conference on Research Methods in Business and Management* (pp. 138–145). Academic Conferences International Ltd. https://doi.org/10.34190/ictr.8.1.3562

International Chamber of Commerce, & ESOMAR. (2008). *ICC/ESOMAR international code on market and social research*. https://www.apodemo.pt/wp-content/uploads/2017/03/ICCESOMAR_Code_English_.pdf

Jacob, S., Schiffino, N., & Biard, B. (2018). The mystery shopper: A tool to measure public service delivery? *International Review of Administrative Sciences*, *84*(1), 164–184. https://doi.org/10.1177/0020852315618018

Jarrahi, M. H. (2018). *Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making.* Business Horizons, 61(4), 577–586. *https://doi.org/10.1016/j.bushor.2018.03.007*

Kuckartz, U., & Rädiker, S. (2019). *Analyzing qualitative data with MAXQDA*. Springer.

Ladhari, R. (2009). A review of twenty years of SERVQUAL research. *International Journal of Quality and Service Sciences*, *1*(2), 172–198. https://doi.org/10.1108/17566690910971445

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science*, *33*(1), 126–148. https://doi.org/10.1287/orsc.2021.1549

Lemon, K. N., & Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey. *Journal of Marketing, 80*(6), 69–96. https://doi.org/10.1509/jm.15.0420

Liu, B. (2012). The problem of sentiment analysis. In *Sentiment analysis and opinion mining*. Springer. https://doi.org/10.1007/978-3-031-02145-9_2

Lobo, G. A., & Shetty, D. (2025). *Transforming mystery shopping with artificial intelligence: A paradigm shift in real-time customer experience evaluation and strategic insights*. *International Journal of Innovation Studies, 9*(1), 179–197.

Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). Sample size in qualitative interview studies: Guided by information power. *Qualitative Health Research*, *26*(13), 1753–1760. https://doi.org/10.1177/1049732315617444

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). SAGE Publications.

Mystery Shopping Professionals Association. (2018). Common codes of professional standards and ethical conduct (Version 5). https://mspa-global.org/files/documents/MSPA%20Common%20Codes%20of%20Professional%20Standards%20and%20Ethical%20Conduct,%20Version%205,%20Adopted%2019%20September%202018.pdf

Mystery Shopping Professionals Association. (2023). *Guide to mystery shopping*. https://mspa-ea.org/files/documents/Guide%20to%20Mystery%20Shopping/MSPA%20Guide%20to%20Mystery%20Shopping%20(010323).pdf

Narsaria, S. (2024, November). *The future of consumer experience: Evolving trends, the impact on mystery shopping companies and how we all need to evolve*. Mystery Shopping Professionals Association Europe/Africa. https://mspa-ea.org/en_GB/news/newsitem/192-the-future-of-consumer-experience-evolving-trends-the-impact-on-mystery-shopping-companies-and-how-we-all-need-to-evolve.html

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1985). A conceptual model of service quality and its implications for future research. *Journal of Marketing, 49*(4), 41–50. https://doi.org/10.1177/002224298504900403

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing, 64*(1), 12–40.

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

Saunders, M. N. K., Lewis, P., & Thornhill, A. (2023). *Research methods for business students* (9th ed.). Pearson.

Stake, R. E. (1995). *The art of case study research*. SAGE Publications.

Suárez-Barraza, M. F., & Huerta-Carvajal, M. I. (2024). Kaizen-mindfulness a twin continuous improvement approach at workplace: A qualitative exploratory study. *The TQM Journal*, *36*(6), 1591–1626. https://doi.org/10.1108/TQM-07-2023-0226

Van der Wiele, T., Hesselink, M., & Van Iwaarden, J. (2005). *Mystery shopping: A tool to develop insight into customer service provision.* Total Quality Management & Business Excellence, 16(4), 529–541. *https://doi.org/10.1080/14783360500078433*

Venkatraman, D. P., & Kurtkoti, M. (2024). Artificial intelligence in the service industry: Transforming operations and enhancing customer experience. *Nanotechnology Perceptions, 20*(S16), 198–201. https://doi.org/10.62441/nano-ntp.vi.3674

VERBI Software. (2023). *MAXQDA 24* [Computer software]. VERBI Software.

Wilson, A. M. (2001). Mystery shopping: Using deception to measure service performance. *Psychology & Marketing*, *18*(7), 721–734. https://doi.org/10.1002/mar.1027

Xu, L., & He, S. (2014). Analysis on the survey method of mystery shopping in hospitality management. In *E-commerce, e-business and e-service*. CRC Press.

Yin, R. K. (2018). Case Study Research and Applications: Design and Methods (6th ed.). Thousand Oaks, CA: Sage Publications

Zikmund, W. G., Babin, B. J., Carr, J. C., & Griffin, M. (2009). *Business research methods (8th ed.). South-Western College.*

APPENDICES

Appendix 1: Interview Protocol

Introduction:

1. Formal welcome
2. Interviewer introduction
3. Participant introduction
4. Presentation of the Research Question

Questions:

Part A (based on Block et al. (2022); Wilson (2001))

1. How does the Mystery Shopping practice create value for your company?
2. In what circumstances/context is the practice most needed today?
3. In which areas or scenarios do you feel limitations/flaws in the method?

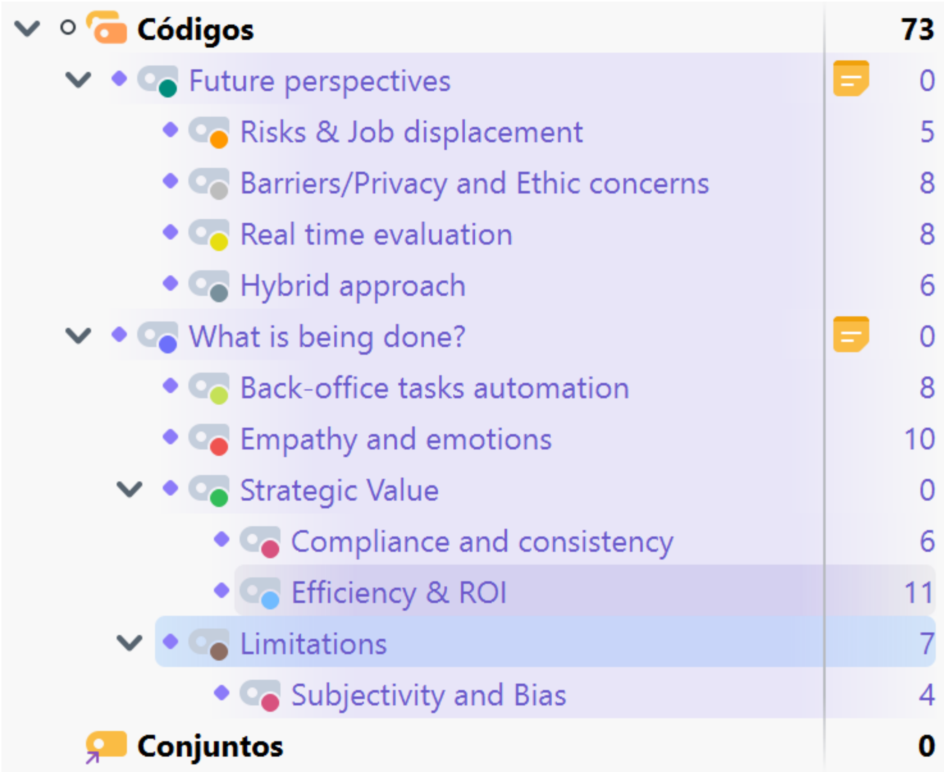
Part B (based on Lobo & Shetty (2025))

1. How do you envision the role of embedded AI in the Mystery Shopping process?
2. What specific tasks/stages of the process do you consider to be the target of this change?
3. What new results/metrics/outputs could come from the use of AI?

Part C (based on Blessing & Natter (2019); van der Wiele et al.(2005))

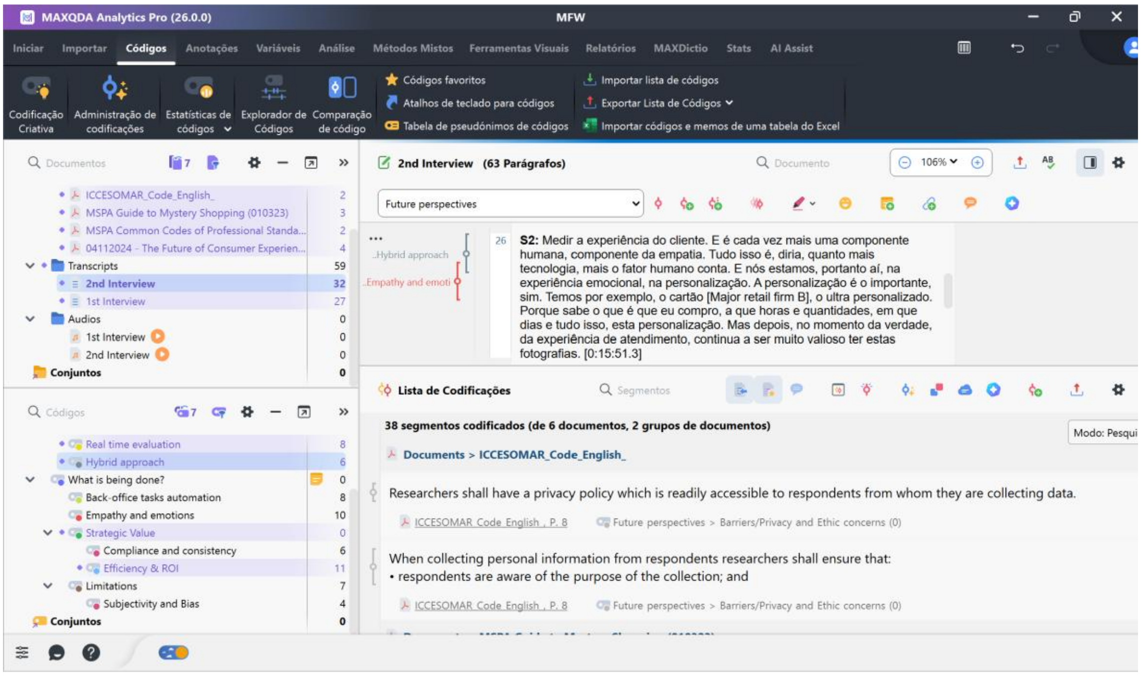
1. How can AI improve the usefulness of this method for predicting satisfaction and performance?
2. What risks/trade-offs do you specifically identify in introducing AI into Mystery Shopping practice (subjectivity, ethics, trust, validity)?
3. Compared to what is already done (without AI), where do you see the greatest risk? And where do you feel more secure?
4. If you had to propose a "hybrid model" (human + AI), what would it look like?

Appendix 2: MAX QDA codes list



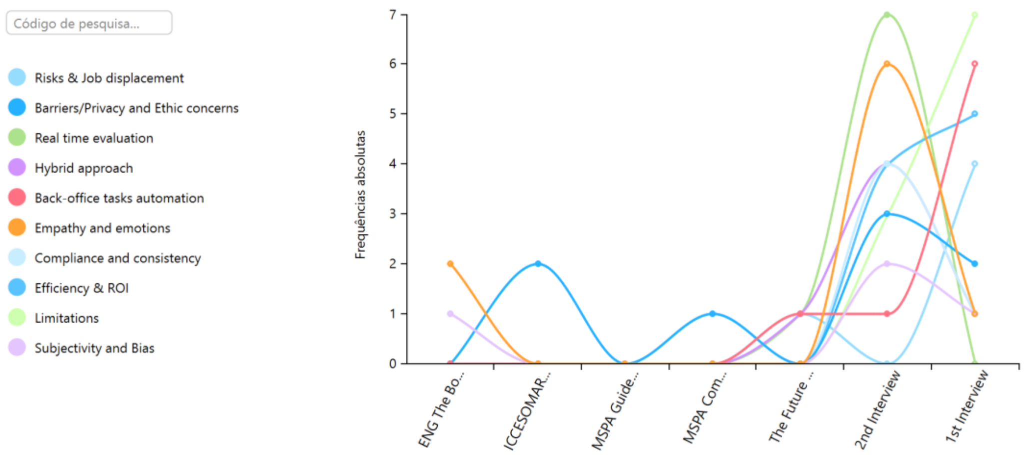
Source**:** Author's elaboration using MAXQDA 24 software

Appendix 3: Software Desktop



Source**:** Author's elaboration using MAXQDA 24 software

Appendix 4: Word trends per document



Source**:** Author's elaboration using MAXQDA 24 software