STOCKHOLM SCHOOL OF ECONOMICS
1351 Master's thesis in Business Management
Academic year 2025-2026

Performance evaluation with algorithms in the loop: How managers evaluate subordinates in AI-supported decision-making contexts

Vinh Thi Le (42819)

Supervisor: Lily Gao
Date examined: May 28th, 2026
Discussant(s): Tianjun Feng
Examiners: Johan Berglund & Viktor Skyrman

Abstract

Organizations are increasingly adopting AI for decision support, in which algorithms provide recommendations but humans remain responsible for final decisions. While prior research has examined human-AI decision structures, responses to algorithmic advice, and responsibility attribution, little is known about how human decision-makers are evaluated in these settings. To help fill this gap, this thesis aims to address: *How do managers evaluate subordinates' performance in AI-supported decision-making contexts?* Drawing on fourteen semi-structured interviews with managers from diverse industries and hierarchical levels, the study examines both the performance dimensions managers emphasize and how they combine these dimensions into overall judgments. The findings show that, while outcomes remain central, AI-specific procedures and AI-related capabilities are becoming increasingly salient in performance evaluation. Furthermore, managers tend to adopt an average-like logic to integrate multiple dimensions to form the final judgment. From a theoretical perspective, the thesis extends the management control and outcome bias literatures by showing how AI reshapes the content and weighting of established control types and how outcomes influence not only evaluative judgments but also the criteria considered during evaluation. From a practical standpoint, the study highlights the importance of designing performance evaluation systems that explicitly integrate results, action, personnel, and cultural controls to support effective human-AI collaboration.

Keywords: artificial intelligence, AI-supported decision-making, human-AI collaboration, performance evaluation, management control systems.

Acknowledgement

I would like to express my gratitude to my supervisor, Lily Gao, for her guidance and support throughout the development of this thesis. Her constructive feedback has been instrumental in strengthening the analytical rigor of this study.

I am also grateful to all interviewees who generously dedicated their time and shared their perspectives. Their thoughtful answers to interview questions provided the empirical foundation for this study and were essential to its completion.

Furthermore, I would like to thank my peers at the Stockholm School of Economics for their constructive discussions and intellectual exchange. Their feedback contributed to refining my ideas and improving the overall quality of this thesis.

Table of Contents

Definitions

Artificial intelligence (AI)	An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments (OECD, 2024). In this study, the terms "AI" and "algorithm" are used interchangeably.
Human-AI collaboration	Human-AI collaboration refers to the arrangement in which humans and AI, through some form of collaboration, together produce the final decision (Aman et al., 2025; Puranam, 2021).
AI-supported decision-making context	An AI-supported decision-making context is an organizational setting in which AI systems provide predictions or recommendations as decision support, while human actors retain formal authority and responsibility for the final decision; this setting is also referred to as "algorithms in the loop" (Green & Chen, 2019; Zeiser, 2024).
Management control systems	Management control systems are all the devices or systems managers use to ensure that the behaviors and decisions of their employees are consistent with the organization's objectives and strategies (Merchant & Van der Stede, 2003).

1. Introduction

Artificial intelligence (AI) is becoming increasingly embedded in organizational practices. A global survey by McKinsey & Company reports that in 2025, 88 percent of surveyed companies had integrated AI into at least one business function, up from 55 percent in 2023, while 79 percent were already using generative AI tools, indicating that algorithmic systems now shape a large share of everyday business decisions (McKinsey & Company, 2025). At the same time, Accenture's Pulse of Change survey found that 86 percent of C-suite leaders planned to increase AI investment in 2026, suggesting that organizations are continuing to expand the integration of AI into business practices (Accenture, 2026).

The increasing adoption of AI in organizations has attracted corresponding scholarly attention. A systematic review by Aman et al. (2025) documents the rapid expansion of research on human-AI collaboration in recent years, showing that the number of publications addressing this phenomenon has grown steadily over time, with a particularly sharp rise after 2022. This growing interest reflects the prevalence of organizational settings in which humans and AI jointly contribute to decision-making processes. Puranam (2021) conceptualizes such arrangements as Human-AI Collaborative Decision-Making (HACD), defined as situations in which *"humans and AI, through some forms of collaboration, together produce a decision that is implemented by a third party."*

In this emerging research stream, AI has been conceptualized not merely as a technology but as a participant in organizational decision-making structures. Shrestha et al. (2019) argue that rapid advances in AI development are gradually establishing algorithmic decision-makers as key organizational actors whose capabilities differ from those of human decision-makers. While AI excels at processing large datasets and generating rapid, replicable decision outcomes, humans retain advantages in handling ill-defined problems and providing interpretable decisions. These complementary capabilities create opportunities for organizations to combine humans and AI in ways that can significantly improve decision quality. A central challenge for organizations, however, lies in determining how to structure such collaboration, as decision quality depends on how decision authority and tasks are allocated between humans and AI (Shrestha et al., 2019). Puranam (2021) echoes this point by framing human-AI collaboration as an *"organizational design problem,"* arguing that organizations must deliberately determine how decision tasks and authority are distributed between human and algorithmic actors.

One prominent design choice concerns the allocation of final decision authority. Even when AI systems provide highly accurate recommendations, many organizations maintain human oversight and responsibility for the final decision. This arrangement allows organizations to benefit from algorithmic analytical capabilities while preserving human discretion and accountability. As Zeiser (2024) notes, most AI systems today operate in this manner, functioning as decision-support systems or *"algorithms in the loop,"* as Green and Chen (2019) put it. However, this structure raises new questions about how human decision-makers should appropriately weigh and use AI recommendations, how responsibility and accountability should be allocated between humans and AI, and how human decision-makers are evaluated when decisions are made with algorithmic support.

Prior research on AI as a decision-support system has focused on the design of effective human-AI decision structures (Puranam, 2021; Shrestha et al., 2019), human's responses to algorithmic advice, including algorithm aversion, appreciation, and reliance (Daschner & Obermaier, 2022; Gill et al., 2024; Kim et al., 2024; Logg et al., 2019; Ning et al., 2024), and concerns of responsibility and accountability (Elish, 2019; Joo, 2024; Passlack et al., 2024; Tsumura & Yamada, 2025; Zeiser, 2024). However, the question of how human decision-makers are evaluated in such contexts remains underexplored.

Against this background, this thesis aims to explore how managers evaluate subordinates' performance when AI is incorporated into decision-making processes. Accordingly, the following research question is formulated:

> **How do managers evaluate subordinates' performance in AI-supported decision-making contexts?**

To address this question, two sub-questions are developed:

1. What performance dimensions do managers emphasize when evaluating subordinates in AI-supported decision-making contexts?
2. How do managers combine different dimensions to form a final judgment of their subordinates' performance?

The rest of this thesis is structured as follows. Chapter 2 presents the theoretical foundations of the study, including literature review and theoretical framework developed to structure the analysis.

Chapter 3 outlines the methodology, including the research approach, data collection methods, data analysis process, and considerations of validity. Chapter 4 presents the empirical findings, followed by a discussion of these findings in Chapter 5. Finally, concluding remarks are provided in Chapter 6.

2. Theory

In this chapter, the theoretical foundations of the study are presented. In section 2.1, a review of the relevant literature is conducted, focusing on AI as decision support, performance evaluation in organizations in general, and responsibility attribution in AI-supported decision-making contexts. In section 2.2, the theoretical framework is developed, integrating management control literature with Information Integration Theory (IIT) to guide the empirical analysis.

2.1. Literature review

2.1.1. AI as decision support in organizations

Shrestha et al. (2019) argue that humans and AI possess distinct decision-making capabilities. They compare the two across five dimensions: the specificity of the decision search space, the interpretability of the process and outcome, the size of the alternative set, decision-making speed, and replicability. According to the authors, humans perform better in ill-defined problems that require contextual understanding and can explain the reasoning behind their decisions, but they tend to be slower and less consistent. AI systems, by contrast, excel in situations involving large alternative sets, offering speed and replicability, yet often lack interpretability and struggle in novel or ambiguous contexts. Based on these differences, Shrestha et al. (2019) argue that organizations should structure decision processes in ways that leverage the complementary strengths of humans and AI as the technology becomes increasingly embedded in organizational decision-making.

To address this design problem, Shrestha et al. (2019) propose three organizational settings through which humans and AI can be integrated: full delegation of decision-making authority to AI systems, hybrid sequential structures in which human and algorithmic decisions occur in sequence, and aggregated structures in which independent human and algorithmic judgments are combined. Selecting the most appropriate structure, according to the authors, requires organizations to consider the relative strengths and limitations of humans and AI across different dimensions of decision-making, as well as their relevance to the task at hand. Complementing this perspective,

Puranam (2021) suggests that the division of labor in human-AI collaboration can be characterized along two dimensions: the nature of interdependence – whether human and algorithmic decisions are produced sequentially or in parallel, and the nature of specialization – whether humans and algorithms perform different or identical decision tasks. Based on these dimensions, Puranam proposes a taxonomy of human-AI collaboration settings. In one configuration, humans and AI perform different tasks in parallel, with the final decision integrating both contributions. In another, humans and AI independently perform the same decision task in parallel, after which their outputs are aggregated. Sequential configurations are also possible, where either the algorithm's output informs human decisions or human decisions feed into the algorithmic process. This taxonomy aligns closely with the structures proposed by Shrestha et al. (2019).

Among the possible structures through which humans and AI can collaborate, a common approach is to use AI as a decision-support system, in which AI systems generate outputs that are subsequently interpreted and acted upon by human decision-makers. Shrestha et al. (2019) conceptualize this configuration as *"AI-to-human sequential decision-making,"* which is particularly common in contexts where firms seek to benefit from algorithmic analytical capabilities while maintaining human oversight and accountability. In such arrangements, humans are technically the final decision-makers (Zeiser, 2024). Zeiser further distinguishes between different ways AI systems can guide human action: they may either provide direct recommendations for action (prescriptive systems) or provide information on the basis of which decisions are made (descriptive systems). Both forms can play an important role in shaping organizational decisions (Zeiser, 2024).

AI as decision support, while creating opportunities for organizations to leverage algorithmic capabilities, also changes the nature of decision-making itself. Zeiser (2024) argues that, in such settings, final decisions reflect not only the reasoning of human decision-makers but also the logic embedded in the AI systems. This creates ambiguity with important implications for performance evaluation: if outcomes are shaped not only by human judgment but also by algorithmic influence, questions arise about how human decision-makers should be evaluated in these contexts.

2.1.2. Performance evaluation in organizations

To situate performance evaluation within the organization's management control system, Otley's (1999) performance management framework is useful. Otley conceptualizes control around five central questions concerning objectives, strategies and plans, performance targets, rewards, and information flows. Among these, the third and fourth questions – what level of performance is required to realize the organization's strategies, and what rewards or penalties follow from achieving or failing to reach those targets – position performance evaluation at the core of the system. In this sense, performance evaluation serves as the mechanism through which desired performance levels are articulated and linked to rewards and incentives, thereby aligning employee behavior with the organization's objectives.

Recent developments in performance management suggest that evaluation practices are evolving. Cappelli and Tavis (2016) document a shift away from long-standing annual performance reviews toward more frequent and informal processes. Specifically, they argue that annual reviews often fail to adapt to flatter structures, team-based work, and rapidly changing roles in contemporary organizations. Accordingly, organizations are adopting practices such as continuous check-ins tied to project cycles and short-term goals, which better support coaching and organizational agility (Cappelli & Tavis, 2016). In such settings, evaluations become more embedded in everyday managerial practice, requiring managers to continuously interpret employees' performance. This development underscores the importance of understanding how managers evaluate subordinates, particularly in emerging contexts where AI is incorporated into decision-making processes.

Research in management control systems suggests that performance evaluation can rely on multiple dimensions. A foundational framework proposed by Ouchi (1979) distinguishes between market, bureaucratic, and clan mechanisms of control. Markets address control problems through the precise measurement and rewarding of individual contributions; bureaucracies rely on formal rules and monitoring combined with a socialized acceptance of organizational objectives; and clans rely on shared values and processes that reduce goal incongruence between individuals (Ouchi, 1979). Building on this idea, Merchant and Van der Stede (2017) categorize management control systems into results controls, action controls, personnel controls, and cultural controls, which broadly correspond to outcome-based, process-based, and norm-based forms of control that Ouchi proposes. Goebel and Weißenberger (2017) examine this framework in contemporary

organizations, highlighting the growing relevance of informal control mechanisms (personnel and cultural controls) compared to formal control instruments (results and action controls). Taken together, these literatures suggest that managers possess multiple lenses through which they can assess their employees' performance.

Despite the availability of multiple dimensions through which performance can be evaluated, research in behavioral decision-making shows that such evaluations are often biased toward outcomes. A foundational work by Baron and Hershey (1988) demonstrates this outcome bias through a series of experiments in which observers are asked to evaluate other people's decisions. Their study shows that identical decision processes were rated as higher quality, more competent, and more worthy of future delegation when they resulted in favorable outcomes. Interestingly, observers acknowledged that outcome should not be used to judge the quality of a decision, yet they relied on it regardless, suggesting that outcome bias is both strong and intuitive (Baron & Hershey, 1988). Gillenkirch and Velthuis (2023) examine this phenomenon in a delegated risk-taking context in which a client evaluates an agent who makes risky decisions on their behalf. Across three experiments, they show that even when evaluators observe both the decision and all relevant information needed to assess its quality, they still allow the realized outcome to dominate their evaluations. Studying this phenomenon in organizational contexts, Ferguson (2025) finds that outcome bias is not uniform: it is stronger in situations where objective performance measures are incomplete, and evaluators must rely heavily on subjective judgment. These findings, taken together, suggest that outcome bias is likely to play a significant role in performance evaluation.

To summarize, performance evaluation plays a vital role in management control systems because it links performance targets with individual incentives and rewards, thereby guiding employee behavior towards the organization's objectives. Prior literature shows that, although multiple lenses for assessing performance are available, managers tend to be biased toward outcomes, creating a risk that results could overshadow qualitative dimensions such as judgment and the appropriateness of the decision process. This has important implications for AI-supported decision-making contexts in which, as Zeiser (2024) argues, outcomes are shaped not only by human judgment but also by the logic embedded in AI systems.

2.1.3. Responsibility attribution in AI-supported decision-making contexts

Crant and Bateman's (1993) study on the assignment of credit and blame for performance outcomes shows a positive relationship between responsibility attributions and reward allocation: higher assigned credit is associated with greater rewards, whereas higher assigned blame is associated with lower rewards. This suggests that responsibility attribution is closely connected to performance evaluation. In AI-supported decision-making contexts, however, the literature is imbalanced: while responsibility attribution has received considerable scholarly attention, performance evaluation remains relatively underexplored. This section therefore examines how responsibility in AI-supported decision-making contexts has been discussed and considers its implications for performance evaluation.

One central question regarding responsibility in human-AI collaboration is who should be held responsible for outcomes. Bleher and Braun (2022) examine AI-driven clinical decision-support systems and argue that, rather than simply creating a *"responsibility gap,"* the use of AI as decision support leads to *"responsibility diffusion"* across multiple actors. They distinguish three dimensions of this diffusion: causal (who contributed causally to the outcome), moral (who is blameworthy or praiseworthy), and legal responsibility (who can be held liable in formal terms). According to the authors, clinicians, AI developers, and the hospital all feature as potential bearers of responsibility along these dimensions, and participants struggle to separate who is responsible for what when AI is tightly integrated into decision processes. Their findings suggest that AI as a decision-support system reorganizes responsibility attribution in complex ways.

Zeiser (2024) extends this line of thought by arguing that AI decision-support *"poses a challenge to responsibility that goes beyond the familiar problem of finding someone to blame."* The author introduces the notion of *"attributability gap,"* suggesting that AI-supported decision-making creates ambiguity around decision ownership. Specifically, Zeiser argues that although humans remain formally responsible for the decision, outcomes may partly reflect the logic embedded in AI systems rather than solely the reasoning of human actors. As a result, it becomes unclear who truly owns the decision. For this thesis, this raises an important question: if employees' decisions are only partially attributable to them, how should their performance be evaluated?

Elish's (2019) concept of *"moral crumple zones"* suggests that responsibility can be misallocated in human-machine systems. Analyzing incidents in aviation, military, and automated driving

contexts, Elish shows that the nearest human operator often *"absorbs disproportionate blame when automated systems fail,"* even when control, information, and design decisions are distributed across a broader sociotechnical network. A subsequent study by Tsumura and Yamada (2025), however, suggests that *"providing prior knowledge of an agent can reverse this dynamic by shifting responsibility away from the human user and toward the system itself."* Their experiments also show that higher perceived importance of the decision shifts responsibility away from human users and toward developers and system providers. The authors therefore conclude that responsibility attribution in human-AI collaboration is dynamic and context-dependent, *"shaped by both cognitive framing (i.e., prior knowledge of the agent) and contextual framing (i.e., the perceived importance of the decision)"* (Tsumura & Yamada, 2025).

Recent empirical work adds further nuance to the topic by showing that managers are not objective when allocating responsibility in AI-supported decision-making contexts. Through several vignette experiments, Passlack et al. (2024) compare how managers assigned credit and blame to themselves and to their AI advisors. The study reveals a hedonic bias, where managers take more personal responsibility for positive outcomes than for negative ones. This asymmetry suggests that responsibility attribution is also shaped by psychological bias rather than reflecting a stable principle for how responsibility should be shared between humans and AI.

To summarize, scholars have made significant progress in exploring AI-supported decision-making contexts. Prior research has shown that incorporating AI into decision-making processes requires redesigning decision structures to leverage the complementary strengths of humans and AI. Additionally, the emergence of AI in organizational settings complicates how responsibility is attributed: responsibility becomes diffused across multiple actors, decision ownership becomes less clear, and attributions of responsibility depend on factors such as outcomes, AI knowledge, and the stakes of the decision. Despite these advances, an important gap remains in the literature. Although performance evaluation is increasingly embedded in managerial practice, as discussed in section 2.1.2, little is known about how managers evaluate subordinates in AI-supported decision-making contexts. This thesis, therefore, seeks to address this gap.

2.2. Theoretical framework

To explore how managers evaluate subordinates' performance in AI-supported decision-making contexts, this study adopts a theoretical framework that combines Merchant and Van der Stede's

(2003) typology of controls with Information Integration Theory proposed by Anderson (1981). While Merchant and Van der Stede's framework helps structure the performance dimensions managers draw upon, Anderson's theory helps explain how these dimensions are combined into an overall judgment. Together, these perspectives address both sub-questions systematically: identifying the performance dimensions managers focus on in their evaluation process (sub-question 1) and analyzing how they integrate these dimensions to form a final judgment (sub-question 2).

2.2.1. Management control alternatives

A first building block of the framework is Merchant and Van der Stede's (2003) typology of management control alternatives. Specifically, the authors distinguish four main types of controls: results controls, action controls, personnel controls, and cultural controls. Results controls influence behavior by specifying desired performance indicators, measuring actual results against those indicators, and rewarding or penalizing individuals based on their performance. Action controls, in contrast, involve ensuring that specific actions are taken (or not taken). Although commonly used in organizations, Merchant and Van der Stede argue that action controls are *"effective only when managers know what actions are desirable (or undesirable) and have the ability to ensure that the desirable actions occur (or that the undesirable actions do not occur)."* Personnel controls, on the other hand, aim to *"increase the likelihood that employees behave appropriately on their own by selecting the right people for a particular job, providing training, and ensuring that jobs are designed to give motivated and qualified employees a high probability of success,"* according to the authors. And finally, cultural controls operate through shared norms, beliefs, values, ways of behaving, and so on (Merchant & Van der Stede, 2003).

Although Merchant and Van der Stede (2003) adopt a broad view of management control, focusing on the guidance of behavior and the execution of strategy rather than solely performance measurement, their typology offers a useful framework for this thesis. Specifically, the four types of control can be understood as different lenses through which managers assess subordinates. Managers may emphasize results controls by focusing on observable outcomes and quantitative performance metrics. They may emphasize action controls by assessing whether subordinates follow prescribed decision processes, such as requirements related to consulting or documenting the use of AI systems. Personnel controls become visible when managers evaluate subordinates'

capabilities, such as competence, prudence, or professional judgment. Finally, cultural controls emerge when managers draw on broader assessments of value alignment, such as adherence to organizational norms regarding AI use.

The study also seeks to explore how the presence of AI influences the types of controls managers emphasize. Specifically, it examines whether AI affects which performance dimensions managers prioritize and how the content of each dimension is defined in AI-supported decision-making contexts. For example, within personnel controls, AI may foreground new skill requirements such as AI literacy and the ability to validate AI outputs, while within action controls, consulting AI may become a required step in the decision process.

2.2.2. Information Integration Theory (IIT)

The second building block of the theoretical framework is Information Integration Theory, developed by Anderson (1981). Information Integration Theory explains how individuals combine multiple pieces of information into a single overall judgment. Specifically, the theory proposes that when evaluating a person, the evaluator does not simply form a vague impression but instead performs an internal calculation. Each observed trait is assigned a value, and these values are then combined using *"cognitive algebra"* to form an overall impression (Anderson, 1981). According to Anderson (2016), this idea has been demonstrated across a wide range of psychological domains, including psychophysics, learning, social attitudes, and moral judgment.

Information Integration Theory identifies three core mathematical laws through which overall judgments are formed: the adding law, the averaging law, and the multiplying law (Anderson, 1981). The *adding law* assumes that each observed trait or cue about a person carries a positive or negative value, and the overall judgment is the sum of all those values. In performance evaluation, this means that a negative result combined with a lack of sound reasoning will lead to a more severe judgment than a negative result alone. The *averaging law*, by contrast, proposes that people form judgments by averaging multiple cues rather than summing them. Each cue may carry a different weight depending on its perceived importance, and the final judgment reflects an overall weighted average. In performance evaluation, for example, if a subordinate has already achieved excellent results, adherence to process may have a limited effect on the final judgment because it is averaged with other cues rather than added on top. Finally, the *multiplying law* applies to

situations in which information is not simply added but interacts, suggesting a multiplicative process rather than an additive one.

Although developed in the psychological realm, Information Integration Theory offers transferable insights for this thesis. Specifically, the theory is used to explore how managers combine different performance cues when forming final judgments of their subordinates in AI-supported decision-making contexts. The underlying assumption is that managers do not rely on a single lens when assessing subordinates but instead integrate multiple considerations into the overall judgment. Moreover, the study also explores how the presence of AI influences this information integration process. For example, AI may alter the subjective value managers assign to evaluation criteria such as outcomes, judgment, and process compliance.

Information Integration Theory of Anderson (1981) is integrated with Merchant and Van der Stede's (2003) typology of controls to create an analytical approach that captures both the dimensions of performance managers attend to and the cognitive logic through which those dimensions are combined into an overall evaluation. In the empirical analysis, this integration is operationalized as follows. First, managers' comments are categorized according to the types of control they focus on. Second, the analysis examines how managers combine those dimensions to arrive at a final judgment. By applying these lenses sequentially, the analysis aims to capture first what managers evaluate, and then how different performance cues are integrated into the assessment of subordinate performance in AI-supported decision-making contexts.

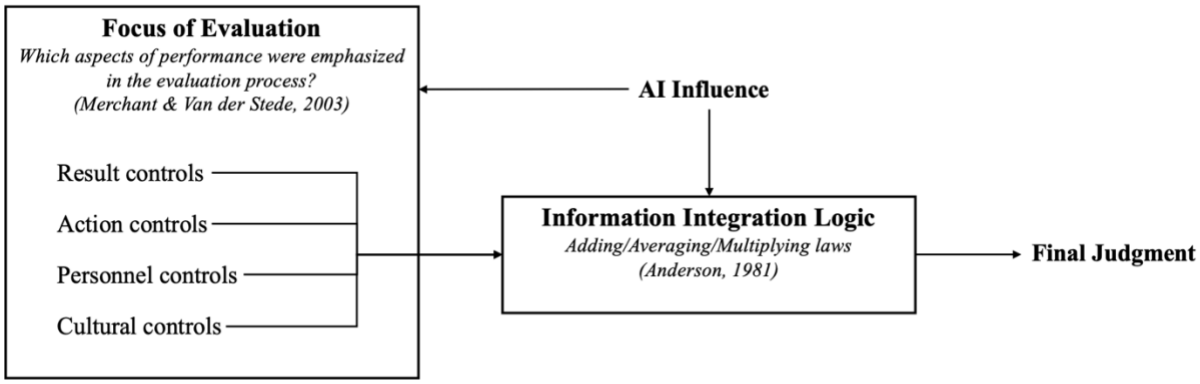

Figure 1: Visual representation of the theoretical framework

3. Methodology

In this chapter, the research methodology employed in the study is outlined. Section 3.1 presents the research strategy and design. Section 3.2 describes the data collection process, followed by section 3.3, which details the data analysis approach. Finally, section 3.4 discusses the validity of the study.

3.1. Research approach

3.1.1. Research strategy

The phenomenon investigated in this study – how managers evaluate subordinates in AI-supported decision-making contexts – is fundamentally interpretive. Specifically, the study aims to uncover not only final assessments of subordinate performance but also how managers arrive at those evaluations, thereby requiring access to participants' reasoning as they retrospectively interpret events. This perspective treats managers as knowledgeable agents who actively construct and articulate their organizational reality through language and interpretation (Gioia et al., 2013).

In addition, answering the research question requires understanding the process through which managers reconstruct subordinates' decision-making process, including when and how references to AI emerge in their narratives. This calls for methods capable of capturing the temporal sequencing and contextual nuance of events. Such process data often involves complex and "messy" patterns that are difficult to capture through variance-based quantitative models (Langley, 1999). Furthermore, the evaluation of human performance in AI-supported decision-making represents a relatively new and evolving phenomenon where theoretical understanding remains limited. In such contexts, research approaches that allow close engagement with empirical observations are particularly valuable for generating insights and developing theory (Pratt, 2017). Taken together, these considerations make a qualitative research strategy appropriate for addressing the research question of this study.

3.1.2. Research design

To operationalize the research strategy, semi-structured interviews were used as the data collection method. Addressing the research question requires both guidance and flexibility: key themes such as focus of evaluation, information integration logic, and the role of AI must be consistently discussed, while room for participants to introduce their own interpretations must be maintained.

Semi-structured interviews provide this balance by enabling a theory-driven pathway while leaving space for participants to surface novel perspectives (Galletta & Cross, 2013). Furthermore, semi-structured interviews allow questions to be adapted and unexpected responses to be probed during the interview, which is essential for capturing the richness, nuance, and complexity of participants' perspectives on the topic (Rubin & Rubin, 2011).

As part of the interview design, a vignette was used as a stimulus for discussion. Vignettes present participants with short, hypothetical scenarios and are particularly useful for eliciting reasoning about complex, emerging, or sensitive phenomena that may otherwise be difficult to articulate directly (Wiegmann et al., 2025). As AI-supported decision-making represents a relatively new and evolving context, the use of the vignette provides participants with a concrete situation to engage with. In addition, discussing performance in organizations can be sensitive, especially when it involves negative outcomes. By presenting the situation through a hypothetical character rather than the participant's own experience, vignettes create a *"protective layer"* or *"distance"* that encourages more open discussion (Bradbury-Jones et al., 2014). Finally, realistic vignette scenarios can function as powerful elicitation devices that stimulate detailed reflection and narrative explanation, generating richer, more nuanced responses that standard interview questions often fail to capture (Sampson & Johannessen, 2020). Together, these characteristics make vignette-based interviewing particularly suitable for this study.

The sampling approach was primarily purposeful, as interviewees were selected based on criteria that they were familiar with AI-supported decision-making contexts and held managerial positions, enabling them to provide *"information-rich cases"* relevant to the research question (Patton, 2015). In addition, the study incorporated a convenience sampling approach, as managers were recruited from roles and organizations where participation could realistically be obtained, a strategy commonly acknowledged in qualitative research (Patton, 2015). At the same time, the study sought variation across key contextual dimensions, including participants' industries and hierarchical levels, consistent with Patton's notion of maximum variation sampling.

3.2. Data collection

3.2.1. Interview sample

Data collection continued until theoretical saturation was reached, in the sense that consistent patterns were detected in the data and additional interviews did not appear to contribute substantively new information relevant to the analysis (Glaser & Strauss, 1967). Specifically, by Interviews 13 and 14, no significantly new themes were emerging; instead, participants largely confirmed patterns already identified in earlier interviews. Data collection was therefore concluded after that, resulting in a final sample of fourteen semi-structured interviews conducted between March 5 and April 22. An overview of all interviews was provided in Appendix 1.

The interviewees were selected in accordance with the sampling approach described in section 3.1.2 and include one country manager, one solution delivery director, one delivery manager, three project managers, one business development manager, one concept leader, one head of data and AI, one business analyst manager, one technical leader, one consultant, and two C-level executives. Together, the interviewees represent a range of organizational roles and span several industries, including e-commerce, truck manufacturing, logistics, strategy consulting, banking, finance, food and beverage, apparel and fashion, IT software, and financial technology.

3.2.2. Interview design

An interview guide (Appendix 2) was used to structure the interviews. Each interview began with a brief ice-breaking phase to ease the pressure of the interview setting, acknowledging the importance of building rapport with interviewees when eliciting qualitative data (Ryan & Dundon, 2008). This was followed by an explanation of the study's purpose, assurance of anonymity and confidentiality, and a request for consent regarding recording and data storage in accordance with GDPR, which all interviewees provided. Specifically, the questions were designed to explore both the performance dimensions that managers focus on and how they integrate these dimensions to form the overall judgment. Follow-up questions were used throughout the interviews to clarify meanings and probe emerging themes.

As explained in section 3.1.2, a vignette was used during the interviews as a stimulus for discussion. In line with Wiegmann et al. (2025), the vignette was based on a common scenario in sales management, designed to be easily understood without requiring specific domain knowledge.

Also, the vignette was written in a clear and concise manner, using language that matched the participants' expected literacy levels (Wiegmann et al., 2025). In addition, following the recommendations of Evans et al. (2015), the vignette was constructed as a narrative with a clear story-like progression to help participants follow the situation step by step and reflect on the decision-making process as it unfolded. The vignette also avoided placing the participant directly inside the scenario, either as a first-person or third-person character, so that participants could engage with it from a neutral position (Evans et al., 2015).

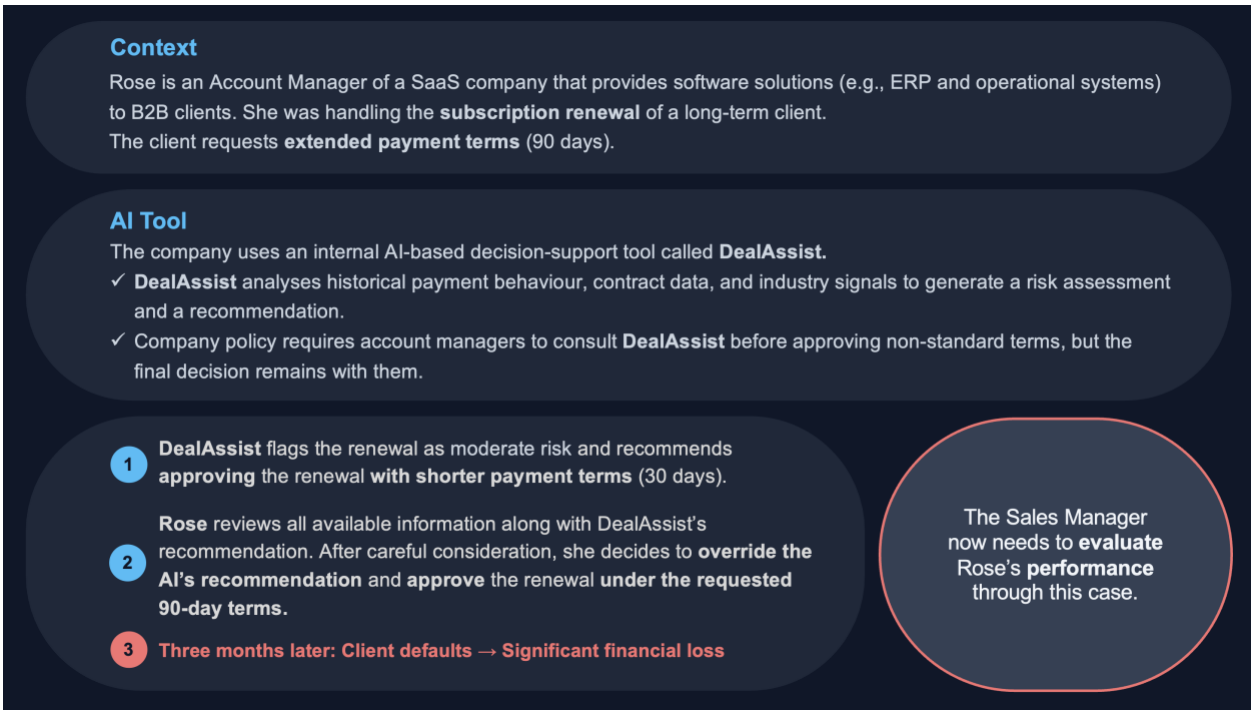

Figure 2: The vignette used in the interviews

Each interview was conducted online via Google Meet and lasted between 30 and 60 minutes. Interviews were carried out in either English or Vietnamese, depending on the participant's language preference. All interviews were transcribed in real time using Voicenotes, an AI-powered transcription tool, and were verified by the interviewer within 24 hours of completion. The transcripts were subsequently anonymized before any further processing was conducted. Interviews conducted in Vietnamese were then translated into English to ensure consistency in coding and interpretation. The translations were performed verbatim, with hesitations, pauses, and conversational nuances retained to preserve the original meaning and tone of the participants' accounts. Translation was carried out using Claude Sonnet 4.6, which was selected after testing

several models due to its stronger ability to retain the original meaning, and was subsequently verified by the author immediately after completion.

3.3. Data analysis

The interview material was analyzed using thematic analysis as described by Braun and Clarke (2006). Thematic analysis was chosen because it offers a systematic approach for identifying recurring patterns in performance evaluation practices across a set of semi-structured interviews, which fits the interpretive orientation of this study. Specifically, a theoretical (deductive) thematic analysis was employed rather than an inductive approach. Following the theoretical framework developed in section 2.2, initial codes were organized around categories such as results controls, action controls, personnel controls, and cultural controls, as well as different information integration logics. At the same time, the analysis remained open to additional codes whenever recurrent patterns emerged that were not anticipated by the framework, for example, specific ways in which AI influences managers' evaluation of their subordinates.

The study broadly followed Braun and Clarke's (2006) six phases of thematic analysis: familiarization with the data, generating initial codes, constructing themes, reviewing themes, defining and naming themes, and producing the report. After the interviews were transcribed and anonymized, the transcripts were read and re-read to develop familiarity with the content and to note initial ideas related to evaluative dimensions and integration logics. During the coding phase, relevant extracts were coded manually, and coding proceeded iteratively: earlier transcripts were revisited as new codes emerged in later interviews. Codes were then examined and grouped into candidate themes representing broader patterns in how managers evaluate subordinates' performance and how AI influences this practice. These themes were subsequently reviewed and refined to ensure internal coherence and clear distinctions between them. Throughout the process, themes were generated through an active, interpretive engagement with the data, acknowledging Braun and Clarke's (2006) notion that the researcher plays an active role in organising and making sense of participants' accounts rather than simply discovering themes that pre-exist in the data. Finally, illustrative extracts were selected for inclusion in the empirical analysis (Chapter 4). Examples of the coding process are provided in Appendix 3.

3.4. Validity

To ensure the validity of the study, the principles of authenticity and plausibility were consistently maintained, acknowledging Lukka and Modell's (2010) notion of validation in interpretive management accounting research.

Authenticity refers to whether researchers provide an account that is genuine to their field experience, such that readers are convinced that the researchers have "been there" (Lukka & Modell, 2010). In this study, authenticity was pursued by providing thick, context-rich descriptions of interviewees' accounts, including direct quotations that preserve participants' own vocabulary. The analysis deliberately retained inconsistencies, ambiguities, and contradictions in the accounts rather than attempting to smooth them out, i.e., managers explicitly state that action controls should play a role in the overall evaluation process but are still biased toward outcomes eventually. Moreover, the interview sample includes managers from different roles and industries, which supports a multifaceted portrayal of how performance evaluation takes place in practice rather than a single, homogenized narrative.

Lukka and Modell (2010) explain plausibility as "*whether an explanation makes sense and can be intersubjectively accepted as a likely one*." According to the authors, *"failing to demonstrate the mechanisms that may have produced a particular phenomenon can lead readers to question whether researchers have penetrated deeply enough into the field to develop an emic understanding."* To maintain plausibility, this study seeks to go beyond providing descriptive accounts and aims to offer thick explanations of those accounts. To do so, the analysis followed a *process of abduction*, as suggested by Lukka and Modell: recurrent patterns in the data were first interpreted through the theoretical framework to generate explanations, which were then further refined in light of subsequent interview data, allowing alternative interpretations to emerge. For example, a statement made by one interviewee may initially have been interpreted as reflecting action controls in Merchant and Van der Stede's (2003) framework. However, when a later interviewee described a similar issue but provided a richer explanation, the earlier interpretation was re-evaluated, leading to the conclusion that it was more appropriately categorized as personnel controls. This movement between empirical data and theory aims at *"inference to the best explanation,"* as Lukka and Modell put it.

4. Empirical analysis

In this chapter, the empirical analysis is presented. Following the theoretical framework developed in section 2.2, the analysis is structured around two main areas: (1) the types of controls managers focus on when evaluating subordinates in AI-supported decision-making contexts, and (2) the integration logics managers use to combine different performance dimensions to arrive at a final judgment, corresponding to sections 4.1 and 4.2.

4.1. Control types in performance evaluation

4.1.1. Results controls

A central pattern across the interviews is that results remain salient in performance evaluation. Several interviewees emphasized that outcomes strongly shape how employees are judged. Interview 6 put it bluntly:

> *"The only way you're going to be viewed positively is if there's a positive outcome. If you make money – positive... If you lose money, I don't care whether you listened to the AI or not; I'm not happy." – Interview 6*

Similarly, Interview 13 expressed this clearly:

> *"I evaluate my people on the value they've brought to the organization, on what they've delivered. That's the effectiveness side, and that's the angle I take." – Interview 13*

When discussing outcomes, however, interview data suggest that managers rarely assess individual decisions in a simplistic manner. Instead, they tend to focus on broader patterns of performance and the extent to which employees achieve performance targets over time. Interview 12 articulated this clearly:

> *"In the review meeting, team members will sit down with me and review the goals I cascaded down at the start of the year that they agreed to. They'll list out their achievements against those goals... Essentially, based on the goals that were cascaded down and agreed upon, I would assess whether their work was effective or not." – Interview 12*

A similar perspective appeared in Interview 6, which highlighted the importance of track record rather than isolated outcomes:

"I cannot manage every individual decision. I don't want to have to do that. What I want is a systematic approach to decision-making. If the system is right most of the time, I expect you to do your due diligence to confirm it and then go with it. Now, if you consistently go against the recommendation and your personal track record is actually better, say you're right 65% of the time compared to the machine's 51%, I can't really hold the occasional failure against you." – Interview 6

This suggests that results are interpreted through cumulative performance rather than a single outcome. When a negative outcome occurs, a strong historical track record tends to shift causal explanations away from internal shortcomings and toward situational or uncontrollable factors. In such cases, a poor result is less likely to be seen as evidence that the decision-maker is generally incompetent. Interview 8 echoed this logic by arguing that one loss-making case should not dominate the evaluation if KPIs remain strong:

"Out of 10 decisions she made, 1 was wrong and 9 were right, and her revenue is still positive. So why would we evaluate her as poor if the KPI is still hit? That doesn't make sense, it's not fair." – Interview 8

Here again, the manager does not treat a single negative result as sufficient proof of poor performance when broader outcome patterns remain positive.

Another important theme emerging across the interviews is that the presence of AI changes how outcomes are interpreted. Interview 3 points to the AI system as a possible cause of the decision outcome:

"What's important is to deep-dive into the specific steps so you can see where the problem lies – whether the problem is that AI provided not enough information, or whether AI provides enough information but the account manager didn't take it seriously, or whether the account manager did take it seriously but was overconfident and still went ahead with her decision." – Interview 3

This illustrates that managers may locate the cause of decision outcomes at different points in the decision process: in the quality of the AI output, in the employee's use of that output, or in the employee's subsequent judgment.

Interview 1 similarly argued that AI introduces *"additional data, additional sources, additional criteria"* into decision-making, meaning that outcomes should be interpreted against a richer informational backdrop as opposed to purely human judgment. Commenting on the case presented in the vignette, the same interviewee noted that when the organization has intentionally incorporated AI into the decision process, responsibility for decision outcomes may also lie partly with the organization: *"In Rose's decision, you can split it 50/50, 50% is her responsibility, and 50% belongs to the company, because the company provided her with this tool."* This suggests that the locus of causality becomes more distributed: the decision outcome is linked not only to the human decision-maker but also to managerial choices about incorporating AI into the decision-making process. The interviewee further elaborated on this point:

> *"Any AI tool that is introduced, analysts have already assessed it to evaluate risk levels. And they accept that the system is just a system and will have certain risks. So if the tool is being deployed and applied broadly, or even enriching the company's knowledge base so that over time the AI tool becomes smarter, producing better outcomes, then the risk here is that for those initial deals when you've just started applying it, there will be a certain percentage of failure. That's acceptable." – Interview 1*

This organizational dimension was also evident in Interview 5, where the interviewee suggested that a poor outcome may indicate process failure rather than individual incompetence: *"I would not blame this person immediately; there could be a review step missing somewhere in the organization."* This indicates that negative results may be interpreted as evidence of missing controls in the organization rather than solely as failures of the human decision-maker. In other words, the cause of failure may be located externally – that is, in organizational structures or processes – rather than internally in the subordinate's judgment.

To summarize, the findings show that results controls remain vital in AI-supported decision-making contexts, but their interpretation becomes more complicated. Outcomes matter, yet they do not map straightforwardly onto the decision-maker's judgment; they are instead evaluated in relation to the broader performance track record. Managers also pay attention to whether causality lies within the subordinate's judgment, within the AI system, or in the broader organizational process design.

4.1.2. Action controls

A recurring theme across the interviews is that, in companies where AI has been adopted as part of the organization's strategies, using AI is becoming an expected behavior. Managers repeatedly expressed the view that AI use is no longer optional, but something employees are expected to incorporate into their everyday work practices. Interview 3 expressed this clearly:

> *"AI has now become mandatory at our company. All roles, from scrum master and business analyst all the way through developers and QAs, are all required to document how they use AI to increase productivity and quality." – Interview 3*

Interview 5 expressed a similar point, noting that AI usage and adoption rates have become formal KPIs across the organization:

> *"As per company requirements, 100% of staff must use AI. The company is developing an internal tool, and 100% of staff are required to use it. There was a trial phase and then mandatory adoption." – Interview 5*

However, simply using AI does not suffice. The interviews also show that managers increasingly evaluate whether subordinates use AI in a manner they consider appropriate. This goes beyond whether AI is consulted and extends to whether the subordinates treat AI as a decision-support tool that informs further analysis or as a substitute for their own judgment. Interview 2 described simply following AI without further assessment as "lazy" and emphasized that the system's output should be understood as a recommendation rather than a final judgment: *"We should be aware that the tool only provides a forecast. It makes a recommendation, not a definitive judgment. The tool is there to help with that, but it's a support, not a substitute."* The interviewee further argued that AI recommendations should trigger deeper analysis, particularly when the recommendation is close to a decision threshold:

> *"I think the tool needs to do more than just give a binary recommendation – it needs to convey context about how close the decision is to the threshold. For example, the model says 51%, but the limit is 50%, so we're right on the edge. That kind of nuance needs to be surfaced by the tool, so the user knows whether to dig further or whether it's clearly within the safe zone… In simple cases where everything is well within range, you can just go with*

the recommendation. But when you're approaching a borderline, you really need to understand the situation more deeply and not simply follow the output." – Interview 2

These accounts reveal an expectation that employees should interpret AI recommendations critically and investigate their outputs further rather than accept them at face value. Interview 6 expressed this expectation even more directly:

"If things go bad, I don't care about the AI. You had better have done your own homework. And if you don't listen to the AI, you had better have done your homework even more thoroughly. Either way, you need to do your own homework. That's the bottom line." – Interview 6

Similarly, Interview 3 warned that employees may over-trust AI and skip necessary checks. In the interviewee's view, AI should be treated as one input among several rather than as a replacement for independent assessment or consultation with others:

"The problem now is this: people trust AI too much, skipping the steps that come afterward, and that's where the danger lies. But if people use AI correctly, it will help them make better decisions – that's certain. So, what's the failure mode here? It's when people start over-trusting AI, treating everything AI says as the truth, and making decisions without any further assessment or evaluation, just locking it in immediately. And that can lead to very severe consequences." – Interview 3

The idea of critically evaluating AI output was also emphasized in Interview 11, where the interviewee argued that because AI is known to produce hallucinations, human users must be able to question its answers and push back when necessary. Without such critical judgment, the interviewee noted, decision-makers risk being led seriously astray by AI.

Beyond the expectation to use AI critically, some interviewees described AI-specific routines that are becoming part of what they consider expected work processes. In Interview 1, for example, the project manager suggested that, in high-risk decisions, human decision-makers should use multiple AI agents configured with different perspectives and allow them to debate before arriving at a final recommendation. This approach, according to the interviewee, would generate a broader set of AI-produced arguments for the human decision-maker to assess:

"With a Deal Assist like this, you could create a Deal Assist 2 to counter the first one's opinion and let the two of them debate to see whose final decision is the most convincing. You could create many different Deal Assists to ultimately get multiple perspectives on this case, and from there make the final decision." – Interview 1

This suggests that proper use of AI may involve actively designing and structuring decision processes to create comparisons across different AI systems rather than relying on a single tool. Interview 12 echoed this view by stating that using multiple AI systems simultaneously is considered the most "advanced" level of AI application in the organization:

"The most advanced level is when they combine multiple AIs. For example, one AI specializes in collecting information and building a repo of domain knowledge. Then they use another AI with greater power and faster automation capabilities to generate user stories from the client's requirements. But they can't just trust 100% of what AI produces without validation. After the second AI generates user stories, they can use the first AI to validate them." – Interview 12

Interview 7 expressed a similar expectation, describing AI not only as a source of recommendation but also, as the interviewee put it, as a potential "devil's advocate." According to the interviewee, good decision-making involves prompting AI to critique its own output as part of the process:

"As for using AI, if it were me, I would have the AI verify its own answers… You need a checker – and that checker is re-analyzing the arguments, the pros and cons that the person proposing put forward, and criticizing them. There should always be a devil's advocate role: playing the bad guy to challenge all of that." – Interview 7

Another key theme emerging across the interviews is that, when negative outcomes occur, managers pay particular attention to whether subordinates followed established procedures. In such situations, procedural compliance appears to function as an attributional filter through which outcomes are interpreted. Interview 2, for example, emphasized the importance of understanding whether the subordinate acted within the formal authority attached to the role and followed the required process when overriding the AI recommendation:

"I would want to know what authority the account manager actually has. Is it within her responsibility to deviate this far from what the tool suggested? Or does the account manager

need to escalate – to bring in one more level of management – before going outside a certain boundary?" – Interview 2

This indicates that unfavorable outcomes are not automatically attributed either to the human decision-maker or to the AI system. Instead, managers consider whether the subordinate acted within the decision space assigned to them and complied with established procedures. Interview 9 reflected a similar logic by stressing that evaluation should begin with whether the broader organizational process had been followed rigorously:

"I would want to know whether the sales process was followed rigorously. Even though in this case the account manager is within her authority, had she followed the proper process? This is very important because the company's sales operations aren't simply "client wants this, we sell this." There are many other business team members involved." – Interview 9

Here, adherence to process functions as a safeguard: when required procedures have been followed, failure is more likely to be interpreted as external, situational, or systemic. By contrast, bypassing the process makes the outcome appear more internally caused and therefore more blameworthy. Interview 6 expressed this logic directly: *"You followed procedures, you did everything right. It's just a one-off; nothing we can do."*

Taken together, these accounts suggest that, when negative outcomes occur, established procedures serve not only as behavioral expectations but also as a lens through which responsibility is reconstructed and performance is judged.

In summary, the findings suggest that action controls remain vital in AI-supported decision-making contexts, but they take new forms as AI becomes embedded in the work processes. Managers evaluate not only whether employees use AI, but whether they use it appropriately – that is, engaging critically with AI recommendations, conducting further analysis to validate AI outputs, and actively interacting with AI in a disciplined process rather than relying on it passively. Procedural compliance becomes particularly important when negative outcomes occur, as adherence to established processes shapes how the situation is reconstructed and whether failure is attributed to internal factors such as judgment or competence, or to situational or organizational factors.

4.1.3. Personnel controls

A consistent pattern across the interviews is that, in AI-supported decision-making contexts, the reasoning human decision-makers put forward becomes particularly salient when their decision quality is evaluated. Interviewees repeatedly emphasized the need to understand the thought process that led subordinates to their decisions, including the information they considered, the logic they applied, and the context they weighed. Interview 5 expressed this explicitly when discussing the case presented in the vignette: *"I would listen to her first to hear what she was thinking and what led her to that decision. I would then assess based on my experience and the company's processes."* Interview 7 explained this view in more detail, emphasizing that evaluation should focus on the reasoning based on information available at the time the decision was made rather than judging the decision in hindsight:

> *"Certainly, I would want to know what argument she had for overriding the AI's recommendation. If I'm trying to identify the root cause, the first thing I would ask is: what did the AI recommend? Was it that the recommendation was wrong, didn't reflect reality, or whatever. And I would evaluate it at the time of the decision, not from now. At that point, with the available information, combined with the analysis, if the reasons she gave, the arguments she put forward, make sense... then honestly there's no right or wrong, it's just risk appetite." – Interview 7*

This view was echoed in Interview 4, which emphasized that reasoning should be documented at the time of the decision rather than reconstructed afterward. According to the interviewee, requiring explicit justification when collaborating with AI, especially in high-stakes decisions, helps ensure that the human decision-maker remains actively engaged rather than merely endorsing the system's output:

> *"What I would really want is that she documents her reasoning when she overrides AI recommendation, recorded at the time of the decision, not reconstructed afterward… If the decision is genuinely critical to the company, if getting it wrong poses a real threat, then requiring the account manager to articulate multiple arguments, rather than simply deferring to AI, is important. It keeps the person responsible meaningfully engaged in the decision rather than just rubber-stamping the output." – Interview 4*

Interview 3 made a similar point by emphasizing that the purpose of managerial follow-up is to understand the subordinate's judgment:

> *"If I had a one-on-one conversation with her, I would ask her this question: if this happened again, what would you do? The purpose of that meeting is to understand her judgment, her reasoning, to see whether, in the process of making the decision, she treated AI as a support tool and still used her own reasoning to make the decision or not." – Interview 3*

Taken together, these accounts illustrate that, in AI-supported decision-making contexts, managers place strong emphasis on whether subordinates can articulate a coherent rationale for their actions and demonstrate sound judgment when interacting with AI outputs.

Besides reasoning, managers also expect that employees demonstrate the right approach to decision-making when collaborating with AI. Rather than relying on intuition, managers expect their subordinates to apply a clear and systematic process for evaluating options and interpreting AI outputs. This includes, among others, defining the problem, assessing available information, considering risks, and consulting relevant stakeholders. Interview 3 articulated this clearly:

> *"Employees should make decisions based on agreed principles and a transparent method. For example, considering risks and consulting relevant stakeholders." – Interview 3*

Interview 7 reinforced this logic by emphasizing that the presence of a methodology is even more important than the final answer itself. From the interviewee's perspective, a structured methodology signals critical thinking, while decisions made without a clear approach are perceived as impulsive and difficult to evaluate:

> *"The first thing I would evaluate is whether you have your own methodology to approach the problem. Because not having a methodology leads to decisions that are very impulsive, very feeling-based rather than reasoned. Things that are feeling-based can't be judged, and they just become a matter of luck." – Interview 7*

Interview 11 expressed a similar idea but framed it as a *"working workflow"* – the structured thought process an employee follows when approaching a problem. The interviewee explained that effective AI use involves a systematic approach built around a sequence of steps: clearly defining the task, planning the work, implementing it, and then reviewing and testing the output.

Interview 10 expanded this notion of having the right approach even further by pointing to a mindset shift that distinguishes stronger from weaker AI users. According to the interviewee, human decision-makers need to see themselves as managers of AI: *"I'm the manager (of AI), and I'm not doing everything directly anymore because AI can handle tedious and research-intensive work much faster and take that burden from me."* This mindset, according to the interviewee, translates into the ability to take the lead in human-AI collaboration – that is, to think abstractly, define the overall framework, direct AI to perform the required tasks, and critically evaluate its output before acting upon it.

Interview 12 echoed this view by emphasizing the importance of effectively guiding AI through prompt design, framing it as a foundational skill for working with AI systems:

> *"For me, the priority is the ability to write prompts. If you have the mindset of using a prompt to make a request in a way that is sufficiently detailed but not excessive, that's actually the hard part. For AI to read that prompt and deliver exactly the result you want, that skill is genuinely not easy, not simple." – Interview 12*

This capability to work with AI is described by several interviewees as part of the expectations attached to professional roles rather than as an optional skill. Interview 3 expressed this clearly: *"AI has now become mandatory at our company… Managers assess not only domain performance but also whether individuals are developing effective AI usage."* Interview 12 similarly described an *"invisible pressure"* on employees to continuously upgrade their AI-related skills:

> *"In terms of performance, those who don't use AI will almost certainly be replaced by those who do. AI doesn't replace people at this point in time. But AI is a driver for people to upskill. If you don't use AI, or if you resist and reject it, you're likely to be replaced by people who embrace technology early, who adopt AI sooner, and who are ready to use any AI tool to increase their work effectiveness – those people will have work." – Interview 12*

This suggests that AI competence is likely to shift from an option advantage to a normal requirement for employability, or in other words, from a bonus dimension to a threshold condition in performance evaluation.

Besides AI-related knowledge and skills, interview data also point to the continued importance of foundational capabilities such as domain expertise and traditional professional skills. While AI

may change how work is performed, interviewees emphasized that it does not remove the need for employees to possess the knowledge required to exercise judgment and evaluate AI outputs critically. Interview 11 expressed this clearly:

> *"We still need to be extremely, extremely detailed, extremely strict about the foundational knowledge of the user, the employee. Meaning AI is still just a tool, and the person using it definitely needs to have foundational knowledge in order to validate AI outputs and use it effectively." – Interview 11*

Interview 12 echoed this view by warning against developing skills in a one-sided manner, focused only on AI skills while neglecting foundational knowledge:

> *"Besides mastering new skills, your understanding of domain knowledge must also be reinforced. If you develop in a completely lopsided way, only toward AI use, without deepening your domain knowledge and industry expertise, at some point, you'll be surpassed. Why? Because you've got an extremely powerful tool and you know nothing else. One, it will give you information you can't verify, and two, because it's so powerful and capable of synthesizing enormous amounts of information, at some point it will replace you entirely. By then you'll probably be too dependent on it, and it will be strong enough that you will no longer have the chance to hold onto that work." – Interview 12*

Taken together, these accounts indicate that managers do not view AI competence as a substitute for foundational capabilities. Rather, effective performance in human-AI collaboration depends on a combination of new AI-related capabilities and enduring traditional knowledge and skills.

In summary, the findings suggest that personnel controls play an increasingly important role in AI-supported decision-making contexts. When AI is incorporated into the process, managers expect their subordinates to articulate sound reasoning, apply a structured approach to decision-making, demonstrate the capabilities to work effectively with AI tools, and actively take the lead in human-AI collaboration. Besides, foundational knowledge such as domain expertise remains vital, as they enable humans to critically validate AI outputs and apply them appropriately. Such capabilities strongly influence how managers perceive their subordinates' performance, as they are often regarded as internal and controllable qualities for which the individual can be held accountable.

4.1.4. Cultural controls

Across the interviews, a strong expectation emerged that AI should be understood as a support tool rather than a decision-maker, and the final decision authority should remain with the human employee. Interview 7 expressed this view clearly:

> *"My thinking is that AI is just an assistant, and the human has the decision-making authority. AI only provides information, recommendations, and arguments. But it's still human judgment when using that information in the process of making a decision. AI is like … something that synthesizes information and hands it back to you, plus a bit of a recommendation." – Interview 7*

Interview 8 echoed this perspective: *"AI is just a support tool. Ultimately, there still have to be other parameters from humans added in to get the right number. AI just supports humans in making certain decisions faster."* And interview 11 made a similar point: *"The final decision was still the account manager. Whatever AI says, it's still her decision."*

These accounts suggest the presence of a shared norm: AI may inform decisions, but it should not replace humans as the final decision-makers. However, the interviews show that managers' views begin to diverge when discussing how responsibility should be distributed in practice. Some respondents maintained a strong view on human responsibility. Interview 3 expressed this position explicitly:

> *"You are the one who bears responsibility, not AI. You can't say, 'because AI said so, therefore I take no responsibility.' If you made the decision, you are responsible for it. AI is just one step. Afterward, you still have to look back, re-evaluate, and possibly get advice from other people before you finalize the decision." – Interview 3*

Interview 13 made a similar point, arguing that: *"In my view, the responsible party is always the human. You can't say AI is responsible… You must be able to control the outcome that AI produces."* Similarly, Interview 12 distinguished between AI's functional role and human accountability for final decisions. While AI may be responsible for carrying out assigned tasks, the interviewee argued that accountability remained with the human decision-maker:

> *"I think AI is responsible for what it's been assigned, while the human is accountable for the outcome at the end. You can't blame AI. If you truly treat that as AI's responsibility, you*

end up asking: who is AI serving, itself or humans? From my perspective, I still see AI as an assistant rather than a colleague. If you put AI on equal footing with a human as a colleague, then at some point, it replacing you becomes unavoidable. I think at this stage, AI still should be, and still is, an assistant." – Interview 12

Taken together, these accounts reflect a strong expectation that accountability cannot be delegated to technology. Even when AI performs substantive analytical tasks or influences decisions significantly, respondents still viewed the human actor as the party ultimately responsible for the consequences.

Other interviewees, by contrast, articulated a more nuanced view of responsibility. Interview 1, for example, acknowledged that legal accountability still rests with humans: *"In all situations, the only one who can go to prison is a human, right? Not a tool, not a machine."* At the same time, the interviewee proposed that some responsibility should be shared with the organization that built and mandated the AI system, particularly because *"there are certain logics under AI recommendation that human decision-makers are not fully aware of."* According to the interviewee, the appropriate distribution of responsibility depends on the context, with a larger share placed on the human decision-maker in high-stakes situations.

Interview 10 expressed a similar tension. The interviewee first emphasized human responsibility: *"I think the final output of this employee is her responsibility. Right? Because you pay her salary, you don't pay the AI salary. It's her responsibility."* Yet the interviewee also pointed to collective responsibility for adopting AI: *"This is company strategy. It comes down to whether leadership believes AI changes the game or not. If they believe it and adopt AI, then it's also the responsibility of the whole organization. When you introduce an AI system into the organization, there may be disruption to the business in the early phase for sure."*

This divergence among interviewees' perspectives reflects an underlying cultural difference, suggesting that cultural dimensions play a role in shaping what managers expect from their subordinates and how they evaluate subordinate performance. Specifically, in more technology-native organizations, where experimentation with digital tools is normalized and the incorporation of AI into everyday work is recognized by leadership as a strategic priority, there appears to be a more nuanced view of responsibility. In such contexts, managers seem more willing to take into consideration the quality of AI systems and the organization's responsibility for designing and

adopting the technology. By contrast, in industries characterized by stricter regulation, higher accountability demands, or lower tolerance for error, managers appear more likely to place full responsibility on the human decision-maker. In these environments, AI may still be used as a support tool, but responsibility is less readily shared with the technology.

This cultural dimension was explicitly acknowledged by several interviewees. Interview 12, for example, noted that: *"When it comes to policies and norms, it's very industry-specific. If you are in insurance or banking, it is very different from, say, military or enterprise information systems."* Interview 13 emphasized a similar point, arguing that AI adoption and expectations depend not only on the industry itself but also on organizational readiness and workforce capabilities:

> *"AI can be applied in many different angles and professions. It is just: do you have the money, is there a business case for it, what specifically do you want to apply it to, and are your employees ready for it? For example, bank employees are typically very well-educated, so AI application is relatively easy. But in retail, say there are 7,500 employees, most of them work in stores. Store staff could be students or even high school graduates, so the level is not high. And if the level isn't high, it is hard to apply or roll out something effectively." – Interview 13*

Cultural dimensions also influence how much weight managers think AI should carry in decisions. Interview 6 argued that in quantitative and relatively objective domains such as credit risk, the AI *"should be king,"* as the interviewee puts it, suggesting a norm that algorithmic recommendations should generally be followed and that overriding them requires strong justification:

> *"But when you put in credit metrics, that's quantitative, measurable, and analyzable data. Credit risk is objective. What's your margin? Is it deteriorating? What's your growth rate? Are accounts receivable growing or shrinking? These are all quantifiable, objective indicators. In that domain, AI should be king." – Interview 6*

By contrast, other managers adopted a more skeptical stance. According to these interviewees, AI is fast and powerful but not always correct, so good practice involves surfacing and scrutinizing its reasoning, analyzing its outputs in the broader business context, and treating it as one of several inputs. Interview 8, for example, argued that:

> *"Actually, at this point, for me, AI is used as something like a consultant. What AI is based on here is all old data; clearly, it's working with historical data, and it will produce a suggestion, a sort of forecast. With this kind of thing, honestly, I don't trust it 100%. For me, it's only about 65%. I usually rely more on its data analysis capability, meaning I would ask it to analyze this data for me. I don't want it to give me a recommendation, because I need external factors to assess it." – Interview 8*

Interview 9 made a similar point: *"If AI doesn't have strategic context about the company, the recommendation may sound good but could actually be bad... It actually doesn't help you make the decision; it only recommends."*

Taken together, these accounts suggest that cultural dimensions and industry characteristics shape the weight granted to AI recommendations. In some settings, AI is treated as a source whose outputs deserve substantial weight, whereas in others it is viewed as a useful but relatively less powerful advisor. This difference appears to stem from the perceived importance of contextual understanding in effective decision-making, the areas in which humans are still seen as stronger than AI at this stage.

To summarize, organizational norms and culture shape managers' perspectives on the role of AI, the extent to which responsibility can be delegated to technology, and the weight that AI recommendations should carry in decision-making. While a dominant view across interviews is that humans remain the final decision-makers and are ultimately accountable for outcomes, the strength and interpretation of these expectations vary across organizational and industry contexts.

4.2. Information integration logic in performance evaluation

4.2.1. Multi-dimensional evaluation

Interview data suggest that managers rarely evaluate subordinates through a single lens. Instead, they combine multiple considerations to form a final judgment. Several interviews illustrate this multi-lens pattern explicitly. Interview 6, for example, initially foregrounds outcomes, stating that *"the only way you are going to be viewed positively by me is if there is a positive outcome."* However, the interviewee also insisted that subordinates must have *"done their own homework,"* whether they followed or overrode AI recommendations, bringing action controls (procedural diligence) and personnel controls (individual conscientiousness) into the evaluation. The same

interviewee later described credit risk as an objective, quantitative domain where *"AI should be king,"* which adds a cultural-control dimension: in this setting, there is a norm that algorithmic advice should generally carry substantial weight.

Interview 2 also illustrates how multiple lenses come together when reflecting on the case presented in the vignette. The interviewee first wanted to know what formal authority the account manager had and whether deviating from AI recommendations should have triggered escalation to higher management, which reflects an action-control perspective focused on compliance with the established process. At the same time, the interviewee criticized simply following AI output without further analysis as *"lazy,"* reflecting a personnel-control lens focused on the employee's diligence. The outcome of the deal matters, but it is evaluated in conjunction with process and capability considerations rather than standing alone.

Interview 14 similarly makes this multi-lens integration particularly explicit. When asked what dimensions he would consider when evaluating subordinates, the interviewee argues that *"just looking at the result isn't enough because AI operates as a black box,"* and instead suggests evaluating *"the entire process of how they used AI – how they asked questions, provided context, validated results, and stayed involved at relevant checkpoints,"* which clearly reflects action controls. At the same time, the interviewee also highlights the importance of using the right AI model and being aware of the hallucination problem, which speaks to personnel controls around AI knowledge.

At the same time, the interviews reveal an asymmetry in how negative and positive outcomes are treated. When the outcome is negative, managers tend to open discussion of the decision process and examine it in depth: they ask whether authority boundaries were respected, whether procedures were followed, whether the subordinate's reasoning was strong, and whether AI was used appropriately. By contrast, when the outcome is positive, the same decision receives considerably less scrutiny. Interview 6 makes this point explicit by stating that *"nobody checks you as thoroughly when things go right as when they go wrong,"* suggesting that success tends to shield the underlying process and reasoning from diagnostic scrutiny, whereas failure triggers retrospective examination. Interview 2 echoes this asymmetry; when asked how the decision would be viewed if the outcome was positive, the interviewee replied that:

> *"Then it wouldn't be an issue, or at least not in our organization as things currently stand. Everything went fine, so we would just carry on." – Interview 2*

The same interviewee, however, acknowledged that assessing process compliance should be independent of the outcome when reflecting on the case presented in the vignette: *"But I actually think there should be symmetry here: if she went outside her scope, it should be addressed for next time, regardless of whether the outcome was good or bad. The compliance piece should be in place independently of the result."* This illustrates a tension between what managers believe should happen normatively and what they think would happen in practice.

That said, a positive outcome does not crowd out all other dimensions. Interview 6 made this point explicit:

> *"The only way you're going to be viewed positively by me is if there's a positive outcome. If you make money – positive. If you make money but you put the firm at high risk – negative, because I don't want unnecessary risk." – Interview 6*

This suggests that while positive outcomes may not trigger the same level of scrutiny as negative ones, they do not eliminate other evaluative lenses. Action controls, for example, still operate as boundary systems when certain behaviors are considered critical to the organization, such as avoiding excessive risk in risk-sensitive industries. In such cases, even successful outcomes can be evaluated negatively if the underlying behavior is deemed inappropriate.

Taken together, these accounts show that managers rarely rely on a single lens when evaluating subordinates. Instead, they bring together information about outcomes, procedural compliance and AI use, individual capabilities and reasoning, and broader organizational norms to form an overall judgment of subordinate performance. The extent to which these dimensions are considered, however, is influenced by the outcome: negative outcomes tend to trigger a more comprehensive evaluation, prompting managers to examine process, reasoning, and AI use in greater depth, whereas positive outcomes often lead to less scrutiny.

4.2.2. Averaging law as the dominant integration logic

Interview data suggest that the way managers integrate information to formulate the final judgment of their subordinates follows the logic that resembles Anderson's (1981) averaging law. When combining multiple performance dimensions to form an overall judgment, managers do not assign

equal weight to each dimension. Interview 14, for example, suggested that around 60 percent of the overall evaluation should focus on the outcomes, while the remaining 40 percent should be divided between the AI system itself and how the employee used it, with each contributing 20 percent. However, the weighting of these dimensions appears to vary across contexts. Interview 10 described an approach in which evaluation is split between traditional KPIs and AI adoption – that is, whether employees adopt AI in the expected manner – with an equal 50/50 weighting during the early phase of AI implementation. This variation suggests that the importance assigned to each dimension, or in Anderson's (1981) terms, the *"subjective value"* of each cue, is contingent on organization strategies and industry contexts.

Regarding how these dimensions are integrated, the interviews indicate that when other dimensions move in the opposite direction from the outcome, they tend to balance out its impact. Commenting on the case presented in the vignette about how the account manager should be evaluated after the negative outcome, Interview 2 and Interview 9 both suggested that evaluation should begin by asking whether the account manager followed the established process; if she did, failure is more likely to be interpreted as the result of external or systemic factors, which mitigates the severity of negative outcomes in the performance judgment. Interview 6 echoes this view by pointing out that if the account manager *"followed procedures and did everything right, have a strong argument for why she did what she did,"* then a bad outcome may be seen as *"just a one-off"* that cannot be entirely avoided, adding that *"I'm not gonna view her negatively because she did what she was trained to do."* This suggests that diligent adherence to process and strong reasoning can balance negative outcomes in forming the overall judgment. On the other hand, several managers emphasized that failing to comply with established processes can reduce the value of positive outcomes. For instance, Interview 6 stressed that *"if you make money but you put the firm at high risk – negative, because I don't want unnecessary risk."* Similarly, in the IT outsourcing context, Interview 5 treated a profitable contract achieved through weak process adherence or missing review steps as evidence of *"process gaps,"* suggesting corrective action even when the financial result was favorable.

Taken together, these accounts indicate that managers approach performance evaluation as a process of balancing multiple sources of information, where each dimension carries a certain weight and the final judgment reflects their combined, averaged influence.

4.2.3. Temporal dynamics of integration logic

Beyond showing that managers combine multiple lenses through the averaging logic, the interviews also suggest that the logic itself is expected to evolve over time as AI becomes more embedded into work processes. Specifically, both the set of dimensions that matter and the relative weights assigned to them are contingent on the AI adoption phase.

Interview 10 provides a clear account of this temporal perspective. According to the interviewee, performance metrics remain important, but evaluating employees solely on KPIs immediately after introducing AI may be counterproductive, as employees can be reluctant to experiment with new tools that do not generate immediate results. The interviewee suggested an even split between performance KPIs (results controls) and AI adoption (action controls) in the early phase of AI implementation, while also pointing out that this weighting should change over time as AI use becomes normalized within the organization:

> *"If the goal is for AI to help the organization improve, then ultimately it still comes down to performance and KPI. But it depends on the phase: if you introduce AI and immediately evaluate only on KPI, you may be out of sync because people are reluctant to try new things, especially things that don't produce results right away. So you need a two-dimensional measurement: in the early phase, split the weight between performance KPIs and metrics for AI adoption – could be 50/50. Gradually increase the performance KPIs weighting as people see the effectiveness and adopt it more on their own." – Interview 10*

Interview 14 similarly points out that these integration weights would change throughout the AI adoption journey. In the current evaluation scheme, the interviewee suggested assigning roughly 60 percent weight to outcomes and 40 percent to the AI system and its use. At the same time, the interviewee also highlighted the temporal dynamic of this balance by comparing AI skills to Microsoft Office skills, arguing that such capabilities will eventually become a basic minimum requirement. Once this occurs, performance evaluation is likely to shift greater emphasis back toward outcomes.

To summarize, the interview data suggest that the integration logic through which managers evaluate their subordinates' performance is not only multidimensional and averaging in nature, but also temporally evolving, with the relative weights among dimensions changing as AI becomes more embedded in decision-making processes.

5. Discussion

In this chapter, the empirical findings are discussed in relation to the research question in section 5.1. Subsequently, theoretical contributions and managerial implications are presented in sections 5.2 and 5.3, respectively. Finally, the chapter ends with a discussion of the study's limitations and directions for future research in section 5.4.

5.1. Answer to research question

This study set out to answer the research question: ***How do managers evaluate subordinates' performance in AI-supported decision-making contexts?*** From this, two more detailed sub-questions were developed to guide the analysis. Accordingly, this section is structured around these two sub-questions.

Sub-question 1: What performance dimensions do managers emphasize when evaluating subordinates in AI-supported decision-making contexts?

Empirical data show that managers draw on multiple lenses when assessing their subordinates. Aligning with Merchant and Van der Stede's (2003) typology, interviewees consistently refer to all four types of control – results, action, personnel, and cultural controls – when discussing the vignette and reflecting on their own experiences. Results controls remain central: managers pay close attention to whether subordinates create value, meet KPIs, and contribute to financial or operational performance, typically over a track record rather than through a one-off case. At the same time, the presence of AI makes action controls particularly salient. Several interviewees explicitly describe expectations that subordinates follow AI-related procedures, such as using AI in their daily work, validating AI outputs, and consulting multiple AI models rather than relying on a single one.

Personnel controls are reframed in AI-supported decision-making contexts: new AI-related capabilities, such as selecting appropriate models for specific tasks, using AI tools methodically, and structuring decision processes to leverage multiple AI models, emerge as important skills. However, these new capabilities do not replace traditional ones. Domain expertise and general competence remain essential as they enable human decision-makers to retain a leading role in human-AI collaboration rather than becoming dependent on AI. Moreover, as AI is gaining ground in tasks such as data synthesis and analysis, human capabilities in reasoning, critical thinking, and

judgment become even more important. Finally, cultural controls surface in the form of shared norms around AI use. These include expectations that AI should support but not replace human judgment, and that organizational principles such as risk appetite and compliance standards must not be compromised, particularly in risk-sensitive industries.

However, the study shows that the dimensions managers prioritize in their evaluation processes are context-dependent. Across industries, different aspects of performance are emphasized. In finance and other highly measurable, risk-sensitive settings, results controls and risk-related action controls tend to dominate, and AI is framed as a "risk machine" whose recommendations carry considerable weight. By contrast, in IT and engineering contexts, managers place greater emphasis on structured workflows, documentation, learning from rare failures, and developing AI literacy within teams, reflecting a stronger orientation toward process and personnel controls.

Sub-question 2: How do managers combine different dimensions to form a final judgment of their subordinates' performance?

The data indicate that managers anchor their evaluations around outcomes. Specifically, negative outcomes tend to bring other dimensions into consideration, prompting deeper examination of how the decision was made, such as whether required AI tools were used, whether procedures were followed, and whether the subordinate can provide a coherent rationale. In such cases, adherence to process, alignment with norms, and strong reasoning can soften the evaluation, leading managers to interpret the failure as bad luck or as evidence of broader process or system flaws rather than as individual incompetence. Conversely, positive outcomes tend to attract less scrutiny, although they do not entirely eliminate other considerations. Certain critical controls, i.e., risk management in risk-sensitive industries, remain salient and still influence the final judgment.

At the same time, the interviews reveal that when other dimensions align with outcomes, they tend to reinforce rather than substantially change the evaluation. For example, when all dimensions are positive – targets are met, AI is appropriately used, employees demonstrate sound judgment, and norms are followed – managers generally view this as strong performance, without assigning significant additional credit to individual dimensions beyond what is already reflected in the outcome. In contrast, when other dimensions move in the opposite direction of the outcome, i.e., when the outcome is negative but the employee demonstrates diligent adherence to procedures and

strong judgment, these dimensions tend to balance the impact of the outcome in the final evaluation.

Finally, the relative weight assigned to different performance dimensions is both context- and time-dependent. Interview data show that managers across industries perceive different levels of importance for each dimension. Moreover, several interviewees explicitly describe how these weights evolve over the AI adoption journey: in the early phases, they advocate placing greater emphasis on AI adoption KPIs (action controls) alongside traditional performance metrics to encourage learning and experimentation. Over time, as AI capabilities become normalized, evaluation is expected to shift back toward a stronger focus on results.

5.2. Theoretical contributions

This thesis offers three main theoretical contributions. First, it contributes to the management control literature by examining the four control types proposed by Merchant and Van der Stede (2003) in the emerging context of AI-supported decision-making. The study shows that all four types of control remain present in the evaluation process, but the incorporation of AI reshapes both the content and the relative weight of each dimension. In this sense, the study extends the management control literature by demonstrating that, in AI-supported decision-making contexts, existing control mechanisms do not disappear but are transformed in how they are interpreted, prioritized, and integrated into performance evaluation.

Second, the study extends behavioral decision-making and outcome bias research. Prior work shows that evaluators tend to judge decision quality more favorably when outcomes are positive and more harshly when outcomes are negative (Baron & Hershey, 1988; Blank et al., 2015; Ferguson, 2025; Gillenkirch & Velthuis, 2023; Marshall & Mowen, 1993; Peecher & Piercey, 2008). This thesis adds nuance to this literature by showing that outcomes not only influence the final judgment but also shape which evaluative criteria are considered in the first place. Negative outcomes tend to trigger broader scrutiny, bringing process, reasoning, and AI use into consideration. By contrast, positive outcomes often reduce attention to process quality and other criteria unless critical boundaries are clearly violated. In other words, whereas prior research has established that outcome bias affects final judgments, this study suggests that it also affects the composition of evaluation itself, manifesting as an asymmetric pattern in which negative results prompt broader assessment while positive results narrow attention around outcomes themselves.

Finally, the thesis contributes to the human-AI collaboration literature. Prior research focuses largely on how tasks and authority are allocated between humans and AI (Puranam, 2021; Shrestha et al., 2019), how individuals respond to algorithmic advice (Daschner & Obermaier, 2022; Gill et al., 2024; Logg et al., 2019; Ning et al., 2024), and how responsibility is distributed when AI is involved (Elish, 2019; Joo, 2024; Passlack et al., 2024; Zeiser, 2024). This thesis extends that literature by shifting attention to the evaluation stage, showing how human decision-makers are evaluated in AI-supported decision-making contexts. Put differently, while earlier work clarifies how humans and AI should collaborate and how responsibility for outcomes is distributed, this study examines how those collaborations are subsequently evaluated, addressing a critical but previously underexplored aspect of human-AI decision systems.

5.3. Managerial implications

The findings of this study offer several managerial implications for companies that are incorporating AI into decision-making processes. The first implication is that organizations should make the multi-dimensional nature of performance evaluation explicit. Empirical data show that, in practice, even though outcomes tend to dominate, managers already combine results, action, personnel, and cultural lenses when assessing subordinates. In AI-supported decision-making contexts, action and personnel controls become particularly salient, with emerging expectations around AI-specific routines and AI-related skills. Without clearly communicating these expectations, employees may remain unaware of what is expected of them and focus excessively on results to the exclusion of other dimensions, potentially leading to issues in process quality or irresponsible AI use. This is particularly important in the early stages of AI adoption, where new ways of working and high levels of uncertainty create ambiguity that can result in inconsistent practices and misaligned behaviors.

A second implication concerns the design of control systems to better align with the reality of AI-supported decision-making. Organizations can, for example, define how decision processes should be structured when AI is incorporated, including when AI must be consulted and when conflicting recommendations require escalation. In addition, clearly specifying what should not be done is equally important to prevent irresponsible use of AI, as Simons (1995) notes, boundary systems act as organizational brakes and *"like racing cars, the fastest companies need the best brakes."* At the same time, organizations should invest in personnel controls by developing employees' AI-

related capabilities. This includes training on prompt design, awareness of AI limitations such as hallucinations and bias, and skills for validating AI outputs or combining multiple models. Importantly, such efforts should complement rather than replace the development of core domain expertise, which remains essential for employees to critically assess AI-generated content. Finally, shared norms around AI use (i.e., AI should be understood as a powerful assistant, not a scapegoat) should be established to ensure consistent practices across the organization. Designing control systems in this way can support a form of human-AI collaboration in which employees remain accountable and engaged, rather than becoming passive *"rubber-stampers"* of algorithmic output, as one interviewee put it.

A third implication is that performance evaluation criteria and their relative weights should evolve over the AI adoption journey. Empirical data suggest that strict KPI-based evaluation in the early phases of AI implementation may discourage experimentation, as new AI tools do not always yield immediate gains and can initially disrupt established routines. In such phases, organizations may benefit from assigning explicit weight to AI-adoption behaviors, such as experimenting with recommended tools, developing human-AI workflows, and contributing to shared practices, alongside traditional performance indicators. Over time, as AI use becomes normalized, organizations can gradually shift the emphasis back toward results while treating AI-related capabilities as baseline requirements. Designing evaluation systems with this temporal perspective helps avoid two potential problems: on the one hand, maintaining an overly outcome-focused system that discourages learning; on the other, overemphasizing AI use in a way that neglects results, which undermines the very objectives that motivated AI adoption in the first place.

5.4. Limitations and future research

5.4.1. Limitations

This thesis has several limitations that should be acknowledged. First, the study is based on a single vignette design. While this approach is effective in eliciting rich reflections and insights that may otherwise be difficult to uncover through standard interview questions, as Sampson and Johannessen (2020) point out, it also means that the analysis is anchored in one specific scenario. The chosen vignette – an account manager making a sales decision – while realistic and useful for prompting managers to reflect on AI use and performance evaluation, is inherently outcome-oriented and may therefore foreground results more strongly than other dimensions. Consequently,

aspects that may be more salient in other settings, such as recruitment, product development, or operational planning – where action and personnel controls could play a more central role – may be underrepresented in the findings. Future research using multiple vignettes across different decision contexts would therefore be valuable for assessing the stability of the evaluation logics identified in this study and whether the relative emphasis placed on results, action, personnel, and cultural controls varies with the nature of the task.

A second limitation concerns the heterogeneity in AI expertise among interviewees. The study includes managers from different hierarchical levels and industries, in line with Patton's (2015) notion of maximum variation sampling. However, given the limited scope of this thesis, differences in AI knowledge among participants (i.e., between those with deep technical understanding and those who simply use AI as a support tool) and how these differences influence performance evaluation are not systematically analyzed. Tsumura and Yamada (2025) have already demonstrated that AI knowledge affects responsibility attribution in human-AI collaboration; a corresponding effect on performance evaluation is therefore worth exploring. Future research comparing managers with varying levels of AI expertise could examine whether deeper technical understanding leads to different evaluation logics and thereby capture this nuance more systematically.

Finally, the study captures how managers say they evaluate subordinates and how they reason about performance evaluation, but it does not observe actual evaluations in HR systems. This means that the findings reveal managers' evaluative judgments but not necessarily how these judgments translate into formal ratings, compensation decisions, or promotion outcomes in practice. A quantitative study could therefore be valuable in testing whether the patterns identified here are also reflected in observable performance evaluation data.

5.4.2. Future research

Given the limitations acknowledged above, several avenues for future research emerge. First, future studies could incorporate multiple vignettes across different decision contexts, such as recruitment, product development, or operations management, to examine whether the evaluation patterns identified in this thesis can be generalized beyond the current setting. Second, future research could systematically explore how different levels of AI expertise shape evaluative logics and control practices. Third, quantitative studies using organizational data could test whether the

patterns observed in this study are reflected in actual performance ratings, compensation, and promotion decisions. Finally, longitudinal research could examine how evaluation criteria and their relative weights evolve over time as AI becomes more embedded in organizational practices.

6. Conclusion

This thesis set out to explore how managers evaluate subordinates' performance in AI-supported decision-making contexts, focusing on both the performance dimensions they emphasize and how these dimensions are combined to form an overall judgment. Drawing on fourteen semi-structured interviews with managers across industries and hierarchical levels, the study examines performance evaluation with "algorithms in the loop."

The findings show that performance evaluation in AI-supported decision-making contexts is multi-dimensional: managers rely on a combination of results, action, personnel, and cultural lenses. While outcomes remain central, AI as a decision support increases the salience of action and personnel controls. Employees are expected not only to deliver results but also to use AI appropriately, follow structured processes, and demonstrate critical judgment. These expectations are shaped by shared norms that AI remains a support tool and should not replace human judgment. When integrating multiple dimensions to form the final judgment, the study finds that negative outcomes trigger broader scrutiny, bringing process, reasoning, and AI use into consideration, while positive outcomes often receive less attention unless critical boundaries are violated. Overall, managers appear to combine multiple evaluative dimensions through the averaging logic, with the relative importance of each dimension being context- and time-dependent.

By uncovering these patterns, the study makes three main theoretical contributions. It contributes to management control literature by showing how AI reshapes the content and interaction of control types, adds nuance to outcome bias research by demonstrating that outcomes influence not only final judgments but also which criteria are considered in the first place, and extends human-AI collaboration literature by showing how human decision-makers are evaluated in this context. From a practical perspective, the findings suggest that organizations need to rethink how they design and communicate performance evaluation systems in AI-supported decision-making contexts. Making the multi-dimensional nature of evaluation explicit, aligning control systems with human-AI workflows, and adjusting evaluation criteria over the course of the AI adoption

journey can help ensure that employees remain accountable, develop relevant capabilities, and use AI responsibly.

Despite its limitations, particularly the use of a single vignette and the inherent shortcomings of qualitative design, this study provides a foundation for future research. Future work could test the insights uncovered in this study using quantitative studies, explore variations across decision domains, and examine how performance evaluation evolves along the AI adoption journey.

References

Accenture. (2026, January 15). *Accenture Pulse of Change*. https://www.accenture.com/us-en/insights/pulse-of-change

Aman, F., Hamid, S. R. A., Isa, A. M., & Mohamad, M. H. S. (2025). A Systematic Review of Trends in Human–AI Collaboration Research. *The International Journal of Technology, Knowledge, and Society*, *21*(1), 189–217. https://doi.org/10.18848/1832-3669/CGP/v21i01/189-217

Anderson, N. H. (1981). *Foundations of Information Integration Theory*. Academic Press.

Anderson, N. H. (2016). Information Integration Theory: Unified Psychology based on three mathematical laws. *Universitas Psychologica*, *15*(3). https://doi.org/10.11144/Javeriana.upsy15-3.iitu

Baron, J., & Hershey, J. C. (1988). Outcome Bias in Decision Evaluation. *Journal of Personality and Social Psychology*, *54*(4), 569–579.

Blank, H., Diedenhofen, B., & Musch, J. (2015). Looking back on the London Olympics: Independent outcome and hindsight effects in decision evaluation. *British Journal of Social Psychology*, *54*(4), 798–807. https://doi.org/10.1111/bjso.12116

Bleher, H., & Braun, M. (2022). Diffused responsibility: Attributions of responsibility in the use of AI-driven clinical decision support systems. *AI and Ethics*, *2*(4), 747–761. https://doi.org/10.1007/s43681-022-00135-x

Bradbury-Jones, C., Taylor, J., & Herber, O. R. (2014). Vignette development and administration: A framework for protecting research participants. *International Journal of Social Research Methodology*, *17*(4), 427–440. https://doi.org/10.1080/13645579.2012.750833

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

Cappelli, P., & Tavis, A. (2016). The performance management revolution. *Harvard Business Review*, *94*(10), 58–67.

Crant, J. M., & Bateman, T. S. (1993). Assignment of Credit and Blame for Performance Outcomes. *The Academy of Management Journal*, *36*(1), 7–27. https://doi.org/https://doi.org/10.5465/256510

Daschner, S., & Obermaier, R. (2022). Algorithm aversion? On the influence of advice accuracy on trust in algorithmic advice. *Journal of Decision Systems*, *31*(sup1), 77–97. https://doi.org/10.1080/12460125.2022.2070951

Elish, M. C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society*, *5*, 40–60. https://doi.org/10.17351/ests2019.260

Evans, S. C., Roberts, M. C., Keeley, J. W., Blossom, J. B., Amaro, C. M., Garcia, A. M., Stough, C. O., Canter, K. S., Robles, R., & Reed, G. M. (2015). Vignette methodologies for studying clinicians' decision-making: Validity, utility, and application in ICD-11 field studies. *International Journal of Clinical and Health Psychology*, *15*(2), 160–170. https://doi.org/10.1016/j.ijchp.2014.12.001

Ferguson, P. J. (2025). When and Why Do Supervisors' Evaluations Overweight Subordinates' Performance Outcomes? Evidence from a Team Setting in the Field. *The Accounting Review*, *100*(2), 133–159. https://doi.org/10.2308/TAR-2021-0672

Galletta, A., & Cross, W. E. (2013). *Mastering the Semi-Structured Interview and Beyond: From Research Design to Analysis and Publication*. NYU Press.

Gill, A., Gillenkirch, R. M., Ortner, J., & Velthuis, L. (2024). Dynamics of Reliance on Algorithmic Advice. *Journal of Behavioral Decision Making*, *37*(4), e2414. https://doi.org/10.1002/bdm.2414

Gillenkirch, R. M., & Velthuis, L. (2023). Delegated risk-taking, accountability, and outcome bias. *Journal of Risk and Uncertainty*, *67*(2), 137–161. https://doi.org/10.1007/s11166-023-09414-2

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology. *Organizational Research Methods*, *16*(1), 15–31. https://doi.org/10.1177/1094428112452151

Glaser, B., & Strauss, A. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Sociology Press.

Goebel, S., & Weißenberger, B. E. (2017). Effects of management control mechanisms: Towards a more comprehensive analysis. *Journal of Business Economics*, *87*(2), 185–219. https://doi.org/10.1007/s11573-016-0816-6

Green, B., & Chen, Y. (2019). The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW), 50:1-50:24. https://doi.org/10.1145/3359152

Joo, M. (2024). It's the AI's fault, not mine: Mind perception increases blame attribution to AI. *PLOS ONE*, *19*(12), e0314559. https://doi.org/10.1371/journal.pone.0314559

Kim, H., Glaeser, E. L., Hillis, A., Kominers, S. D., & Luca, M. (2024). Decision authority and the returns to algorithms. *Strategic Management Journal*, *45*(4), 619–648. https://doi.org/10.1002/smj.3569

Langley, A. (1999). Strategies for Theorizing from Process Data. *Academy of Management Review*. https://doi.org/10.5465/amr.1999.2553248

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Lukka, K., & Modell, S. (2010). Validation in interpretive management accounting research. *Accounting, Organizations and Society*, *35*(4), 462–477. https://doi.org/10.1016/j.aos.2009.10.004

Marshall, G. W., & Mowen, J. C. (1993). An Experimental Investigation of the Outcome Bias In Salesperson Performance Evaluations. *Journal of Personal Selling & Sales Management*, *13*(3), 31–47. https://doi.org/10.1080/08853134.1993.10753956

McKinsey & Company. (2025, November 5). *The state of AI in 2025: Agents, innovation, and transformation*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai#/

Merchant, K., & Van der Stede, W. (2003). *Management Control Systems: Performance Measurement, Evaluation and Incentives*. Pearson Education.

Ning, X., Lu, Y., Li, W., & Gupta, S. (2024). How transparency affects algorithmic advice utilization: The mediating roles of trusting beliefs. *Decision Support Systems*, *183*, 114273. https://doi.org/10.1016/j.dss.2024.114273

OECD. (2024). Explanatory memorandum on the updated OECD definition of an AI system. *OECD Artificial Intelligence Papers*. https://doi.org/10.1787/623da898-en

Otley, D. (1999). Performance management: A framework for management control systems research. *Management Accounting Research*, *10*(4), 363–382. https://doi.org/10.1006/mare.1999.0115

Ouchi, W. G. (1979). A Conceptual Framework for the Design of Organizational Control Mechanisms. *Management Science*, *25*(9), 833–848. https://doi.org/10.1287/mnsc.25.9.833

Passlack, N., Hammerschmidt, T., & Posegga, O. (2024). It was(n't) me: Vignette experiments on managers' responsibility attribution in AI-advised decision-making. *Behaviour & Information Technology*, *0*(0), 1–30. https://doi.org/10.1080/0144929X.2024.2431050

Patton, M. Q. (2015). *Qualitative research & evaluation methods: Integrating theory and practice.* (4th ed.). SAGE.

Peecher, M. E., & Piercey, M. D. (2008). Judging Audit Quality in Light of Adverse Outcomes: Evidence of Outcome Bias and Reverse Outcome Bias. *Contemporary Accounting Research*, *25*(1), 243–274. https://doi.org/10.1506/car.25.1.10

Pratt, M. G. (2017). From the Editors: For the Lack of a Boilerplate: Tips on Writing Up (and Reviewing) Qualitative Research. *Academy of Management Journal*. https://doi.org/10.5465/amj.2009.44632557

Puranam, P. (2021). Human–AI collaborative decision-making as an organization design problem. *Journal of Organization Design*, *10*(2), 75–80. https://doi.org/10.1007/s41469-021-00095-2

Rubin, H. J., & Rubin, I. S. (2011). *Qualitative Interviewing: The Art of Hearing Data*. SAGE.

Ryan, P., & Dundon, T. (2008). Case research interviews: Eliciting superior quality data. *International Journal of Case Method Research & Application*, *20*(4), 443–450.

Sampson, H., & Johannessen, I. A. (2020). Turning on the tap: The benefits of using 'real-life' vignettes in qualitative research interviews. *Qualitative Research*, *20*(1), 56–72. https://doi.org/10.1177/1468794118816618

Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Organizational Decision-Making Structures in the Age of Artificial Intelligence. *California Management Review*, *61*(4), 66–83. https://doi.org/10.1177/0008125619862257

Simons, R. (1995). Control in an age of empowerment. *Harvard Business Review*, *73*(2), 80–88.

Tsumura, T., & Yamada, S. (2025). Effects of knowledge and importance on responsibility in human-AI decision making. *Scientific Reports*, *16*(1), 2670. https://doi.org/10.1038/s41598-025-32513-w

Wiegmann, L., Conrath-Hargreaves, A., Guo, Z., Hall, M., Kober, R., Pucci, R., Thambar, P. J., & Thiagarajah, T. (2025). Methodological Insights: This is not an experiment: using vignettes in qualitative accounting research. *Accounting, Auditing & Accountability Journal*, *38*(1), 418–440. https://doi.org/10.1108/AAAJ-10-2023-6704

Zeiser, J. (2024). Owning Decisions: AI Decision-Support and the Attributability-Gap. *Science and Engineering Ethics*, *30*(4), 27. https://doi.org/10.1007/s11948-024-00485-1

Appendices

Appendix 1. List of interviews

Interview	Participant Role	Industry	Length of Interview	Date of Interview
1	Project Manager	e-Commerce	51 minutes	2026-03-05
2	Business Development Manager	Truck Manufacturing	30 minutes	2026-03-16
3	Country Manager	Logistics	52 minutes	2026-03-20
4	Concept Leader	Truck Manufacturing	35 minutes	2026-03-20
5	Delivery Manager	e-Commerce	45 minutes	2026-03-25
6	Consultant	Management Consulting	40 minutes	2026-04-02
7	Head of Data & AI	Finance	40 minutes	2026-04-03
8	Project Manager	Food & Beverage	53 minutes	2026-04-03
9	Solution Delivery Director	IT Software	60 minutes	2026-04-04
10	Chief Technical Officer	Financial Technology	45 minutes	2026-04-16
11	Technical Leader	IT Software	42 minutes	2026-04-17
12	Business Analyst Manager	Banking	49 minutes	2026-04-18
13	CxO – Business Transformation	Apparel & Fashion	35 minutes	2026-04-18
14	Project Manager	IT Software	30 minutes	2026-04-22

Appendix 2. Interview protocol

Section 1. Opening

Purpose	*Example prompts*
Introduce the project and obtain informed consent from participants.	• Briefly state the purpose of the study: to explore how managers evaluate subordinate performance when AI is incorporated into the decision-making process. • Address anonymity and confidentiality of data. • Ask for permission to record the interview.

Section 2. Participant background

Purpose	*Example prompts*
Understand the participant's background, identify their evaluation criteria, and explore how they perceive the role of AI in decision-making contexts.	• Could you briefly describe your current role and main responsibilities? • When evaluating subordinates' performance, what kinds of information do you usually draw upon? • Have you, or people in your team, used AI or other support system in decision-making process? If yes, how is it typically used? • How do you think AI influence human performance?

Section 3. Performance evaluation in AI-supported decision-making contexts (using a vignette as a stimulus for discussion)

Theme	*Example prompts*
First reaction and framing	• What stands out immediately to you after reading the case? • If you put yourself in the role of the Sales Manager, what would you want to understand first in order to evaluate Rose's performance?
Focus of evaluation	• To form a view about Rose's performance, what would be the main things that you would consider? • What aspects of the situation matter to you when deciding whether Rose made good or bad decision? - Why is it important? - How does that information help you evaluate the situation? • What would Rose have done to make you think that she did make a good decision, even though the outcome was negative? • Imagine Deal Assist had recommended approving on 90-day terms, and Rose's decision was completely in line with that. How would you view the situation in that case?
Information integration logic	• Assuming Rose strictly follow the process that she's supposed to follow, carefully validate AI's output and consult other sources of information, what do you think could be the reason for the negative outcome? - In that case, how would you evaluate her performance? • Imagine you have a one-on-one conversation with Rose and she can demonstrate strong valid reasoning why she made that decision, how would you evaluate her performance in that case?

Section 4: Closing

Purpose	*Example prompts*
Summarize reflections, clarify broader views, and allow participants to raise additional points.	• Do you think the presence of AI changes how human performance is evaluated, or not really? If yes, in what way? • Is there anything about this topic (evaluating performance when AI is involved) that we haven't touched on, but that you think is important?

Appendix 3. Example of coding process

Themes	**Codes**	**Quotes**
Results controls	Outcome bias; negative results anchor judgment	*"The only way you're going to be viewed positively is if there's a positive outcome. If you make money – positive… If you lose money, I don't care whether you listened to the AI or not – I'm not happy." – Interview 6*
Results controls	Comparing actual performance to pre-agreed targets	*"In the review meeting… they'll list out their achievements against those goals… I would assess whether their work was effective or not based on the goals that were cascaded down and agreed upon." – Interview 12*
Results controls	Emphasis on delivered value	*"I evaluate my people on the value they've brought to the organization, on what they've delivered. That's the effectiveness side, and that's the angle I take." – Interview 13*
Action controls	Attribute blame to process/model when procedures were followed	*"If she did what the sales process says – checked credit, used the tool, escalated if needed – then I'd see it more as a problem with our process or the model than just her." – Interview 2*
Action controls	Mandatory AI use	*"We require 100% of staff to use the internal AI tool… and we expect around 30% productivity improvement per person." – Interview 5*
Action controls	Structured workflow for using AI; multi-step decision process	*"You need a method, a workflow… defining the goal, then planning, then implementation, then reviewing and testing… instead of just telling [AI] one sentence and hoping it answers correctly." – Interview 11*
Personnel controls	Decision-making methodology; structured thinking	*"The first thing I would evaluate is whether you have your own methodology to approach the problem… Not having a methodology leads to decisions that are very impulsive… Things that are feeling-based can't be judged, and they just become a matter of luck." – Interview 7*

Personnel controls	Foundational domain knowledge as prerequisite for effective AI use	*"We still need to be extremely, extremely detailed… foundational knowledge of the user… AI is still just a tool, and the person using it definitely needs to have foundational knowledge in order to use it effectively." – Interview 11*
Personnel controls	AI literacy and tool competence as part of job requirements	*"Now, experience using AI tools is a mandatory requirement in job descriptions. The productivity difference between using and not using it is really quite significant." – Interview 11*
Personnel controls	Prompt design as a key skill when working with AI	*"For me, the priority is the ability to write prompts… For AI to read that prompt and deliver exactly the result you want, that skill is genuinely not easy, not simple." – Interview 12*
Cultural controls	Norm that humans remain accountable despite AI support	*"In our culture AI supports you, but you still own the decision… you can't just say 'the system said so'." – Interview 9*
Cultural controls	Industry shaping norms about AI use	*"When it comes to policies and norms, it's very industry-specific. If you are in insurance or banking, it is very different from, say, military or enterprise information systems." – Interview 12*
Integration logic	Profitable but high-risk decisions evaluated negatively	*"If you make money but you put the firm at high risk – negative, because I don't want unnecessary risk." – Interview 6*
Integration logic	Combining outcome, risk, and AI-related process to form the final judgment	*"But if you do a good job, listen to the AI, and make money – I'm happy. If you don't listen to the AI but still make money – I'm happy. If you lose money, I don't care whether you listened to the AI or not – I'm not happy." – Interview 6*
Integration logic	Evaluation asymmetry	*"Sometimes AI makes people lazy. If the system says approve and it works out, no one checks you as thoroughly as when something goes wrong." – Interview 6*
Integration logic	Weighting between outcomes and AI-adoption metrics changes over time	*"If the goal is for AI to help the organization improve, then ultimately it still comes down to performance and KPI. But… in the early phase, split the weight between performance KPIs and metrics for AI adoption – could be 50/50. Gradually increase the performance KPIs weighting as people see the effectiveness and adopt it more on their own." – Interview 10*
Integration logic	Weighting between outcomes and AI-related metrics	*"I'd say maybe 60% is about outcome and 40% about how they use the AI and whether they follow the framework we set up." – Interview 14*

Appendix 4. AI disclosure

AI tools employed	The thesis made use of ChatGPT 5.3, Claude Sonnet 4.6, Gemini 3.1, Perplexity, and Voicenotes (an AI-powered transcription tool).
Contribution to the thesis	AI tools help improve both the speed and quality of the research process. • Speed: AI was used to assist with mechanical and time-intensive tasks, including, among other, transcribing interviews, translating interview transcripts from Vietnamese to English, correcting grammatical errors and typos in the text, rephrasing and smoothing transitions between ideas. • Quality: AI served as a "thinking partner" during the research process: AI tools were asked to criticize ideas, challenge assumptions, and explore alternative perspectives. This process supported deeper analytical thinking and helped refine the arguments. Furthermore, by delegating routine tasks to AI, the author was able to allocate more time and cognitive effort to higher-level reasoning and synthesis of findings, which ultimately improve the thesis quality.
Risks and mitigation measures	Key risks included the potential for overreliance on AI-generated outputs, which could negatively affect the quality of the thesis and reduce authorial integrity. To mitigate this risk, AI was consistently treated as a support tool whose outputs were considered as one input among others in the research process and were not used without critical evaluation. All AI-generated outputs were carefully reviewed, validated, and revised by the author where necessary. The author retains full responsibility for the content and conclusions presented in this thesis.
Insights gained from the process	The process of using AI for this thesis demonstrates that AI is particularly useful in synthesizing large amounts of information as well as performing clearly defined tasks. At the same time, the process also underscored the importance of continuous oversight: AI-generated contents require careful review and validation as they may contain incorrect information, weak reasoning, or hallucinations. AI should be used to augment, rather than substitute, human judgment. Overall, the experience suggests that AI can meaningfully enhance the research process when used thoughtfully.