LUNDS UNIVERSITET
Ekonomihögskolan
Institutionen för informatik

AI in the Shadows:

Governance Mechanisms Organisations Use to Manage the Risks of Unauthorised AI Tools

Kandidatuppsats 15 hp, kurs SYSK16 i Informatik

Författare: David Bergh
Amelie Hörnfeldt
Peggy Schnitzer

Handledare: Umberto Fiaccadori

Rättande lärare: Blerim Emruli
Markus Lahtinen

AI in the Shadows: Governance Mechanisms Organisations Use to Manage the Risks of Unauthorised AI Tools

SVENSK TITEL: AI i skuggorna: Styrningsmekanismer som organisationer använder för att hantera riskerna med icke-godkända AI-verktyg

FÖRFATTARE: David Bergh, Amelie Hörnfeldt and Peggy Schnitzer

UTGIVARE: Institutionen för informatik, Ekonomihögskolan, Lunds universitet

EXAMINATOR: Osama Mansour, Docent

FRAMLAGD: maj, 2026

DOKUMENTTYP: Kandidatuppsats

ANTAL SIDOR: 156

NYCKELORD: Shadow AI, Organisational Risk, AI Governance Mechanisms, AI Policy, AI Adoption

ABSTRACT:

The rapid adoption of generative AI in organisations has given rise to a growing phenomenon known as Shadow AI, which refers to the informal and unsanctioned use of AI tools by employees outside organisational oversight. This study examines what governance mechanisms organisations in Sweden, using Microsoft Copilot, implement to manage risks associated with Shadow AI. A qualitative research approach was adopted, combining semi-structured interviews with fourteen representatives from seven organisations and documentary analysis of internal AI policies. The findings show that organisations implement governance mechanisms across both technical and organisational dimensions, and that neither category is sufficient on its own. The primary risks identified in relation to unsanctioned AI tools were data leakage, output quality and over-reliance, and vendor and supply chain exposure. The study further suggests that Shadow AI is best understood as a structural outcome rather than individual non-compliance, and that restrictive governance approaches risk pushing unsanctioned use further from organisational oversight rather than reducing it.

Content

Figures

Tables

1. Introduction

In the introduction, a background is presented for the research topic, followed by the problem formulation and the research question, together with its sub-questions. The chapter also outlines the study's purpose and its delimitations.

1.1 Background

Artificial intelligence has, over the last few years, moved from something discussed mostly in technology circles to something ordinary employees encounter in their daily work (Batool et al., 2025; Dwivedi et al., 2023).

A central focus of AI research has long been the replication of human intelligence, and while the concept itself has existed for over seventy years, what AI looks like today is fundamentally different from earlier versions (Gbadegeshin et al., 2021). What has particularly driven the most recent shift is the emergence of a category called *generative AI*. Generative AI refers to algorithms and models that resemble human-like creativity and adaptability. In practical terms, such systems are capable, among other things, of producing written reports, drafting emails, summarising meetings and responding to questions in fluent, contextually relevant language (He et al., 2025).

A major turning point came with the launch of ChatGPT by the American company OpenAI in November 2022. Shortly after its release, it was adopted across a remarkably diverse range of users for a broad variety of tasks (Dwivedi et al., 2023). The pace of AI adoption and integration into modern organisations has accelerated considerably, changing the way work is done across industries (Shao et al., 2025).

Microsoft Copilot has become among the most widely adopted enterprise AI systems in use today, with nearly 70 percent of Fortune 500 companies having implemented it by late 2024 (Shaw, 2024). Built directly into the Microsoft 365 software suite, Copilot works from within the tools that many organisations already rely on daily, allowing users to ask questions or assign tasks in natural language. What makes Copilot particularly powerful, and also particularly sensitive from an organisational perspective, is that it not only draws on general knowledge but also reads and draws on the organisation's internal data, specifically files, emails and documents that the individual user already has permission to access (Microsoft, 2026d).

As a result of the rapid development of AI tools and their increased accessibility, a new phenomenon has been growing in workplaces: Shadow AI. This term refers to the informal and unsanctioned use of AI tools by employees within organisations, outside the knowledge and control of IT departments or management (Raizada et al., 2026). Silic et al. (2025) explain Shadow AI as an extension of Shadow IT, meaning the broader practice of employees using technology that has not been approved by their organisation. What makes Shadow AI particularly concerning is the vast capacity of AI systems to handle sensitive information, produce complex outputs and influence decisions well beyond what earlier unsanctioned tools could do. Because these AI tools have not been reviewed or configured by the organisation, there is limited control over what information is being shared or where it ends up, exposing organisations to data breaches, regulatory non-compliance and security vulnerabilities (Puthal et al., 2025).

This creates a significant governance challenge for organisations. Traditional IT governance frameworks were not built with generative AI in mind and struggle to address the specific risks it introduces, including data leakage, inaccurate outputs and accountability gaps (Crocco et al., 2026). It is against this backdrop, one of rapid AI adoption, the specific demands of tools like Microsoft Copilot and the parallel spread of informal AI usage among employees, that this study is situated.

1.2 Problem Formulation

The governance challenges introduced by generative AI are increasingly recognised in both academic literature and regulatory practice. International organisations, including the European Union and the National Institute of Standards and Technology, have developed frameworks to guide responsible AI deployment. It reflects a broad consensus that the risks associated with AI-generated outputs, such as factual inaccuracy, data exposure and accountability gaps, require active organisational management (Batool et al. 2025). Yet, despite this growing regulatory attention, research examining how organisations actually govern these risks in practice remains limited, with existing literature predominantly addressing governance at national and international levels rather than the organisational (Batool et al., 2025).

This gap is particularly evident with respect to Shadow AI. While the phenomenon has received growing scholarly attention, existing frameworks tend toward normative and policy-level guidance, overlooking the decentralised, behaviour-driven nature of informal, user-initiated tool use in practice (Anthuvan et al., 2025). The studies that do exist are largely confined to specific national or sectoral contexts. Lamchek and Trieu (2026), whose study is confined to the deployment of Copilot in Australian government agencies, leave the question of how governance mechanisms operate in private sector and international organisational contexts open, a gap that this study seeks to address. Wang and Cao (2025) similarly base their findings on Chinese firms, and the applicability of their conclusions to other national and organisational contexts remains to be examined.

The Swedish organisational context of this study adds further relevance. AI adoption among Swedish organisations is advancing rapidly while governance frameworks are still being developed and implemented. At the same time, Microsoft Copilot has become one of the most widely deployed enterprise AI tools in Microsoft (Shaw, 2024), yet there is a gap in research of studies examining how organisations using it manage the specific governance challenges it introduces in Sweden, including the parallel spread of unsanctioned AI tools among employees.

Ross et al. (2025) state that "*Further research is essential to deepen understanding of Shadow AI's operational dynamics, organisational responses, and broader societal impacts.*" There is therefore a need to examine what governance mechanisms organisations implement to manage the risks connected to Shadow AI, how these mechanisms address the specific properties of tools like Microsoft Copilot, and whether a relationship exists between the strictness of those mechanisms and the extent to which Shadow AI is perceived to occur. This study addresses that need.

1.3 Research Question and Sub-questions

In order to understand the connection between the extent of governance mechanisms and the participants' perceived Shadow AI usage within their organisation, this study is guided by the following research question and accompanying sub-questions:

Research question: What governance mechanisms do organisations using Microsoft Copilot implement to manage risks connected to Shadow AI?

Sub-question 1: What risks are identified with the use of unauthorised AI-tools?

Sub-question 2: Which governance mechanisms appear to be most effective in managing Shadow AI risks?

1.4 Purpose

The purpose of this study is to explore and develop an understanding of how Swedish organisations manage the governance challenge that arises from Shadow AI in the context of Microsoft Copilot adoption. Perspectives from senior and specialist roles across multiple organisations, spanning IT, legal, strategy and AI leadership functions, will be examined through qualitative interviews, complemented by documentary analysis. This study aims to contribute to the limited existing research on Shadow AI governance and provide practical insights for organisations navigating the balance between AI accessibility and risk control.

1.5 Delimitations

The scope is limited to organisations that have adopted Microsoft Copilot as one of their sanctioned systems, allowing the study to remain focused and manageable within the time available.

The study does not include employees at all levels of the organisations investigated. Participants were selected based on their roles in relation to AI and information security, as these individuals were considered most likely to possess direct insight into organisational governance practices. The perspectives of general end-users without governance responsibilities consequently fall outside the scope of this study.

The study does not examine the technical properties, architecture or model behaviour of Microsoft Copilot or the AI systems underlying it in depth. Where technical aspects of the tool are described, this is done briefly and only to the extent necessary to establish an understanding of the governance challenges the tool introduces. The analytical focus remains strictly organisational, meaning how organisations manage, regulate and respond to the use of AI tools at a policy and process level rather than a technical one.

Finally, the empirical material is drawn from a Swedish organisational context and the data collection took place during a defined and relatively short period in the spring of 2026. The study therefore captures a snapshot of organisational governance practice at a specific moment, and the findings may not reflect developments that occur as the technology and its regulatory environment continue to evolve.

2. Literature Review

This chapter reviews existing research and theoretical concepts relevant to the study. Through an examination of prior literature, four interconnected areas are addressed around the topics: Artificial intelligence and Shadow AI, AI governance and regulatory frameworks, Organisational risk management and resilience, and Human-AI interaction and automation bias. Together, these areas form the conceptual foundation for the study's analytical framework.

2.1 Artificial Intelligence

2.1.1 Defining Artificial Intelligence and Generative AI

Artificial intelligence can be defined in many ways, with no single, universally agreed upon definition. The term artificial intelligence has been discussed for over seventy years, and what AI was like before differs from what it is today, reflecting the rapid pace of constant technological development. A central focus of AI research has been the replication of human intelligence (Gbadegeshin et al., 2021). To be concrete, we can base our definition of AI on Gbadegeshin et al.'s (2021) study, defining it as a system created and initially trained by humans that learns on its own and uses that knowledge to perform assigned tasks.

AI enables systems within organisations to expand rapidly. It has transformed businesses and manufacturers and has been used in ways that only humans could do before (Dwivedi et al., 2021). Of particular relevance to this study is the emergence of generative AI. Generative AI is a part of Artificial Intelligence, referring to the algorithms and models that resemble human-like creativity and adaptability (He et al., 2025). Unlike earlier AI technology, generative AI tools are designed to be accessible to a broader range of users, as they can operate through natural language rather than technical expertise (He et al., 2025).

2.1.2 Large Language Model behaviour

The most widely adopted generative AI tools are built on a specific architecture known as large language models (LLMs). A defining characteristic of large language models, the architecture that underlies generative AI tools, is that they generate outputs by predicting statistically probable tokens rather than by retrieving or verifying facts. Each word in a response is selected because it represents the most likely continuation of what precedes it, meaning the model is optimising for linguistic coherence rather than factual accuracy (Lamchek & Trieu, 2026). This is why a response can read as fluent, contextually

appropriate, and confident while nonetheless containing incorrect or fabricated information, a property referred to in the literature as *hallucination*. For organisations deploying generative AI in knowledge-intensive work, this carries significant governance implications, as users may be unable to distinguish a correct output from an incorrect one without independently verifying the content against authoritative sources (Lamchek & Trieu, 2026).

2.1.3 Growth of Artificial Intelligence in the Workplace

Artificial Intelligence has been present in organisations for a long time and has progressively transformed how people and businesses operate (Dwivedi et al., 2023). In its earlier stages, AI was much more restricted to processing structured data, such as spreadsheets and numerical datasets. This was limiting both its flexibility and availability. Over time, advances in deep learning and machine learning enabled AI to process unstructured data, such as images and audio. In recent years, AI has improved through continuous learning and has increasingly been capable of interpreting and responding to information in ways that resemble human reasoning (Dwivedi et al., 2023).

The launch of OpenAI's LLM ChatGPT in November 2022 significantly expanded the capabilities of earlier AI tools by delivering human-like responses across a wide range of tasks. ChatGPT reached over one million users within the first five days and became widely adopted across workplaces globally, supporting tasks such as writing business letters, essays and developing software (Dwivedi et al,. 2023). Shortly after its release, it was adopted by a remarkably diverse range of users, illustrating how generative AI, compared with earlier technology, could be utilised easily without prior technical expertise (Dwivedi et al., 2023).

The adoption and integration of AI into modern organisations and workplaces have accelerated considerably in recent years, changing the way work is done across industries (Shao et al., 2025). The growing use of AI tools has actively supported employees in daily work tasks such as decision-making, knowledge support and insight generation (Shao et al., 2025). This integration has been shown to bring benefits such as improving efficiency and work progress while simultaneously causing uncertainty and confusion in leadership and responsibilities (Shao et al., 2025). Adding to this complexity, the broad adoption has raised concerns about the generation of inaccurate content from incorrectly reported sources, which has raised the question of the organisational need for careful oversight. (Dwivedi et al., 2023).

2.1.4 Microsoft Copilot

Microsoft Copilot is a generative AI assistant developed by Microsoft, integrated directly into the Microsoft 365 application suite. Rather than functioning as a standalone tool, Copilot operates within the applications employees already use, enabling tasks such as drafting documents, summarising meetings and retrieving information through natural language

prompts. Users enter a prompt and Copilot provides real-time assistance by working alongside Microsoft 365 apps and services (Microsoft, 2026d).

Microsoft Copilot operates within an organisation's adopted Microsoft 365 environment and is governed by role-based access controls, meaning it can only retrieve data that the individual user already has permission to access (Microsoft, 2026a). This makes Copilot's performance directly dependent on the quality of the organisation's underlying permissions structure.

Copilot can search the web to supplement its responses. When enabled, it generates a short search query that is sent to the Bing search service, which retrieves content from its pre-indexed archive of web pages rather than accessing live websites in real time (Microsoft, 2026c). User and tenant identifying information is removed from the query before it is sent, and the query is treated as customer confidential information (Microsoft, 2026c).

Copilot runs on large language models developed by OpenAI, hosted within Microsoft's Azure cloud environment, which keeps data within the organisation (Microsoft, 2026a). Interaction data, including audit records for prompts, responses and referenced content, is stored within Microsoft 365 services, where it can be discovered, audited and retained. This data is encrypted and is not used to train the foundation large language models used by Copilot (Microsoft, 2026d). This distinguishes it from unsanctioned personal tools, where data entered by users may be used to further train the underlying model and is therefore potentially exposed beyond the organisation's control (Lamchek & Trieu, 2026).

Administrators manage access and configuration through the Copilot Control System, which provides controls to monitor, detect and respond to AI-related risks. Critically, these mechanisms require deliberate administrative configuration, meaning that the effective governance posture of a Copilot deployment is determined not by the platform alone but by the governance maturity of the deploying organisation (Microsoft, 2026c; 2026b).

2.1.5 Shadow AI

2.1.5.1 Defining Shadow AI

Shadow AI refers to the informal and unauthorised use of artificial intelligence tools within organisations which are not explicitly approved, monitored, or governed from established IT, security, or compliance structures (Raizada et al., 2026). Puthal et al. (2025) further state that such use often occurs when employees or teams adopt easily accessible AI solutions in order to tackle particular problems or to automate processes without permission, thereby placing these activities outside normal governing frameworks. Silic et al. (2025) position shadow AI as a contemporary evolution of shadow IT, but one which introduces enhanced governance concerns due to the vast capacity of AI systems to handle sensitive information, produce complex outputs and influence work processes and decisions beyond organisations' oversight. Therefore, shadow AI may be understood not only as unapproved AI technology use, but also

as a sociotechnical governance problem associated with privacy, cybersecurity, compliance and accountability risks (Kwan, 2024; Puthal et al., 2025).

2.1.5.2 Shadow AI Occurrence

Anthuvan et al. (2025) express that employees who are independently using AI tools such as ChatGPT, Microsoft Copilot and Gemini through personal accounts and often on personal devices to complete work tasks, do so in attempts to work more efficiently, avoid slow processes and compensate for vague AI guidelines. Anthuvan et al. (2025) issued a survey of 345 participants which targeted digitally active professionals across sectors where integration of generative AI is most prominent. This survey reported that 41.4% accessed AI through personal accounts, 27.0% through a combination of personal and official accounts, whereas only 26.4% relied exclusively on organisation-provided access. The same study showed that these tools were most frequently used for email drafting, creating presentation slides, research and summarisation, brainstorming ideas and data analysis, all used by over 40% of respondents. Adding to this, Silic et al. (2025) note that employees turn to unsanctioned tools not simply out of convenience but also because they perceive official tools as misaligned with their work requirements. A finding that the transition from Shadow IT to Shadow AI suggest remains applicable in the current concept.

Beyond individual behaviour, organisational culture plays a significant role in shaping the prevalence of Shadow AI. Ross et al., (2025) argue that organisations with cultures of innovation and continuous learning act as enablers for informal AI adoption. Employees in such environments are often empowered to experiment with new technologies, which can sometimes lead to bypassing formal processes in doing so. Therefore, a highly accepting and open approach can work against governance, as responsible AI initiatives can fall flat when organisational culture does not align with ethical AI principles (Ross et al., 2025). The authors further note that misalignment between stated values and actual practices, such as unclear accountability or lack of communication, can create an environment where Shadow AI goes largely undetected.

2.1.5.3 AI and Shadow AI Risks

The use of AI tools in organisational settings introduces a range of risks that go beyond technical malfunction. Such tools may produce content that is inaccurate, biased, or misleading. Because such errors are not shown within the output itself, they can disperse undetected into business decisions, reports and client-facing communications (Raizada et al., 2026). Hallucination is a particularly significant example of this, as outputs can appear authoritative and well-formed while containing fabricated information that users have no immediate way of identifying. Beyond output quality, the broad deployment of AI tools across organisational workflows raises concerns about data exposure, intellectual property and regulatory compliance (Anthuvan et al., 2025).

Puthal et al. (2025) argue that unsanctioned AI use occurs outside adequate security, accountability and transparency mechanisms. Therefore, it exposes organisations to data breaches, regulatory non-compliance, malicious threats, unauthorised computation of sensitive data, safety vulnerabilities in unmonitored models and model poisoning. Raizada et al. (2026) similarly characterise Shadow AI as a source of regulatory, operational and reputational risk and identify four central risk domains: data leakage, model hallucination, compliance breaches and shadow process automation. Anthuvan et al. (2025) reinforce the significance of these concerns empirically, reporting that 42% of survey respondents expressed moderate to high concern about risks such as data leakage and AI hallucinations.

A major risk category is data leakage and confidentiality loss. Puthal et al. (2025) state that Shadow AI often involves the unauthorised use of sensitive data and complicates compliance with data protection requirements because unsanctioned tools operate outside formal controls. Raizada et al. (2026) develop this further by arguing that employees may input protected corporate information into public LLMs for tasks such as summarising reports or generating code, thereby creating a compliance blind spot around both data privacy and intellectual property. In the same article, the authors note that the lack of transparency of public models makes it difficult for auditors to establish where enterprise data is stored or how it is used. Their results section identifies data leakage as the most frequent risk category in their dataset, accounting for 61% of recorded risk instances. Ross et al. (2025) extend this further by highlighting that data leakage risks are amplified when third-party platforms are involved. The study notes that leakage may occur not only through user prompts submitted to external models but also through data retained on third-party platforms after use and through the untracked sharing of AI-generated content.

A second category is output risk, which refers to the possibility of AI tools generating incorrect, misleading, sensitive or discriminatory outputs that are then incorporated into organisational work. Raizada et al. (2026) emphasise hallucinations as one of the most pressing risks of unmonitored LLM use, especially because such systems can produce factually incorrect but confident-sounding information. They note that, in professional settings such as legal and medical work, these inaccuracies may lead to liability and documentation errors and may also contaminate the organisation’s internal knowledge base when errors go undetected. The same article further states that LLMs may generate sensitive information or spread discriminatory or biased outputs, with potential reputational, legal and security consequences if those outputs shape business decisions. In their reported risk frequencies, hallucinated content accounts for 53% of cases. Puthal et al. (2025) likewise note that Shadow AI creates safety vulnerabilities in unmonitored models and increases the difficulty of maintaining trustworthy and controlled use of AI.

A third risk category is compliance and legal exposure. Puthal et al. (2025) explicitly connect Shadow AI to regulatory non-compliance and explain that unsanctioned AI systems make it more difficult to uphold data protection and compliance obligations. Raizada et al. (2026) state that LLMs used in regulated sectors may breach directives such as The General Data Protection Regulation (GDPR) or internal audit protocols. They also argue that the legal risk

of Shadow AI is heightened by the lack of traceability. When a business decision is based on an unrecorded AI interaction, the lack of an audit trail makes it difficult to defend the rationale behind that decision during an audit or legal challenge. Their results further identify compliance violations in 47% of recorded cases, underscoring that this is a central risk of Shadow AI use.

A fourth risk category is cybersecurity exposure and attack-surface expansion. Puthal et al. (2025) argue that Shadow AI broadens the threat landscape by introducing unmonitored systems into organisational infrastructures, thereby increasing the number of interfaces available for adversaries to exploit. They specifically identify data and security breaches, malicious threats, model poisoning and vulnerabilities associated with unmonitored models. Raizada et al. (2026) add an LLM-specific cybersecurity perspective by highlighting prompt injection and leakage attacks as important threats in unmanaged interfaces. Such interfaces may lack the security layers required to prevent malicious prompts from manipulating outputs or extracting sensitive system instructions. Their reported results list prompt injection exposure in 22% of cases, indicating that this risk is less common than data leakage or hallucination, but still significant (Raizada et al., 2026).

Finally, risks related to governance, accountability and operations are also identified. Raizada et al. (2026) argue that unmonitored AI use can result in “untraceable decision-making,” which threatens corporate governance and accountability because organisations cannot reconstruct how AI-supported decisions were reached. The same article also identifies shadow process automation as a distinct risk domain, indicating that the issue is not limited to isolated prompts but may become embedded in workflows and internal processes. Anthuvan et al. (2025) complement this by showing that risk perception was the only statistically significant predictor of AI tool avoidance in their survey, whereas policy awareness and role level were not significant predictors.

2.2 Organisational Risk Management and Resilience in AI Adoption

2.2.1 Organisational Risk

Risk, in an organisational context, is understood as a composite of two elements: consequences and uncertainty. Consequences refer to the potential outcomes of an event that affect organisational value, such as data integrity, operational continuity, legal standing, or reputation. Uncertainty concerns the incomplete knowledge of whether, when or how severely those consequences will become reality (Aven, 2016). This distinction matters because, as Aven (2016) argues, reducing risk to probability alone yields an incomplete picture that fails to account for surprises that fall outside modelled scenarios.

The International Organization for Standardization (2018) describes in ISO 31000:2018 risk management as the coordinated set of activities through which an organisation systematically identifies, analyses, evaluates and treats risks to protect its objectives and operations. They position risk management not as a periodic compliance exercise but as an ongoing organisational practice integrated into decision-making at every level (ISO, 2018). In the context of AI, National Institute of Standards and Technology (NIST) adopts the same definition of risk, drawing directly on ISO 31000:2018 as its source (NIST, 2023).

The implementation of AI in organisations has contributed to both significant opportunities and considerable risks (Crocco et al., 2026). AI adoption has created many competitive advantages in the workplace, while simultaneously opening new categories of risk that organisations must actively manage. These risks can be broadly categorised as conceptual risks related to moral and legal harm and practical risks, such as unpredictability of AI outputs and unregulated use of AI tools within an organisational setting (Crocco et al., 2026).

Risk management can, in general, serve as an effective tool for organisations, enabling managers to identify and mitigate risks before they become significant threats (Crocco et al., 2026). An important observation in this context is the tendency for employees to use AI tools more independently with personal accounts and devices outside the formal organisation's oversight to complete tasks (Anthuvan et al., 2025). Effective risk management requires more than restricting compliance-based approaches. It requires governance frameworks that are responsive to real-world use and position employees as responsible in AI integration, rather than as potential rule-breakers (Anthuvan et al., 2025).

2.2.2 Organisational Resilience as Adaptive Capacity

Organisational resilience encompasses the capabilities needed to both proactively prepare and reactively recover from disruptive events (Shojaee et al., 2025). Organisational resilience is a concept that has gained increasing relevance as AI adoption reshapes how organisations operate. While the term is often used broadly, its meaning is more complex than it first appears. Phillips and Chao (2024) note that resilience is commonly described in terms of recovery, adaptability and homeostasis, yet it goes beyond simply returning to a prior state. Resilience involves learning and transforming in order to realign with organisational goals following a disruption (Phillips & Chao, 2024). Furthermore, Francis and Bekera (2014) define resilience, similar to Phillips and Chao (2024), as a system’s capacity to absorb, adapt and recover from disruptive events. However, a central part of their study is the emphasis that resilience is less about predicting and preventing and more about flexibility and adaptive response. This reflects a mindset that disruptions are inevitable and that the priority is to maintain function and recover effectively when they happen.

Emerging technologies generally strengthen firm resilience according to Wang and Cao (2025). Simultaneously, AI adaptation makes resilience more demanding and important. AI’s technical characteristics, such as variability, complexity, ambiguity and uncertainty, impose greater demands on an organisation's ability to prepare and respond to disruptions than

traditional technologies (Wang & Cao, 2025). Furthermore, AI adoption creates new forms of information asymmetry within organisations, which can complicate knowledge sharing and hinder clear communication about how AI is being used and where potential risks arise (Wang & Cao, 2025).

2.3 AI Governance

2.3.1 AI Governance: Definition and Challenges

Governance, in this context, refers to the rules, policies, processes and technical controls that organisations put in place to manage how technology is used, determining who can use what, under what conditions and with what safeguards in place (Batool et al., 2025).

AI governance encompasses regulations, methods, procedures and technological mechanisms to ensure that an organisation's development and deployment of AI technologies align with its strategies, principles and goals. As AI becomes embedded in organisations and drives innovation, it simultaneously introduces risks that must be mitigated. In response, governmental and international organisations, including the European Union, ISO and NIST, have published ethical AI principles and guidelines (Batool et al., 2025).

However, governing AI in practice is considerably more complex than principle-based definitions suggest. Zaidan and Ibrahim (2024) highlight that national regulatory efforts have been criticised for failing to establish consistent standards, sufficiently address the core challenges of AI and ensure compliance with rules. The governance of these fields is, as the authors conclude, highly complex and likely to remain so, as governments continue to navigate the difficult balance between enabling innovation and establishing adequate societal protection through law (Zaidan & Ibrahim, 2024).

2.3.2 Accountability as a Governance Concept

Within the complex regulatory landscape, a central question concerns who is responsible for ensuring that AI systems are developed and used appropriately. Horneber and Laumer (2023) address this through the concept of algorithmic accountability. Algorithmic accountability focuses on the question of who bears the obligation to justify the design, use and outcomes of machine learning systems and who assumes responsibility for negative consequences. Algorithmic accountability functions as a governance mechanism through which organisations can either proactively avoid the negative impacts of machine learning systems or reactively sanction accountable actors when negative effects have occurred (Horneber & Laumer, 2023).

Horneber and Laumer (2023) identify four types of accountability in relation to AI systems. Social accountability concerns the ability of users and affected individuals to demand

justifications and impose consequences on organisations through collective or individual action. Institutional accountability regards the role of legal and administrative institutions, such as regulators and certification bodies, in enforcing accountability through legislation and auditing. Organisational accountability refers to the internal governance measures through which organisations align their AI practices with their values and external requirements, including governance instruments such as development guidelines, internal audits and impact assessments. Technical accountability concerns the interpretability and controllability requirements that developers must address through design, in particular design transparency and auditability (Horneber & Laumer, 2023).

A significant observation from Horneber and Laumer (2023) is that most organisations currently rely on a principle-based approach to govern their AI systems, which has been criticised as insufficient due to the difficulties involved in translating ethical principles into practice. The authors argue that, by building on such guidelines, organisations can specify standardised processes and rules for developing, testing, deploying and operating their AI systems and use them as the basis for internal audits.

2.3.3 Regulatory Frameworks

The most consequential regulatory development affecting AI governance in the European context is the EU Artificial Intelligence Act (Regulation (EU) 2024/1689) , which entered into force in August 2024. The Act establishes a four-tier risk classification system: unacceptable risk, high risk, limited risk and minimal risk, with compliance obligations scaled accordingly. Bolgouras et al. (2025) examine how the AI Act interacts with the General Data Protection Regulation and the Network and Information Security Directive (NIS2), identifying convergence points around transparency, accountability and security. Their analysis finds that aligning these regulations is pivotal for building user trust and mitigating risks associated with data breaches, algorithmic bias and privacy violations (Bolgouras et al., 2025).

At the international level, the NIST AI Risk Management Framework (NIST, 2023) provides a voluntary governance structure that organisations can apply regardless of where they operate. Unlike regulatory instruments that impose compliance obligations, the NIST framework is designed to support organisations in building internal processes for identifying, assessing and responding to AI-related risks (Dotan et al., 2024).

2.3.4 Data Protection and Compliance in AI Deployment

Puthal et al. (2025) identify regulatory non-compliance as a central risk of Shadow AI, noting that unsanctioned tools complicate compliance with data protection laws, including GDPR. This has direct implications for organisational accountability. GDPR states that organisations remain legally responsible as data controllers for how personal data is processed, regardless of whether that processing was authorised (Regulation (EU) 2016/679) .

GDPR establishes the legal framework for the processing of personal data within the European Union and applies to all organisations that handle the personal data of EU residents, regardless of where those organisations are based (Regulation (EU) 2016/679). The regulation is built on several core principles that are directly relevant to AI governance. These principles are: data must be processed lawfully, transparently and for specified purposes; only the minimum data necessary for a given purpose should be collected; and individuals retain rights of access, correction and erasure over their personal data. Organisations acting as data controllers bear primary accountability for ensuring these principles are upheld throughout their operations. Where employees submit personally identifiable or sensitive data to unsanctioned AI tools, these principles are violated without the organisation's knowledge. This makes Shadow AI not only a security risk but a legal liability under EU law.

2.4 Human-AI Interaction and Automation Bias

2.4.1 Defining Automation Bias

Automation bias refers to the tendency of people to rely excessively on automated systems, preferring automated recommendations over their own judgment (Romeo & Conti, 2026), failing to notice errors or inconsistencies, as well as failing to intervene when necessary (Dula et al., 2024).

2.4.2 Automation Bias in Generative AI

Romeo and Conti (2026) argue that in the context of artificial intelligence, automation bias refers to the fact that users may perceive AI-generated information as particularly reliable, and therefore fail to verify it before making decisions. Automation bias is a central challenge in human-AI collaboration, especially as AI systems are increasingly incorporated into high-stakes domains (Romeo & Conti, 2026). In relation to generative AI specifically, Wingerter et al. (2025) state that generative AI, particularly LLMs, can significantly augment human decision-making, but also introduces automation bias. In their study, participants who received faulty AI support performed significantly worse than their control group, which the authors state shows that users often uncritically accept AI outputs.

2.4.3 Contributing Factors

Romeo and Conti (2026) state that traditional perspectives explain automation bias through over-trust in automation or attentional constraints, which results in a perception of AI-generated outputs as reliable. Their own review further identifies additional interacting factors, including AI literacy, level of professional expertise, cognitive profile, developmental trust dynamics, task verification demands and explanation complexity. They also state that, across the studies included in their review, trust and task complexity were confirmed as

prevalent factors in automation bias, and that overreliance is reinforced by high workload, time pressure and task complexity.

Regarding user knowledge and experience, Romeo and Conti (2026) state that higher levels of AI literacy enable appropriate reliance on automation, and that professional experience and domain expertise are among the most protective factors, although experts can still be affected under conditions of high cognitive load. They further state that verification-related cognitive engagement is critical as a debiasing mechanism and that higher verification intensity helps counteract errors. In the personnel-selection study cited in their review, Kupfer et al. (2023) found that information about possible system errors increased verification intensity and that less aggregated data was important for reducing automation bias and enhancing decision quality. They also state that individual differences significantly influence the manifestation of automation bias, including need for cognition, familiarity with AI, prior exposure to AI and conceptions about how AI works (Romeo & Conti, 2026).

2.4.4 Governance Implications and Regulatory Context

Romeo and Conti (2026) state that output explanations and transparency mechanisms are often insufficient to improve decision accuracy or mitigate automation bias. They further state that user engagement is the most attainable and impactful point of intervention, and that increased verification effort has been shown to reduce complacency toward AI-driven misrecommendations. In connection with this, they recommend explanation design strategies which actively promote critical engagement and independent verification.

In the regulatory context, Laux and Ruschemeier (2025) state that the AI Act sanctions human oversight for high-risk AI systems and requires providers to enable awareness of automation bias. They further state that the AI Act creates an asymmetric division of responsibility between providers and deployers for mitigating automation bias, and that the focus on providers does not adequately address design and context as causes of automation bias.

Laux and Ruschemeier (2025) also state that automation bias can apply significant influence when legal and institutionalised decision-making structures are formally based on a human decision, while the direction of that decision is in fact provided by a machine. They therefore propose that harmonised standards should reference the state of research on automation bias and human-AI interaction, holding both providers and deployers accountable. They further conclude that more empirical research on human-AI interaction is needed for effective safeguards.

3. Methodology

The following chapter presents the methodological choices made throughout the study. It covers the overall research approach, the literature search process, the empirical data collection through semi-structured interviews and documentary analysis, as well as the approach to thematic analysis. The chapter also addresses the research quality and ethical considerations.

3.1 Choice of Method

The study aimed to gain a deeper understanding of how organisations using Microsoft Copilot implement governance mechanisms to control and minimise the risk of Shadow AI, and how this is established and communicated within the organisation. To get a deeper understanding, an initial review of existing research was conducted. This enabled a clearer view of what has previously been studied in the fields of Shadow AI, AI governance and enterprise AI adoption and risks. This provided a foundation for identifying gaps in the literature, and given the aim of the study, a qualitative research approach was chosen to address the research questions. Rather than measuring the governance mechanisms, this study explores and discusses how organisations using Microsoft Copilot implement mechanisms to manage the risks associated with Shadow AI, and how aware the organisations are of these risks. A qualitative approach was therefore well-suited to capturing the perspectives within organisations (Bryman, 2016).

Data was collected through semi-structured interviews with employees at Swedish organisations that use Microsoft Copilot. This method was chosen to provide structured guidance throughout the interview while maintaining an open mind and flexibility, allowing respondents to elaborate on their own experiences and perspectives without overly targeted questions that would control the outcome (Bryman, 2016). Additionally, where available, internal organisational AI policies were also analysed. The combination supports triangulation, which strengthens the validity and credibility of the study's findings.

3.2 Literature Search

The term 'Shadow AI' began to gain traction during 2023-2024, in line with the rise in AI adoption within enterprises. As the term is relatively new, the majority of the sources included in our literature review are from 2024-2026, with some from 2021-2023, and a few dating back to 2014. Sources older than this have exclusively been used in the method chapter, and are not drawn upon in the literature review or analysis. All sources from before

2021 have been used to explain underlying concepts applicable to several scenarios and not exclusively connected to Shadow AI, though they are strongly connected to the topic.

In order to identify relevant previous research, we utilised primarily Lund University's online academic library, Finn, as well as Legimus and physical libraries. We filtered all searches to peer-reviewed sources to ensure we used only sources considered reliable, valid and of good quality. Through these sources, we have also reviewed their references to find additional relevant sources. In addition to peer-reviewed literature, we have drawn upon the EU AI Act, GDPR and the NIST AI Risk Management Framework as well-established and authoritative frameworks for AI governance and risk management. We have also included Microsoft documentation solely to provide factual context on Microsoft 365 Copilot, which is central to the scope of this study. To find our sources, we used certain search words, both individually and in combination. The following words were the primary search words used to find relevant sources:

- Shadow AI
- Shadow AI Risks
- Automation Bias
- Human-AI Collaboration
- Cybersecurity Risks with Shadow AI
- Artificial Intelligence within Organisations
- Microsoft Copilot
- AI risk organisations
- AI governance
- AI Risk Management
- Responsible AI
- Ethical AI
- Organisational Resilience
- Risk regulation
- Risk management
- AI governance regulatory
- AI Hallucination
- Generative AI

The search process was iterative, meaning that as new concepts and themes emerged through the reading of initial sources, additional search terms were introduced to ensure that the literature review remained responsive to the evolving scope of the study. Where searches gave limited results, particularly among emerging concepts such as Shadow AI, the reference lists of relevant sources were reviewed to identify foundational and related works that may not have appeared directly in database searches.

3.3 Empirical Data Collection

3.3.1 Qualitative Semi-structured Interviews

3.3.1.1 Selection of Participants

For this study, a total of fourteen respondents were selected across seven organisations based in Sweden. The organisations were identified through existing contacts and chosen due to the limitation of having Microsoft Copilot implemented. The respondents were identified through contacts within these organisations who suggested relevant employees with insights into the organisation's AI adaption, governance practices and use of Microsoft Copilot. Respondents were selected from each organisation with the aim of capturing multiple perspectives where possible, ranging from one to three representatives per organisation depending on availability and the roles accessible within each. The respondents hold a selection of roles considered particularly relevant given the study's focus on governance, risk management and AI policy. While only a few respondents within the same organisation cannot represent the full picture of a larger organisation, the selected employees were chosen for their direct involvement with and knowledge of the study's topic.

Respondent	**Organisation**	**Role**	**Interview Length (min)**	**Date**	**Appendix**
Respondent 1	Organisation A	Chief Digital and Information Officer	40:55	28-04-2026	3
Respondent 2	Organisation A	IT Infrastructure Architect	23:02	29-04-2026	4
Respondent 3	Organisation A	Legal Counsel and Compliance Officer	39:43	04-05-2026	5
Respondent 4	Organisation B	Senior Manager Business Enabling Services	24:54	29-04-2026	6
Respondent 5	Organisation B	Strategy Manager, Research & Innovation	26:17	29-04-2026	7
Respondent 6	Organisation C	Head of IT	22:39	29-04-2026	8
Respondent 7	Organisation C	Chief Information Security Officer	48:34	28-04-2026	9

Respondent 8	Organisation D	Head of AI	25:49	04-05-2026	10
Respondent 9	Organisation D	IT Service Manager	24:36	08-05-2026	11
Respondent 10	Organisation E	Digitalisation and Cybersecurity Manager	24.14	12-05-2026	12
Respondent 11	Organisation F	VP & Head of R&D Services	24:56	12-05-2026	13
Respondent 12	Organisation G	Head of Legal	22:23	08-05-2026	14
Respondent 13	Organisation G	Chief Information Security Officer	45:55	08-05-2026	15
Respondent 14	Organisation G	AI Transformation Lead	21:09	11-05-2026	16

Table 3.3: Interview Information

3.3.1.2 Interview Setting

All interviews were individual and conducted via digital meetings, which were recorded with the participant's consent. To allow participants to express themselves as naturally and freely as possible, they were given the option to conduct the interview in their preferred language. As all organisations are based in Sweden, interviews were in either Swedish or English, depending on the respondent’s preference. The decision was based on making the setting as good as possible for the interview, reducing any language barrier that could limit the depth or nuance of the responses.

3.3.1.3 Interview Process

Semi-structured interviews were chosen as they provide enough structure to ensure relevant topics are covered, while still allowing respondents to speak freely and shaping the direction of the conversation.

The interviews were guided by different topics to align with the research inquiry and theoretical framework. However, the questionnaire was formulated as open-ended questions to allow respondents to answer freely in their own words without being steered towards predetermined answers (Bryman, 2016). The interview questions were developed in accordance with the guidelines outlined by Bryman (2016), with the primary aim of allowing participants to share their perspectives freely while ensuring that the relevant research areas were covered. Rather than constructing an unchanged, structured questionnaire, the process

began by reflecting on what needed to be understood in order to answer the research questions, and then identifying which topics and areas the interview questions needed to address. The questions were further grounded in the existing literature to ensure the topics were theoretically informed. As suggested by Bryman (2016), background information about each respondent, such as their role and organisational context, was also collected as this contextualises their responses and adds depth to the analysis. The interview questions are found in Appendix 2.

A selection of questions were designated as crucial, meaning they were prioritised in situations where time was limited or the conversations drifted from the central research focus. This ensured that the important areas were consistently covered across all interviews, regardless of how individual conversations developed.

To present the reason for the interview, the study's content and who we are, the interview started with a short introduction. Regarding ethical considerations related to respondent privacy, consent to record the interview was obtained before proceeding.

The interviews were conducted through the following topics:

Topic 1: AI usage within the organisation

Topic 2: Risks with AI

Topic 3: AI policies

Topic 4: AI and Shadow AI

Topic 5: Regulatory frameworks

Topic 6: Control and follow-ups

Topic 7: Challenges

The topics were not followed by a strict order, instead they were there to make sure we covered all the important topics in our research. Each interview began with a question about the respondent's role and ended with asking whether the respondent wished to add anything further and what they hoped to gain from the study.

3.3.2 Documentary Analysis

A documentary analysis was done as a complementary data collection method alongside the qualitative data collection of semi-structured interviews. Documentary analysis is a systematic way of analysing and assessing documents (Bowen, 2009). Like semi-structured interviews, this method in qualitative research requires active analysis to gain, understand and develop empirical knowledge from non-self-explanatory information written without a researcher's intervention (Bowen, 2009). Unlike primary data collection methods, such as

interviews, documentary analysis is more about data selection rather than data collection. The researcher identifies, accesses and interprets data from existing documents rather than collecting new data through interaction (Bowen, 2009). For this reason, the documentary analysis serves as a valuable addition to the semi-structured interviews conducted in this study.

Internal AI governance documents were requested from all seven participating organisations. Organisations A, C, D and F provided documentation, whilst the remaining three either did not have a formal policy in place at the time of data collection or were unable to share one. The documents received varied considerably in nature, scope and maturity. Organisation A provided a single comprehensive and extensive AI directive, developed jointly by the Digital & Tech and Legal functions and approved at the group management level. It covers ethical principles, governance structures, risk classification and approval processes. Organisation C provided two documents: a brief existing policy oriented around general principles, alongside an incoming revised policy, whilst incomplete in several sections, outlined a considerably more structured governance framework currently under development. Organisation D provided an AI manifesto, concise in nature but reflective of the organisation's stated values and strategic orientation toward AI enablement. Organisation F provided two documents: a Group AI Use Directive and a Group Responsible AI Policy, together covering both operational rules and broader ethical principles, both adopted at the board level in April 2025.

This variation in the nature and maturity of the documents is itself analytically relevant, reflecting the differing stages of governance development across the organisations studied. Where documents were available, they were examined alongside the interview data from the corresponding organisation, with points of alignment and divergence between documented governance intentions and respondent accounts identified as part of the analytical process.

3.4 Data Analysis

Thematic analysis (TA) is a common method in qualitative research for identifying, analysing and reporting patterns in a data set (Braun & Clarke, 2023). These patterns, also referred to as themes, are useful to make sense of shared meanings and experiences. It is a commonly used approach in qualitative research because of its flexibility and is not tied to a specific theoretical framework, making it applicable to a wide range of research contexts (Braun & Clarke, 2023).

The interviews were recorded using the built-in recording system in Microsoft Teams. These recordings were then transcribed using the tool Sunet Scribe, which is a software tool for digitally transcribing research data from video files. The recorded interviews were downloaded as MP4-files, uploaded to Sunet Scribe and exported as text files (.txt). The tool offers features such as language, number of speakers and type of speech patterns in order to

ensure accuracy. The text files then allow editing, enabling the transcriptions to be manually monitored and corrected based on the video recording's content. This step was revealed as very important, as the automatic transcriptions were, in some circumstances, inaccurate by mixing speakers and misinterpreting words. Each interview recording was therefore listened back to in full and cross-referenced against the transcript, with any inconsistencies corrected prior to analysis.

In cases where the interviews were held in the native language (Swedish), the interview transcripts were translated from Swedish to English using the AI tool Claude Sonnet 4.6 by Anthropic, with the prompt in Appendix 17. All translations were reviewed by the authors to verify accuracy and faithfulness to the original.

The thematic analysis was conducted based on the six-phase framework developed by Braun and Clarke (2023), which consists of *(1) Data Familiarisation, (2) Data Coding, (3) Generating Initial Themes, (4) Reviewing and developing themes, (5) Refining, Defining and Naming Themes* and *(6) Producing the Report.* To make this process more effective, the same AI tool used to translate the interview transcripts was also used to assist the thematic analysis.

Data Familiarisation: This is a critical step to become familiar with the data and create a foundational understanding of the responses by reading and rereading the transcripts and rewatching the recordings until getting comfortable with the content. This is not about reading the surface meaning of the words, but instead analysing and critically getting the meaning of the content (Braun & Clarke, 2023). This was initially done by reading the transcript while watching the recording and taking notes. Further rereading the transcript and comparing it with the notes to actively get a sense of the content.

Data Coding: Working with the data in iterative stages is a part of the thematic analysis process (Braun & Clarke, 2023). The initial coding focused on a grounded line-by-line coding to extract interesting quotes from the interviews using these relevant codings. To improve the efficiency of the coding process, Claude was used as an analytical aid during the initial coding stage. The prompt that was used is shown in Appendix 18. The use of AI tools supports coding in thematic analysis by autonomously detecting keywords, phrases and patterns within qualitative data (Christou, 2024), with Claude in particular showing a strong capacity for this type of analytical work in a qualitative research context (Anjos et al., 2024). Given the large dataset from the interviews, using AI assistance at this stage enabled a broader initial read-through before applying systematic human review. Following this, each code was reviewed and assessed against the data and research questions through independent human analysis of the dataset. This initial stage provided a clear overview of interesting reflections and shared meanings. Data coding is a way of identifying features of the data relevant to the research questions (Braun & Clarke, 2023). The coding was organised in an Excel file with the respondent, their organisation, the code name and the corresponding quote.

A	B	C	D
Organisation	**Respondent**	**Code Name**	**Quote**
Organisation A	Respondent 1	Friction-by-design principle	It should be easy to do the right thing but it should also be hard to do the wrong thing

Figure 3.4: Thematic Analysis Coding

This structure enables a systematic linking of codes to the original context, making it easier to follow where each statement comes from. It also helps with comparing across interviews and supports the identification of patterns between codes (Bryman, 2016). While it supports structure and transparency, it also requires careful attention to preserving the original context of the statement, as coding can risk cutting out the contextual meaning of the data (Bryman, 2016).

Creating Themes: Themes are developed in an active process in thematic analysis. The initial generation of themes involves clustering codes that share similarities (Braun & Clarke, 2023). Claude was used as an analytical aid in this initial clustering stage, generating a preliminary grouping of codes into themes, with the prompt in Appendix 19. However, as Christou (2024) emphasises, Claude was used as a complementary aid rather than a replacement in thematic analysis. A thorough evaluation and validation of the AI-generated outcome was done to ensure the quality and precision of the analysis. Accordingly, the AI-generated theme clusters were subsequently reviewed and revised through independent human reflection on the dataset, the codes and the relationships between them. The themes were adjusted to ensure they were meaningfully grounded in the data and aligned with the research questions.

Twelve initial themes were developed from the codings by clustering the most similar codes. These themes were then reviewed and developed through a repeated process, whereby each theme was evaluated against the coded data to assess whether it meaningfully captured the codes and quotes assigned to it and whether the distinctions between themes were clear and coherent (Braun & Clarke, 2023). Where themes were found to overlap or lack sufficient support in the data, they were relocated and themes were merged or redefined until a coherent set of themes were established. During this phase, some codes were found to be less relevant to the overall analysis and were therefore excluded. The focus is to analyse the data and report it in relation to the research questions, not to describe everything that was brought up (Braun & Clarke, 2023). The themes were therefore subsequently evaluated against the entire dataset to ensure they captured the most relevant patterns and shared meanings across all interviews in relation to the research questions. Following this review process, each theme was refined, defined and named to clearly capture its essence and distinguish it from the other themes (Braun & Clarke, 2023). The refinement process was conducted through independent human analysis. The original dataset and codes were re-evaluated to ensure each theme was firmly grounded in the empirical findings. The final analysis resulted in five themes, each reflecting patterns identified in the empirical findings that were particularly relevant to the research questions.

3.5 Research Quality

The question of quality in research is fundamentally one of trustworthiness, whether the findings of a study are worth taking into account and whether confidence in them is justified (Guba, 1981; Lincoln & Guba, 1985). Within the naturalistic paradigm, conventional criteria drawn from positivist tradition, namely internal validity, external validity, reliability and objectivity, are regarded as unsuitable to the cognitive assumptions that support qualitative studies (Guba, 1981). Naturalistic study operates under a different set of assumptions regarding the nature of reality and the relationship between the inquirer and the object of inquiry, and it therefore demands criteria that are equivalent with those assumptions rather than borrowed from an incompatible practice (Guba, 1981; Lincoln & Guba, 1985).

To address this, Guba (1981) and Lincoln and Guba (1985) propose a parallel set of criteria specifically designed to assess the trustworthiness of naturalistic studies. These criteria serve the same methodological purpose as their traditional counterparts whilst remaining philosophically consistent with the naturalistic paradigm. The framework consists of four dimensions: *credibility*, which corresponds to internal validity and concerns the extent to which the findings accurately represent the question under study; *transferability*, the naturalistic correspondent to external validity, which addresses the applicability of findings to other contexts; *dependability*, corresponding to reliability, which concerns the consistency and auditability of the study process; and *confirmability*, the correspondent to objectivity, which relates to the degree to which the findings are grounded in the data rather than the bias of the inquirer (Guba, 1981; Lincoln & Guba, 1985).

In order to critically examine the methodological choices made throughout this study, each of the four dimensions of trustworthiness will be addressed in turn in the following sections.

3.5.1 Credibility

Credibility concerns the extent to which the findings of a study accurately represent the question under study, and serves as the naturalistic correspondent to internal validity in conventional research (Guba, 1981; Lincoln & Guba, 1985). Several measures were taken throughout this study to strengthen the credibility of its findings.

Triangulation was employed in two forms. Denzin (1978) identifies multiple types of triangulation, of which data triangulation and method triangulation were applied in the present study. Data triangulation was achieved through the deliberate selection of participants from seven different organisations and across a range of roles directly connected to artificial intelligence in the workplace, including positions such as Head of AI, CISO and IT Infrastructure Architect. By drawing on perspectives from varied organisational contexts and professional backgrounds, the study reduces the risk of findings being dependent on any single source or viewpoint. Where conclusions merge across these diverse participants, confidence in their truth-value is correspondingly strengthened (Denzin, 1978). Method triangulation was additionally achieved by combining semi-structured interviews with

documentary analysis of the participating organisations' own AI policies. This allows for a direct cross-examination of what participants describe as their organisational approach and AI governance against what has been formally documented, enabling connections and inconsistencies to be identified and strengthening the credibility of the findings (Denzin, 1978; Lincoln & Guba, 1985).

The relevance and firsthand experience of participants further support credibility. All interviewees hold roles with direct professional involvement in the question under study, meaning the data is grounded in lived occupational experience rather than marginal or secondhand accounts. This strengthens the likelihood that the findings reflect an accurate picture of the question as it is experienced in practice (Guba, 1981).

Member checking was conducted by returning the quotes attributed to each participant for their review and confirmation prior to inclusion in the final analysis. This ensures that participants' contributions are represented accurately and that no misinterpretation of their accounts has occurred (Lincoln & Guba, 1985).

Finally, peer debriefing was employed as a systematic practice throughout the analytical process. Each author reviewed the others' sections with the explicit aim of identifying potential bias and ensuring consistency in interpretation across the study. This provided an ongoing external check on individual analytical judgements and contributed to a more balanced and reflective analysis (Lincoln & Guba, 1985).

3.5.2 Transferability

Transferability concerns the extent to which the findings of a study may be applicable to other contexts or settings, and serves as the naturalistic correspondence to external validity in conventional research (Guba, 1981; Lincoln & Guba, 1985). However, as Lincoln and Guba (1985) make clear, it is not the naturalist's task to provide an index of transferability, but rather to provide the database that makes transferability judgments possible on the part of the potential appliers. The responsibility for determining whether findings are applicable to another context therefore rests with the reader, not the researcher.

To support such judgements, purposive sampling was employed as the primary means of ensuring contextual richness in the database. Participants were deliberately selected on the basis of their direct professional involvement with the question under study, holding roles closely connected to artificial intelligence in the workplace. The deliberate selection ensures that the database is informed by firsthand occupational experience, providing contextually relevant material from which potential appliers can assess the applicability of the findings to comparable settings (Lincoln & Guba, 1985). Furthermore, the documentary analysis of the participating organisations' own AI policies contributes an additional layer of contextual richness beyond the interview accounts alone, offering further material from which to assess transferability.

Nevertheless, several limitations constrain the thickness of the description that can be provided in the present study. In the interest of participant confidentiality, both participants and their organisations remain anonymous, with only professional titles disclosed. This restricts the contextual detail available and limits the ability to assess the degree of similarity between the present study's setting and their own. Additionally, all participants are based in Sweden, which may reflect particular regulatory, organisational or cultural conditions that are not necessarily present in other national contexts, and which readers should take into consideration when assessing the relevance of the findings to their own setting.

Finally, it should be acknowledged that the question under study, what governance mechanisms organisations use in order to minimise unauthorised artificial intelligence, is subject to rapid and continuous development. The findings of this study reflect the practices, policies and conditions present at a specific moment in time, and it is likely that the approaches described by participants will evolve as the technological landscape continues to change. The findings should therefore be understood as contextually and temporarily situated, and potential appliers are encouraged to consider the extent to which the conditions described remain applicable to their own context at the time of reading (Lincoln & Guba, 1985).

3.5.3 Dependability

Dependability concerns the consistency and auditability of the research process, and serves as the naturalistic correspondence to reliability in conventional research (Guba, 1981; Lincoln & Guba, 1985). The primary technique for establishing dependability is the inquiry audit, in which an external auditor examines the process of the inquiry to assess whether the methodological decisions made throughout the study are consistent, transparent and defensible (Lincoln & Guba, 1985). However, as Lincoln and Guba (1985) also acknowledge, a demonstration of credibility is in itself sufficient to establish dependability, since there can be no validity without reliability. The measures taken to establish credibility in the preceding section therefore contribute simultaneously to the dependability of the present study.

Several additional measures were taken to support dependability throughout the research process. The methodological choices made in this study, including the selection of research design, data collection methods and analytical approach, have been documented and accounted for transparently throughout this chapter. This allows the reasoning behind each decision to be traced and evaluated by the reader. The transparency serves a function comparable to that of an audit trail, enabling an external party to follow the logic of the inquiry as it is developed (Lincoln & Guba, 1985).

Furthermore, the use of exclusively peer-reviewed or official sources throughout the study ensures that the theoretical and analytical foundations upon which the study rests are grounded in academically validated knowledge, reducing the risk that methodological decisions were informed by unverified or unreliable material. An exception is made for official Microsoft documentation, which is used exclusively to describe the technical operation of Copilot, for which no peer-reviewed equivalent exists and no analytical claims in

the study depend. Finally, the ongoing supervision process provided a form of external oversight throughout the study, in which methodological decisions and analytical judgements were continuously reviewed and challenged, contributing to the consistency and defensibility of the study process as a whole.

3.5.4 Confirmability

Confirmability concerns the extent to which the findings and interpretations of a study are grounded in the data rather than the dispositions or assumptions of the researcher, and serves as the naturalistic parallel (Guba, 1981; Lincoln & Guba, 1985). Several measures were taken to ensure that the conclusions of this study reflect the empirical material rather than the researchers' own preconceptions.

Throughout the analysis and discussion, interpretations are consistently anchored in specific empirical statements, with findings attributed to respondents and organisations. This allows the reader to trace analytical claims directly back to the data from which they derive, fulfilling the core purpose of the confirmability audit principle (Lincoln & Guba, 1985). The transparent documentation of methodological choices throughout this chapter further supports this, enabling an external party to assess not only what conclusions were reached but also on what basis.

The exclusive use of peer-reviewed sources for all analytical and theoretical claims ensures that interpretations are anchored in academically validated knowledge rather than unexamined assumptions. The single exception is Microsoft's official product documentation, used solely for descriptive contextual purposes and not in support of any analytical claim.

Finally, the analysis demonstrates a conscious effort to follow the data rather than confirm expectations, explicitly acknowledging findings that complicate or extend existing literature rather than simply confirming it. This willingness to engage critically with the empirical material, including where it challenges or adds nuance to prior research, reflects the core principle of confirmability as Lincoln and Guba (1985) describe it.

3.5.5 Methodological Limitations

Despite the measures taken to ensure the trustworthiness of this study, several methodological limitations must be acknowledged.

The study is based on fourteen participants across seven organisations, representing a relatively small sample. Whilst the participants were purposely selected for their direct professional involvement with AI governance, the views expressed remain those of a limited number of individuals within each organisation and cannot be assumed to fully represent the broader organisational reality. This should be kept in mind when considering the weight of the findings. Furthermore, the distribution of participants across organisations was uneven, with some organisations represented by up to three respondents, whilst two organisations

contributed only a single respondent each. This imbalance means that certain organisational perspectives inevitably carry greater weight in the findings than others, which may have influenced the patterns identified in the analysis. Similarly, the number of quotes included per respondent in the empirical findings varies considerably, with some respondents contributing significantly more to the empirical material than others. Whilst this reflects the natural variation in the richness and relevance of individual responses, it should be acknowledged that the findings may disproportionately reflect the perspectives of those whose statements were most extensively drawn upon.

Additionally, a significant portion of the responses gathered through the interviews were based on perception and individual opinion and professional speculation rather than verifiable facts. This is particularly relevant when respondents reflect on broader phenomena such as the drivers of Shadow AI, employee motivations or organisational behaviour, where accounts necessarily represent informed professional judgement rather than empirically established facts. This should be considered when interpreting the empirical material, particularly where findings make claims about causation or behavioural patterns.

Limitations in transparency must be acknowledged as some respondents may have been restricted in what they could share due to confidentiality requirements or organisational sensitivity around AI governance practices and unauthorised AI usage across the firm. This may have impacted the depth and openness of certain responses. Additionally, respondents were selected because of their involvement with AI governance, meaning they likely hold a more informed and clear view of how AI is managed within their organisation compared to other employees. This could be seen as a risk of knowledge bias and may not fully reflect how AI tools and policies are experienced or understood across other roles within the organisation. While interviewing respondents with different roles across various functions was intended to broaden the perspective and reduce bias, it cannot be guaranteed that this fully captured the organisational reality.

The documentary analysis was intended to complement and cross-examine the interview data by drawing on the participating organisations' own AI policies. However, not all organisations had a formal AI policy in place and time of data collection, and not all of those that did were willing or able to provide it. This limits the extent to which the documentary analysis could be applied consistently across all cases, and should be considered when interpreting findings that draw on policy documentation.

All interviews were conducted remotely, which whilst practical and enabling access to participants across different organisations, may have limited the depth of understanding and the richness of contextual observation that in-person interviews can create. The majority of interviews were conducted in Swedish, which is the native language of both the researchers and most participants. However, two interviews were conducted in English, which is not the researchers' first language. Whilst every effort was made to ensure accurate interpretation and translation for all responses, the possibility of nuance being lost in the process cannot be entirely discounted.

As mentioned in the preceding section, the geographical scope of this study is limited to respondents operating in Sweden, which may reflect particular regulatory, organisational, or cultural conditions not present in other national contexts. Furthermore, given the rapid pace of AI development, the findings should be understood as temporarily situated and likely to evolve over time.

Finally, as this study was conducted by bachelor's students engaging in qualitative research for the first time, the relative inexperience of the researchers in conducting and analysing qualitative inquiries should be acknowledged as a limitation. Whilst systematic measures were taken throughout the study to ensure accuracy and minimise the influence of researcher bias, the possibility that greater experience would have yielded additional depth or insight cannot be ruled out.

3.6 Ethical Considerations

The following section outlines the ethical standpoints and responsibilities that have guided the conduct of this study. Bell and Bryman (2007) argue that ethical considerations should be understood as central to the entire research process rather than as a separate procedural component relevant only at certain points, and it is in this spirit that the following reflections are offered.

3.6.1 Informed Consent and Anonymity

Prior to each interview, participants were informed of the purpose of the study and the context in which their responses would be used. Both participants and their respective organisations have been rendered anonymous throughout the study, with only professional titles disclosed. This was communicated to all participants at the outset of each interview, before asking permission to record the interview.

Bell and Bryman (2007) draw an important distinction between confidentiality and anonymity, noting that whilst confidentiality concerns the protection of information supplied by research participants from other parties, anonymity involves protecting the identity of individuals or organisations by concealing their names or other identifying information. Both principles are recognised as near-universal obligations across social research ethics codes. In the present study, both were observed: participants were assured that their identities and those of their organisations would not be disclosed, and all data were handled and presented in accordance with this commitment.

Anonymity in this study carries particular ethical weight beyond that of a standard procedural safeguard. The subject matter: governance of Shadow AI and the risks associated with unsanctioned AI tool usage in organisational settings, is sensitive. Participants were on several occasions asked to reflect on and acknowledge potential vulnerabilities, governance

gaps, or informal practices within their organisations. Bell and Bryman (2007) note that the protection of anonymity in management research may enable the study of unofficial or sensitive organisational practices that would otherwise remain inaccessible. This observation is directly relevant to the present study. For professionals whose role is precisely to manage AI-related risks, speaking candidly about governance shortcomings requires a degree of trust that anonymity actively enables. Anonymity is therefore understood in this study not only as an ethical obligation towards participants, but as an active condition for the validity and depth of the findings (Bell & Bryman, 2007; Lincoln & Guba, 1985).

3.6.2 Researcher Influence and Interview Ethics

Semi-structured interviews were employed as the primary method of data collection, allowing for open and exploratory conversations with participants. Whilst this format was chosen for its flexibility and capacity to generate rich, contextually grounded data, it also introduces an essential ethical responsibility around researcher influence.

Bell and Bryman (2007) observe that in management research, the relationship between the researcher and participant is often characterised by a power imbalance that favours the research subject rather than the researcher. Senior professionals and organisational experts tend to have considerably higher social status and domain knowledge than the researchers conducting the study, and management researchers are frequently in a position from which they must navigate access and secure engagement on the participant's terms. This observation is applicable to the present study, in which all participants hold senior professional roles with specialist knowledge of AI governance that considerably exceeds that of the researchers. It is therefore not believed that participants responded in ways they felt were expected, or that their accounts were anything other than truthful.

The ethical concern is more subtle. As relatively inexperienced researchers navigating technically complex professional territory, there is an acknowledged risk that conversations were at times steered in particular directions in order to ensure that all research questions were adequately covered. This steering, whilst not intentional in a manipulative sense, may have introduced unconscious framing effects that shaped the direction of certain responses. Bell and Bryman (2007) emphasise that honesty and transparency are amongst the most widely recognised ethical obligations in social research, and it is in accordance with this principle that this limitation is openly acknowledged here. The measures taken to establish credibility, including peer debriefing and member checking, served in part to identify and mitigate such effects during the analytical phase (Lincoln & Guba, 1985).

3.6.3 Representation and Translation

Given that participants and their organisations are anonymous, they have no public means of correcting misinterpretation should their views be inaccurately portrayed. Bell and Bryman (2007) note that misinterpretation: the misleading, misunderstanding or false reporting of

research findings, is recognised as an ethical concern across social research codes. This obligation is heightened in studies where participant anonymity prevents individuals from publicly contesting inaccurate accounts of their contributions. This places an elevated ethical responsibility on the researchers to represent participant contributions as faithfully and authentically as possible. Member checking, conducted by returning attributed quotes to participants for verification prior to inclusion in the final analysis, was employed as the primary means of upholding this responsibility (Lincoln & Guba, 1985).

The translation of citations from Swedish to English introduces the risk that nuance, intention, or meaning may be partially lost in the process. The researchers have approached this responsibility with care, prioritising authentic representation of participants' intended meaning over translation, and have cross-checked translations between authors to ensure consistency and accuracy. This process is understood not only as a methodological precaution but as an ethical obligation towards the participants whose voices are being represented.

4. Empirical Findings

This chapter presents the empirical findings from the study's data collection. The findings are presented through the themes which were created during the thematic analysis. The structure of this chapter is designed to reflect on the different governance mechanisms which are practised by the organisations in question, as well as their perceived risks which grow from primarily unauthorised AI use. The themes are presented in the order they emerged from the data.

4.1 Theme 1: How Organisations Approach AI Governance: Philosophy and Strategic Positioning

A common pattern across the organisations was the discussion of the fundamental strategic orientation they adopt towards AI governance. This theme captures how organisations decide between a more culture- and trust-based approach and a more strict, control-heavy approach, and the reasons behind these choices. It was described as a tension between implementing sufficient governance to manage security risks while maintaining enough openness to support innovation and AI adoption. While all organisations acknowledged the need for some degree of structure, they differed considerably in where they drew the line and how they justified their position.

4.1.1 Enablement over restriction

Across the organisations, a preference for enablement over restriction was commonly expressed. Several respondents described a governance approach centred on providing employees with sufficient and capable sanctioned tools, with the underlying reasoning that good alternatives reduce the motivation to seek out unauthorised ones.

Respondent 8 (Org D) framed this most directly, arguing that an anti-AI policy is not a real governance choice but rather an inherent endorsement of Shadow AI:

> *"[...] if you try to have an anti-AI policy you're just basically saying, telling your employees use shadow AI which is a very bad policy thing to have so I think you have to embrace reality."*

The same respondent described the proactive effort their organisation had made to ensure employees had access to powerful sanctioned tools:

> *"But we've worked hard to bring in the most advanced tools to our employees. So I don't think we and we have all the, not all of them, but we have options that are, that cover all the*

most advanced use cases that we can think of. So we're taking it, we're taking a very pro AI policy in the company."

Respondent 13 (Org G) echoed the same principle, framing it as a question of control rather than permission:

"[...] if we say no, people will do it anyway, full stop. So we have to say yes, we have to take control of it."

Respondent 10 (Org E) described an open approach to AI governance, stating that it is difficult to govern and prohibit AI use completely:

"I personally think that you simply cannot fully govern the use of AI. And it is probably difficult to prohibit something. Then again we are a much smaller organisation so it might be easier to govern and see and keep track of everyone and educate everyone that one may not use it."

Respondent 4 (Org B) described a related strategy to tolerating a small degree of unsanctioned tool use among developers, reasoning that making sanctioned tools sufficiently integrated and effective would naturally draw employees back to approved alternatives:

"Given that it's such a small percentage it's not something we chase, rather we chase how we optimise those who use the central ones. Like for example if you have your own then you'll both pay for the licence and you'll pay for all the tokens you use, which adds up to quite a lot. [...] It becomes very limited and if we then rather accelerate all those parts for the approved tools then it will eventually make it painful enough for those using something else that why would they bother? Why not just use one of the approved ones?"

Respondent 2 (Org A) summarised the underlying logic shared across these organisations, framing agility and security not as opposites but as interdependent:

" [...] I think you as an organisation have to balance. Security and being agile. Because if you're too slow then you get quite widespread Shadow IT."

Respondent 2 (Org A) also expressed the view that their organisation's open approach was contributing to lower levels of Shadow AI, though acknowledged that this was his own assessment rather than a verified finding:

"I think our strategy with being fairly open and agile means we have less Shadow AI. I say that and then in a year or so I'll have to eat my words."

Similarly, Respondent 4 (Org B) described an organisational strategy of allowing testing of new tools and only formalising control when a tool gains traction:

"[...] our strategy at [organisation] is rather to allow that type of testing and when we see it gaining traction we try to control it more and say okay but then we should probably have an official agreement with them and try to bring them into our portfolio."

A related concern raised by several respondents was that overly strict governance risks hindering innovation. Respondent 4 (Org B) noted:

"[...] when you talk about Shadow IT it's often about control. And I think the other side of this is innovation. So balancing control against innovation. If you control too tightly you'll miss out on the innovation side."

Respondent 13 (Org G) reinforced this, describing the challenge of governance as fundamentally one of finding ways to say yes rather than defaulting to restriction:

"[...] it's bloody easy to say no, bloody hard to say yes."

Respondent 11 (Org F) articulates the same underlying, noting that complete elimination of Shadow AI is neither achievable nor desirable:

"[...] it's unavoidable that we have that and it's also not our goal to completely eliminate it. Because it will help on the innovation, otherwise also."

Respondent 14 (Org G) describes the organisation's response to employees using ChatGPT on personal accounts as a decision to provide a sanctioned alternative rather than restrict:

"In November we made the decision to purchase ChatGPT Enterprise for all employees. That was one way of giving everyone an AI licence to create a channel that was approved for the need that existed."

4.1.2 Guardrails over prohibition

At the same time, organisations acknowledged that some level of structure was necessary, leading to what several described as a preference for guardrails over prohibitions. This was described as having technical and procedural controls that make correct behaviour easier without explicitly forbidding incorrect behaviour.

Respondent 1 (Org A) articulated this most directly:

"[...] rather than telling people what they're not allowed to do, I want them to work with what I call different types of guardrails, technical protections [...]."

"It should be easy to do the right thing but it should also be hard to do the wrong thing."

"[...] I'd rather we build and work in a way so that it's possible to really develop the business and scale with this without having to follow the ten commandments every time."

Respondent 6 (Org C) reflected the same logic from a contractual perspective, noting that the goal of governance is not to stop tool usage but to ensure it stays within the boundaries that protect client relationships:

"No, I think nobody wants to stop Claude and those types of services, it's more about we have to make sure we don't commit contract breaches against our clients [...]."

Respondent 8 (Org D) adds on to this, further reinforcing the wish to have control without hindering innovation:

"If you try to start from the side of policies first and then only add in controls where necessary, I think that's the better approach."

4.1.3 Organisational alignment as a prerequisite

Several respondents noted that regardless of which governance philosophy an organisation adopts, its effectiveness depends heavily on consistent and aligned messaging from leadership. Inconsistent signals from management were identified as a risk in their own right, creating confusion among employees about what is and is not permitted and undermining whatever governance structures are in place.

Respondent 7 (Org C) was the most direct on this point, mentioning that rules alone are insufficient and that management alignment is a prerequisite for any governance approach to work:

"[...] the only thing that'll help is blocking and enabling things. Rules won't help outright, you have to have an alternative and a management that shows the way. And if you then have a split brain as management, then you have a big problem. So if one says yes and one says no, massive problem."

Respondent 6 (Org C) described the practical consequences of this misalignment at the employee level, where unclear leadership signals in employees acting without understanding the consequences of their behaviour:

"And information from leaders is unclear too. Where one leader says go ahead. And one leader says no absolutely not. As an employee, I understand that employees make mistakes. They don't do it consciously or out of ill will but you don't really know the consequences of things."

4.2 Theme 2: Technical Governance Mechanisms

Several technical mechanisms were identified across the organisations as ways of governing AI usage. These ranged from information classification systems and monitoring tools to access controls and human-in-the-loop requirements. While the specific mechanisms differed between organisations, a common underlying logic was observed: rather than attempting to prohibit unsanctioned tool use entirely, most organisations focused on controlling what data could flow into and out of AI systems, and on detecting when boundaries were crossed.

4.2.1 Information classification

The most commonly observed technical mechanism was information classification, present in some form across the majority of the organisations. The core principle was to label data according to its sensitivity level, determining which tools it could be processed by, rather than restricting which tools employees could access.

Respondent 2 (Org A) described this as a deliberate strategic choice, preferring data protection control over tool prohibition:

"We don't prevent, some organisations do that you can't use Gemini on your computer at all. We don't do that, instead we see it as good for employees to use other AI services privately in order to learn and so on. So we work more with protecting the information than with preventing people from being able to use the tools."

Respondent 5 (Org B) described a four-tier classification system used across all documents, with the most sensitive tier explicitly excluded from all AI tools including sanctioned ones:

"And we classify all our documents with Secret, Confidential, Internal or External. That's standard on all our documents and then it's been said that Secret documents you're not allowed to share with ChatGPT or Copilot."

Respondent 9 (Org D) described a similarly structured system, where the classification level directly determines which external services data may be entered into:

"You're also not allowed to feed in sensitive information classified as C2 to C4 into external AI services. I don't know if you know information classification. There's C1 to C4 and C1 is public information and C2 is like information that maybe C4 is super classified that only management has access to for example, so it's a classification. C1 is what's public that anyone can find. That's the only thing you're allowed to feed into external AI services."

Respondent 8 (Org D) highlighted an additional benefit of classification within the Microsoft Copilot environment specifically, noting that the system respects existing access permissions so that the tool cannot surface information a user would not otherwise be able to see:

"[...] the tool can only see, can see things based on classification levels and it can distinguish between what I can see versus what you can see. So if we were both in the company, then it can only see what you can see in the system. It can't see what I can see there. That way, it doesn't leak information across the company to other people. So that's a very valuable feature of the Copilot suite."

4.2.2 Monitoring and detection

A second technical mechanism observed across several organisations was active monitoring and detection of AI tool usage. Respondents described how organisations monitor network traffic and device activity to identify which external services company data is being sent to, enabling them to detect unsanctioned usage and respond accordingly.

Respondent 6 (Org C) described how this monitoring is already operational and yields visible results:

"We can see which computers are accessing the Claude API and things like that and we see how much data goes back and forth from our computers to various services. It's fairly easy for us to see."

Respondent 7 (Org C) described a more comprehensive approach, mapping tool usage across both installed processes and web traffic:

"But when it comes to the technical side, what we do is look at and map who uses which tools using both what processes we find on the computer. But also which websites are visited by our employee on their company computer. And then we look at what we know are places where AI models are downloaded, where the data flows from and so on."

The same respondent described a risk-based response mechanism currently being implemented, where detecting an unsanctioned tool on an employee's device does not result in an immediate block, but instead eventually requires the employee to repeatedly verify their identity in order to continue to access company data:

"So an example of things we're currently implementing is if we find Claude Code on someone's computer, then we'll raise the risk level on that specific computer. And if you exceed a certain level, then you have to continuously re-authenticate to access our data for example."

Respondent 7 (Org C) also described planned enhancements to monitoring infrastructure, including Data Loss Prevention (DLP) solutions and a Cloud Access Security Broker:

"We'll implement information classification, we'll implement DLP solutions beyond what we already have today, we'll have a Cloud Access Security Broker. And in this case it's Defender for Apps, Defender for Cloud Apps and Defender for Cloud, in which we map where and from where our data comes and then gets sent off."

Respondent 9 (Org D) described a similar DLP project currently in progress, which will automatically filter data before it leaves the organisation's environment:

"We're currently running a Data Loss Protection project that will mean all data we feed in will go through this solution before it goes further out on the web and it will include for example if it finds personal data then you'll get… the data will land in a portal where it's then filtered and if you've fed in personal data then it'll probably either mean that what you fed in just disappears from the text or that you get warned that we can see you've written personal data."

Respondent 11 (Org F) describes how their organisation conducted a traffic-based study to identify unsanctioned tool usage:

"[...] we did a big study based upon our internet traffic that we're having from a corporate perspective. And we identified what are the other common used AI tools."

Following this, the organisation decided to block approximately 200 identified AI tools, though the respondent acknowledges the limitations:

"[...] it's not perfect because, you know, new AI tools are popping up every day, but at least the most common ones we now have blocked for usage."

4.2.3 Access controls and licence-gating

A third mechanism described across the organisations was the use of access controls and licence-gating to restrict how and by whom AI tools can be used. These controls operated at different levels, from restricting what data Copilot can search within an organisation's environment, to requiring approval before new AI use cases can be deployed.

Respondent 9 (Org D) described a configuration that prevents Copilot from freely accessing all organisational data, limiting its search scope to sites a user already has membership access to:

"[...] we've restricted Copilot so that it doesn't search for content in SharePoint sites that are locked. So only if you're a member of a site does it search that site for information. It doesn't search freely everywhere in our data."

Respondent 8 (Org D) described a use-case level approval requirement, where creating a new AI-enabled process requires a cross-functional review:

"[...] if you're using AI to help out with an existing use case, that's one thing, but if you're creating a new use case, you have to get approved. Which includes checks through legal and compliance and risk and IT security, et cetera, depending on how complex the use case is."

Respondent 4 (Org B) described a complementary approach focused on controlling where data is processed geographically rather than which tools are being used:

"*It's difficult to control the answers. So it's hard to ensure that it always answers correctly. So we don't do that. On the other hand we look at ensuring that it can't share information that isn't okay so that we can filter out personal information or sensitive information or things like that. And we look at where it shares the information so that it isn't stored in geographical areas that we're not permitted to distribute the information to.*"

4.2.4 Human-in-the-loop requirements

Finally, a mechanism mentioned by several organisations was the human-in-the-loop requirement. This is a technical-procedural control which mandates human review before AI-generated outputs can be acted upon, ensuring that AI cannot independently execute decisions without a person first reviewing and confirming the proposed action.

Respondent 2 (Org A) described how this operates in practice within agentic AI systems, where the tool explicitly requests approval before taking any action:

"[...] every time this AI is going to do something, it will ask the question before it does it. And it will describe what it's going to do. So you simply get a question: is it okay if I do that. And in some cases you answer in text and in some cases it's a checkbox, like I approve. Or a button then, deny or approve."

Respondent 8 (Org D) described the same principle:

"[...] we require a human-in-the-loop, so an output from an AI can't be used to make a decision directly. It needs to be reviewed by a human first."

Respondent 2 (Org A) also noted the broader reasoning behind this argument, pointing to the risk of agentic AI systems making changes to live business systems without adequate oversight:

"[...] it doesn't just provide information, it can actually do things for you and there it's incredibly important that we have steps with human-in-the-loop where a person has to approve, like can I do this, that is the agent, can it make this change in our business system or change production data or similar."

4.3 Theme 3: Human and Organisational Governance: Policies, Training and Awareness

Across the organisations, several human and organisational mechanisms were identified as means of governing AI usage. These mechanisms span formal policy frameworks, training and awareness programmes, structured approval processes and an expectation of individual

employee responsibility. While all organisations described some combination of these mechanisms, they varied considerably in their maturity, formality and perceived effectiveness.

4.3.1 Formal AI policies and directives

A commonly observed mechanism was the existence of formal AI policies or directives, which were present in almost all organisations to varying degrees. These policies were generally described as guidelines covering what tools employees are permitted to use, how data should be handled and the ethical principles expected of employees when using AI. Some organisations mentioned having well-established and multi-tiered policy structures, while others had policies that were still being drafted or revised at the time of the interviews.

Respondent 8 (Org D) described a layered policy structure that goes from board level down to operational guidance:

"[...] we have the high level things we have, we have board level, management level, and then operational level guidelines, we have… there's different names for those internally like instructions and policies etc. that determine high level how we use AI ."

By contrast, Respondent 13 (Org G) described a deliberately minimal policy designed to encourage usage rather than restrict it:

"Yes, we do. And it's simple. Use AI, just go for it. If you're unsure, ask the CISO. That's basically the policy, that's how it is. And really it's, yeah, use the tools you've been given and then just go for it. That's really it. Then there are more detailed things, but it's like a one, two pages max. And it's been updated. I think that over the last month it's been four, five times. Which means absolutely no one in the history of the world reads it. It's more for those of us who sit and think about this. Is it important? And when we communicate it becomes a PowerPoint. Where it just says go."

Respondent 12 (Org G) highlighted that policies also serve a client-facing function, governing not just internal behaviour but the commitments made to external parties:

"And then we have internal policies on what's okay based on what we also promise our clients in turn. So we have internal processes to sort of capture and make these various types of assessments that are required then to be able to approve tools [...]."

However, the existence of a policy was widely acknowledged as insufficient on its own. Respondent 3 (Org A) noted that developing a policy is the easy part, and that familiarity with its contents requires sustained communication and training effort:

"Because just developing a policy. That's fairly easy to do. But it applies to everyone… Being familiar with this and the content too. And that requires quite a lot of information and training."

Respondent 10 (Org E) mentions that they do not have a formal AI policy, however, they do have an IT policy where they briefly mention basic recommendations and risks associated with AI use:

"In the IT policy there are a few sentences and words about what one should bear in mind when using AI and the risks associated with it all. But there are no restrictions as such. And it says more about precautionary principles that one should keep in mind when using AI."

Respondent 10 (Org E) continues to argue that leaving AI use to AI policy alone will not change employees' behaviours:

"[...] policies will never change behaviours in themselves. [...] policies are important for setting the framework. But I think what is important here is education."

4.3.2 Training and awareness programmes

Another mechanism which was mentioned was training and awareness programmes. Several organisations described running regular training sessions, ranging from general AI literacy to tool-specific guidance. In some organisations, the training was described as mandatory and tied to licence access, while in others it remained voluntary. Training sessions were also described as being supplemented by internal networks, newsletters and within intranet pages where employees can find guidance on permitted tools and best practices.

Respondent 4 (Org B) described one of the most formalised approaches, where completing and passing training is a verified prerequisite for receiving a licence:

"[...] before you get access to any type of AI tool you're required to complete what we call an AI literacy training where you very generally get to go through what it is I can and can't do with an AI tool and what data should I not share? And what type of questions should I not ask? Because there are such things too. So that's one part and the other part is that for the specific tool you also need to complete a training session. [...] So those two you have to have completed and it's verified before you're assigned a licence that you've actually completed and passed those knowledge tests at the end before you can receive a licence."

Respondent 7 (Org C) described a less formal but organisation-wide approach, using annual company conferences as a forum for communicating what is permitted:

"These two things have been mentioned at annual conferences where we've explicitly gone out at the annual conference saying that now we're releasing this functionality. And then it's also clarified that yes, but what are you allowed to do, how can you do it and so on."

Respondent 11 (Org F) describes a mandatory training requirement tied to policy updates:

"We also have a mandatory training on our AI policies. [...] I think, yeah, you need to renew it every second year or something like that. But of course, when we make an update, we reset

it. Like I said, we made a change in the policies, so now everybody needs to redo the training."

Respondent 10 (Org E) offers a candid statement of training stalling due to leadership scepticism:

"We did that in the beginning of last spring. We had some training and education and more, and we planned a few more. But since then our CEO has been a little sceptical of AI and all sorts of things like that. We have not quite had the resources to push on further with training."

They continue to note the importance of continuous training in order to learn, something not yet broadly organised throughout the organisation:

"[...] it is also very easy to have training for one hour or two hours. But I don't think that is always where you learn. [...] I think you need to actively work with AI yourself and learn by doing. [...] we have nano learning for cybersecurity and we have also discussed whether we should have nano learning, five minutes per week or five minutes every other week, where you receive one of these training pieces about what you should bear in mind. [...] But it becomes sort of, five minutes comes out and it is just one such session, so it is once a year that that part comes around."

Despite these efforts, several respondents acknowledged that reaching all employees remained a significant challenge. Respondent 1 (Org A) noted that voluntary training creates a widening gap between advanced and non-users:

"[...] those who are occasional users or who have never used it. Then the barrier to entry becomes quite high. It's always easier to get better at something you're already reasonably good at. So those who are into this and working with it. They're easy and you can push out a ten-minute training for them. They'll go through it. But if you do it for someone who has never been in Copilot. Even if they have a licence. Yes, you understand, they're not just going to go from non-usage to champion, you really have to work with that change management."

This challenge of engagement was also illustrated by Respondent 1 (Org A) when reflecting on the limits of policy communication more broadly:

"People see road signs and still drive too fast."

Respondent 9 (Org D) offered a similarly candid observation about the limits of documentation-based training, noting that many employees treat required reading much like software terms and conditions:

"I think many get a licence. They're told to go in and read the documentation and then they think yes, whatever. It's a bit like when you're going to install a program on the computer. You scroll through the terms, click next. I think many do that and it's a problem."

Respondent 13 (Org G) went further, claiming that training should not be relied upon as a governance mechanism, and that the focus should instead be on designing environments where incorrect behaviour becomes structurally difficult:

"If it turns out we find that this wasn't quite right, then we have to look at... how can we support, like, training doesn't work, full stop. Instead it's, how can we actually with other elements make it so you can't do wrong, but you have full freedom. That's it, and how do you do that? Not a bloody clue, but so far it seems to be working what we're doing."

Most organisations continued to invest in training as a primary governance mechanism, but it anticipates the turn toward technical controls discussed in the preceding theme.

4.3.3 Formal approval processes for new AI tools

The use of formal approval processes for introducing new AI tools was a further mechanism mentioned across the organisations. These were described as evaluation committees or review processes that assess new tools from legal, technical and security perspectives before approving their use. The purpose was described as a process to ensure that any tool that handles company data meets the organisation's security and compliance requirements before being widely adopted.

Respondent 3 (Org A) described involvement in a cross-functional evaluation group that brings together legal and technical perspectives:

"[...] we have a special process for how we, how we evaluate various AI tools before we're allowed to use them at [Org A]. So we've evaluated Copilot and decided or said okay to using Copilot. So I was involved and looked at the actual product itself. But then I'm not skilled in the technical side, rather we have an evaluation group with various competencies you could say. Where I provide the legal perspective [...]."

Respondent 8 (Org D) described a more formalised version of this, built on an existing new product approval process that had been extended to cover AI use cases:

"[...] we've extended the existing procedure we have which is called the new product approval process or NPAP [...] this basically is the series of steps that require it's for approvals for lots of new things [...] that process is for anything from a use case to a a new external system we're acquiring [...]."

4.3.4 Individual responsibility as a governance layer

Several respondents also noted that despite formal mechanisms, governance ultimately depends to a significant degree on individual employees applying their own judgement. Some respondents expressed uncertainty about how effectively this worked in practice, particularly where policies were not yet clearly defined or consistently communicated.

Respondent 6 (Org C) was direct about the limitations of relying on individual responsibility in the absence of a clear policy:

"So right now it's up to each individual employee to understand themselves and that never really works. And we already have problems with it."

They then highlighted the organisational paralysis that can result when leadership does not take clear decisions:

" [...] nobody wants to take the decision to block anything even though it should be done until you have, if we take Claude as an example since that's what's very hot right now at least for us and certainly for many others, quite a few people are running it [...]. There will be a controlled pilot. But nobody wants to block those who are running it outside [...]. It becomes very strange. And that's because some people think the business value is important enough that people should still be able to test it. In my view it would have been better to bring them into an extended pilot phase where you have control over those people."

Respondent 5 (Org B) offered a more optimistic perspective, noting that the act of agreeing to training and policy documentation at least establishes a formal basis for accountability:

"I've in some way approved them when I applied for access, that I've read and agreed to them. So if I were to use a tool in an incorrect way then at least as a licence user I've somewhere approved that I've read through the documents and trainings that exist."

Respondent 12 (Org G) similarly framed individual responsibility as a deliberate and communicated expectation rather than a fallback, noting that employees are actively encouraged to take ownership of staying informed:

"[...] it's also an individual responsibility that you ensure yourself and stay updated, so both verbally and there's also information available on various platforms that we have for our guidelines and policies that you can always go in and find them."

Respondent 14 (Org G) notes the absence of technical monitoring controls, describing the approach as trust-based instead:

"[...] we have no systems that control or audit, not as far as I know at least. Rather it is a very trust-based culture."

Observed in the data, not all respondents shared the view that training and awareness are sufficient as governance mechanisms. A respondent argued that technical controls are a more reliable means of ensuring correct behaviour, as discussed in the previous theme.

4.4 Theme 4: Identified Risks of AI-generated outputs and unsanctioned tools

Several categories of risk were identified across the organisations in connection with both authorised and unauthorised AI tool usage. These risks are connected to three main dimensions: data and confidentiality risks, output quality risks and vendor and supply-chain risks. A notable observation across the interviews is that these risks were not seen as exclusive to unsanctioned tools, several respondents were careful to point out that even approved tools carry significant risk if used without sufficient care or awareness.

Respondent 4 (Org B) brings up a risk connected specifically to the use of unsanctioned tools:

"When it comes to non-approved tools it's very much that we don't have oversight of what data is shared and where it potentially ends up."

4.4.1 Data leakage and confidentiality breach

The most consistently identified risk across all organisations was data leakage and confidentiality breaches. Respondents described concerns about sensitive or customer data being entered into unauthorised AI tools where the organisation has no control over how it is stored or processed. This was particularly noticeable in Org C, where the consultancy context meant that client data carried contractual as well as security implications:

“The biggest risk we have with AI that we don't have control over is client data. [...] employees perhaps don't know what they're allowed to do and what they're not allowed to do. And then you just go ahead and then maybe there are very clear contract breaches” - Respondent 6 (Org C)

Respondent 12 (Org G) echoed this concern, framing data leakage as one of the most fundamental risks regardless of which tool is used, emphasising the importance of owning your data and not letting AI train on it:

“That is in the tools we have, it's always the Enterprise version, that is we have control over it and the AI doesn't train on our data. [...] We don't want any type of AI to train on the data we put in. And then of course also, where does this data end up? [...] Is it stored in the US? Can you sort of limit the storage to Sweden, or within the EU?”

Related to this was the risk of being non-compliant to GDPR. Respondent 1 (Org A) noted that:

“I mean, no agent in the world yet manages to distinguish what is GDPR compliant or not. You can naturally trigger it and say that if you find personal data you can't pick it up. We can prompt that in. But if we don't do that then I don't know if there are any built-in protections, not that I'm aware of at least.”

Respondent 3 (Org A) reinforced this from a legal perspective, highlighting how the broad data access of tools like Copilot creates a heightened responsibility for employees to consider what they do with the output:

"Copilot has access to all information that I have access to if I take this workspace then. Or Copilot work or whatever it's called, work, so it accesses all the information that I can recall from my computer. So it has access to an enormous amount of information and of course if I then use Copilot and it generates a result that I then… since it has access to so much sensitive information then I really have to think carefully when I use the result afterwards in what context since the risk is otherwise that I share something that I'm not allowed to or should share with someone."

4.4.2 Output quality and over-reliance

Another aspect of risk concerned output quality and the tendency toward uncritical over-reliance on AI-generated content. Respondents across all organisations noted the risk of employees trusting AI-generated outputs without enough critical review, and emphasised that the use of AI does not shift responsibility for the quality or accuracy of work away from the employee.

"[...] we see quite a lot of risks with people trusting the answers too much." - Respondent 2 (Org A)

"It doesn't matter whether you've used AI or not, you're still equally responsible for what you hand in." - Respondent 4 (Org B)

"It's super important that you don't use AI in a way that you reveal a lot of trade secrets without being aware of it." - Respondent 3 (Org A)

Respondent 10 (Org E) illustrated the risk of over-reliance with a concrete example:

"I have heard horror stories or read about horror stories where companies have gone for several months based on sales analyses and what is it called, opened themselves up to new markets in other countries based on AI recommendations and figures saying that this market looks good, and then it turns out that someone in the IT department realises that it has simply made up the data."

Respondent 10 (Org E) also paints a picture of imagining the AI tool being a new employee and analysing the output as if it were this new employee speaking:

"[...] to treat AI like a colleague. Think of it as a newly hired person at the office who just asks a load of questions and doesn't understand anything themselves [...] But you also have to think it through yourself, if a new hire had said this to me, would I have questioned it, or yes, maybe that is a good idea."

Respondent 6 (Org C) noted that hallucination is a particular challenge in specialist domains, where employees without deep subject knowledge may be unable to identify when an AI has produced a believable but incorrect answer:

"AI is a master at making things up. I think it's gotten worse. So you have to have basic knowledge in the subject you want to work on and understand what ChatGPT or Copilot is good at and not good at. It's like this, if you struggle with a specialist area it will probably struggle with a specialist area. But if it's a broad question where there's a lot of information on the internet then it will be accurate."

Respondent 9 (Org D) observed that this risk is difficult to govern technically, as there is currently no reliable control function for detecting when an employee has acted on incorrect AI output:

"There's no control function. We've said that we assume that if you use the data then you should ensure it's correct beforehand, or that in uncertainty you've cross-checked another way so that you don't rely on it. It still hallucinates quite a lot, we notice. There's probably no good control function for that, I think, unfortunately."

Respondent 14 (Org G) describes a deliberate approach to managing hallucination risk by selecting use cases where errors are tolerable:

"And then a third risk is, as the question may be getting at, the quality of the output and hallucinations and so on. And our approach there is mainly that we primarily start with use cases where an inaccuracy is not a catastrophe. So that we sort of start… Either in areas where a mistake is completely fine. Or in areas where we are very clear that this may not be correct."

4.4.3 Vendor and supply-chain risks

A third category concerns vendor and supply-chain risks, particularly in relation to third-party tools built on top of foundational AI models. Respondent 1 (Org A) noted that smaller vendors building such tools may lack the resources to invest adequately in security, creating a new attack surface when employees use unapproved niche tools:

"The three of us could build one and sell it. But we might not have the money to spend on the security parts so there we get a new attack vector for something. So that's trickier, so everything that comes in has to go through what we call first something called the AI Council where everything is reviewed and they check what is the vendor, what's their background."

Respondent 6 (Org C) raised a related concern about the traceability and ownership of outputs generated through third-party tools, particularly in the context of code development:

"And then the lack of traceability of information is also like this. Who actually has access to it? Well, Anthropic has access to it, if we feed something in. If we write our own code or

applications, who actually owns it when we write it, Claude Code and so on? Also an interesting question."

Respondent 11 (Org F) raises data residency as a specific governance concern when evaluating providers:

"[...] providers like Anthropic with Clouds, they currently don't have an offering for EU data residency. So that definitely, yeah, again, add some complexity on implementing this in our organisation."

4.5 Theme 5: Drivers of Shadow AI: tool gaps, habit, and hype

A recurring and cross-cutting theme concerns why employees turn to unsanctioned tools in the first place. Respondents identify a consistent set of drivers: familiarity with a tool used elsewhere or privately; a perception that the sanctioned tool is inferior or less capable for a specific task; and the influence of external hype around particular models. Several respondents also point to governance bottlenecks, such as overly slow approval processes, as a structural driver that pushes motivated employees toward Shadow AI.

4.5.1 Tool performance and task specialisation

One of the drivers mentioned across several interviews is the perception that unsanctioned tools are simply better suited to specific tasks than the sanctioned alternative. This was not framed by any respondent as reckless or defiant behaviour, but rather as rational problem-solving by employees seeking the most effective means of completing their work.

Respondent 1(Org A) was the most explicit in naming specific tools for specific purposes, noting that certain models outperform Copilot in clearly defined areas:

"Gemini is better at Computer Vision, looking at images, generating images, that kind of thing. Anthropic's Claude, which I use privately myself, is the only tool that is really sharp at making presentations for example. If you're coding then Claude Code is far superior, truly far superior."

Respondent 10 (Org E) adds on to this, noting the possibility of employees wanting certain functions in specific AI tools:

"Claude and certain other tools, there is a certain functionality that people are looking for that those tools are a bit better at than others."

Respondent 2 (Org A) mentioned the possibility of employees attending external lectures and concluded that the sanctioned tool cannot replicate what they have seen:

"It could be that you've been to a lecture where someone shows something really impressive and you think there's absolutely no way I can do this with the existing tools."

Respondent 3 (Org A) offered a simpler version of the same point, noting that dissatisfaction with Microsoft Copilot is a key driver:

"I think people think ChatGPT is better. You might have used it privately and then you think this was really great. I've heard many who are dissatisfied with Copilot. Who think it's simply much worse. So I think that's the reason."

This view is shared across organisations. Respondent 6 (Org C) acknowledged that different AI tools genuinely excel at different things and attributed the Shadow AI pressure at their organisation to consultants recognising this:

"Regardless of what kind of services they are, different things are better at different things. And right now Claude is very hot."

Respondent 7 (Org C), while more critical in tone, similarly accepted that *"certain services are better than others,"* though they were careful to add that superiority does not grant permission to use a tool outside approved channels. Respondent 13 (Org G) took the same position, acknowledging performance differences while emphasising that the right response is to surface those needs through legitimate channels rather than bypass governance structures:

"If you find a tool you'd really like to use, well then we do a demo. Then we do a pilot of the whole thing."

Several respondents nuanced this by suggesting that the perception of a tool being insufficient is something rooted in lack of prompting skill rather than an actual tool performance gap. In other words, the respondents argue that employees may turn to sanctioned tools, believing the sanctioned ones are worse, when in reality the output quality is closely tied to how the tool is prompted.

"*As they've always said, garbage in, garbage out. So it's incredibly important for us to get better at working with prompting techniques or similar.*" - Respondent 2, Organisation A

"Garbage in, garbage out we usually talk about. Sometimes it might not be about needing a different tool but just needing to get better at prompting" - Respondent 5, Organisation B

"*So correct prompting is something people need to get better at and not assume it knows what you mean, because it doesn't always.*" - Respondent 9, Organisation D

4.5.2 Habit and the blurry line between private and professional AI use

A second driver identified across multiple organisations is the natural blurring of private and professional AI use. Several respondents described a mechanism whereby employees develop

habits with AI tools in their personal lives and then default to those same tools at work, often without a conscious decision to violate policy.

Respondent 9 (Org D) offered a concrete illustration:

"I think the big problem is rather that people maybe use ChatGPT at work because they use it privately and they have a bookmark, they click on it. They take what they're used to. And that applies especially to those who might not be super on board with what IT security means and don't understand data security and so on." They continue to state that this driver is rooted in habit rather than there being strict technical restrictions: *"So I don't think technical restrictions are the big problem. I think it's people's approach and habits."*

Respondent 5 (Org B) reflected this dynamic from personal experience, describing how intensive AI use at work had made private life feel inefficient in comparison:

"I'm fairly keen privately to get such a license just because I feel I use AI so much in my job now that privately I feel like God it's so inefficient."

Respondent 1 (Org A) was notably candid in acknowledging this from their own experience as a senior governance figure:

"[...] there are probably those, just like me, who have a private laptop with both ChatGPT and Claude on it that they pay for and use. But that generates things, something that I then use in my job."

4.5.3 Insufficient awareness of sanctioned tools and their capabilities

Several respondents pointed to a gap in employee awareness, that employees sometimes turn to outside tools not because the sanctioned alternative is genuinely inferior, but because they do not know what it can do or that it exists at all.

Respondent 8 (Org D) listed this explicitly as one of the primary drivers:

"[...] probably one could be familiarity with something else, so if you don't know what's available internally but you know something externally, you can use that."

Respondent 9 (Org D) observed that many employees who had been allocated a Copilot licence had not yet begun using it:

"[...] quite a few have gotten a licence and haven't even started it because they probably think, what am I going to use this for?"

Respondent 2 (Org A) noted that this gap is compounded by the rapidly shifting landscape of available tools, making internal communication about what is sanctioned an ongoing challenge:

"Thanks to Microsoft we got both GPT and Anthropic's models given that they collaborate with both. So that's been an advantage, but it's hard to navigate now when so much is happening all the time."

Respondent 7 (Org C) mentions working on an information campaign in order to raise awareness and speak on how guidelines are communicated to employees, as well as bringing up challenges connected with this:

"So one of the things we're currently doing is both putting out an information campaign [...]. So we have the problem that some people think that an inspiration session done by a specific partner or stakeholder leads to something being an approved application. No, it's not. We've previously said that only approved applications may be used and so on. But people have started stretching the rules and tried to say 'yes but this is approved'. No, it's not approved. You considered that everyone uses it and incorrect behaviour, but that doesn't mean something was approved just because people do it wrong. So we need to be clearer about that and communicate further, that we go back to what's expected and so on."

4.5.4 Governance processes as a structural driver

Another driver concerns the governance process itself. Several respondents argued that overly slow or inconvenient approval procedures can become a structural cause of Shadow AI, as motivated employees bypass processes they perceive as barriers rather than safeguards.

Respondent 3 (Org A) mentioned this most directly:

"The biggest challenge, well it's precisely that AI is everywhere and it's easy if you have such a process that it might become a bottleneck. [...] So that simply you can't wait, you can't wait for that approval. Instead you ignore the process because it becomes too inconvenient and takes too long."

Respondent 8 (Org D) described how their organisation are working on developing fast-track approval pathways for lower-risk use cases:

"So our biggest challenge for… on this kind of process is that the existing process, [...] it's based on the idea that we need to approve new use cases that take three to six months or longer to go through [...]. But AI speeds up everything and so if an employee can add a new use case in an hour by building an AI agent, then we have to have a process that is as fast as that right. And so our biggest challenge now is try to figure out how to speed up our existing processes. [...] when certain kinds of use cases or fit certain guidelines that they're lower risk for some reason, then we can fast track and not go through every process and make a process much faster. So that's still a work in progress making that faster."

These two perspectives both tie into the challenge of staying on top of AI. With AI being such a fast-paced technology, having slow and inconvenient processes could act as a driver to

Shadow AI as employees want to work with the best technology and work as efficiently as possible.

4.5.5 External hype and social influence

Several respondents identified external hype and social influence as a driver of interest in unsanctioned tools, though they interpreted this differently.

Respondent 7 (Org C) named the underlying psychological mechanism directly:

"Hype, curiosity and everyone else is doing it. [...] People's desires. Again FOMO (Fear Of Missing Out), 100% FOMO."

Respondent 6 (Org C) described this more as contagious enthusiasm rather than individual evaluation:

"Some of it is certainly hype, that you experience or someone says that something is better than something else. Saab is better than Volvo or BMW is better than Mercedes. Then you believe it."

Respondent 4 (Org B) illustrated how hype can take on a more principled dimension, describing a situation where an ethical concern about a sanctioned tool's provider drove interest in an unsanctioned alternative:

"There's been some discussion about whether Claude Code is an ethically better alternative given that OpenAI approved the agreement with the US defence forces, which Anthropic did not. And when things like that happen there's always a small spike in that there are some who think that this new tool, this one we haven't approved centrally, is something I want to stand behind personally. But I think there will always be such people and there will always be that type of reasoning."

Respondent 4 (Org B) also adds that they think external influence always will be an apparent attitude from certain employees by saying that "*There will always be some who think now I've seen a new test and in that test that tool was a bit more effective so therefore I should use it*".

Respondent 2 (Org A) noted that the pace of change in the AI market makes hype a persistent governance challenge:

"[...] if you look at the market right now, it's three-month cycles. Where there are new leaders every quarter in the AI industry. Which means that those who are really constantly following this market always want a new tool every three months, it's become very clear now, Claude with Anthropic has been at the forefront now, then a week or so ago the new GPT model comes out which is the quarter's hype and then you want that."

4.6 Documentary Analysis

To complement the interview data, AI Policy documents were collected from four of the seven participating organisations. One organisation does not currently have an AI policy and therefore has not provided one. Another organisation, a relatively new company, has not yet developed a formal AI policy, but notes that their current policies touch on the subject of what to be aware of when using AI tools. Those policies however were confidential and have therefore not been provided. The remaining organisation did not respond to the request of providing their AI policy. Where available, documents were used to contextualise interview findings rather than verify respondent accounts.

Organisation A provided a formal AI directive (version 2.0), developed jointly by the Digital & Tech and Legal functions and approved at the group management level. This document establishes a three-tier governance model consisting of an AI Council, an AI Committee and an acquisition function, each with defined responsibilities for risk assessment and approval of AI functionality. This confirms Respondent 3's description of a cross-functional evaluation group bringing together legal and technical competencies before any tool is approved for use, and Respondent 1's reference to an AI Council through which all new tools must pass. The directive also clearly requires human oversight of AI outputs and requires that results always must be reviewed and validated by the user, directly reflecting the human-in-the-loop emphasis described by Respondent 2. On data handling, the directive instructs employees to carefully consider the risk of sharing confidential information or personal data with AI systems, aligning with the GDPR concerns raised by Respondents 1 and 3. The directive also addresses output quality directly, warning that AI-generated outputs may be incorrect, misleading or legally non-compliant, consistent with the over-reliance concerns raised across multiple respondents from Organisation A.

Organisation C provided an existing and an incoming revised policy. The existing policy is brief and principle-oriented, establishing general expectations around data protection and ethical use without detailed operational guidance, consistent with Respondent 6's description of governance currently relying too heavily on individual judgment. The incoming policy is considerably more structured, introducing formal approval processes, a user register, business-area usage restrictions and defined non-compliance consequences. The organisation will be using option B, meaning AI use is permitted but restricted to company-approved and licensed tools only, with unauthorised AI explicitly prohibited. This trajectory corroborates both respondents' accounts of an organisation that recognises its governance gaps and is actively working to address them. Notably, however, several sections of the incoming policy remain incomplete, including the list of approved AI technologies and specific data input restrictions. This is itself consistent with Respondent 7's observation that inconsistent messaging from leadership, where some permit what others would block, creates significant governance problems, and Respondent 6's description of the difficulty in reaching unanimous decisions about which tools to formally approve or restrict.

Organisation D provided an AI Manifesto rather than a formal policy document. The document is principles-based and technology-neutral, organised around strategic pillars rather than rules or approval processes. This is consistent with Respondent 8's description of a governance philosophy that prioritises enablement and starts from principles rather than controls. The manifesto states that human accountability applies to all AI-generated output, that sensitive data must be protected within approved systems and that a human-in-the-loop is required for outputs affecting customers, compliance, or production systems. These were all described as organisational requirements by Respondent 8 and 9. The document frames AI as a productivity amplifier rather than a replacement, and includes an explicit commitment to investing in training and role-specific enablement, consistent with Respondent 8's statement of the organisation's efforts to ensure employees have access to capable sanctioned tools. The manifesto makes no reference to consequences for non-compliance or formal approval processes for new tools, whereas Respondent 8 separately described such processes as being actively developed.

Organisation F provided two documents: one Group AI Use Directive and one Group Responsible AI Policy, both adopted in April 2025 and approved at the board level. The AI Use Directive explicitly requires that all AI-generated outputs must be verified for correctness before use and prohibits automated decision-making in matters which affect employees. The directive also addresses data classification directly, permitting sensitive information to be processed only in sanctioned services following a prior risk assessment. This connects to Respondent 11's statement of how the organisation has moved from a communication-based approach, allowing employees to use AI as long as it was solely with public information, to the technical blocking of unsanctioned tools. The directive's instructions to keep trial use of public AI services to a minimum also align with Respondent 11's description of approximately 200 tools being blocked, while simultaneously negotiating enterprise contracts with major providers to ensure sanctioned alternatives were available. The Responsible AI Policy covers seven principles, with one being training, consistent with Respondent 11's statement of training being reset whenever policy updates are made.

5. Analysis and Discussion

This chapter analyses and discusses the empirical findings presented in the previous chapter. The findings are discussed in relation to the study's research question and sub-questions and are discussed in relation to the five themes identified in the thematic analysis. While the empirical findings chapter presented the themes in the order they emerged from the data, this chapter opens with the theme of identified risks, as it most directly speaks to the study's research questions and provides the foundation upon which the subsequent discussion of governance approaches, mechanisms and Shadow AI drivers is built.

5.1 Risk Perception vs Risk Reality

The rise of AI tools has created both significant opportunities and considerable risks within organisations. Crocco et al. (2026) describe these as conceptual risks, related to moral and legal harm and practical risks, such as unregulated use and the unpredictability of AI outputs. This is also seen in the empirical findings of this study. A recurring pattern across the organisations is that AI-related risks are framed, both in organisational policies and in employee awareness, primarily as a question of which tool is being used. An implicit assumption is that sanctioned tools are safe and unsanctioned tools are risky. Even though unauthorised AI tools may expose data leakage and lack transparency, the empirical findings show that risks also arise from how employees use these AI tools. This makes the boundary between safe and unsafe less about which tool is used and more about the behaviour surrounding it.

The risk of output quality and over-reliance supports this point further. Respondents across the organisations noted employees' tendency to trust AI-generated outputs without critically reviewing them. As several respondents noted, the quality of AI output is determined not only by which tool produces it but also by how the user prompts it. The phrase *"garbage in, garbage out"* appears several times throughout the interviews, indicating that the quality of prompting directly shapes the output quality. This shows that many governance mechanisms have a limited impact on this type of risk, with the data suggesting it is left primarily to training and education. Access controls or technical restrictions will not help detect whether an employee has poorly prompted a tool, acted on a hallucinated output or failed to verify a result before incorporating it into their work. Respondent 9 (Org D) acknowledges this as a significant gap, noting that there is currently no reliable control function to detect when an employee has acted on incorrect AI output and that organisations must rely on employees to take personal responsibility for verification. This is reflected in the policy documents as well, with Organisation A's directive explicitly warning that AI-generated outputs may be incorrect, misleading or legally non-compliant, and Organisation D's manifesto requiring human

accountability for all AI-generated outputs regardless of which tool produced it. This connects to a broader accountability concern identified by Raizada et al. (2026), arguing that unmonitored AI use results in untraceable decision-making, where organisations cannot reconstruct how AI-supported decisions were reached. It also relates to the kind of risk Anthuvan et al. (2025) identify when arguing that the informal adoption of AI tools has created a risk that traditional IT governance frameworks are not designed to address.

This is directly connected to the concept of automation bias discussed in the theoretical framework. Romeo and Conti (2026) define automation bias as the tendency to rely too much on automated systems without independent judgment and therefore failing to detect errors. The empirical findings point to a remarkable risk of automation bias in both sanctioned and unsanctioned AI tools. One respondent noted that hallucination is particularly hard to detect in specialist domains, where users have little expertise within the subject. This may result in the user being unable to identify when an AI has produced a confident but incorrect answer, consistent with Romeo and Conti's (2026) finding that lower domain expertise increases vulnerability to the uncritical acceptance of AI outputs. The findings acknowledge that there is currently no reliable control function to detect when an employee has acted on incorrect AI output, leaving organisations dependent on individual judgment. This reflects on a governance gap that Romeo and Conti (2026) address, arguing that transparency mechanisms alone are insufficient to mitigate automation bias and that the most effective way is to design a system that actively promotes critical engagement and independent verification from users. Confirmed in Organisation A and D’s policies, as well as mentioned by their respective respondents, they have adopted human-in-the-loop requirements in order to require user verification, though most organisations have not yet operationalised this, leaving the risk of over-reliance largely ungoverned in practice.

Another risk identified in the empirical findings concerns vendor and supply chain exposure, spanning both primary AI providers and third-party tools built on their underlying models. While Ross et al. (2025) briefly acknowledge that data leakage risks are increased when third-party platforms are involved, the empirical findings suggest this deserves recognition as a distinct and broader governance concern. Respondent 1 (Org A) highlighted the security posture of smaller vendors building tools on foundational models, noting that niche tools may lack sufficient security investment and thereby introduce new attack vectors into the organisation. Beyond third-party tools, concerns were also raised about primary providers themselves. Respondent 11 (Org F) described how certain providers’ lack of EU data residency options creates a concrete barrier to adoption and adds a significant complexity to governance, while Respondent 4 (Org B) noted that unsanctioned tool use leaves organisations with limited visibility into where data ends up. Additionally, Respondent 6 (Org C) raised questions about output ownership, particularly regarding who owns content generated through tools such as Claude Code. This raises a distinct but related accountability concern, where the lack of transparency of third-party tools and their providers makes it hard to trace where the data goes and who ultimately controls it. This complements Raizada et al.’s (2026) argument about untraceable decision-making, extending beyond the internal workflows to the vendor relationship. Taken together, these findings suggest that vendor and

supply-chain risks represent a governance gap that few existing studies have addressed, despite emerging as a notable and consistent concern across the organisations studied. This leaves a valuable space for future research to investigate how organisations can govern third-party AI exposure in practice.

These findings point to a meaningful gap between how AI risk is perceived and where it actually concentrates. The governance approach of treating sanctioned tools as safe and unsanctioned tools as risky is understandable but not entirely valid. Addressing this gap requires more than restricting compliance-based approaches. As Anthuvan et al. (2025) argue, effective risk management in this context will require governance frameworks that are responsive to real-world use and position employees as responsible participants in AI use, rather than potential rule-breakers. The empirical findings support this and suggest that these kinds of risks can only be meaningfully addressed if employees have the right knowledge and habits to engage critically with AI tools, regardless of which tool they use.

5.2 The Governance Spectrum and its Paradox: How restrictions can drive the behavior it seeks to prevent

Based on the empirical findings, organisations do not approach AI governance uniformly. It is rather a position along a spectrum, ranging from control-heavy, restriction-oriented governance at one end, to open, enablement-oriented governance at the other. Above all, where an organisation sits on this spectrum is rarely accidental. The organisation's approach reflects beliefs about innovation, risk tolerance and trust in employee behaviour. The approach represents a foundational governance choice that shapes the mechanisms discussed in the subsequent sections. This makes it the primary frame through which the findings of this study are understood and analysed.

Across all organisations, a consistent preference for enablement over restrictions was expressed. Respondents broadly questioned whether restrictive governance reliably prevents the use of Shadow AI. Several noted that employees tend to find ways around controls rather than comply and that prohibition does not remove the motivation to use tools that they believe better meet their needs. This aligns with Anthuvan et al. (2025), discussed in section 2 of the study, who found that employees use unsanctioned tools to work more efficiently, bypass slow processes and navigate vague AI guidelines. This supports the notion that underlying motivations for Shadow AI use can remain even if an AI policy exists. This suggests that governance approaches primarily focused on restriction may be treating the symptom rather than the cause. The empirical findings provide clear grounds for this: Respondent 13 (Org G) argued that if employees are told no, they will use unsanctioned tools anyway, and Respondent 8 (Org D) similarly framed an anti-AI policy as an implicit endorsement of Shadow AI. If employees turn to unsanctioned tools because sanctioned alternatives are perceived as insufficient, approval processes are too slow or awareness of

available tools is lacking, then the more effective governance response is one that addresses those conditions directly, rather than one that simply prohibits the behaviour they produce.

This paradox can further be understood through the lens of organisational resilience, as discussed in previous research. Francis and Bekera (2014) argue that resilience is less about predicting and preventing disruption, and more about promoting flexibility and adaptive response. The governance approach described in the empirical findings closely reflects this logic. Rather than attempting to eliminate Shadow AI through strict prohibition, several organisations advocated governance structures that are sufficiently flexible to absorb and redirect informal AI usage into sanctioned systems. This is possibly most clearly illustrated by Respondent 11 (Org F), who described how the discovery of around 10,000 employees accessing ChatGPT without authorisation led the organisation to obtain an enterprise licence, bringing that usage into sanctioned channels. This approach aligns with a broader understanding of resilience, not as the ability to avoid disruption entirely, but as the capacity to adapt and respond when it occurs. Several respondents framed this as a preference for guardrails over prohibition, with technical and procedural controls designed to indirectly influence behaviour in desired directions without explicitly restricting it. In this sense, governance is constructed as a flexible barrier and an enabling framework that guides and shapes organisational behaviour under conditions of uncertainty. Wang and Cao (2025) further reinforce why this adoptive orientation is particularly necessary in the context of AI. Their findings show that AI's inherent variability, complexity and uncertainty place higher demands on organisational resilience than traditional technologies do. It suggests that reactive responses alone are insufficient and that organisations benefit from governance mechanisms that proactively build adaptive capacity in addition to reactive responses. This is something that the empirical findings of this study reflect in how organisations implement proactive mechanisms while still enabling AI to the best of their ability.

A significant finding within this theme is that an organisation's position on the governance spectrum does not appear to determine the extent to which Shadow AI occurs. Rather, it shapes how organisations perceive and respond to it. In organisations with a more open, enabling-oriented governance approach, Shadow AI is not treated as a violation to be eliminated. Across these organisations, restrictions were broadly seen as neither a viable nor desired response to Shadow AI. These organisations tend to allow employees to explore and test AI tools, viewing this as a way to gather knowledge and make informed decisions about which tools to formally adopt and have a license for. This focus leans towards ensuring that usage is controlled, that processes are being tested and no sensitive data is put at risk. The existing literature positions Shadow AI primarily as a sociotechnical governance problem centred on risk containment (Silic et al., 2025), though the findings of this study suggest a more nuanced picture, reframing how Shadow AI governance should be viewed. If the volume of Shadow AI is not meaningfully reduced by where an organisation sits on the governance spectrum, then the primary governance question is not how to eliminate it, but rather how to make it visible and manageable. This gives a new perspective on how existing literature frames the problem. Instead of focusing on detection, restriction and enforcement as the primary response, the findings of this study suggest that restrictive governance does not

remove the underlying motivation to use sanctioned tools. It rather pushes the usage out of sight, away from organisational oversight.

An additional observation from this theme is that governance design alone is insufficient. Several respondents highlight that how governance is communicated and embedded across the organisation is as important as the specific rules it contains. A policy that exists but is not understood, internalised, or supported is unlikely to influence employee behaviour or reduce the risk it was designed to address. Respondent 10 (Org E) directly captures this, arguing that policies alone will never change behaviour and that education is what matters. Respondent 9 (Org D) mentions this in another way by comparing to employees treating AI policies like software terms and conditions, scrolling through and clicking next without genuine engagement. In earlier research, Horneber and Laumer (2023) argue that most organisations currently rely on principle-based approaches to govern their AI systems. This has been criticised as insufficient due to the difficulties involved in translating ethical principles into practice. The findings of this study suggest that the same challenge applies to AI governance more broadly. There is a gap between what formal governance structures require and how this is reflected in employees' behaviour in practice, as clearly shown in the observed organisations in this study.

5.3 Technical and Organisational Mechanisms as Complementary

The empirical findings reveal a broad range of governance mechanisms implemented across the organisations, including both technical and organisational dimensions. The primary technical mechanisms discussed in the empirical findings chapter were information classification, monitoring and detection, access controls and licence-gating and human-in-the-loop requirements. These mechanisms operate directly on systems and data flows in order to control what information can enter or leave AI tools and to detect when organisational boundaries are crossed. Human and organisational mechanisms, such as formal policies, training programmes and approval processes, work through people rather than systems, seeking to influence behaviour through guidance, accountability and awareness. While these two categories of mechanisms are distinct in nature and function, empirical findings consistently point to the same conclusion: neither is sufficient on its own.

A repeated observation across the organisations was that technical controls can restrict what is possible but do not govern how employees think, interpret or act on information. Equally, policies and training establish what is expected of employees, but the effectiveness relies entirely on employees actually reading, understanding and internalising them, something the findings show is far from guaranteed in practice. Rather than relying on a single mechanism, the empirical findings suggest that these mechanisms do not function in isolation but require each other to be effective. The technical mechanisms should be used to set boundaries and the organisational mechanisms to build the awareness and culture needed to operate within them responsibly.

Looking at the technical mechanisms, the empirical findings show that the primary approach lies in reducing risk rather than preventing unsanctioned usage entirely. Having information classification, for instance, does not hinder employees from accessing external AI tools, but it limits the damage that incorrect usage can cause by controlling which data is permitted to be used in these tools. This reflects a logic that several respondents point out regarding the fact that protecting information is a more realistic and effective goal than restricting access to tools. This is consistent with Puthal et al. (2025), who identify data leakage and unregulated handling of sensitive information as among the most significant risks with Shadow AI. The policy documents reflect the same logic, with Organisation A's directive focusing on the risk of sharing confidential data rather than restricting tool access, Organisation C's incoming policy introducing data classification restrictions and Organisation F's directive permitting sensitive information only in sanctioned alternatives. Drawing from this, the findings suggest that restricting tool access without controlling data flows may address the wrong problem, shifting the relevant question from which tools employees use to what information those tools can access.

Monitoring and detection mechanisms similarly do not prevent Shadow AI from occurring, but rather make it visible when it does. This can be connected back to the earlier argument in the preceding section (5.1), that an important governance question is not how to eliminate Shadow AI but rather how to surface it and bring it under control. As described in the empirical findings, Respondent 7 (Org C) explained a feature of their future approach. When they detect an unsanctioned tool on an employee's device, it does not result in an immediate block. Instead, they will gradually increase the device's risk level, eventually requiring repeated re-authentication to access company data. This reflects an approach that prioritises awareness and a structured response over prohibition.

While the technical mechanisms address the structural dimensions of governance, they say little about the behavioural ones. They cannot ensure that employees understand why those boundaries exist, what the organisational policies are, or how to use AI tools responsibly within them. This is where human and organisational mechanisms become essential. The empirical findings describe these mechanisms as operating across different dimensions. The most commonly mentioned were training and awareness programmes aimed at building the knowledge and understanding employees need to act within those expectations. Further, the findings mentioned an approval process to ensure that new AI tools are evaluated for potential risks before adoption, while simultaneously providing employees with clear guidelines for introducing new tools, rather than resorting to unsanctioned use. Additionally, formal policies were mentioned as a common mechanism for establishing a shared understanding of expectations, yet the findings revealed the limitations of these mechanisms on their own. There was a clear acknowledgement that policies often are unread and that individual responsibility as a governance layer is unreliable without clear organisational direction. This aligns with Horneber and Laumer (2023), who argue that principle-based governance approaches are insufficient when organisations struggle to translate stated principles into consistent practice. The gap between what the policy says and how employees

actually behave remains one of the most significant challenges in governance across organisations.

When looking at the findings and comparing the two mechanisms, the analysis suggests that the most effective governance approach is one that combines both technical and organisational mechanisms, designed with an awareness of each other's limitations. Technical mechanisms provide structural safety without individual awareness or intention, while organisational mechanisms build the understanding and culture that make employees active participants in governance.

This is an important implication of how AI governance is designed in practice. Connecting back, the empirical findings make it clear that neither technical mechanisms nor organisational mechanisms can ensure safety across the organisation on their own, they are interdependent. Based on these findings, it suggests that governance effectiveness is not only a question of which mechanisms are in place but of how clearly they are designed and communicated as a whole. An organisation with strong technical controls but poorly understood policies, or well-established training but no technical safeguards, creates gaps that neither mechanism can close independently.

5.4 Shadow AI as a structural outcome

The empirical findings show that Shadow AI is recognised and widely reflected upon as a phenomenon across the organisations studied. The reasons employees reach for unsanctioned tools beyond the organisations' sanctioned alternatives have been shown to be varied. It ranges from perceived gaps in tool performance and external hype to ingrained habits formed through private AI use.

A primary driver of Shadow AI identified across the organisations studied was the perception that unsanctioned tools simply perform better at certain tasks. Respondents described this not as disregard for governance or defiant behaviour, but as a reason to work effectively and avoid falling behind. Therefore, they turn to other tools where the sanctioned tools fall short. This directly supports Anthuvan et al. (2025), who argue that employees use unsanctioned tools primarily to work more efficiently, bypass slow processes and compensate for vague guidelines. This is further consistent with Silic et al. (2025), who identify misalignment between official tools and job requirements as a central driver of unsanctioned tool adoption, a pattern that the findings of this study suggest extends into the context of Shadow AI. The empirical findings extend this further by addressing that the perceived performance gap is not always about the tools themselves. Several observations show that employees may underestimate the capabilities of sanctioned tools due to insufficient prompting, as discussed earlier. This suggests that part of what drives Shadow AI is the knowledge gap rather than a tool gap. By addressing this through training and awareness, organisations could narrow the perceived gap between sanctioned tools and employee needs, reducing the motivation to seek alternatives beyond the implemented systems.

A second and closely related driver is the indistinct boundary between private and professional AI use. Several respondents suggested that employees who use AI tools in their personal lives may, out of habit and familiarity, default to those same tools at work, often without a conscious decision to violate policy. This is not framed by respondents as a deliberate non-compliance but rather as an unremarkable extension of existing routines, where the tools an employee reaches for at home are simply the same ones they reach for at their desk. This reflects a broader shift in how people incorporate technology into their routines, in which the boundary between personal and professional use has become increasingly blurred. Anthuvan et al. (2025) capture a related dynamic, finding that a significant share of employees access AI tools through personal accounts at work. However, the habitual and largely unconscious dimension underlying this behaviour, the idea that Shadow AI can emerge not from intent but from routine, is not explicitly addressed in the existing literature, representing a potentially valuable direction for future research.

A third driver concerns the governance process itself. Some respondents argued that too slow or inconvenient approval procedures can become a structural cause of Shadow AI. Zaidan and Ibrahim (2024) identify at the regulatory level that governance frameworks struggle to keep pace with innovation and the time required to establish adequate oversight. At the organisational level, this becomes a structural driver of Shadow AI rather than only a regulatory challenge. An example of this was the observation that existing approval processes could take between three and six months, while an employee could build a new AI-enabled use case in under an hour. This gap between governance speed and technological pace is significant. It points to Shadow AI being not only a product of employee behaviour, but also of institutional structures that have failed to adapt to the conditions under which AI is now being used. This is consistent with what Silic et al. (2025) identify in the transition from Shadow IT to Shadow AI, noting that employees turn to unsanctioned tools not only out of convenience but because official processes are perceived as misaligned with their job requirements and too slow. The findings of this study add further weight to both of these points, with multiple respondents identifying slow approval processes as a structural driver that pushes motivated employees toward Shadow AI rather than waiting for formal authorisation. These observations suggest that the governance process itself, not just the approval process for individual tools, can lag behind the pace of AI adoption. This is visible in the documentary analysis, where Organisation C’s incoming policy remains incomplete in several key areas despite representing a clear governance ambition.

Finally, external hype and social influence were identified as notable drivers of Shadow AI. The empirical findings point to several distinct aspects of this, including peer influence, fear of missing out and the rapidly shifting landscape of AI tools, where different models rise to prominence on a near-quarterly basis. This manifests as employees feeling pressure to keep up with what colleagues or the broader industry are using, as well as a constant pull toward whichever tool is currently leading the market. Another, more distinctive mechanism, raised within one organisation, was that an ethical concern about a sanctioned provider's partnership drove interest in an unsanctioned alternative. The respondent’s example regarded OpenAI’s approved agreement with the US defence force, which led some employees to seek

alternative tools based on ethical rather than performative grounds. This last example is particularly worth noting, as it illustrates that Shadow AI cannot always be understood as a policy breach driven by convenience or dissatisfaction. The motivation here was a value-based judgement that falls outside the scope of governance mechanisms, suggesting that the boundaries of Shadow AI as a concept may be broader than existing literature acknowledges.

5.5 Summary

Across this chapter, consistent patterns are shown. The risks organisations associate with AI use, whether stemming from unsanctioned tools, over-reliance on outputs, or vendor exposure, shape the governance philosophy they adopt and the mechanisms they implement in response. Yet the drivers of Shadow AI identified in the findings reveal that governance responses do not always address the underlying conditions that drive unsanctioned use in the first place. Tool gaps, habits, slow approval processes, and social influence are not only drivers of Shadow AI but also reflect limitations in the design and communication of governance. Together, these observations point to a governance challenge that is as much cultural and structural as it is technical.

6. Conclusion and Recommendations

The following chapter presents the conclusion drawn from the study. It opens with a summary of the key findings in relation to the research question and sub-questions, before writing recommendations for future research.

6.1 Conclusion and Key Findings

This study set out to examine the governance mechanisms that organisations using Microsoft Copilot implement to manage risks associated with Shadow AI, and what risks are identified with the use of unsanctioned AI tools.

The primary risks identified are centred around the three key concerns: data leakage, output quality and over-reliance, and vendor and supply chain exposure. A key insight is that these risks are not exclusive to unsanctioned tools. The boundary between safe and unsafe AI use is less about which tool is used and more about how it is used.

A notable finding is that organisations position themselves along a governance spectrum, ranging from restriction-oriented to enablement-oriented approaches. The position does not determine how much Shadow AI occurs, but shapes how organisations perceive and respond to it. The study suggests that a restrictive governance approach risks pushing unsanctioned use further out of sight, while enabling approaches tend to make it more visible and manageable.

The findings show that organisations implement governance mechanisms across both technical and organisational dimensions. Technical mechanisms include information classification, monitoring and detection, access controls and human-in-the-loop requirements. Organisational mechanisms include formal AI policies, training programmes and structured approval processes. A central finding is that neither technical nor organisational mechanisms are sufficient on their own. They are interdependent and most effective when designed and communicated in awareness of each other. However, the data also suggest that each of these mechanisms has limitations, and that the overall effectiveness of technical and organisational governance is heavily dependent on consistent communication, leadership alignment, and employee engagement.

A further finding is that Shadow AI is best understood as a structural outcome rather than individual non-compliance. The identified drivers, being perceived tool gaps, habitual use, slow approval processes and societal or ethical influence, suggest that unsanctioned tool use is largely a product of governance structures that have not kept pace with the speed of AI adoption and development.

In conclusion, the findings suggest that effective governance of Shadow AI requires an approach that addresses the conditions that drive it in the first place, combining technical safeguards with cultural awareness, trust and adaptability. Governance frameworks that position employees as responsible participants, rather than potential rule-breakers, and that are designed to evolve alongside the technology they seek to govern, appear best suited to the conditions under which AI tools are now used in organisations. Equally, the findings point to the importance of training individuals in how to use AI tools responsibly, as awareness and knowledge at the individual level appear to be a necessary complement to organisational governance structures.

6.2 Recommendations for Future Research

Several directions for future research emerge from this study. Vendor and supply chain risks, spanning data residency, output ownership and the security of third-party tools built on foundational models, represent a governance gap that few existing studies have addressed. Future research could investigate what frameworks and contractual mechanisms are best suited to governing third-party AI exposure in practice.

The relationship between organisational size, industry context and governance effectiveness justifies further attention. Scale and sector appear to shape both the risks organisations experience and the available mechanisms. Comparative studies across different organisational contexts could result in more differentiated guidance for governance design. Additionally, the habitual dimension of Shadow AI, where employees default to personally familiar tools without deliberate intent to violate policy, remains underexplored and represents a valuable direction for future study.

This study examined governance through the perspectives of primarily those responsible for designing it. Future research could complement this by exploring employee perspectives more directly, examining how governance is experienced, interpreted and acted upon by those it is designed to influence.

Appendix 1: AI Contribution Statement

In the completion of this bachelor's thesis, the following AI-based tools were used:

1. Tools used:
    - Claude (Anthropic)
    - Grammarly
    - Sunet Scribe

2. Degree of usage:
    - Claude was primarily used as a support tool for text formulation and linguistic refinement, including rephrasing, improving sentence structure, and enhancing overall flow. Claude was also used as an analytical aid in the thematic analysis, contributing to the structuring and development of themes derived from the collected interview material, which concerned the methodology chapter and the empirical findings chapter, prompts for these can be found in Appendix 18 and 19. Since most of the interviews were held in Swedish, Claude was also used for translating the interviews to English, through the prompt in Appendix 17.
    - Grammarly was used throughout the thesis as a complement for grammatical proofreading and linguistic quality assurance.
    - Sunet Scribe was used for the transcription of the semi-structured interviews conducted as part of the data collection.

Appendix 2: Interview Guide

Background & Role

1. What do you do for work?
2. For how long have you had this role?

Follow-up: Has your role changed with the growth of AI, and if so, how?

AI-usage within the organisation

1. How is Microsoft Copilot currently used within the organisation? In which process and by which function?

   Follow-up: Has usage changed since the tool was introduced?

2. Do you find that employees use AI tools beyond those provided by the organisation? In what contexts does this occur?

Risks with AI

1. What risks has the organisation identified in connection with the use of AI-generated responses and outputs in daily work?

   Follow-up: How does the organisation manage risks related to incorrect or misleading information generated by AI tools?

2. How does the organisation manage questions of data integrity and information security in connection with employees using Copilot? For instance, when sensitive information is entered into the tool?

   Follow-up: Are there any technical restrictions in the organisation, or does it primarily rely on trusting the employees to follow guidelines, etc?
   Follow-up: What do you define as sensitive information?

AI-policy

1. Does the organisation have a formal AI policy or guidelines for the use of AI tools? How is it designed and what does it cover?
   Follow-up: How is the policy communicated to employees and how do you ensure compliance?

2. What concrete governance mechanisms has the organisation implemented to manage risks related to AI use?

   Follow-up: What works well and what has proven more difficult to implement?

View on AI + Shadow AI

1. Is your organisation familiar with the concept of Shadow AI?
2. To what extent does the organisation experience Shadow AI occurring?

   Follow-up: What do you believe drives employees to use other AI tools when Copilot is available?

Regulatory Frameworks

1. How do you (the organisation) relate to external regulatory frameworks such as the EU AI Act or international standards such as the NIST AI Risk Management Framework in your AI governance work?

    Follow-up: How does the organisation work with questions of accountability and transparency related to AI-generated decisions or outputs?

Control & Follow-up

1. How does the organisation follow up to ensure that AI usage guidelines and policies are actually adhered to in practice?

    Follow-up: Are there formal incident management processes if an AI tool causes an error or a security incident?

Challenges

1. What do you see as the greatest challenge in governing and controlling AI use within the organisation today?

    Follow-up:: Do you find that overly strict rules risk driving employees toward Shadow AI rather than preventing it?

2. Does the organisation have upcoming new AI-governance in order to change the current guidelines and governance mechanisms?

Conclusion

1. Is there something we haven't asked about which you find important for our study?
2. What do you hope to get out of our study?

Appendix 17: Prompt for Translation – Claude

You are assisting with a qualitative research project as part of an academic bachelor's thesis. Your task is to translate the following interview transcript from Swedish to English.

Please follow these guidelines when translating:

1. Accuracy over fluency, prioritise faithfully conveying the meaning of what was said, even if it results in slightly informal English phrasing.
2. Preserve natural speech, retain hesitations, repetitions, incomplete sentences, and colloquial language where present, as these are analytically significant in qualitative research.
3. Maintain speaker labels, keep all speaker identifiers (e.g. "Interviewer:", "Respondent 1:") exactly as they appear in the original.
4. Preserve formatting, maintain the original structure, line breaks, and any transcription conventions (e.g. [...] for pauses or omissions).
5. Do not interpret or summarise, translate the full transcript verbatim. Do not omit, paraphrase, or add any content of your own.
6. Cultural and idiomatic expressions, where a direct translation would obscure meaning, provide the closest English equivalent and add a brief translator's note in square brackets (e.g. [Swedish idiom, closest equivalent: ...]).

The transcript to be translated follows below. Output only the translated transcript, with no preamble or commentary.

Appendix 18: Prompt for Initial Coding – Claude

I am conducting a qualitative thematic analysis for my bachelor's thesis. The study explores how Swedish organisations manage governance challenges related to Shadow AI in the context of Microsoft Copilot adoption. Below is an interview transcript. Please read through it and suggest initial codes for meaningful sections of text. For each code, provide the relevant quote, a short code name, and a brief explanation of what it captures. Do not generate themes yet, only initial codes.

Transcript:

Appendix 19: Prompt for Clustering Codes – Claude

I am conducting a qualitative thematic analysis for my bachelor's thesis. The study explores how Swedish organisations manage governance challenges related to Shadow AI in the context of Microsoft Copilot adoption.

Below is data exported from an Excel file containing initial codes generated from interview transcripts. Each row contains a quote from an interview and its corresponding code.

Your task is to cluster these codes into groups based on similarities and patterns, this is NOT the final themes, but an intermediate step to identify candidate theme clusters. Use both the codes and their corresponding quotes to inform your clustering, as the quotes provide important context for what each code captures.

For each cluster:

- Give the cluster a working title that captures what the codes have in common
- List which codes belong to it
- Write a brief explanation (2-3 sentences) of what the cluster represents
- Include 1-2 illustrative quotes that best represent the cluster

Do not reduce the data too aggressively, it is better to have more clusters at this stage than too few. Some codes may belong to more than one cluster if relevant.

Here is the data

References

Anjos, J., de Souza, M. G., Serrano, A. & Campello de Souza, B. (2024) An analysis of the generative AI use as analyst in qualitative research in science education, *Qualitative Research Journal*, vol. 12, no. 30, pp. 01–29, doi: 10.33361/RPQ.2024.v.12.n.30.724

Anthuvan, T., Prabhuram, S., Raju, G., Maheshwari, K., & Mishra, A. (2025). Bring Your Own AI (BYOAI) in the Workplace: Patterns, Pitfalls, and the BYOAI-Gov™ Framework for Behavior-aware Governance. *Lex Localis: Journal of Local Self-Government*, vol. 23, no. S6, pp. 2693–2716. https://doi.org/10.52152/mmrgh422

Aven, T. (2016). Risk assessment and risk management: Review of recent advances on their foundation. *European Journal of Operational Research*, vol. 253, no.1 https://doi.org/10.1016/j.ejor.2015.12.023

Batool, A., Zowghi, D. & Bano, M. (2025). AI governance: a systematic literature review. *AI and Ethics*, vol. 5, pp. 3265-3279, https://doi.org/10.1007/s43681-024-00653-w

Bell, E. & Bryman, A. (2007). The Ethics of Management Research: An Exploratory Content Analysis. *British Journal of Management*, vol. 18, pp. 63-77, DOI: 10.1111/j.1467-8551.2006.00487.x

Braun, V., & Clarke, V. (2023). 'Thematic analysis', In Cooper, H., Coutanche, M. N., McMullen, L. M., Panter, A. T., Rindskopf, D., & Sher, K. J. (eds.), *APA handbook of research methods in psychology: Research designs: Quantitative, qualitative, neuropsychological, and biological*, vol. 2, pp. 65–81.Washington, DC: American Psychological Association. https://doi.org/10.1037/0000319-004

Bryman, A. (2016). Social research methods. 5th ed. Oxford University Press.

Bolgouras, V., Zarras, A., Leka, C., Stylianou, I., Farao, A. & Xenakis, C. (2025). EU regulatory ecosystem for ethical AI. *AI and Ethics*, vol. 5, pp. 5063-5080, https://doi.org/10.1007/s43681-025-00749-x

Bowen, G. A. (2009). Document Analysis as a Qualitative Research Method. *Qualitative Research Journal*, vol. 9, no. 2, pp. 27–40, https://doi.org/10.3316/QRJ0902027

Christou, P.A. (2024). Thematic Analysis through Artificial Intelligence (AI), *The Qualitative Report*, vol. 29, no. 2, pp. 560–576, DOI: 10.46743/2160-3715/2024.7046

Crocco, E., Ballesio, E., Broccardo, L., & Alghafes, R. (2026). Welcome to the dark side: unveiling the potential risks related to the implementation of Artificial Intelligence. *Journal of Risk Research*, pp. 1-21, https://doi.org/10.1080/13669877.2025.2611954

Denzin, N. K. (2009 [1978]) The research act: A theoretical introduction to sociological methods. Routledge. https://doi.org/10.4324/9781315134543

Dotan, R., Blili-Hamelin, B., Madhavan, R., Matthews, J. & Scarpino, J. (2024) Evolving AI Risk Management: A Maturity Model based on the NIST AI Risk Management Framework. *ArXiv*, DOI: 10.48550/arXiv.2401.15229

Dula, I., Berberena, T., Keplinger, K. & Wirzberger, M. (2024). Conceptualizing Responsible Adoption of Artificial Agents in the Workplace: A Systems Thinking Perspective. *In: Agrifoglio, R., Lazazzara, A., Za, S. (eds) Navigating Digital Transformation.* vol. 73, pp. 229-249. https://doi.org/10.1007/978-3-031-76970-2_15

Dwivedi, Y.K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P.V., Janssen, M., Jones, P., Kar, A.K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., Medaglia, R., Le Meunier-FitzHugh, K., Le Meunier-FitzHugh, L.C., Misra, S., Mogaji, E., Sharma, S.K., Singh, J.B., Raghavan, V., Raman, R., Rana, N.P., Samothrakis, S., Spencer, J., Tamilmani, K., Tubadji, A., Walton, P., Williams, M.D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, vol. 57, article 101994, https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Dwivedi, Y.K., Kshetri, N., Hughes, L., Slade, E.L., Jeyaraj, A., Kar, A.K., Baabdullah, A.M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M.A., Al-Busaidi, A.S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., Carter, L., Chowdhury, S., Crick, T., Cunningham, S.W., Davies, G.H., Davison, R.M., Dé, R., Dennehy, D., Duan, Y., Dubey, R., Dwivedi, R., Edwards, J.S., Flavián, C., Gauld, R., Grover, V., Hu, M.C., Janssen, M., Jones, P., Junglas, I., Khorana, S., Kraus, S., Larsen, K.R., Latreille, P., Laumer, S., Malik, F.T., Mardani, A., Mariani, M., Mithas, S., Mogaji, E., Horn Nord, J., O'Connor, S., Okumus, F., Pagani, M., Pandey, N., Papagiannidis, S., Pappas, I.O., Pathak, N., Pries-Heje, J., Raman, R., Rana, N.P., Rehm, S.V., Ribeiro-Navarrete, S., Richter, A., Rowe, F., Sarker, S., Stahl, B.C., Tiwari, M.K., van der Aalst, W., Venkatesh, V., Viglia, G., Wade, M., Walton, P., Wirtz, J., Wright, R. (2023). So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, vol. 71, article 102642, https://doi.org/10.1016/j.ijinfomgt.2023.102642

Francis, R., & Bekera, B. (2014). A metric and frameworks for resilience analysis of engineered and infrastructure systems. *Reliability Engineering & System Safety*, vol. 121, pp. 90–103, https://doi.org/10.1016/j.ress.2013.07.004

Gbadegeshin, S.A., Natsheh, AI., Ghafel, K., Tikkanen, J., Gray, A., Rimpiläinen, A., Kuoppala, A., Kalermo-Poranen, J., Hirvonen, N. (2021). What is an artificial intelligence (AI): A simple buzzword or a worthwhile inevitability? *Kajaani University of Applied Sciences* DOI:10.21125/iceri.2021.0171

Guba, E.G. (1981). Criteria for Assessing the Trustworthiness of Naturalistic Inquiries. *Educational Communication and Technology*, vol. 29, no. 2, pp. 75-91, https://doi.org/10.1007/BF02766777

He, R., Cao, J. & Tan. T. (2025). Generative artificial intelligence: a historical perspective. *National Science Review*, vol. 12, no. 5, https://doi.org/10.1093/nsr/nwaf050

Horneber, D. & Laumer, S. (2023). Algorithmic accountability, *Business & Information Systems Engineering*, vol. 65, no. 6, pp. 723-730, https://doi.org/10.1007/s12599-023-00817-8

International Organization for Standardization (ISO). (2018). *ISO 31000:2018 Risk management — Guidelines*. Geneva: ISO. https://www.iso.org/obp/ui/en/#iso:std:iso:31000:ed-2:v1:en [Accessed 14 May 2026]

Kupfer, C., Prassl, R., Fleiß, J., Malin, C., Thalmann, S. & Kubicek, B. (2023). Check the box! How to deal with automation bias in AI-based personnel selection. *Frontiers in Psychology*, vol. 14, article 1118723, https://doi.org/10.3389/fpsyg.2023.1118723

Kwan, T. W. (2024). Navigating the Risks of Shadow AI, *ITNOW*, vol. 66, no. 2, pp. 42–43, https://doi.org/10.1093/itnow/bwae054

Lamchek, J. & Trieu, V-H. (2026). Risk regulation of generative AI: a case study of Microsoft Copilot in the Australian government, *Technology and Regulation*, vol. 2026, pp. 18-37, https://doi.org/10.71265/esg7cf11

Laux, J. & Ruschemeier, H. (2025). Automation Bias in the AI Act: On the legal Implications of Attempting to De-Bias Human Oversight of AI, *European Journal of Risk Regulation*, vol. 16, no. 4, pp. 1519-1534, DOI: 10.1017/err.2025.10033

Lincoln, Y. S. & Guba, E. G. (1985). Naturalistic Inquiry. Beverly Hills: Sage

Microsoft. (2026a). Data, privacy, and security for Microsoft 365 Copilot. https://learn.microsoft.com/en-us/microsoft-365/copilot/microsoft-365-copilot-privacy [Accessed 30 April 2026]

Microsoft. (2026b). Microsoft 365 Copilot app settings for IT admins. https://learn.microsoft.com/en-us/microsoft-365/copilot/microsoft-365-copilot-app-admin-settings [Accessed 30 April 2026]

Microsoft. (2026c). Data, privacy, and security for web search in Microsoft 365 Copilot and Microsoft 365 Copilot Chat. https://learn.microsoft.com/en-us/microsoft-365/copilot/manage-public-web-access [Accessed 30 April 2026]

Microsoft (2026d). Microsoft 365 Copilot overview. https://learn.microsoft.com/en-us/microsoft-365/copilot/microsoft-365-copilot-overview [Accessed 30 April 2026]

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://doi.org/10.6028/NIST.AI.100-1 [Accessed 21 April 2026]

Phillips, F. Y. & Chao, A. (2024). Rethinking resilience: definition, context, and measure. *IEEE Transactions on Engineering Management*, vol. 71, pp. 12289-12296, https://doi.org/10.1109/TEM.2021.3139051

Puthal, D., Mishra, A. K., Mohanty, S. P., Longo, A. & Yeun, C. Y. (2025). Shadow AI: Cyber Security Implications, Opportunities and Challenges in the Unseen Frontier. *SN Computer Science*, vol. 6, no. 5, https://doi.org/10.1007/s42979-025-03962-x

Raizada, A., Shakya, A. & Rahul, M. (2026). Shadow AI: Mapping the Risks of Unmonitored LLM Use in Enterprise Workflows. *Advances in Consumer Research*, vol. 3, no. 1,

pp. 1136-1141. https://acr-journal.com/article/shadow-ai-mapping-the-risks-of-unmonitored-llm-use-in-enterprise-workflows-2517/ [Accessed 19 April 2026]

Regulation (EU) 2024/1689. Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). European Parliament & Council of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

Regulation (EU) 2016/679. On the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). European Parliament & Council of the European Union. https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

Romeo, G. & Conti, D. (2026). Exploring automation bias in human-AI collaboration: a review and implications for explainable AI. *AI & Society*, vol. 41, no. 1, pp. 259-278, https://doi.org/10.1007/s00146-025-02422-7

Ross, J. A. J., Hibbert, L. & Moss, E. J. (2025) Shadow AI: Governance, Risk, and Organisational Resilience. *2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications* (ACDSA), pp. 1-9, DOI: 10.1109/ACDSA65407.2025.11166415

Shojaee, A., Rashed, M., Islam, M. F. & Vasa, L. (2025). Innovative configurations for organizational resilience: Bridging the proactive and reactive capability in volatile environments. *Sustainable Futures*, vol. 10, article 101236, https://doi.org/10.1016/j.sftr.2025.101236

Silic, M., Silic, D. & Kind-Trüller, K. (2025). From Shadow It to Shadow AI-Threats, Risks and Opportunities for Organizations. *Strategic Change*. DOI: 10.1002/jsc.2682.

Shao, Y., Huang, C., Song, Y., Wang, M., Song, Y. H., & Shao, R. (2025). Using Augmentation-Based AI Tool at Work: A Daily Investigation of Learning-Based Benefit and Challenge. *Journal of Management*, vol. 51, no. 8, pp. 3352–3390. https://doi.org/10.1177/01492063241266503

Shaw, F. X. (2024). Ignite 2024: Why nearly 70% of the Fortune 500 now use Microsoft 365 Copilot, Microsoft Blog Post, https://blogs.microsoft.com/blog/2024/11/19/ignite-2024-why-nearly-70-of-the-fortune-500-now-use-microsoft-365-copilot/ [Accessed 30 April 2026]

Wang, D. & Cao, X. (2025). Artificial intelligence adversity event, inter-organisational trust, and firm resilience: the moderating effect of responsible innovation. *Technology Analysis & Strategic Management*, vol. 37, no. 12, pp. 3051-3064. https://doi.org/10.1080/09537325.2024.2390075

Wingerter, T. L., Straub, T. & Schweitzer, S. (2025). Mitigating automation bias in generative AI through nudges: a cognitive reflection test study. *Procedia Computer Science*, vol. 270, pp. 2106-2114, https://doi.org/10.1016/j.procs.2025.09.331

Zaidan, E. & Ibrahim, I. A. (2024). AI governance in a complex and rapidly changing regulatory landscape: a global perspective. *Humanities and Social Sciences Communications*, vol. 11, article 1121, https://doi.org/10.1057/s41599-024-03560-x