Privacy Threats in MCP and A2A: Agentic AI Communication Protocols

A LINDDUN-Based Privacy Threat Analysis of MCP and A2A in a Financial Fraud Detection Context

Master Thesis

Willem Elshout

Delft University of Technology

TU Delft

avanade

Privacy Threats in MCP and A2A: Agentic AI Communication Protocols

A LINDDUN-Based Privacy Threat Analysis of MCP and A2A in a Financial Fraud Detection Context

Master thesis submitted to Delft University of Technology
in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in Complex Systems Engineering and Management

Faculty of Technology, Policy and Management
Delft University of Technology

Student name: Willem Elshout
Student number: 5347718
Thesis defence date: 22 June 2026

Graduation committee

Prof.dr.ir. G.A. (Mark) de Reuver *Section of Information and Communication Technology, TU Delft*	Chair & First supervisor
Prof.dr. M.E. (Martijn) Warnier *Section of Multi-Actor Systems, TU Delft*	Second supervisor
Valentijn Bieger *Avanade*	External advisor

TU Delft avanade

Preface

This thesis marks the end of my master's in Complex Systems Engineering and Management at Delft University of Technology. It would not have come together without the people who supported me, and I would like to thank them here.

First and foremost, I want to thank my main supervisor and chair of my graduation committee, Mark de Reuver, for all our meetings and email exchanges and for the care he put into guiding me. I also thoroughly enjoyed his teaching, in particular Digital Platform Design, which proved more relevant than I expected: MCP and A2A are open standards underpinning the platform ecosystems forming around agentic AI, and the very openness that makes them valuable is also the source of the privacy risks I analyse. Viewing them through the lens of platform governance, rather than as purely technical standards, is something his teaching prepared me for, and it was courses like this, at the heart of the Information and Communication track, that made me truly come to enjoy CoSEM.

I would also like to thank my second supervisor, Martijn Warnier, whose expertise in complex systems engineering, and especially in multi-agent systems, was a great help in sharpening the ideas in this thesis, and whose Complex Systems Engineering course I greatly enjoyed. I must admit that it is only now, having worked in this field myself and with AI agents in particular, that I have come to see its true value: agentic AI makes it tangible, with individually simple agents interacting to produce emergent properties that no single component displays on its own.

To both Mark and Martijn I owe particular thanks for the early stage of this project. Writing the research proposal was a struggle at times, but it was precisely because they were so demanding then that the thesis ended up far sharper in its framing. Their insistence on getting the problem definition right early on saved me time and effort later and kept me focused on the actual problem, and I am grateful that they pushed me when it mattered most.

My thanks also go to Valentijn Bieger, my supervisor at Avanade, who was always willing to spar with me and whom I could always turn to with my questions. I am also especially grateful to Kim Wuyts, a leading privacy engineering expert who led the development of the LINDDUN framework at KU Leuven, for taking the time to be interviewed even after moving on from the research world, and for letting me return to her with further questions; her guidance was invaluable. I would further like to thank the seven experts at Avanade who contributed to the validation interviews, whose input led me to new ideas and helped me understand my own work far better, and Avanade more broadly for hosting me as an intern. Finally, I thank my family and friends for their support and encouragement throughout this year.

On a personal note, this thesis is the culmination of a path I am glad to have chosen. I was delighted with the Information and Communication track, which led me to choose CoSEM and Technische Bestuurskunde, and working on this topic has only deepened that enthusiasm. I intend to stay within the technology sector, and this research has only strengthened my motivation to build my career there. I hope you enjoy reading this thesis.

W.H.J. Elshout
Delft, June 2026

Executive Summary

Agentic artificial intelligence is moving from research prototypes to production deployments in domains that process sensitive personal data. Two open communication standards have emerged at the centre of this transition: the Model Context Protocol (MCP), which connects an agent to its tools and data sources, and the Agent-to-Agent protocol (A2A), which connects agents to one another. Both protocols were recently introduced as open standards and are being adopted at speed, yet their privacy properties have so far received considerably less attention than their security properties. This thesis examines which privacy threats are inherent to MCP and A2A when they are deployed in a multi-agent system that processes personal data.

The main research question is: *What privacy threats emerge from the use of MCP and A2A communication protocols, as identified through a LINDDUN-based analysis of a financial fraud detection scenario?* A representative cross-organisational fraud detection use case was selected as the analytical instrument. It was chosen because it realises all four protocol configurations within a single workflow: MCP and A2A, each used both within and across organisations. This combination makes it well suited to isolating the privacy properties of each protocol. The use case was modelled as a single Data Flow Diagram comprising three agent processes, five data stores, four external entities, eighteen data flows, and two organisational trust boundaries. LINDDUN was applied to every data flow and trust boundary in the diagram, and the resulting threat inventory was then validated through seven structured interviews with privacy, security, and fraud-detection experts.

The analysis identifies fifty-four privacy threat entries across the eighteen data flows. Eight of these were selected for detailed description on the basis of protocol specificity, configuration coverage, and analytical significance. The eight fall into three groups. Four are protocol-inherent: they arise directly from design decisions in MCP and A2A and would appear in any faithful deployment of these protocols, regardless of the sector in which they are used. Two are use-case-inherent, reflecting the structural exclusion of the cardholder from a processing chain that concerns them directly. The remaining two sit between these poles.

The four protocol-inherent threats are the central finding. The first is session identifier linkability: MCP's stateful session architecture persistently links all tool calls within a session to the same data subject. The second is data aggregation at the orchestrator: MCP delivers full responses from all queried data stores to the agent simultaneously, producing a combined profile that exceeds the sensitivity of any individual store. The third is bilateral non-repudiation: A2A's task lifecycle model leaves permanent records on both sides of every inter-agent interaction that neither party can unilaterally remove. The fourth is a data minimisation violation: A2A's task context object carries the full assessment record to the receiving agent by default, with no built-in mechanism to restrict what is sent. Each of these is a property of the protocol specification itself, not of any particular implementation.

Expert validation confirmed that all eight threats are recognised as realistic, and the protocol-inherent threats received the broadest confirmation.

The findings carry a significance that extends beyond the protocols and the use case analysed. The protocol-inherent threats are properties of the specifications themselves and will be reproduced in every system built on these standards unless they are addressed at the governance level; a technical fix applied by a single deploying organisation cannot correct a structural property of the protocol layer. Two findings sharpen this picture. Severity is driven less by which protocol is used than by whether data crosses an organisational boundary, so the boundary, rather than the protocol choice, is the primary amplifier of privacy risk; and the threats form a cross-protocol chain, running from MCP data aggregation through A2A cross-organisational transmission to re-identification at the receiving agent, which is visible only when both protocols are analysed within a single system model. The central lesson is that privacy in agentic AI is as much a governance problem as an engineering one, requiring coordination between protocol designers, deploying organisations, and regulators across multiple levels and jurisdictions. On this basis the thesis makes three contributions: the first structured, data-subject-centred privacy threat analysis of MCP and A2A; the demonstration that the organisational boundary rather than the protocol choice is the primary amplifier of threat severity; and the identification of a cross-protocol threat chain that no single-protocol study could surface, together with a validated, reusable method and a set of transferable structural patterns for analysing privacy in agentic systems.

Contents

List of Figures

List of Tables

List of Abbreviations

A2A	Agent-to-Agent protocol
ACL	Agent Communication Language
AI	Artificial Intelligence
AML	Anti-Money Laundering
ANP	Agent Network Protocol
API	Application Programming Interface
DFD	Data Flow Diagram
FIPA	Foundation for Intelligent Physical Agents
FIPA-ACL	FIPA Agent Communication Language
GDPR	General Data Protection Regulation
GenAI	Generative Artificial Intelligence
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
IoA	Internet of Agents
JSON	JavaScript Object Notation
JSON-RPC	JSON Remote Procedure Call
KYC	Know Your Customer
LINDDUN	Linkability, Identifiability, Non-repudiation, Detectability, Data Disclosure, Unawareness & Unintervenability, Non-compliance
LLM	Large Language Model
MAESTRO	Multi-Agent Environment, Security, Threat, Risk, and Outcome
MAS	Multi-Agent System
MCP	Model Context Protocol
NIST	National Institute of Standards and Technology
PET	Privacy-Enhancing Technology
PRIAM	Privacy Risk Analysis Methodology
PSD2	Payment Services Directive 2
STRIDE	Spoofing, Tampering, Repudiation, Information disclosure, Denial of service, Elevation of privilege
SWIFT	Society for Worldwide Interbank Financial Telecommunication
URI	Uniform Resource Identifier
URL	Uniform Resource Locator

1979

“A computer can never be held accountable,
therefore a computer must never make a management decision.”[1]

Attributed to an IBM internal training presentation

2023

“Soon you will live surrounded by AIs.
They will organize your life, operate your business,
and run core government services.”[2]

Mustafa Suleyman, *The Coming Wave* [2]

[1] This statement is widely attributed to an internal IBM training presentation from 1979. No primary archival source has been located; the original document is reported to have been lost, and the quotation entered wider circulation only after the image of the slide was shared online. It has nonetheless been reproduced by IBM itself [1]. It is used here for its thematic relevance to the questions of accountability and autonomous decision-making raised in this thesis, not as a verified historical document.

[2] Mustafa Suleyman co-founded DeepMind in 2010 (acquired by Google in 2014) and has been Chief Executive Officer of Microsoft AI since March 2024. The juxtaposition with the preceding IBM statement is deliberate: Microsoft has natively integrated the Model Context Protocol, one of the two agentic communication protocols whose privacy threats this thesis analyses, into Copilot Studio, Azure AI Foundry, Windows 11, and GitHub Copilot. The voice prophesying that AIs will “operate your business” speaks from within the firm most actively building the infrastructure to make that prophecy true.

1 Introduction

1.1. Problem Introduction

Organisations across the financial sector are increasingly deploying autonomous AI agents to automate complex workflows. These agents do not operate in isolation. They query internal databases, invoke external services, and communicate with agents operated by other organisations, all in real time and with minimal human oversight. A bank's fraud detection system, for example, may rely on one agent to retrieve a customer's transaction history, a second to query an external sanctions screening service, and a third to exchange risk scores with an agent operated by a card network. Each of these interactions involves the transfer of sensitive personal data across system and organisational boundaries.

What makes this shift significant from a privacy perspective is not the use of AI agents as such, but the communication infrastructure that connects them. Two protocols have emerged as de facto standards for this infrastructure. A **protocol** is a set of rules that defines how two systems exchange messages: it specifies what a message looks like, in what order messages are sent, and what each party is expected to do in response. The Model Context Protocol (MCP), introduced by Anthropic in 2024, governs how an agent accesses external tools and data sources. The Agent-to-Agent protocol (A2A), introduced by Google in 2025, governs how agents communicate directly with one another across organisational boundaries. Together, these protocols define what data flows where, under what conditions, and with what degree of control.

The reason these protocols are being adopted so rapidly is also the reason they are analytically significant. MCP and A2A were designed to be fast and frictionless. Modern financial fraud operates in milliseconds, and effective detection must therefore operate under correspondingly tight real-time constraints [3, 4]. Detecting it reliably, however, is not only a question of speed but of breadth. Fraud signals are distributed across heterogeneous data sources, and accurate detection depends on combining transaction history, identity records, device data, and network-wide behavioural and relational signals rather than examining any single source in isolation [5, 6]. Bringing these dispersed signals together in real time, across multiple databases and increasingly across organisational boundaries, is precisely the kind of cross-source orchestration that MCP and A2A are designed to enable [7]. The financial sector is already actively adopting AI for fraud detection and risk assessment [8], which makes the privacy properties of the protocols that such systems may be built on a timely concern. Industry adoption of MCP in the financial sector is already under way: payment and financial-platform providers including Stripe, Block, and Alipay have begun exposing payment and data-processing capabilities through MCP [7]. Where querying each data source previously required its own bespoke, separately maintained integration, MCP lets an agent reach any number of sources through a single standardised interface within one working session [7]. A2A makes cross-organisational collaboration between agents possible through a standardised message exchange, in the same way that a shared language allows two people who have never met to communicate immediately.

This speed and interoperability is the core functional value of both protocols. It is also the direct source of the privacy risks this thesis analyses. The same session that allows an agent to query three databases in one operation also links all three queries to the same cardholder in the system logs, collapsing the separation that the organisation deliberately built between those databases. The same interaction that allows Bank X to delegate a risk assessment to Card Network Y creates a permanent record of that exchange that neither party can unilaterally remove. The privacy risks are not incidental side effects of poor implementation. They are structural consequences of the design choices that make MCP and A2A effective at their primary purpose. This is the fundamental trade-off the thesis examines: the same properties that make these protocols fast and powerful are the properties that make them privacy-threatening.

This creates a concrete and unresolved problem. Neither MCP nor A2A was designed with privacy as a primary concern. Both protocols operate in open, cross-organisational environments where agents discover capabilities dynamically, message contents are largely unstructured, and the personal data being transmitted includes transaction records, behavioural profiles, and identity information. The **data subjects**, that is, the individuals whose personal data is being processed, have no visibility into these flows and no mechanism to influence them.

This thesis adopts the *contextual integrity* conception of privacy [9]: privacy is preserved when personal information flows match the norms appropriate to the context in which the information was originally shared, and is violated when those norms are breached. The structured operationalisation of this notion

used in this thesis is the LINDDUN threat modelling framework, introduced in Section 1.3 and presented in detail in Chapter 2.

It should be noted that, in a fraud detection context, the collection and processing of transaction data is legally authorised and operationally necessary. The privacy risks of concern in this thesis are therefore not about whether data should be collected at all, but about what structural properties of the protocols govern how that data flows, to whom it is disclosed, and whether the data subject can remain identifiable across interactions that cross organisational boundaries.

The problem is not only technical. MCP and A2A are governance artefacts as much as they are technical specifications. They determine which actor controls which data, where organisational responsibility begins and ends, and which data subjects are included in or excluded from the system's accountability structures. In a multi-agent financial workflow, no single organisation designs or controls the full system: Bank X designs its internal agent, Card Network Y designs its own, and the combined system emerges from their interaction through the protocol. Privacy risks in such a system are therefore not solely an engineering problem. They are a distributed governance problem that requires coordinated decisions across organisational boundaries, protocol specifications, and regulatory frameworks. This sociotechnical character places the research squarely within the domain of Complex Systems Engineering and Management.

Existing research has begun to examine the security properties of MCP and A2A, identifying weaknesses in authentication and access control [7, 10]. However, a security analysis and a privacy analysis answer different questions. A system can be fully secured against external attackers while still exposing personal data through excessive collection, unintended disclosure, or the ability to link interactions back to a specific individual [11]. Anbiaee notes that no standardised threat modelling framework yet existed for these protocols, and proposes a security-oriented risk assessment that begins to address that gap [10]; the absence of a *privacy*-specific, data-subject-centred analysis is a further gap that follows from what this security-oriented literature does and does not examine, rather than a claim made by those studies themselves. Recent work has begun to document compositional privacy risks in multi-agent collaboration, showing that combining individually non-sensitive data flows can produce privacy violations at the system level [12]. However, no study has applied a structured privacy threat modelling methodology to MCP or A2A from the data subject's perspective within a concrete, sector-specific deployment scenario. This thesis addresses that gap.

1.2. Literature Exploration

The problem introduced in Section 1.1 sits at the intersection of two research areas: the study of multi-agent systems and their communication protocols, and the study of privacy in software architectures. This section reviews the relevant scientific literature in both areas and identifies the specific research gap that motivates this thesis.

The literature search was conducted primarily using **Scopus**, due to its broad coverage of peer-reviewed academic journals. To complement this, **ScienceDirect** was used to identify additional peer-reviewed articles on AI architecture, privacy, and governance. Both **backward and forward snowballing** were applied throughout. Backward snowballing, following a paper's reference list to find the earlier foundational work it builds on, was used to trace established research on topics such as privacy in multi-agent systems, agent coordination, and distributed AI architectures. Forward snowballing, identifying newer papers that cite a given work, was used to locate the most recent protocol-level literature: because MCP and A2A emerged only in 2024–2025, several of the relevant analyses postdate any single starting paper and were reached by following citations forward from the core protocol papers. Together these strategies connected current developments to the established scientific literature in these areas while keeping pace with a rapidly evolving field.

Search terms were applied in two stages. In the first, orienting stage, broader combinations were used to map the overall landscape, including terms such as *"Agentic AI," "Multi-agent systems," "LLM-based agents," "AI agent communication," "Privacy in multi-agent systems,"* and *"Agent coordination and trust."* In the second, more focused stage, searches were narrowed to the specific protocols and privacy dimensions relevant to this thesis, using terms such as *"Model Context Protocol," "Agent-to-Agent protocol," "A2A"* (also written *"Agent2Agent"*), *"Agent Network Protocol (ANP)," "Agora protocol," "agent interoperability protocols," "MCP security," "MCP privacy," "A2A privacy," "agentic AI security," "LINDDUN threat modeling," "privacy threat modelling for agents," "compositional privacy,"* and *"cross-organisational agent communication."* The protocol names ANP, Agora, and the umbrella term "agent interoperability protocols" were added at this stage because the comparative protocol literature that

this thesis builds on is indexed under these newer terms rather than under the broader agentic-AI vocabulary used in the orienting stage. This two-stage approach prevented the initial search from being diluted by overly broad terms while ensuring that protocol-specific and privacy-specific literature was retrieved systematically.

Because MCP and A2A are protocols that emerged from industry in 2024 and 2025, little peer-reviewed academic literature exists on these specific protocols yet. For this reason, a limited number of preprints from **arXiv.org** were included where they directly address the architecture or security properties of these protocols. These sources are used with caution: they have not undergone peer review, and some may contain errors or never reach publication. They are therefore used only to describe protocol-specific technical details, and not as the basis for broader theoretical claims.

The search was primarily limited to publications from **2020–2026** for recent developments in agentic AI and communication protocols, and extended to older foundational work through backward snowballing, particularly for multi-agent systems theory and privacy research. The combination of these strategies resulted in a collection of sources that covers both the established scientific foundations of the topic and the most recent protocol-level developments.

A source was included if it satisfied three conditions: (i) relevance to at least one of the two research areas defined above, namely agentic AI communication protocols or privacy in multi-agent and software-architecture settings; (ii) a focus on the protocol, privacy, or threat-modelling concepts central to the research question, rather than on an adjacent application domain in which agents only incidentally appear; and (iii) for the most recent material, direct treatment of MCP, A2A, or a comparable agent-communication protocol such as ANP or Agora. Peer-reviewed journal and conference publications were preferred, and preprints were admitted only for protocol-specific technical detail under the caution described above. A source was excluded if it addressed AI agents or privacy only in general terms without bearing on protocol-level data flows, if it merely duplicated a finding already covered by a stronger source, or if it fell outside the 2020–2026 window without being a foundational work reached through backward snowballing. Applying these criteria across the database searches and the backward and forward snowballing yielded approximately forty sources judged directly relevant to the research question, which form the basis of the review that follows. The thesis as a whole draws on around eighty references; the remainder consists largely of methodological and protocol-specification material introduced in later chapters.

1.2.1. AI and LLMs

Artificial intelligence (AI) is broadly understood as the field concerned with building computational systems that can perceive information, reason over it, and support or automate decision-making. The study of intelligent agents, systems that perceive their environment and take actions to achieve goals, is a central strand of AI research with decades of theoretical and applied development [13, 14, 15].

A more recent and consequential development within AI is the emergence of **large language models (LLMs)**: systems trained on vast collections of text to understand and generate natural language [16, 17]. LLMs derive from the transformer architecture, a neural network design introduced by Vaswani et al. that processes entire sequences of text in parallel rather than word by word, enabling models to learn relationships between words and concepts across very large datasets [16]. Unlike earlier AI systems that required explicit rule specification or narrowly scoped training data, LLMs develop broad language competence that can be applied flexibly across diverse tasks, including summarisation, question answering, code generation, and multi-step reasoning [18].

The flexibility of LLMs has made them the foundation of a new generation of AI applications. Sapkota et al. note that LLMs now function as the cognitive core of many agentic systems because they provide the language understanding and reasoning capabilities that allow agents to interpret instructions, plan actions, and interact with both humans and other systems through natural language [19]. This ability to connect LLM reasoning with external tools, databases, and other agents is what distinguishes the current generation of AI systems from earlier generations and makes communication protocols such as MCP and A2A technically feasible [7].

It is important, however, to acknowledge the limitations of LLM-based systems. They are known to produce outputs that are plausible in form but factually incorrect, a phenomenon known as **hallucination** [20, 21]. Their reasoning can also be manipulated through carefully crafted inputs, a vulnerability known as **prompt injection** [21]. In a prompt injection attack, malicious instructions are hidden inside data that the agent reads, causing it to take actions it was not intended to take. Both limitations are directly relevant to agent systems that process external data in automated pipelines, and they mean

that LLM-based agents should not be treated as fully reliable autonomous reasoners, especially in high-stakes environments such as financial services [21].

1.2.2. AI Agents, Agentic AI, Multi-Agentic AI, and Multi-Agent Systems

The concept of an autonomous agent has been studied in computer science since the 1990s. Wooldridge and Jennings define an intelligent agent as a computational system situated in an environment that is capable of *autonomous action* in order to meet its design objectives [13]. Four properties characterise such an agent: autonomy (acting without direct human intervention), reactivity (responding to changes in the environment), proactiveness (taking initiative to achieve goals), and social ability (interacting with other agents or humans) [13]. This definition has remained foundational in the literature and continues to underpin how agents are understood today.

When multiple agents interact within a shared environment, the result is a **multi-agent system (MAS)**. The study of MAS focuses on the challenges that arise from this interaction: how agents coordinate actions, negotiate over resources, allocate tasks, and resolve conflicts without centralised control [14]. Luck, McBurney and Preist established MAS as a foundational paradigm for next-generation computing precisely because many real-world problems involve multiple actors with different roles, capabilities, and goals whose interactions cannot be fully anticipated in advance [15]. The field of MAS therefore developed a rich body of work on coordination mechanisms, trust, and communication protocols, all of which remain directly relevant to the systems studied in this thesis.

More recently, the term **Agentic AI** has emerged in the literature to describe a broader design paradigm in which AI systems, typically powered by large language models, exhibit agency through planning, memory, tool use, and adaptive decision-making [19]. Wang et al. synthesise this literature into a unified framework for LLM-based autonomous agents, identifying four core components (profile, memory, planning, and action) that together distinguish modern LLM-driven agents from earlier rule-based architectures [22]. Sapkota et al. draw an important distinction: where a classical AI agent is one operational unit with fixed capabilities, Agentic AI refers to the architectural approach that gives a system its agentic character, often spanning multiple components and interaction layers [19]. When multiple such LLM-based agents interact and coordinate on shared or interdependent tasks, the result is referred to as **Multi-Agentic AI** [19].

Multi-Agentic AI and classical MAS are related but distinct. Both paradigms are concerned with the same foundational problem: multiple autonomous entities must coordinate, communicate, and share information to solve problems that exceed the capacity of any single agent alone [13, 14]. The coordination challenges studied in classical MAS, including trust between agents, task allocation, and the risks of emergent collective behaviour, do not disappear in LLM-based systems; they reappear in new forms, amplified by the scale and openness of modern deployments [23].

The differences between the two paradigms are equally significant. Classical MAS agents typically communicate through structured, formally specified message formats, designed so that the meaning of every message is unambiguous and machine-verifiable [14]. Multi-Agentic AI agents communicate through rich natural-language messages mediated by protocols such as MCP and A2A, across organisational boundaries that classical MAS architectures were not designed for [24]. Classical MAS was developed and evaluated primarily in controlled, often single-organisation environments, where the boundaries of the system and the identities of the participating agents were known in advance [15]. Multi-Agentic AI operates in open, cross-organisational ecosystems where agent identities, capabilities, and data flows may not be fully visible to any single party [24]. Finally, the failure modes differ: where classical MAS agents fail through miscoordination or logical inconsistency, LLM-based agents introduce new risks including hallucination, prompt injection, and emergent collusion between agents from different organisations [23].

1.2.3. Cooperative AI

A structural shift is underway in how AI systems are deployed. Where earlier AI systems operated in isolation, modern deployments increasingly involve multiple agents interacting continuously across organisational and technical boundaries [14, 19]. This shift is not merely a scaling of existing architectures; it represents a qualitative change in the nature of AI systems, from tools that assist individual users to ecosystems in which agents coordinate, negotiate, and act jointly on shared or interdependent objectives.

The scientific study of such ecosystems is the subject of **Cooperative AI**, a research programme articulated by Dafoe et al. in *Nature* [25]. They argue that as AI agents play an increasingly significant role in consequential decisions, it becomes essential to equip them with the competencies required to

cooperate, not only with humans, but also with other agents. Problems of cooperation, in which agents have opportunities to improve their joint welfare but face barriers to doing so, arise at all scales, from individual workflows to cross-organisational systems [25]. Dafoe et al. identify three forms of cooperation relevant to AI: cooperation between AI agents, cooperation between AI agents and humans, and the use of AI to facilitate cooperation between humans [25].

Cooperative AI is not simply about making agents behave beneficially in isolation. Dafoe et al. show that when agents interact, new failure modes emerge that are absent in single-agent settings [25]. Agents may fail to coordinate due to incomplete information, pursue locally rational strategies that produce globally harmful outcomes, or exploit situations where one party knows more than another in ways that undermine the trust necessary for effective collaboration. These risks are compounded when agents from different organisations interact, because no single party controls or monitors the full system. The challenges of coordination and trust in multi-agent systems are well established in the classical MAS literature [13, 14], but they take on new urgency when agents operate at internet scale across organisational boundaries that were not anticipated in earlier architectures.

Effective cooperation therefore depends critically on how agents exchange information, form commitments, and handle data that may be sensitive, incomplete, or contested. Ferrag et al. demonstrate that the communication layer of LLM-powered agent workflows is a significant source of attack-surface risk, with documented vulnerabilities ranging from prompt injection to protocol-level exploits [21]. The communication protocol is therefore not merely a technical convenience: it is the mechanism through which cooperative behaviour between agents is either enabled or undermined. Whether agents share too much data, too little, or the wrong kind of data with the wrong parties depends directly on how the protocol structures information exchange. This makes the privacy properties of communication protocols a question of fundamental importance for the governance of cooperative AI systems [26, 11].

In summary, the shift toward multi-agent ecosystems raises cooperation challenges that are well-studied in classical MAS theory but inadequately addressed in current LLM-based deployments. The communication protocol is the infrastructure through which these challenges are either managed or exacerbated, which motivates the protocol-level privacy analysis that follows.

1.2.4. Communication Protocols

The need for standardised communication protocols in multi-agent systems is not a recent concern. Jennings argued in *Artificial Intelligence* that for agents to engage in the rich, high-level social interactions required in complex distributed systems, they must share a common communication framework: one that conveys not just data, but the meaning, intent, and context of agent messages [27]. This insight motivated the development of dedicated **agent communication languages (ACLs)** as early as the 1990s. The most influential was the FIPA Agent Communication Language (FIPA-ACL), ratified by the Foundation for Intelligent Physical Agents as an open standard in 1997 and refined thereafter [14]. FIPA-ACL defined a structured set of communicative acts, including *inform*, *request*, and *propose*. Each act had a precisely defined meaning: sending an *inform* message meant the sender believed the content to be true and believed the receiver did not; sending a *request* meant the sender wanted the receiver to perform some action. This precision allowed agents to reason formally about what other agents believed and intended, which was essential in the controlled environments for which FIPA-ACL was designed [14, 27]. These early standards established a principle that remains central today: that interoperability between agents from different systems requires explicit, shared agreement on how messages are structured, what they mean, and under what conditions they may be sent.

The emergence of LLM-based agents has renewed this challenge at a different scale and with different technical requirements. LLM agents must communicate not only within a single platform but across different vendors, organisations, and cloud environments. They must interact with external tools, databases, and services as well as with other agents, often without prior knowledge of each other's capabilities. Yang et al. survey the current generation of AI agent protocols, noting that this landscape now includes MCP, A2A, and several competing proposals, each making different trade-offs between expressiveness, security, and interoperability [28].

This thesis focuses on two protocols that address complementary communication layers.

The Model Context Protocol (MCP) was introduced by Anthropic in 2024 as a standard for connecting LLM-based agents to external tools, databases, and services. In MCP, an agent acts as a client that sends structured requests to an MCP server, which exposes specific tools or data sources. The protocol governs tool discovery, input and output formatting, and context passing between the agent and the external system [7]. MCP is therefore a protocol for *tool invocation and data retrieval* rather than for

agent coordination: it governs how an agent accesses information from systems it does not internally control.

The Agent-to-Agent protocol (A2A) was introduced by Google in 2025 as a standard for direct communication between agents. Where MCP connects an agent to a tool, A2A connects an agent to another agent. It defines how one agent can discover the capabilities of another, delegate tasks, exchange structured messages, and receive results across organisational and platform boundaries [29, 30].

The difference between FIPA-ACL and the current generation of protocols is instructive. FIPA-ACL was designed for controlled, single-organisation environments with known agent populations and precisely defined shared vocabularies (**ontologies**) that gave every concept in a message an unambiguous meaning. MCP and A2A are designed for open, cross-organisational ecosystems where agent identities and capabilities are discovered dynamically, and where message contents consist of rich, unstructured natural language rather than formally specified communicative acts. This shift brings new expressive power but also new risks: where FIPA-ACL enforced strict constraints on what could be communicated and how, MCP and A2A impose far fewer constraints on the content of data flows [7, 29]. The security implications of this openness have begun to be studied [29, 10], but the privacy implications, what personal data flows where, under what conditions, and with what controls, remain unexamined. This gap motivates the analysis in this thesis.

In summary, communication protocols for multi-agent systems have a decades-long history, from FIPA-ACL in the 1990s to MCP and A2A today. The shift from structured, closed environments to open, cross-organisational deployments introduces new risks that the classical protocol literature did not anticipate. The privacy properties of MCP and A2A are the focus of the remainder of this thesis.

1.2.5. Previous Work

The relevant previous work spans two largely separate bodies of literature that have, until now, developed independently of each other. The first concerns security and risk in modern agentic AI communication protocols; the second concerns privacy specifically in multi-agent systems, drawing on a longer research tradition.

Security in Agentic AI Communication Protocols

Since the emergence of MCP and A2A in 2024–2025, several studies have analysed their security properties. Louck, Stulman and Dvir provide a comparative security evaluation of agentic AI communication protocols, identifying weaknesses in authentication and authorisation that create opportunities for unauthorised access and data interception [29]. Anbiaee applies structured threat modelling to MCP, A2A, Agora, and ANP, focusing on authentication, authorisation, and communication integrity [10]. Ferrag et al. map attack-surface threats across LLM-powered agent workflows at the protocol level, from prompt injection to authentication exploits, demonstrating that the communication layer of such systems constitutes a significant and exploitable attack surface [21]. Leo et al. assess risks across the layers of agentic AI systems more broadly, mapping threat categories to system-level trust properties [31].

These studies collectively establish that the protocol layer of modern multi-agent AI systems carries meaningful risk. However, their analytical framework is security-oriented: they examine privacy primarily as a property of data assets to be protected against external attackers, not as a property of information flows from the perspective of the data subject. Ferrag et al. and Leo et al. do address privacy, but in the sense of keeping data confidential and access controlled, not in terms of what the data flows themselves reveal about an individual, whether a person can be singled out or linked across organisations, even when every party is correctly authorised. Neither applies a data-subject-centred methodology such as LINDDUN, and neither analyses linkability or identifiability of the data subject across the cross-organisational flows that MCP and A2A create. They do not apply a structured methodology to ask what personal data flows where, to whom, and with what consequences for the individual whose data is being processed. This is a fundamentally different analytical question from the ones these studies answer.

As Wuyts, Scandariato and Joosen, and subsequently Sion, Van Landuyt and Wuyts, demonstrate, security and privacy threat modelling address fundamentally different concerns [11, 32]. Security analysis is concerned with protecting data assets with respect to the organisation: it focuses on preventing unauthorised access by external attackers. Privacy analysis is concerned with protecting personal data with respect to the *data subject*: it must address not only external attackers but also the consequences of legitimate access, and its goal is to limit what can be done with personal data once access has been correctly granted [11, 32]. A system can therefore be fully secure, with all access controls in place and all authentication mechanisms functioning correctly, while still producing serious privacy

violations through unintended disclosure or the ability to link interactions back to an individual [11, 32]. The security analyses reviewed above do not examine these properties.

Privacy in Agentic AI Communication Protocols

A small number of very recent studies have begun to look at privacy, rather than security, specifically in the context of MCP. They represent the work that comes closest to the subject of this thesis, and for that reason it is important to be precise about what they do and do not address. Yao, Wang and Cheng present IntentMiner, an *intent inversion attack* that reconstructs a user's underlying intent by analysing the sequence of tool calls an agent issues over MCP [33]. Their work is a valuable empirical demonstration that MCP interactions leak privacy-relevant information even when individual messages appear innocuous. It is, however, a single attack technique against a single protocol, evaluated at runtime; it does not provide a structured inventory of the privacy threats that the protocol design produces, nor does it adopt the data subject as its unit of analysis or consider A2A. Wang et al. address privacy more directly across both protocols: they present PrivacyChecker, a contextual-integrity-based mitigation method, and PrivacyLens-Live, a dynamic benchmark that recreates MCP and A2A environments to measure how much private information agents leak in practice [34]. Their study is among the closest in scope to this thesis in that it foregrounds privacy in both MCP and A2A, but its orientation is empirical and mitigation-driven rather than analytical: it measures and reduces leakage at runtime rather than enumerating, from the protocol design, which categories of privacy threat arise. It does not apply a structured privacy threat modelling framework such as LINDDUN, does not represent the system as a data flow diagram with explicit trust boundaries, does not assess the relative severity of the threats from the data subject's perspective, and does not ground its analysis in a concrete sector-specific deployment.

These two studies confirm that privacy in MCP is beginning to attract attention, but they also delineate, by what they leave undone, the specific contribution of this thesis: neither provides a systematic, data-subject-centred threat analysis that spans both MCP and A2A, grounds the analysis in a concrete deployment, and prioritises the resulting threats by severity.

Privacy in Multi-Agent Systems

The scientific study of privacy in multi-agent systems predates the current wave of LLM-based agents by more than two decades. Jennings established that effective agent cooperation requires a shared communication framework that conveys not just data but meaning and context [27]. Yu and Cysneiros formalised the privacy dimension of the resulting tension: agents must share information to collaborate effectively, but information sharing creates the risk of disclosing data that individuals or organisations did not intend to reveal [35]. Piolle, Demazeau and Caelen showed that this tension can be addressed by embedding privacy policies as operational constraints in agent architectures, so that agents reason about privacy as part of their decision-making rather than treating it as an external requirement [36].

Such, Espinosa and Garcia-Fornes provide the most comprehensive survey of privacy challenges in MAS, identifying the recurring tension between information sharing and privacy protection as the central concern. They organise privacy in MAS around Solove's three information-related activities, information collection, processing, and dissemination, within which several recurring privacy concerns surface: **linkability** (the ability to connect separate interactions or records back to the same individual), the unintended disclosure of private information through inter-agent communication, the inference of sensitive attributes from observable agent behaviour, and the difficulty of enforcing **data minimisation** (collecting and sharing only as much data as is strictly necessary) when agents must exchange contextually rich information [26]. Warnier and Brazier developed anonymity services that allow agents to interact without revealing their identities, directly addressing the linkability problem [37]. Maliah, Brafman and Shani demonstrated that deliberately reducing the amount of information exchanged between agents during planning can improve privacy at the cost of some coordination efficiency, a finding directly relevant to the design of MCP and A2A interactions [38]. A recent survey by Kossek and Stefanovic confirms that privacy-preserving mechanisms for cooperative multi-agent systems remain an active research area, categorising ongoing work into differential privacy and encryption-based protocols [39].

More recently, Patil, Stengel-Eskin and Bansal identify a **compositional privacy risk** specific to multi-agent collaboration: even when each individual agent shares only non-sensitive information, the combination of data across multiple agents can produce privacy violations that would not have been possible from any single agent's output alone [12]. This finding is directly relevant to the MCP and A2A context, where agents from different organisations exchange partial views of the same data subject, and demonstrates that privacy risks in multi-agent systems cannot be evaluated by examining each data flow in isolation.

In summary, the MAS privacy literature identifies several threat categories that are structurally relevant to MCP and A2A: linkability, unintended disclosure, inferability, and compositional aggregation. However, this body of work was developed for a fundamentally different technical context, which is the subject of the research gap analysis below.

1.2.6. Research Gap

The two bodies of literature reviewed above define a precise and unaddressed research gap.

As the review in Section 1.2.5 established, the security literature on MCP and A2A examines only security properties, and none of these studies applies a structured privacy threat modelling methodology from the perspective of the data subject. The absence of a data-subject-centred privacy analysis is therefore a separate gap that follows from what the security literature does and does not examine, not a claim made by those studies themselves. The compositional privacy risks documented by Patil, Stengel-Eskin and Bansal illustrate what is at stake: information that is innocuous in isolation can become disclosive when composed across multiple agent interactions, a risk invisible to analyses that examine individual data flows or components on their own [12]. This is precisely the cross-protocol, cross-boundary risk that the combination of MCP and A2A creates in the use case examined here.

The MAS privacy literature presents a second and separate dimension of the gap: it has produced a rich body of theory, threat categories, and privacy-enhancing mechanisms [26, 39, 37, 35, 36, 38]. However, this work was developed for classical agents that communicated through formally specified, typed message structures [14, 27]. Because every message in a classical MAS had a precisely defined structure, privacy analysts could reason about bounded, predictable data flows: they could determine in advance exactly what information would be exchanged in a given interaction. MCP and A2A operate on fundamentally different principles. Their messages contain unstructured natural-language content whose boundaries cannot be statically defined, and agents operate across organisational boundaries that can span multiple jurisdictions [24, 7]. In the fraud detection setting examined in this thesis, the personal data processed across these boundaries includes transaction histories, behavioural profiles, and identity records, transmitted in interaction patterns that the classical MAS literature did not anticipate. The existing MAS privacy theory is therefore insufficient as a direct basis for analysing MCP and A2A: the threat categories it identifies remain relevant, but the methods it developed for formal, bounded environments cannot be applied directly to the open, language-driven context of current agentic deployments. A different analytical approach is needed, one that operates at the level of individual data flows and trust boundaries, accounts for unstructured payloads (the free-form application data a message carries) and cross-organisational dynamics, and maintains the data-subject perspective that the security literature lacks.

No study has applied a structured, data-subject-centred privacy analysis to MCP or A2A at the protocol level within a concrete deployment scenario. This is a precise gap, not a broad one.

As Section 1.2.5 discussed, the closest privacy-focused studies, by Yao et al. and Wang et al., do not apply a structured threat modelling methodology across both protocols from the data subject's perspective [33, 34]. The privacy threat categories identified in the classical MAS literature remain applicable, and the LINDDUN framework provides the methodological bridge needed to apply them to the data flows and trust boundaries that MCP and A2A create. The applicability of LINDDUN to communication protocol analysis has already been demonstrated in adjacent contexts: Mirzamohammadi et al. apply LINDDUN to the Solid protocol in a financial data analytics scenario, confirming both its suitability for protocol-level privacy threat analysis and that such analyses consistently reveal privacy threats that security-only analyses miss [40]. What is missing is the application of this approach to MCP and A2A: a systematic analysis that asks, for each data flow these protocols produce, which categories of privacy risk arise for the individual whose data is being processed, how severe those risks are, and whether they are structural properties of the protocol design or consequences of specific implementation choices.

This thesis fills that gap through a LINDDUN-based analysis of a financial fraud detection use case in which both protocols operate simultaneously. The financial sector is a particularly important context because it concentrates large volumes of sensitive personal data, involves autonomous cross-organisational agent interactions, and is a sector where AI is already being adopted for fraud detection and risk assessment [8], with multi-agent systems beginning to appear in customer-facing functions [41]. This thesis makes three scientific contributions, developed in full in Chapter 9. First, it provides the first structured, data-subject-centred privacy threat analysis of MCP and A2A, distinct from and complementary to the existing security literature. Second, it demonstrates that the organisational boundary, rather than the protocol choice, is the primary determinant of threat severity in multi-agent AI systems.

Third, it identifies a cross-protocol threat chain, running from MCP data aggregation through A2A cross-organisational transmission to re-identification at the receiving agent, that is only visible when both protocols are analysed within the same system model.

1.3. Research Question

The literature review establishes a precise and unaddressed gap: no data-subject-centred, protocol-level privacy analysis of MCP or A2A exists, and the established MAS privacy theory was developed for a fundamentally different technical context. The financial sector provides a deployment context where both the practical urgency and the analytical richness of this gap are greatest.

This leads to the following main research question:

> **"What privacy threats emerge from the use of MCP and A2A communication protocols, as identified through a LINDDUN-based analysis of a financial fraud detection scenario?"**
>
> LINDDUN (Linkability, Identifiability, Non-repudiation, Detectability, Data Disclosure, Unawareness & Unintervenability, Non-compliance) is a structured privacy threat modelling framework that analyses how personal data flows through a system and where privacy risks arise. It is described in detail in Chapter 2.

The main research question is decomposed into three sub-questions that structure the analysis:

- **SQ1:** How do MCP and A2A function within the use case, and how can the system be represented as a Data Flow Diagram?
- **SQ2:** What privacy threats are present in the system, and what is their relative severity?
- **SQ3:** To what extent are the identified threats recognised and validated by domain experts, and which additional threats do experts identify?

SQ1 corresponds to the modelling phase (Chapters 3 and 4), SQ2 to the elicitation and prioritisation phase (Chapters 5 and 6), and SQ3 to the expert-validation phase (Chapter 7). The synthesis and conclusion (Chapters 8, 9, and 10) then integrate the answers to all three sub-questions into the answer to the main research question.

It is important to be precise about what this question takes as its object of study. The object is the *protocols*, MCP and A2A, and the privacy threats that follow from how they are designed. The financial fraud detection scenario is the *instrument* through which those protocol properties are made observable: a concrete setting in which all four protocol configurations occur simultaneously and can be analysed against the data flows and trust boundaries they produce. The use case is therefore a means, not the end. The aim is to characterise the privacy properties that any system inherits from using MCP and A2A as specified, rather than to audit one specific fraud-detection system. This distinction matters for interpreting the findings: a threat such as the data subject's structural unawareness of the processing is reported as a central finding precisely because it is a property of the protocols themselves, which provide no machine-to-data-subject interface by design, rather than an artefact of the fraud detection setting. The use case makes the threat visible; it does not create it.

The methodology used to answer this question, including the justification of the analytical framework, the selection of the financial use case, and the design of the expert interviews, is presented in Chapter 2.

1.4. Relevance to the MSc CoSEM Programme

The MSc Complex Systems Engineering and Management (CoSEM) programme prepares engineers to analyse and govern complex sociotechnical systems, in which technology, organisations, and regulations interact to produce challenges that no single actor or discipline can resolve alone. This thesis is a CoSEM thesis because the central problem it identifies cannot be understood or addressed from within any single discipline or from the perspective of any single actor.

The privacy problems this thesis identifies are not engineering failures in the conventional sense: no component malfunctions and no developer makes a mistake. They arise because five actors with different interests, governance responsibilities, and technical capabilities, the deploying organisations (Bank X and Card Network Y), the protocol designers (Anthropic and Google), the regulator, and the cardholder, are connected through protocols designed for speed and interoperability without privacy as a requirement. No actor acting alone can resolve the resulting governance problem: Bank X has no authority over re-identification at Card Network Y, the protocols offer no data-subject-facing interface through which the cardholder's unawareness could be addressed, the protocol designers bear no legal

responsibility for how their standards are deployed, the regulator lacks a framework for multi-agent distributed processing, and the cardholder has no role in the system at all. This is the kind of system-level problem that CoSEM is designed to analyse: one that is invisible from any single component's perspective and requires coordinated intervention across actors and governance layers.

Three CoSEM concepts are directly instantiated. First, **emergence**: the re-identification of the cardholder at Card Network Y (Chapter 6) is a property of neither the profile Bank X transmits, nor the dataset Card Network Y holds, nor the A2A protocol that connects them, but one that emerges from their combination, structurally invisible to any single-component analysis [42]. Second, **distributed governance**: the four structural trade-offs identified in Chapter 8, setting speed, auditability, cross-organisational collaboration, and agent autonomy against privacy and data-subject control, require coordinated decisions across the protocol, deployment, and regulatory layers simultaneously. Third, **boundary management**: the most significant privacy risks arise at the trust boundaries where one organisation's governance ends and another's begins, which sociotechnical systems theory has long treated as a central governance concern [43].

The thesis also engages a structural conflict between public- and private-domain values. Privacy is a public-domain value, a fundamental right and a matter of regulatory concern; the efficiency, auditability, and fraud-prevention that MCP and A2A enable are private-domain values serving the deploying organisations' commercial interests. The same protocol mechanisms that deliver the private-domain benefits are the source of the public-domain risks, and the costs fall on the cardholder while the benefits accrue to the institutions, as the stakeholder analysis in Chapter 8 sets out. Weighing a public value against a private one where no single actor can strike the balance is exactly the kind of value conflict a CoSEM analysis is meant to surface.

Finally, the thesis connects to the Information and Communication track, in particular I&C Architecture Design, I&C Service Design, and Digital Platform Design. MCP and A2A are digital platform infrastructure: open standards defining the communication rules of a large actor ecosystem, where the platform designers capture the value of adoption while the governance consequences are borne by those who build on and are affected by the platform. The platform governance analysis in Chapter 8 applies this lens to Anthropic and Google as protocol designers whose architectural choices create governance obligations they do not currently acknowledge. In sum, this thesis uses privacy threat modelling to make visible a class of governance problems structural to current agentic AI communication protocols, demonstrating that privacy in multi-agent AI systems is fundamentally a sociotechnical governance problem that the analytical tools of CoSEM are needed to surface and address.

1.5. Reading Guide

Figure 1.1 provides an overview of the chapters of this thesis and their relation to the research phases. The thesis is structured into five parts, described as follows.

- *Phase 0: Introduction and Methodology.* Chapter 1 introduces the research context, reviews the relevant literature, identifies the research gap, and presents the main research question. Chapter 2 describes the research methodology, justifies the selection of the LINDDUN privacy analysis framework, presents the financial use case, and outlines the research design.
- *Phase 1: Model.* Chapter 3 describes MCP and A2A technically, covering their components, roles, and how they function within the selected use case. Chapter 4 constructs the Data Flow Diagram (DFD) of the use case: a visual representation of the system showing which data flows between which components, across which organisational boundaries. The DFD identifies all external entities, processes, data stores, data flows, and trust boundaries.
- *Phase 2: Elicit.* Chapter 5 applies the LINDDUN threat categories to each DFD element, elaborates relevant threats using threat tree patterns, and documents the identified privacy threats. Chapter 6 prioritises the identified threats based on likelihood and impact and provides detailed descriptions of the most significant threats.
- *Phase 3: Validate.* Chapter 7 presents the design and results of the semi-structured expert interviews conducted to assess whether the identified threats are recognised as realistic and relevant by domain experts, and to surface any additional threats that experts identify from their practical experience.
- *Phase 4: Synthesis.* Chapter 8 interprets the findings across all phases: it positions the results against the existing literature, discusses unexpected findings, interprets the expert validation, and identifies implications for protocol designers and future research. Chapter 9 reflects on the research process and describes the scientific and practical contributions of the thesis. Chapter 10

presents the conclusion, providing the explicit answer to the main research question, followed by the limitations of the study and directions for future research.

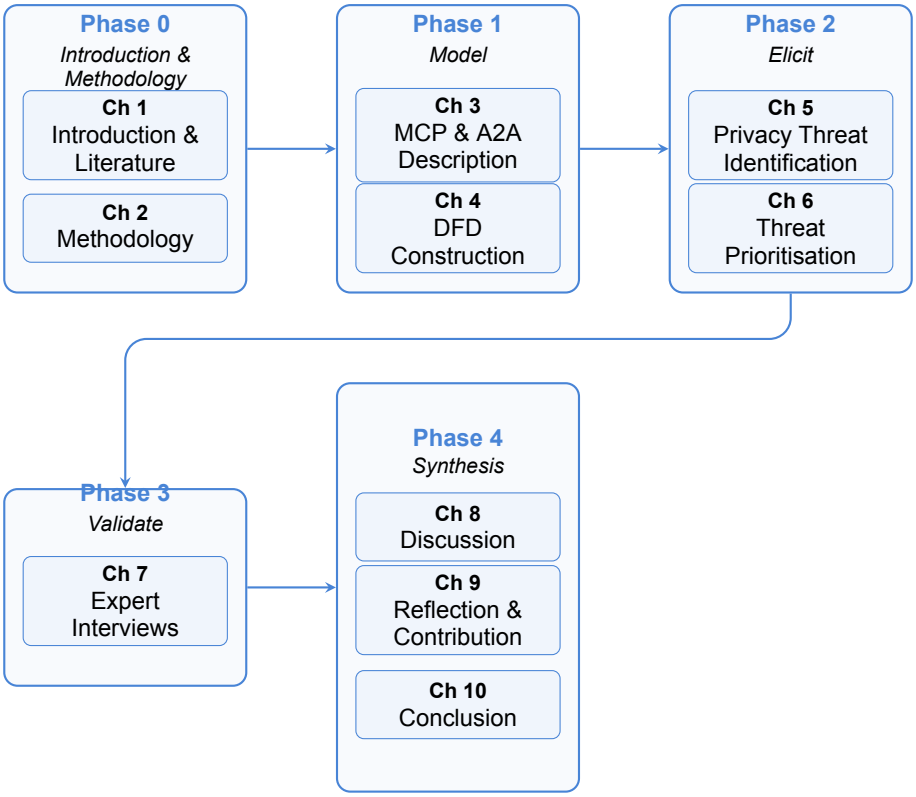

Figure 1.1: Reading guide: thesis structure and chapter overview. (Author's own illustration.)

2 Methodology

Chapter 1 introduced the societal context of this research, reviewed the relevant scientific literature, and identified the research gap that motivates the main research question. This chapter explains how the research is designed and carried out. It presents the overall research approach, justifies the selection of the privacy analysis framework, describes the financial use case including an analysis of the actors involved, and outlines the four phases through which the analysis is conducted.

2.1. Selection of the Research Approach

This thesis examines what privacy threats emerge from the use of MCP and A2A communication protocols in a financial multi-agent AI system, and to what extent these threats are recognised and validated by domain experts. The nature of this question determines the appropriate research approach. The goal is not to test a hypothesis or measure a predefined variable. It is to identify and understand the privacy threats that arise from the architectural and operational properties of the two protocols in a concrete deployment, and then to assess how plausible and complete those findings are through expert knowledge. This requires an approach that supports detailed, interpretive analysis of a complex and relatively unexplored phenomenon.

For this reason, a **qualitative research approach** is adopted. Qualitative research is appropriate when the aim is to develop an in-depth understanding of a phenomenon in its real-world context, particularly when that phenomenon is new, complex, or not yet well-defined in the literature. Both of these conditions apply here: MCP and A2A are protocols that emerged from industry in 2024–2025, no structured privacy analysis of either protocol exists, and the privacy implications of their deployment in financial multi-agent systems have not previously been examined.

Within the qualitative approach, a **single use case** is examined: a real-time anti-fraud transaction monitoring system in the financial sector, involving a cross-organisational interaction between Bank X and Card Network Y. The use case is hypothetical but constructed to reflect a scenario that could plausibly occur in practice. It is designed to realise all four analytically distinct protocol configurations (the distinct ways MCP and A2A are deployed and combined in the system), so that the full range of privacy dynamics the two protocols produce is accessible within a single coherent system. Why a single hypothetical case is appropriate, and why this particular case was chosen, is set out in Section 2.3.2; the use case itself is described in full in Chapter 4.

The LINDDUN privacy threat modelling framework is applied to the use case to systematically identify privacy threats. The selection and justification of LINDDUN as the analytical framework are presented in Section 2.2.

Expert interviews are included as a validation step. They serve to assess whether the identified threats are recognised as realistic and relevant by practitioners with direct experience of agentic AI systems in organisational contexts, and to surface any additional threats that experts identify from their own experience. The design and role of the expert interviews are described in Section 2.4.3.

The overall research design is summarised in Table 2.2, which shows how the four phases of the study relate to the sub-questions and to the main research question.

2.2. Selection of the Privacy Analysis Framework

Having established the qualitative research approach, this section explains which analytical framework is used to identify privacy threats within the selected use case and why this framework was chosen over available alternatives.

2.2.1. Candidate Frameworks

Several structured frameworks exist for analysing privacy and security risks in software systems. Three candidates were considered for this thesis: LINDDUN, MAESTRO, and the NIST Privacy Framework version 1.1. These three were selected for evaluation because they represent distinct but complementary approaches to privacy analysis. MAESTRO is a threat modelling framework developed specifically for agentic AI systems, addressing security threats including those with privacy implications. The NIST Privacy Framework version 1.1 is well suited for establishing a high-level organisational privacy programme. LINDDUN is a privacy-specific threat modelling framework that operates at the level of individual data flows, providing the engineering-focused, in-depth analysis needed for the protocol-level examination carried out in this thesis. Each framework is evaluated in full in Appendix F; this section

summarises that evaluation and states the resulting choice.

Table 2.1 summarises the evaluation of the three frameworks against the criteria most relevant to this thesis.

Table 2.1: Comparison of candidate privacy analysis frameworks. Evaluation based on framework documentation [44, 45, 46].

Selection criterion	LINDDUN	MAESTRO	NIST PFW 1.1
Unit of analysis	Individual data flows & trust boundaries	Architectural layers of AI agent stack	Organisational functions
Protocol-level analysis	Fully met	Partially met	Not met
Privacy-specific (not security)	Fully met	Not met, security framework	Partially met, governance level only
Scopeable to single protocol interaction	Fully met	Partially met	Not met
Peer-reviewed & academically established	Fully met, since 2011	Not met, CSA 2025, not peer-reviewed	Government standard, public consultation
Cross-organisational trust boundary coverage	Fully met	Fully met, security focus	Partially met, policy level only
Addresses AI autonomy	Partially met	Fully met	Partially met

Other privacy methodologies, including PRIAM, GDPR Data Protection Impact Assessments, and the STRIDE family of security frameworks, were also considered but did not match the protocol-level, privacy-specific needs of this thesis; they are discussed in Appendix F. LINDDUN meets the four criteria most critical to this thesis: it is privacy-specific rather than security-oriented, it operates at the level of individual data flows and trust boundaries, it can be scoped to a specific protocol interaction, and it rests on a peer-reviewed academic foundation. MAESTRO addresses AI autonomy more directly, but its security orientation and lack of peer review make it unsuitable as the primary framework, though it is retained as a supplementary cross-check during threat identification (Chapter 5). The NIST Privacy Framework is governance-oriented and cannot be applied at the protocol level. LINDDUN is therefore the most appropriate choice for the research question addressed in this thesis.

2.2.2. The LINDDUN Framework

LINDDUN is a privacy threat modelling methodology developed by Deng et al. to support the systematic elicitation of privacy threats in software system architectures [44]. It was designed as a privacy counterpart to the STRIDE security threat modelling framework, applying a comparable structured approach to privacy rather than security [44]. The framework has since been extended and refined through subsequent peer-reviewed work, including LINDDUN PRO, which performs interaction-based threat elicitation per “source, data flow, destination” triple, and LINDDUN GO, which provides a lightweight threat-tree-based variant for more efficient privacy threat elicitation [47].

The framework organises privacy threats into seven categories, one per letter of the acronym: Linkability, Identifiability, Non-repudiation, Detectability, Data Disclosure, Unawareness & Unintervenability, and Non-compliance. Full definitions of each category, with examples drawn from the present setting, are given in Appendix F (Table F.1). Two of these require a distinction that is central to this analysis. **Unawareness** is the absence of knowledge: the data subject does not know that processing is occurring, which data is involved, or which parties have access to it. **Unintervenability** is the absence of a mechanism to act: even an informed data subject cannot access, correct, or contest the data being processed about them. Unawareness removes the informational precondition for control; unintervenability removes the practical one.

It is important to note that LINDDUN is a privacy threat modelling framework, not a security framework. The existing literature on MCP and A2A addresses predominantly security threats: prompt injection, token theft, supply chain attacks, and authentication gaps [7, 29]. These are threats to the deploying organisation. LINDDUN addresses threats to the data subject: the individual whose personal data is processed. This distinction determines the analytical perspective of this thesis. A threat that is benign from a security standpoint, such as a correctly functioning session identifier (the persistent token MCP assigns to all calls within one working session) that produces cross-store linkability, is a significant privacy threat from the data subject’s perspective. The contribution of this thesis therefore lies precisely

in the application of a data-subject-centred framework to protocols that have until now been analysed almost exclusively from an organisational security perspective.

LINDDUN's DFD-based elicitation also maps onto MCP and A2A at the level this thesis requires: in a high-level exchange, a principal developer of the framework indicated that the session and task models of the two protocols correspond to the data flow and trust boundary elements that LINDDUN uses [48].

2.2.3. How LINDDUN is Applied: Three Steps

The LINDDUN methodology consists of three sequential steps: Model, Elicit, and Manage [44]. A full procedural description of each step is provided in Deng et al. [44]. The following paragraphs summarise how each step is applied in this thesis. Figure 2.1 gives an overview of the methodology.

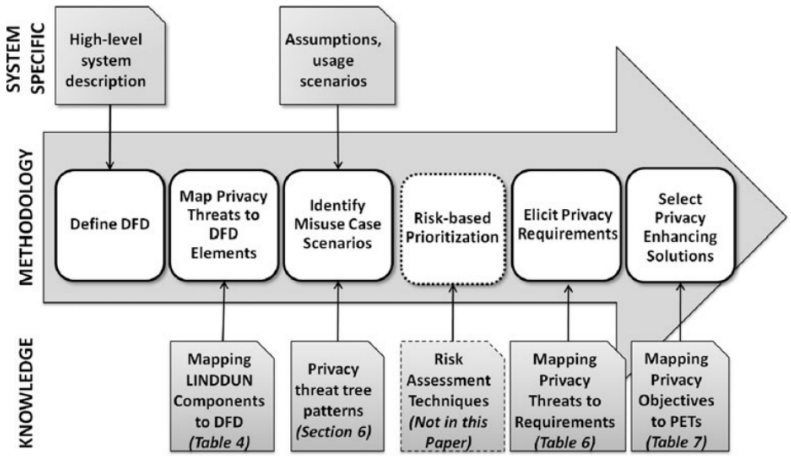

Figure 2.1: LINDDUN methodology overview. Source: [44].

Step 1: Model

The first step is to construct a **Data Flow Diagram (DFD)** of the system: a structured model of how data moves between external entities, processes, and data stores, and where it crosses **trust boundaries**, which in this thesis correspond to organisational boundaries. A key modelling requirement is that data flows are specified by the type of personal data transmitted rather than as generic channels [44]. The DFD notation and the full diagram, covering all four protocol configurations of MCP and A2A in the use case, are presented in Chapter 4.

A methodological consequence of this design follows directly from how LINDDUN works. Because LINDDUN operates on a data flow diagram of a concrete system, it cannot be applied to a protocol in the abstract: a use case is not an optional addition but a requirement of the method. The analysis therefore necessarily produces threats of two kinds. Some follow from the design of MCP and A2A themselves and would be present in any faithful deployment of these protocols; others follow from the broader configuration of the fraud detection system and would arise under any communication technology. Rather than treating the second kind as noise to be discarded, this thesis reports all identified threats and then classifies them by the extent to which they are protocol-inherent or use-case-inherent. This classification is itself a result. Distinguishing what the protocols contribute to privacy risk from what the surrounding system contributes is only possible because the analysis runs over a concrete use case, and it is what lets the protocol-level research question be answered without overstating which threats the protocols actually cause. The classification is applied threat by threat in Chapter 8.

Step 2: Elicit

The second step systematically maps the seven LINDDUN threat categories to each DFD element using the LINDDUN mapping table, and then elaborates each identified threat using privacy threat tree patterns. A **threat tree** is a structured diagram that breaks down how a particular threat can materialise: it shows the conditions that must hold and the steps an attacker or system would need to follow for the threat to occur. The full set of LINDDUN threat trees used in this thesis is reproduced in Appendix A. Each elaborated threat is documented as a **threat description**, which records the threat summary, the assets and data subjects at risk, the type of actor causing the threat, the attack flow, and the preconditions derived from the threat tree [44]. In this thesis, the Elicit step is carried out in Chapter 5.

Step 3: Manage

The third step prioritises the identified threats by their likelihood and impact, and in the full methodology also proposes privacy-enhancing measures. LINDDUN supports manual categorical assessment rather than quantitative scoring [44], which suits a design-level analysis of a hypothetical system where no operational data is available. How likelihood and impact are defined for this study, and how they combine into four priority levels (Critical, High, Medium, and Low), is set out in detail in Chapter 6. The Manage step is applied here in its first sub-activity only: threats are prioritised and the most significant are described in detail; the selection of privacy-enhancing measures is outside the scope of this study (Section 2.3.1).

2.3. Scope, Use Case Selection, and Description

2.3.1. Scope and Delimitations

This thesis is bounded by several explicit delimitations that follow directly from the research question and the analytical framework selected in Section 2.2. These delimitations define the boundaries of the analysis as methodological choices; their implications for the generalisability and limitations of the findings are reflected on in Chapter 8 (Discussion).

Two protocols only. The analysis covers MCP and A2A exclusively. Other emerging agent communication protocols, such as ANP and Agora, are outside the scope of this study. MCP and A2A were selected on three grounds. First, they are the most widely adopted protocols for LLM-based multi-agent systems: MCP was introduced by Anthropic in 2024 and adopted by OpenAI and Google DeepMind within a year; A2A was introduced by Google in 2025 and immediately attracted broad industry support [29, 10]. Both protocols have since been placed under the vendor-neutral governance of the Linux Foundation, A2A through the Agent2Agent project in June 2025 [49] and MCP through the Agentic AI Foundation in December 2025 [50]. ANP and Agora have not reached comparable production adoption. Second, MCP and A2A address complementary communication layers: MCP governs vertical communication between an agent and external tools and data sources, while A2A governs horizontal communication between agents. Analysing both protocols within a single use case makes it possible to examine how personal data flows across both layers simultaneously, which is the configuration most likely to arise in real deployments. Third, MCP and A2A are the only two protocols for which sufficiently complete and stable public specifications exist to support a structured LINDDUN analysis. ANP and Agora lack the specification maturity needed to construct a reliable DFD at the level of detail LINDDUN requires.

Protocol design as specified, not implementation. The analysis focuses on privacy threats that arise from the *protocol design* of MCP and A2A: their specified architecture, communication roles, session and task models, and trust boundary structures. Privacy leaks that arise from specific software implementation errors, configuration mistakes, or deployment vulnerabilities are outside the scope of this study. This distinction is methodologically important: the design-level threats identified in this thesis are structural properties of the protocols as specified. Removing them would require changes to the protocol specifications themselves, not merely to individual implementations. Implementation-level vulnerabilities require different methods, such as code audits or penetration testing, and are identified as a direction for future research.

Technical and structural analysis only. The analysis focuses on the privacy threats that arise from the architectural and operational properties of MCP and A2A. Regulatory frameworks, including the General Data Protection Regulation (GDPR) [51], the Payment Services Directive 2 (PSD2) [52], and the EU Artificial Intelligence Act (AI Act) [53], are outside its scope. Assessing them requires a legal rather than a technical methodology, and LINDDUN operates at the level of data flows and trust boundaries rather than compliance obligations. The AI Act is a particular case, since for much of this research its application to multi-agent AI systems was still being put into practice, making a legal analysis premature; this is identified as future work (Section 10.3). This exclusion concerns the *analysis*, not the discussion. Regulatory and legal considerations are still referred to as background for interpreting threats, most notably the statutory legal basis for fraud detection in Chapters 7 and 8, but no judgement is made about whether the use case complies with any regulation.

Single, hypothetical, financial-sector use case. The analysis is conducted within one hypothetical use case in the financial sector rather than across multiple deployments or domains, because the goal is to understand the privacy properties that follow from the protocols' structural design rather than to compare sectors. The rationale for this choice, and the extent to which the findings transfer to other domains, are set out in Section 2.3.2 and Chapter 9. Multi-case and multi-sector comparison is

identified as future research.

Threat identification and validation only; no privacy-enhancing measures. This thesis identifies and validates privacy threats; it does not propose or evaluate privacy-enhancing measures or mitigations. The selection of appropriate countermeasures would require a separate design-oriented study and is identified as a direction for future research.

Modelled flows only; no agentic autonomy or inference risks. The LINDDUN analysis is applied to the data flows that are explicitly modelled in the DFD. In practice, LLM-based agents may autonomously initiate additional tool calls or task delegations that are not specified in advance and therefore not representable in a static DFD. Privacy threats arising from such autonomous behaviour, including inference risks where agents derive sensitive attributes from individually innocuous data, and model memory risks where prior session content influences future responses, are outside the scope of this analysis. These categories of risk are acknowledged as analytically relevant but require methods beyond LINDDUN to address.

Specific protocol versions. The analysis is based on the MCP specification revision 2025-03-26 and the A2A specification version 0.2.5. Both protocols are actively developed and have received significant updates during the period of this research. Findings tied to specific protocol mechanisms may be affected by future specification changes. Their applicability to future versions requires verification against updated specifications.

DFD completeness as a bounding condition. The scope of the threat analysis is bounded by the completeness of the Data Flow Diagram constructed in Chapter 4. Flows or system components that are not represented in the DFD are not analysed. The DFD was constructed from the use case description and the protocol specifications; it is a modelling artefact rather than an empirical observation of a live system.

2.3.2. Use Case Selection

The selection of an appropriate use case is a critical methodological step. The use case must satisfy three criteria that follow directly from the research question and scope delimitations stated above.

First, the use case must involve both MCP and A2A in active roles, so that the privacy properties of both protocols can be analysed within a single coherent system. Second, all four analytically distinct protocol configurations must be present. These are MCP used by an agent to reach tools or data sources within its own organisation or in another, and A2A used for communication between agents of the same organisation or of different organisations. Each configuration involves a qualitatively different trust boundary structure, and therefore a different privacy risk profile. Third, the use case must involve the processing of sensitive personal data in a cross-organisational setting, so that the data flows central to the research question are present and analytically accessible.

The anti-fraud use case introduced in Section 2.1 satisfies all three criteria. Both protocols are used in complementary roles, all four protocol configurations are realised within a single workflow, and the system processes highly sensitive personal data, including transaction histories, behavioural profiles, and identity records, across an organisational trust boundary separating two independent financial institutions.

The three criteria are necessary conditions, but they do not by themselves single out the financial sector: other domains, such as a healthcare scenario in which a clinical agent retrieves patient records and delegates a referral, could in principle satisfy the same conditions. Financial fraud detection was selected from among these candidate domains for three reasons that make it the analytically strongest choice. First, it is one of the few settings in which all four protocol configurations arise naturally within a single workflow, rather than having to be constructed artificially. A card-issuing bank and a card network operator are operationally independent organisations that must nonetheless cooperate in real time on the same transaction. This makes the cross-organisational A2A configuration inherent to the scenario rather than imposed on it. Second, financial transaction data supports large-scale behavioural inference, as illustrated by industrially deployed fraud-detection systems [54], which places it among the most privacy-sensitive categories of personal data and maximises the privacy stakes of the analysis. Third, the financial sector is demonstrably among the leaders in adopting multi-agent AI for exactly this class of task [41, 8]. The financial use case is therefore best understood as the domain that renders the privacy properties of MCP and A2A most sharply observable, not as the only domain in which they apply; the extent to which the findings transfer to other sectors is examined in Chapter 9.

The use case is hypothetical rather than an audit of a live system, and this is a deliberate methodolog-

ical choice rather than a concession. Two considerations justify it. First, production fraud detection systems operated by banks and card networks are proprietary and are not accessible for external academic study, so analysing a specific live deployment was not feasible in any case. Second, and more fundamentally, the objective is to expose the structural privacy properties that follow from the design of the two protocols, not to audit one particular implementation of them. A hypothetical use case that is grounded in verified current practice makes it possible to realise all four protocol configurations within one coherent system, which a single real deployment would rarely exhibit all at once, while keeping the analysis at the conceptual-design level defined in the scope delimitations in Section 2.3.1. The grounding of the use case in documented practice is set out in detail in Chapter 4. In keeping with the framing introduced with the research question in Chapter 1, the protocols are the object of study and the use case is the instrument: it is the controlled setting in which the protocols' privacy properties are made observable, not the subject of the analysis in its own right.

Figure 2.2 brings this together: the protocol-level question is the object of study, the use case is the instrument through which the protocols' privacy properties are made observable, and the threat classification introduced above is what separates the protocols' contribution from that of the use case.

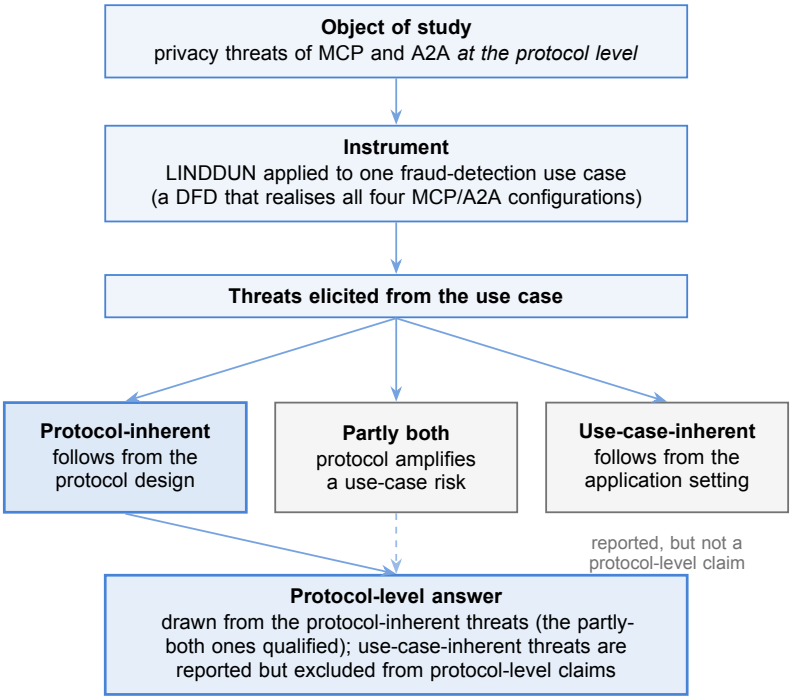

Figure 2.2: The use case as instrument. Because LINDDUN cannot be applied to a protocol in the abstract, the protocol-level question is answered through a single concrete use case. Every elicited threat is then classified by attribution, and the protocol-level answer is drawn from the protocol-inherent threats, with the use-case-inherent ones reported separately. Author's own analysis.

2.3.3. Use Case Description

The selected use case is a hypothetical real-time fraud detection workflow in the financial sector. The workflow involves two independent organisations: Bank X, a card-issuing bank, and Card Network Y, a card network operator. When a payment triggers elevated fraud risk signals, a coordinated sequence of agent interactions runs automatically across both organisations to determine whether the transaction should be approved, declined, or escalated to a human analyst. All four protocol configurations are realised within this single workflow.

The actors, the processes they perform, the two trust boundaries, and the complete set of data flows are presented in full as the use case in Chapter 4. This chapter introduces the use case only to the extent needed to motivate the methodological choices set out above.

2.4. Research Design

The research is structured into four phases. Phases 1 and 2 correspond directly to the Model and Elicit steps of the LINDDUN methodology and constitute the core analytical work of the thesis. Phase 3

validates the findings through expert interviews. Phase 4 synthesises all outputs into a direct answer to the main research question. The foundational scoping work, including the literature review, framework selection, and use case description, is reported in Chapter 1 and in this chapter.

The research is guided by the following main research question:

> *"What privacy threats emerge from the use of MCP and A2A communication protocols, as identified through a LINDDUN-based analysis of a financial fraud detection scenario?"*

The four phases presented below collectively answer this question through a structured set of sub-questions. Table 2.2 provides an overview of all phases, including their sub-questions, deliverables, research methods, and data requirements.

Table 2.2: Research design overview. Author's own elaboration.

Phase	Sub-question	Deliverable	Method	Data req. & sources
Phase 1 *Model*	SQ1. How do MCP and A2A function within the use case, and how can the system be represented as a DFD?	Data Flow Diagram with all five LINDDUN elements: entities, processes, data stores, data flows, and trust boundaries	Document analysis of protocol specifications; DFD modelling following the LINDDUN Model step	MCP and A2A protocol documentation; use case description; actor roles; system and trust boundaries; data flow descriptions
Phase 2 *Elicit*	SQ2. What privacy threats are present in the system, and what is their relative severity?	Privacy threat inventory per DFD element and trust boundary; prioritised threat register (the threat inventory ranked by severity) with detailed descriptions of the most significant threats	LINDDUN threat mapping, threat tree analysis, and risk-based prioritisation following the Elicit and Manage steps	LINDDUN threat categories and threat tree patterns; DFD from Phase 1; trust boundary crossings; protocol interaction points
Phase 3 *Validate*	SQ3. To what extent are the identified threats recognised and validated by domain experts, and which additional threats do experts identify?	Structured summary of expert interview findings; comparison of expert assessments with the threat inventory; documented additional threats	Semi-structured expert interviews	Outputs of Phases 1–2; interview guide; expert knowledge of agentic AI and financial sector privacy
Phase 4 *Synthesis*	Synthesise all findings and answer the main research question.	Explicit MRQ answer; reflection on contributions and limitations; directions for future research	Interpretive synthesis; cross-phase integration	Validated outputs of all phases; expert interview findings; literature on privacy design and agentic AI

2.4.1. Phase 1: Model

Phase 1: Model

Phase 1 corresponds to the first step of the LINDDUN methodology: the Model step [44]. The purpose of this phase is twofold. First, MCP and A2A are described technically, covering their architectures, components, communication roles, and how they function within the selected use case. Second, the system is formalised as a Data Flow Diagram following the LINDDUN Model step, identifying all external entities, processes, data stores, data flows, and trust boundaries. Phase 1 is reported in Chapter 3 and Chapter 4.

Sub-question. SQ1: How do MCP and A2A function within the use case, and how can the system be represented as a DFD?

Deliverable. A Data Flow Diagram with all five LINDDUN elements, serving as the direct structural input to Phase 2.

Method. The technical descriptions of MCP and A2A in Chapter 3 are produced through systematic document analysis of the official protocol specifications and the academic literature reviewed in Chapter 1. Each protocol section identifies which specification source a given architectural property is drawn from, and distinguishes between properties that are explicitly stated in the specification and properties that are analytical interpretations derived from the specification's structural implications. The DFD in Chapter 4 is then constructed by mapping the use case actors, processes, data stores, and data flows onto the LINDDUN Model step elements. The protocol descriptions from Chapter 3 provide the basis for identifying which data flows are present and

what personal data each one carries.

Data requirements and sources. MCP and A2A protocol documentation; use case description; actor roles; system and trust boundaries; data flow descriptions.

Reported in: Chapter 3 (Communication Protocols: MCP & A2A), Chapter 4 (DFD Construction).

2.4.2. Phase 2: Elicit

Phase 2: Elicit

Phase 2 corresponds to the Elicit and Manage steps of the LINDDUN methodology [44]. All privacy threats arising from the DFD are systematically identified, prioritised by likelihood and impact, and the most significant threats are described in detail. Phase 2 is reported in Chapter 5 and Chapter 6.

Sub-question. SQ2: What privacy threats are present in the system, and what is their relative severity?

Deliverable. A complete privacy threat inventory per DFD element and trust boundary; a prioritised threat register; and detailed self-contained descriptions of the most significant threats serving as primary input for Phase 3.

Method. LINDDUN threat mapping, threat tree elaboration, and risk-based prioritisation following the Elicit and Manage steps [44]. Each threat is assessed on likelihood and impact using the four-level categorical scale defined in Section 2.2.3 and assigned a risk level accordingly.

Data requirements and sources. DFD from Phase 1; LINDDUN threat categories, mapping table, and threat tree patterns; technical descriptions of MCP and A2A from Chapter 3.

Reported in: Chapter 5 (Privacy Threat Identification), Chapter 6 (Threat Prioritisation and Descriptions).

2.4.3. Phase 3: Validate

Phase 3: Validate

Phase 3 validates the threat inventory produced in Phase 2 through semi-structured expert interviews. The purpose is to assess whether the identified threats are recognised as realistic by practitioners with direct experience of agentic AI systems, and to surface any additional threats from practical knowledge. Phase 3 is reported in Chapter 7.

Sub-question. SQ3: To what extent are the identified threats recognised and validated by domain experts, and which additional threats do experts identify?

Deliverable. A structured summary of expert interview findings; a validation matrix comparing expert assessments with the threat inventory; documented additional threats.

Method. Semi-structured interviews using purposive sampling, targeting three expertise profiles: MCP/A2A/agentic AI deployment, privacy by design in AI systems, and financial sector AI. Responses are coded as confirmed, refined, or challenged per threat and summarised in a structured validation matrix.

Data requirements and sources. Threat inventory and descriptions from Phase 2; interview guide; expert knowledge of agentic AI system design and financial sector AI practices.

Reported in: Chapter 7 (Expert Interviews).

2.4.4. Phase 4: Synthesis

Phase 4: Synthesis

Phase 4 synthesises the outputs of all previous phases into an explicit answer to the main research question. This phase does not introduce new empirical analysis but integrates and interprets the complete set of findings from Phases 1 to 3. Phase 4 is reported in Chapters 8, 9, and 10.

Deliverable. An explicit answer to the main research question integrating all sub-question answers; a reflection on the scientific and practical contributions; and an identification of limitations and directions for future research.

Method. Interpretive synthesis and cross-phase integration.

Data requirements and sources. All outputs of Phases 1–3; literature on privacy design and agentic AI.

Reported in: Chapter 8 (Discussion), Chapter 9 (Reflection and Contribution), Chapter 10 (Conclusion).

2.5. Chapter Summary

This chapter has established the methodological foundation for the rest of the thesis. The research adopts a qualitative approach examining a single hypothetical use case, justified on the grounds that the goal is to understand privacy properties inherent in the structural design of two specific protocols rather than to generalise across deployments. LINDDUN was selected as the analytical framework over MAESTRO and the NIST Privacy Framework because it is the only candidate that is both privacy-specific and operates at the level of individual data flows and trust boundaries, which is the level at which MCP and A2A function. The selected use case is a real-time fraud detection workflow between Bank X and Card Network Y, which realises all four analytically distinct protocol configurations simultaneously. The actors have been identified, with the Cardholder as the central data subject who has no visibility into the automated processing that concerns them. The analysis is explicitly bounded to the design of the protocols as specified, excluding implementation-level vulnerabilities and regulatory compliance analysis. Likelihood and impact are used as the two assessment dimensions because they capture the independent variables needed to prioritise threats at a design level without requiring access to a live system. The analysis is carried out through a four-phase research design that models the system, elicits privacy threats, validates them through expert interviews, and synthesises the results into an answer to the main research question. The limitations arising from these scope choices are reflected on in Chapter 8.

3 Communication Protocols: MCP and A2A

3.1. Introduction

When a fraud detection agent retrieves a cardholder's transaction history through MCP and forwards a behavioural profile to an external risk analysis agent through A2A, the architecture of those two protocols determines precisely what happens to the cardholder's personal data, what is transmitted, to whom, along which path, and what either party can do with it afterwards. This chapter explains how those protocols work and why their design creates privacy risks that the organisations deploying them cannot fully resolve on their own.

The argument proceeds in three steps. The first part explains what a communication protocol is and why design choices at the protocol level have direct and irreversible consequences for privacy: once a protocol is deployed, the organisation using it cannot undo the structural properties it carries. The second part describes MCP and A2A in sufficient technical detail to identify precisely where those structural properties arise. The third part brings the two protocols together and explains why their combination creates privacy risks that neither protocol produces individually. The protocol components described here map directly to the Data Flow Diagram elements in Chapter 4 and serve as the basis for the LINDDUN threat analysis in Chapter 5.

Analytical approach. The descriptions of MCP and A2A are based on systematic reading of the official protocol specifications: the MCP specification revision 2025-03-26 [55] and the A2A specification version 0.2.5 [30], supplemented by the academic literature reviewed in Chapter 1. Each architectural property is grounded in a specific part of the specification or a specific academic source, which is cited at the point of use. Trust assumptions (Sections 3.2.6 and 3.3.6) are identified by reading the specification for what it explicitly defines, and noting what is structurally absent or left to the deploying organisation.

3.1.1. What a Communication Protocol Is

For two systems to exchange information reliably, they must agree in advance on a common set of rules governing how that exchange takes place. Such a set of rules is called a **communication protocol**, as introduced in Section 1.1. The standard textbook definitions formalise this in two complementary ways. Forouzan describes a protocol as a set of rules governing data communication, determining what is communicated, how, and when [56]. Tanenbaum and Wetherall add that a protocol is an agreement between the communicating parties as to how communication is to proceed [57]. Together, these definitions capture two essential aspects of what a protocol does. It is, first, a technical specification that constrains the form of messages, and, second, a social contract that must be adopted by all parties before meaningful communication can occur.

A protocol is typically characterised by three constituent elements, namely syntax, semantics, and timing [56], summarised in Table H.2 (Appendix H).

For the purposes of this thesis, two aspects of the protocol definition are particularly important. First, although both MCP and A2A are application-layer protocols in the networking sense, operating on top of the lower-level networking and transport infrastructure that simply moves bytes across the network [57], this thesis refers to them as the **protocol layer**: the layer at which MCP and A2A define how LLM-based agents interact with tools and with each other. The **application layer**, in the remainder of this thesis, denotes the agent software that operates on top of this protocol layer, while the *model reasoning layer* denotes the language model that the agents invoke above it. This stack is illustrated in Figure 3.1. The privacy properties of MCP and A2A are therefore determined by their semantic rules and payload constraints (a payload is the application data a message carries), not by the underlying network infrastructure.

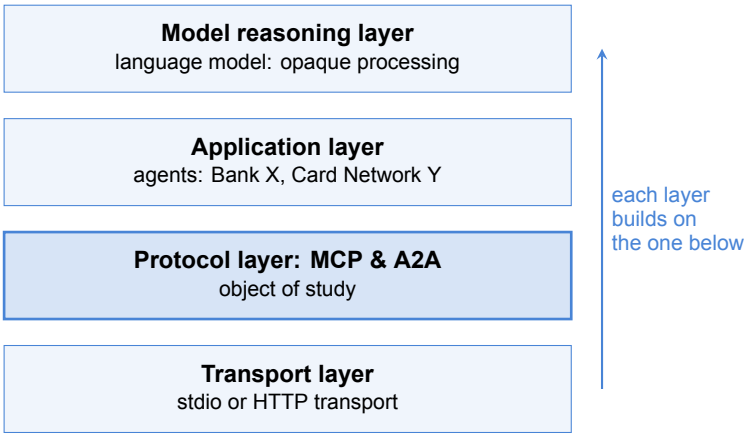

Figure 3.1: The position of MCP and A2A within the communication stack. Both are application-layer protocols in the networking sense, but within this thesis they are referred to as the *protocol layer*: the object of study, situated above the transport layer (stdio or HTTP) and below the application layer at which the agents operate. (Author's own illustration.)

Second, both protocols are *de facto* standards, meaning they have become standard in practice through widespread industry adoption rather than through a formal standardisation process run by an official standards body [56]. This means their specifications may be incomplete, inconsistently implemented, or subject to revision, which is a relevant consideration when assessing the robustness of any trust assumptions the protocols impose.

3.1.2. Why Protocols Matter for Privacy

Privacy violations in communication systems do not arise only from unauthorised access to stored data. They arise equally from how information flows between parties: who sends what to whom, under what conditions, and along which paths. Building on the contextual integrity conception of privacy introduced in Section 1.1 [9], a flow of personal data is privacy-respecting when it conforms to the norms of the context in which it occurs, and privacy-violating when it does not, regardless of whether the data is technically accessible or technically protected. A communication protocol directly governs these flows. It specifies which parties may initiate a data exchange, what information must be included in a request, what the receiving party is permitted to do with the response, and across which boundaries data may travel. Protocol design therefore encodes norms about information flow, whether those norms are made explicit in the specification or not.

Spiekermann and Cranor draw a foundational distinction between two approaches to engineering privacy into systems [58]. *Privacy by policy* relies on notices, consent mechanisms, and contractual commitments between parties to protect personal data. *Privacy by architecture* embeds protection into the structure of the system itself, minimising the collection of identifiable data and ensuring that privacy-violating flows cannot occur regardless of the behaviour of individual parties. These two approaches differ fundamentally in their robustness. Policy-based protections depend on all parties adhering to their commitments, which may be difficult to enforce across organisational boundaries. Architectural protections, by contrast, make certain privacy violations structurally impossible rather than merely contractually prohibited. Communication protocols belong to the architectural layer: they determine what data can be transmitted at all, not merely what parties have agreed to transmit. A protocol that places no semantic constraints on message payload content, as both MCP and A2A do, places the entire burden of privacy protection on the policy layer, which offers substantially weaker guarantees than architectural protection.

Building on the LINDDUN privacy threat categories [44] and the distinction between privacy by architecture and privacy by policy introduced above [58], this thesis identifies three specific design choices at the protocol level that create structural privacy risks directly relevant to this analysis.

Mandatory identifiers. A protocol that requires a sender to include a persistent identifier in every message creates the preconditions for linkability regardless of application intent. Once an identifier is present in every interaction, an observer can link multiple exchanges to the same originating party without knowing the content of those interactions.

Unconstrained payloads. A protocol that places no restrictions on what may appear in a message body creates conditions under which sensitive personal data can be transmitted without any structural safeguard.

Cross-organisational data flow by design. A protocol specifically designed to operate across organisational boundaries means that data flows will routinely cross **trust boundaries**. A trust boundary is a point in a system where data passes from one zone of control into another that is governed by different actors, policies, or assumptions about who can be trusted, for example the boundary between one organisation's systems and an external party's. Such crossings, which in earlier architectures would have constituted natural barriers, become routine. The privacy risks that arise at such crossings are therefore not incidental to how these protocols are used, but inherent to their intended function.

All three characteristics apply to MCP and A2A, as the technical descriptions in Sections 3.2.1 and 3.3 will establish.

A further consequence of these structural properties is that privacy risks introduced at the protocol layer cannot be corrected at the application layer without redesigning the protocol itself. As Tanenbaum and Wetherall establish, protocols are layered such that each layer builds on the services provided by the layer below, and a property introduced at a lower layer is inherited by all layers above it [57]. If a protocol transmits more personal data than is necessary, or routes data across a trust boundary that should not be crossed, an application built on top of it cannot undo those flows after the fact. Organisations deploying MCP or A2A can add application-level access controls and contractual agreements, but they cannot retroactively remove structural properties of the protocol that create linkability or disclosure risks. Identifying and addressing these properties therefore requires analysis at the protocol level. This is the precise motivation for the LINDDUN-based protocol-level analysis carried out in this thesis.

This structural-versus-policy distinction is also why the security analyses of MCP and A2A reviewed in Section 1.2.5 cannot serve as a basis for privacy analysis: a protocol can be fully secured against external attackers while still structurally enabling excessive data collection, unintended disclosure across organisational boundaries, or linkability of individual interactions [58]. Such properties are invisible to a security-oriented analysis and are revealed only by examining the protocol at the level of its data flows and trust boundaries, as the LINDDUN methodology does.

3.1.3. Scope of the Protocol Analysis

Several application-layer protocols have emerged to standardise agent interoperability. Anbiaee identifies four as the current state of the art: MCP, A2A, the Agent Network Protocol (ANP), and Agora [10]. They divide into **vertical** protocols, which govern communication between an agent and an external tool or data source (MCP), and **horizontal** protocols, which govern communication between peer agents (A2A, ANP, and Agora); the two create different trust-boundary structures and therefore different privacy risk profiles. Of the four, only MCP and A2A have the stable, complete, and publicly available specifications and the production adoption needed to support a structured privacy analysis, so the remainder of this chapter focuses on these two. The full four-protocol comparison is given in Table H.1 (Appendix H), and the selection criteria that exclude ANP and Agora are set out in the scope delimitation in Chapter 2.

In summary, communication protocols are not neutral containers for data. Their design choices directly determine what personal data flows where, to whom, and under what conditions. MCP and A2A, as application-layer protocols operating across organisational boundaries without semantic constraints on payload content, concentrate privacy risk at the protocol level. The following two sections describe these protocols in technical detail, establishing the precise architectural properties that the LINDDUN analysis in Chapter 5 will examine.

3.2. The Model Context Protocol (MCP)

The description of MCP in the following sections is based on systematic reading of the official MCP specification revision 2025-03-26 [55], supplemented by the academic literature cited at each point of use. Where a claim goes beyond what the specification states directly, this is identified as an analytical interpretation. The goal is to establish precisely which architectural properties of MCP produce the privacy risks examined in Chapter 5.

3.2.1. Origin and Purpose

The Model Context Protocol (MCP) was introduced by Anthropic in November 2024 as an open standard for connecting LLM-based agents to external tools, data sources, and services [59]. Its development was motivated by a structural limitation in agentic AI systems: before MCP, each integration between an AI application and an external tool or data source required a custom, one-off implementation. As the number of tools and data sources grew, this created what Anthropic termed an $M \times N$

integration problem, in which M AI applications each required separate connectors to N external systems, resulting in fragmented, difficult-to-scale architectures [7]. Figure 3.2 contrasts tool invocation with and without MCP.

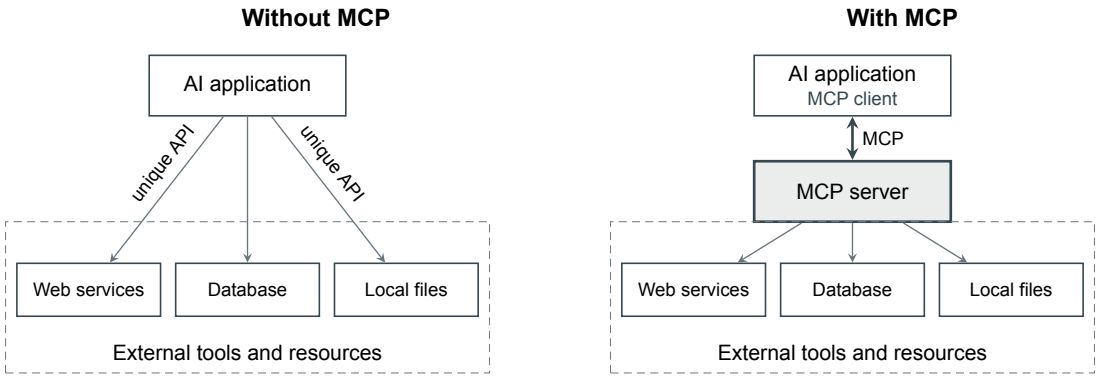

Figure 3.2: Tool invocation with and without MCP. Without MCP, an AI application requires a separate unique API connection for each external tool or data source. With MCP, the AI application functions as an MCP client that communicates with a single MCP server using the Model Context Protocol, providing a unified and standardised interface for all tool access. Author's own illustration, based on [7].

MCP addresses this by providing a universal, standardised interface that any LLM-based agent can use to discover and invoke external capabilities, regardless of the underlying tool or data source. Rather than building point-to-point integrations, developers implement MCP once on the agent side and once on the tool side, and the two can interoperate through the shared protocol. Hou et al. describe MCP as defining a unified, bidirectional communication and dynamic discovery protocol between AI models and external tools or resources, enabling interoperability and reducing the overhead of integration [7].

The protocol was released as open-source software with SDKs (**Software Development Kits**: pre-packaged code libraries that let developers integrate MCP into their applications without writing low-level networking code from scratch) for Python, TypeScript, Java, and C#, together with a reference implementation and a set of pre-built server integrations for common enterprise systems including GitHub, Google Drive, and PostgreSQL [59]. Adoption was rapid: by March 2025, OpenAI had officially adopted MCP across its product ecosystem, and Google DeepMind had announced integration support, signalling broad industry convergence around the protocol as a de facto standard for agent-to-tool communication [7]. In December 2025, Anthropic donated MCP to the Agentic AI Foundation, a directed fund under the Linux Foundation, formalising its status as a community-governed open standard [50].

From the perspective of this thesis, MCP is significant not because of its technical novelty but because of its role in the data flow architecture of agentic AI systems. Every interaction between an agent and an external data source in the use case examined in this thesis is mediated by MCP. As Hou et al. establish, MCP governs what data is requested, what is returned, and under what conditions, making it a primary locus of privacy risk in systems that process sensitive personal data [7].

3.2.2. Architecture and Roles

The most privacy-relevant design choice in MCP is that one component controls everything: all requests, all data flows, and all access decisions pass through a single central point. Understanding this architecture is what makes the privacy risks described in Chapter 5 traceable to a specific structural cause.

MCP defines a three-component architecture consisting of a **Host**, one or more **Clients**, and one or more **Servers** [7, 55]. Each component has a distinct role, and the boundaries between them are a primary locus of privacy risk in this thesis.

Host The application that the end user or orchestrating agent interacts with. In an agentic AI system, the host is the agent runtime itself: it manages the LLM context window, decides when to invoke external capabilities, and is responsible for authorisation decisions on behalf of the user. The host contains and manages all client instances [7].

Client A lightweight component embedded within the host. Each client maintains a dedicated, **stateful** (meaning it remembers the history of a session, rather than treating each message as independent) one-to-one connection with a single MCP server. One host can run multiple

clients simultaneously, each connected to a different server. The client translates the host's requests into **JSON-RPC 2.0** messages (a standard format for calling functions on a remote computer over a network, where JSON is a widely used text-based data format), manages the session lifecycle, and processes responses returned by the server [55].

Server An external process that exposes a specific set of capabilities to the client. It acts as a gateway to an underlying tool, data source, or service. Communication in MCP is always client-initiated: the server is passive and responds only to requests. Each server exposes capabilities through three primitive types: **Tools** (executable functions the agent can invoke), **Resources** (readable data the agent can access), and **Prompts** (templated instructions the agent can use) [7, 55].

A key architectural property is that the host is responsible for mediating what the client may request. MCP does not define a fine-grained access control mechanism at the protocol level: the specification states that hosts should implement appropriate access controls, but leaves the design and enforcement of these controls to the implementor [55]. This architectural openness has direct privacy implications, as discussed in Chapter 5.

Figure 3.3 illustrates the generic multi-server MCP architecture: a host runs one or more clients, each maintaining a stateful session with a separate server that exposes capabilities as tools, resources, and prompts. The multi-server configuration is the one relevant to this thesis. Bank X's Fraud Detection Agent realises this pattern as a host running multiple clients in parallel: one connected to an internal data source within Bank X's organisational boundary, and one connected to an external sanctions screening service outside that boundary. Each client-server pair operates within its own session, and the trust boundary between the host and each server is crossed independently.

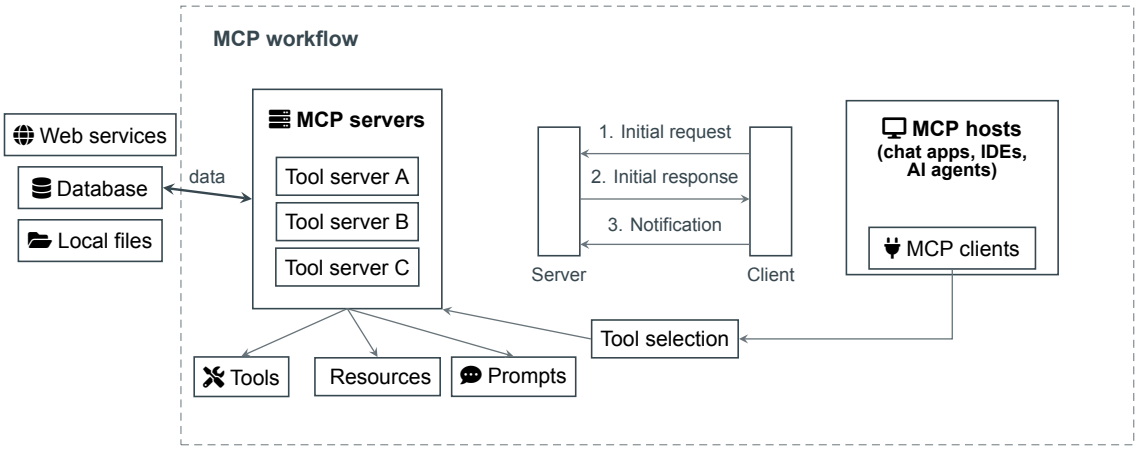

Figure 3.3: Generic multi-server MCP architecture. An MCP host (such as a chat application, IDE, or AI agent) runs one or more MCP clients, each maintaining a stateful session with a separate MCP server. Servers expose capabilities as tools, resources, and prompts, and act as gateways to external data sources. Communication within each session follows a request, response, and notification pattern. Author's own illustration, based on [7].

3.2.3. Session Model

Every MCP interaction takes place within a **session**: a persistent, stateful connection between a client and a server that is established once and maintained across multiple requests. The session begins with a short exchange in which the client and server declare their respective capabilities to each other before any operational data is transmitted. In a cross-organisational deployment, this mutual disclosure occurs across an organisational boundary, and the capability information exchanged at the start of a session may itself reveal details about the system's internal structure [55].

Beyond the initialisation exchange, the session connection persists from the first request to the last. As a result, all requests sent by the same client to the same server within a session can be linked to one another by any party that observes the session connection. A server that logs session activity accumulates a record of every tool invoked and every resource accessed during that session, all attributable to the same originating client. Hou et al. confirm that this session-level linkability is a structural consequence of MCP's stateful design and is present in every standard deployment [7]. The trust assumptions underlying this property are set out in Section 3.2.6, and its privacy implications are analysed in Chapter 5.

3.2.4. Message Content and Unconstrained Payloads

Each request that a client sends to a server contains two fields that are directly relevant to privacy: the input arguments passed to a tool (what the agent asks for), and the result returned by the server (what data the agent receives back). The critical privacy property of MCP's message design is that neither of these fields has any structural limit on what personal data it may contain [55, 7]. A tool that retrieves a customer's transaction history returns that history in the result field without any protocol-level constraint on retention, further use, or disclosure. Building on Forouzan's account of protocol semantics, a protocol that places no semantic constraints on the content of its message fields leaves that content unregulated at the protocol level [56]; under such conditions sensitive personal data can be transmitted without structural safeguards. In MCP, this is a deliberate design choice that maximises flexibility for developers at the cost of placing the entire burden of data governance on the application layer.

Listing H.1 (Appendix H) illustrates this concretely. The request asks for transaction history for a specific bank account. The response returns two transactions. Both the account identifier in the request and the transaction records in the response are personal data. Neither the request nor the response contains any field that records why this data is being accessed, who is accessing it, or how long it should be retained.

3.2.5. Capabilities: Tools and Resources

Tools and Resources are the two mechanisms through which personal data actually enters and leaves an MCP server. Everything discussed in Chapter 5 about what data flows where, and with what consequences, can be traced back to how these two capability types work.

Section 3.2.2 introduced Tools, Resources, and Prompts as the three primitive capability types that an MCP server exposes to a client. This section describes Tools and Resources in technical detail, as these two types are the primary channels through which personal data flows in the use case examined in this thesis. Prompts are templated instructions used to guide model behaviour and do not directly carry personal data; they are therefore not discussed further here.

Tools are executable functions that the agent can invoke on the server. Each tool is described by a name, a human-readable description, and an input schema defined in JSON Schema [55]. The input schema specifies which parameters the tool accepts, their types, and which are required. When an agent wishes to invoke a tool, it sends a `tools/call` request containing the tool name and the argument values, as illustrated in Listing H.1. The server executes the corresponding function and returns the result. Tools can have side effects: they may write to a database, trigger an external API call, or modify system state [7]. This distinguishes them from Resources, which are read-only. Sapkota et al. characterise this pattern as the core mechanism through which agentic AI systems acquire the ability to act on the world, rather than merely respond to queries: **tool invocation is what transforms a language model into an agent capable of producing effects beyond its own output** [19].

Before invoking a tool, the client must discover which tools the server exposes. This is done through a `tools/list` request, to which the server responds with a list of tool descriptors. Each descriptor includes the tool name, description, and input schema. The description field is particularly relevant for privacy: it is a free-text string that the host LLM reads to decide which tool to invoke for a given task. As Hou et al. observe, LLM agents rely on natural-language descriptions of tools to make invocation decisions, and these descriptions are not subject to any validation or access control at the protocol level [7]. A description may therefore name the categories of personal data the tool processes, or carry text crafted to influence agent behaviour, and it does so independently of whether the tool is ever actually invoked.

Resources represent data sources that the server exposes for the agent to read. Each resource is identified by a **URI** (Uniform Resource Identifier: a standardised string that uniquely identifies a resource, similar to a web address) and accompanied by metadata describing its name, type, and **MIME type** (a label that tells the receiver what kind of data a file or response contains, such as `application/json` for structured data or `text/plain` for plain text) [55]. Discovery works through a `resources/list` request; access works through a `resources/read` request specifying the target URI. Unlike tools, resources are read-only and do not have side effects. However, they may expose large volumes of structured personal data in a single read operation, such as a complete transaction history or a customer profile record.

The privacy-relevant property of resources lies in the URI structure. A URI such as `customer://NL91ABNA0417164300/transactions` encodes an account identifier directly in the address

of the resource. Any party that observes the URI, whether in a log, a network trace, or a notification message, can infer the identity of the data subject without reading the resource content itself. Identifiers embedded in protocol-level addressing fields create linkability conditions independently of payload content, because the addressing information is visible to any observer of the communication channel [55]. In the context of MCP, this means that an observer who sees which URIs are requested, and when, can learn which accounts are being accessed even without access to the transaction data returned. The privacy significance of such access patterns is examined in Chapter 5.

Table H.3 (Appendix H) summarises the key properties of Tools and Resources along the dimensions most relevant to privacy analysis, derived from the MCP specification [55] and Hou et al. [7].

3.2.6. Trust Model and Security Assumptions

MCP was not designed with privacy as a primary goal. Reading its specification carefully reveals a set of built-in assumptions about who is trusted and who is not: assumptions that make MCP practical to deploy but that leave the cardholder's personal data without architectural protection once it enters the system.

The architecture and message structures described in the preceding sections rest on a set of trust assumptions. These assumptions are identified by reading the MCP specification for what it explicitly defines, and inferring what is structurally absent or delegated to the implementor [55]. Where a claim goes beyond what the specification states directly, this is noted. The four assumptions below are the ones most relevant to the privacy analysis in Chapter 5.

The first and most significant assumption is that **the host is trusted by design**. The MCP specification places all authorisation responsibility on the host: it is the host that decides which servers clients may connect to, which tools may be invoked, and on whose behalf requests are made. The client defers entirely to the host's decisions and has no independent mechanism to verify whether those decisions are appropriate [55]. This design is deliberate: it mirrors the trust model of operating systems, in which applications are granted capabilities by a trusted runtime environment [55]. A structural consequence of this design, derived from the specification rather than from external literature, is that if the host is compromised or misconfigured, no protocol-level mechanism exists to contain the consequences: all connected servers and all data that passes through the host are exposed. Hou et al. confirm this as a systematic structural risk in their security analysis of MCP [7].

The second assumption is that **server identity is not verified at the protocol level**. The MCP specification does not define a mechanism by which a client can cryptographically verify the identity of the server it connects to before transmitting data. MCP supports two transport mechanisms. In the **stdio transport** (where the client and server exchange messages over the standard input/output streams of a locally launched program), the server is a local **subprocess** (a separate program started and controlled by the host on the same machine), and identity is implicitly established by the host's decision to launch it. In the HTTP transport, the specification recommends the use of **TLS** (Transport Layer Security: the standard encryption protocol that protects data in transit on the web, used for example in HTTPS connections) but does not mandate it. Even where TLS is used, it authenticates the transport channel rather than the MCP server as an application entity [55, 7]. A client connecting to a remote MCP server therefore has no protocol guarantee that the server is operated by the organisation it believes it to be. Ferrag et al. note that this absence of application-level identity verification is a systematic weakness across LLM agent protocols, enabling impersonation and man-in-the-middle attacks (where an attacker secretly relays or alters traffic between two parties) that are invisible at the transport layer [21].

The third assumption concerns **the treatment of tool output as trusted content**. When a server returns a result to the client, that result is passed into the host LLM's context window. The MCP specification does not define any mechanism for the host to validate or sanitise tool output before it reaches the model. A server that returns text containing instructions formatted as system prompts can therefore attempt to manipulate the model's subsequent behaviour, a class of attack known as indirect prompt injection [21, 7]. In the financial fraud detection context of this thesis, this means that a compromised or malicious external server connected via MCP could inject instructions into the agent's reasoning process through the content of its responses, without any interaction with the host's access control mechanisms.

The fourth and analytically most significant assumption for this thesis is the trust model that applies in **cross-organisational deployments**. When the MCP server is operated by a different organisation from the host, the client-side trust boundary extends beyond the deploying organisation's control. The organisation operating the host has no technical means to verify how the remote server processes,

stores, or retransmits the data it receives in tool call parameters. While the specification states data-privacy principles such as user consent and access control, it explicitly notes that it cannot enforce these at the protocol level and leaves them to the implementor; it defines no enforceable data minimisation, retention, or purpose limitation mechanisms in the protocol itself [55]. As Hou et al. observe in their security analysis of MCP, this structural property means that data transmitted to a remote MCP server is, from the sending organisation's perspective, effectively uncontrolled once it crosses the organisational boundary [7].

Table 3.1 summarises the four trust assumptions and their implications for privacy analysis.

Table 3.1: MCP trust assumptions and their privacy implications for cross-organisational deployments. Derived from the MCP specification [55] and Hou et al. [7].

Trust assumption	Basis in specification	Privacy implication
Host is trusted by design	Authorisation delegated entirely to host	Host compromise exposes all connected servers and data
No server identity verification	TLS recommended but not mandated; no application-level auth	Client cannot confirm data recipient before transmitting
Tool output treated as trusted	No output sanitisation defined	Injected instructions can manipulate agent behaviour
Cross-org. data flow uncontrolled	No data handling obligations in specification	Personal data transmitted to remote servers is uncontrolled beyond the trust boundary

In summary, MCP governs every interaction between an agent and an external data source or tool. The four structural trust assumptions described in this section, centralised host authority, absent server identity verification, trusted tool output, and uncontrolled cross-organisational data flow, are not implementation oversights. They are deliberate design choices that make MCP practical and interoperable across diverse environments. The privacy risks they produce are therefore present in every deployment that uses MCP as specified, regardless of how carefully the system is built. An organisation cannot remove these risks by writing better code; it can only add controls on top of a protocol that does not enforce them. A2A, described in the next section, works on a fundamentally different principle: instead of a central host controlling everything, two independent agents communicate as equals. This difference in architecture produces a different set of privacy risks, ones that are distributed rather than centralised and that are equally impossible to resolve at the application layer.

3.3. The Agent-to-Agent Protocol (A2A)

The description of A2A in the following sections is based on systematic reading of the official A2A protocol specification [30], supplemented by the academic literature cited at each point of use. As in the preceding analysis of MCP, the specification is read for what it explicitly defines, and the privacy-relevant trust assumptions are derived by identifying what the protocol leaves unconstrained or delegates to the implementor. Where a claim goes beyond what the specification states directly, this is identified as an analytical interpretation. The goal is to establish precisely which architectural properties of A2A produce the privacy risks examined in Chapter 5.

3.3.1. Origin and Purpose

The Agent-to-Agent Protocol (A2A) was introduced by Google in April 2025 as an open standard for enabling direct communication and task delegation between AI agents across organisational and platform boundaries [60]. To understand why A2A was needed, it is necessary to consider the structural challenge it was designed to address.

Multi-agent systems coordinate autonomous agents to solve complex problems by subdividing them into smaller tasks, each handled by a specialised agent [61]. Dorri, Kanhere and Jurdak identify communication and coordination between agents as one of the central challenges of multi-agent system design: without a shared communication standard, agents built on different platforms cannot reliably interoperate, exchange information, or coordinate their actions [61]. Wooldridge similarly identifies interoperability as a foundational requirement for multi-agent systems operating across organisational boundaries, noting that agents must be able to communicate meaningfully with other agents whose

internal architecture and reasoning mechanisms may be entirely different [14]. In practice, before A2A, each multi-agent framework such as LangChain, AutoGen, and Google's own agent infrastructure solved this problem in its own way, creating siloed ecosystems where agents from different vendors or frameworks could not interoperate [19]. A2A addresses this by defining a single protocol through which any two agents, regardless of their underlying framework or vendor, can discover each other's capabilities, delegate tasks, and exchange results.

The gap A2A fills is distinct from what MCP provides. MCP standardises how an agent connects to external tools and data sources, but it does not define how one agent delegates a task to another agent acting as an autonomous peer. As Yang et al. establish in their survey of AI agent protocols, MCP governs vertical communication between an agent and its tools, while A2A governs horizontal communication between agents acting as peers [28]. The two protocols are therefore complementary: in a typical deployment, an agent may use MCP to retrieve data from a local database and A2A to delegate a subtask requiring specialised capabilities to a remote agent operated by a different organisation.

The protocol was released as open-source software under a permissive open-source licence and launched with the support of more than fifty technology partners including Atlassian, Salesforce, SAP, and Workday [60]. In June 2025, Google donated A2A to the Linux Foundation, formalising its governance as a vendor-neutral, community-governed open standard [49]. This mirrors the governance trajectory of MCP and signals that A2A is intended as a durable infrastructure standard rather than a product-specific solution.

3.3.2. Architecture and Roles

Where MCP concentrates control in a single central host, A2A distributes it: two independent agents interact as equals, and neither has authority over the other. This difference is not merely technical. It means that once a client agent sends personal data to a server agent, it has no architectural mechanism to influence what that agent does with it.

A2A defines a two-role architecture consisting of an **A2A Client** and an **A2A Server** [30]. Both roles are occupied by autonomous agents with their own LLM-based reasoning capabilities, internal state, and the ability to initiate actions. The protocol defines how these two agents communicate; it does not constrain what either agent does internally [30].

A2A Client Also referred to as the client agent. The agent that initiates communication and delegates tasks. It discovers remote agents, selects an appropriate agent for a given task, and submits task requests. The client provides the task description, any required input data, and optionally a **callback endpoint**: a web address the client provides so the server can notify it when the task is complete, rather than the client having to repeatedly check. This is called **asynchronous** communication: the client does not wait for an immediate response but continues other work and receives the result later [30].

A2A Server Also referred to as the remote agent. The agent that receives task requests, processes them using its own reasoning and tool access, and returns results. The server exposes an HTTP endpoint that implements the A2A protocol interface. The A2A specification is designed so that agents interact without needing to share internal memory, tools, or proprietary logic: the server's internal architecture is fully **opaque** (meaning hidden and inaccessible) to the client [30].

Opacity is a deliberate design principle of A2A: agents collaborate on the basis of their declared capabilities without exposing their internal reasoning, tools, or proprietary logic [30]. This protection comes at a cost for the party on the other side of the exchange. Such, Espinosa and García-Fornes establish that in multi-agent systems, once data is shared with another agent the sending party loses the ability to verify how that data is handled after it crosses the communication boundary [26].

A further consequence of this architecture concerns the loss of oversight in delegation chains. As Such, Espinosa and García-Fornes establish, in multi-agent systems operating across organisational boundaries a delegating agent has no structural guarantee that the receiving agent applies the same data handling constraints as the delegating organisation [26]. In the A2A architecture, this property is inherent by design: the server is explicitly described as an opaque agent, and the specification provides no mechanism for the client to audit the server's internal data flows. Dorri, Kanhere and Jurdak further note that in decentralised multi-agent systems the absence of a central trusted authority makes verifying the identity of agents and establishing trust between them inherently difficult [61].

A key mechanism that connects the two roles is the **Agent Card**, a JSON metadata document published by each A2A Server that describes its identity, capabilities, supported interaction modalities,

authentication requirements, and service endpoint [30]. The Agent Card is the sole source of information a client uses when deciding whether to delegate a task to a given server. It is examined in detail in Section 3.3.3. Figure 3.4 shows the two-role client–server architecture.

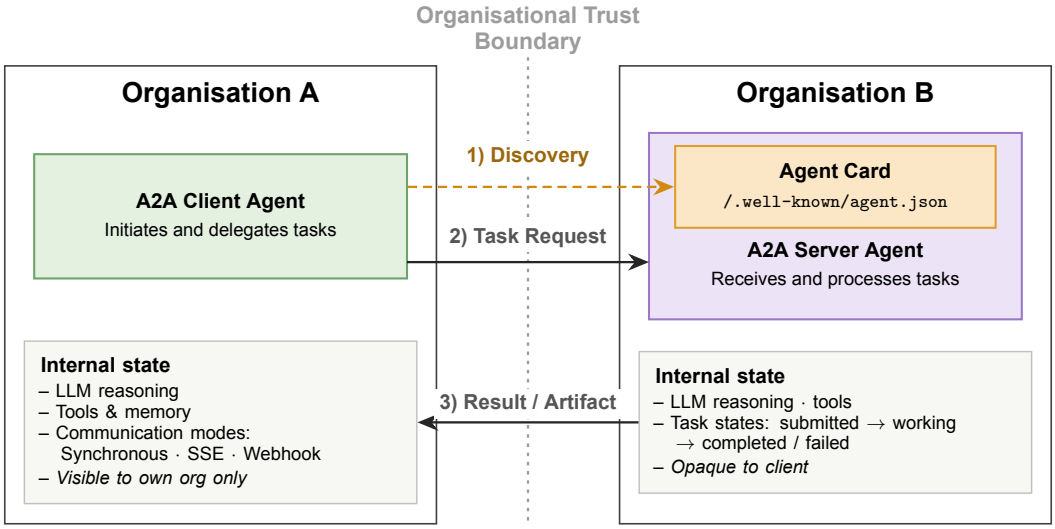

Figure 3.4: The A2A two-role architecture. The A2A Client Agent initiates task delegation by retrieving the Agent Card from a well-known URL on the server's domain (step 1), then transmitting a task request across the organisational trust boundary (step 2) and receiving the resulting artifact (step 3). The server agent's internal state, including its LLM reasoning, tools, and task lifecycle management, is fully opaque to the client by protocol design. Author's own illustration based on [30].

3.3.3. Agent Cards and Capability Discovery

The Agent Card is A2A's mechanism for making agents discoverable. It is also the first point in the A2A architecture at which personal information is exposed, before any task has been delegated and before any data has been exchanged.

Before an A2A Client can delegate a task to a remote agent, it must know that the agent exists and what it is capable of. A2A solves this through the **Agent Card**, a JSON metadata document that each A2A Server publishes and that clients use to determine whether a given server is appropriate for a given task [30].

An Agent Card is served at a **well-known URL** (a standardised, predictable web address that any client can construct and query without prior arrangement) on the agent's domain, by convention at the path `/.well-known/agent.json`. Clients can therefore discover any A2A-compliant agent by querying this path on a known domain, without requiring any prior registration or negotiation [30]. The Agent Card declares the agent's name, description, and version; the URL of its task endpoint; and the default input and output modalities it supports, expressed as media (MIME) types such as `text/plain` or `application/json`. It also declares the authentication scheme it requires, using standard security scheme types (API key, HTTP authentication, OAuth 2.0, or OpenID Connect) or none at all, together with a list of skills, each describing a specific capability the agent offers and any input and output constraints [30].

The Agent Card is a publicly accessible document that can be retrieved by any party, including parties that have no intention of delegating a task and no authorisation to do so. This accessibility creates a structural privacy exposure that is independent of any task delegation. Because the Agent Card declares the agent's skills, the data modalities it accepts, and its authentication scheme, any party that retrieves it learns what kinds of tasks the agent is designed to perform, what categories of data it processes, and what authentication it requires, all without any interaction with the agent itself [30]. The disclosure of this capability metadata therefore constitutes an information flow in its own right, occurring before any task has been delegated.

Nissenbaum's framework of contextual integrity provides a useful lens for analysing this property. A disclosure respects privacy when the information flow matches the norms of the context in which it occurs [9]. The Agent Card is designed to be publicly accessible for the purpose of enabling dynamic discovery. However, the same document may disclose information that is contextually appropriate only within an established relationship between specific parties. When an Agent Card reveals what

categories of personal data an agent accepts, or what authentication scheme it uses, that information may not be appropriate for unrestricted public disclosure. The structural publication of this information through a publicly accessible document can therefore create a contextual integrity violation regardless of whether any harmful use is made of the information.

A further privacy-relevant property of Agent Cards concerns the absence of cryptographic verification. Any party can publish a JSON document at the well-known path of a domain they control and populate it with arbitrary capability claims. The A2A specification does not require Agent Card content to be **cryptographically signed** (digitally sealed in a way that proves it was produced by a specific party and has not been altered) or verified against an authoritative registry [30]. Because such verification is optional rather than mandated, the absence of verifiable identity and capability claims means that a delegating agent cannot confirm that the receiving agent is what it claims to be, creating conditions for impersonation and capability misrepresentation [30]. Because the client cannot confirm who controls the endpoint, personal data transmitted in a task may reach an unintended recipient [44].

Table H.4 (Appendix H) summarises the Agent Card fields most relevant to privacy analysis and the nature of the exposure each creates.

3.3.4. Task Lifecycle and Message Structure

In A2A, personal data does not travel as a single message and disappear. It lives inside a Task: an object that persists on the server for the entire duration of the work, accumulating messages, intermediate results, and final outputs. This is the mechanism through which A2A enables powerful multi-step collaboration, and it is also the mechanism through which personal data can be retained at a remote organisation without the sending agent having any control over it.

The central interaction unit in A2A is the **Task**, a stateful object that represents a discrete unit of work delegated from the A2A Client to the A2A Server. Unlike a stateless function call, a Task persists on the server side throughout its execution and retains its state, messages, and intermediate results until it reaches a terminal state [30]. This stateful design enables A2A to support long-running, multi-turn interactions in which the client and server exchange multiple messages before the task is complete. In multi-agent systems, agents decide their actions partly from the history of their prior actions and interactions [61]; A2A's stateful Task design supports exactly this kind of continuity across a multi-step exchange, rather than requiring the context to be re-established on every message.

Each Task is identified by a unique identifier generated by the server when the task is created, returned to the client in the server's response. A Task progresses through a defined set of states during its lifetime, from `submitted` and `working` through terminal states such as `completed`, `failed`, or `canceled` [30]; the full set of nine states is listed in Table H.5 (Appendix H).

Within a Task, communication between client and server is structured through **Messages**. Each Message has a role field set to either `user` (sent by the client) or `agent` (sent by the server), and contains one or more **Parts** [30]. Parts are the smallest unit of content in A2A and come in three types. A **TextPart** carries plain text. A **FilePart** carries file content together with a MIME type. A **DataPart** carries structured data as an arbitrary JSON object [30].

The DataPart type is the most privacy-relevant of the three. Because it accepts any valid JSON object without schema enforcement at the protocol level (the protocol does not verify that the data matches any predefined structure or type), it allows structured personal data of any kind to be transmitted as part of a Task message. The specification defines no structural constraints on message payload content, which means that sensitive personal data can be transmitted without any protocol-level safeguard, and that the responsibility for limiting what is transmitted rests entirely with the application layer rather than the protocol [30]. Such, Espinosa and García-Fornes further establish that in multi-agent systems, data transmitted during task execution is subject to the privacy policies of the receiving agent rather than the sending agent, and that the sending agent has no structural mechanism to enforce how the received data is used after transmission [26].

When the Task reaches a terminal state, the server returns results through **Artifacts**: structured output objects composed of Parts that represent the agent's response to the task [30]. Artifacts persist on the server side until they are explicitly retrieved or deleted according to the server's retention policy. Data retention in this setting is a structural property of the protocol. Personal data transmitted during task execution may remain stored on the server long after the task is complete, without the sending agent having any visibility into or control over the retention period. The specification defines no retention constraints for Tasks or Artifacts, leaving this entirely to the server's implementation [30].

A further privacy-relevant structural property is that Tasks accumulate a history of the Messages exchanged during their execution. This history is retained on the server side and can be requested by the client through the `historyLength` parameter when querying the task [30]. Because this history is retained, the stored sequence of messages can reveal patterns of interaction, the frequency and timing of task delegation, and the categories of data requested, even when the content of individual messages is not sensitive in isolation [44]. The privacy significance of this accumulation is examined in Chapter 5.

Listing H.2 (Appendix H) illustrates the structure of an A2A task request containing a DataPart with structured data.

3.3.5. Task Persistence and Data Retention

A2A tasks are stateful objects that persist on the server throughout their execution. This means that personal data transmitted in a task request is held on the server for as long as the task remains active, and potentially longer if the server retains completed task records [30]. The protocol defines no constraints on how long a server may retain task data, who may access it, or whether it may be used for purposes other than the original task. Storage limitation and purpose limitation are foundational data-protection principles: personal data should not be retained beyond the point at which it is needed, and should not be used for purposes other than those for which it was collected. A2A provides no architectural mechanism to enforce either requirement.

One specific property of A2A task communication introduces an additional privacy exposure: when a client cannot maintain a persistent connection while the server processes a long-running task, it can provide a callback address to which the server sends the completed result. This means the client must disclose its own network address to the remote server agent at task creation time, before the task has been completed or the server's trustworthiness has been verified [30]. The disclosure of a client-side endpoint to an external agent is a form of infrastructure exposure that has no equivalent in synchronous task communication.

3.3.6. Trust Model and Security Assumptions

A2A's trust model produces a situation that is uncomfortable for the cardholder: personal data crosses an organisational boundary to an agent the sending organisation has never met before, operating under different rules, with no contractual obligations defined at the protocol level. The four assumptions below explain why this is a structural property of A2A, not a deployment oversight.

The trust model of A2A is characterised by the absence of a structural superior: both the A2A Client and the A2A Server are autonomous agents, and neither party holds authority over the other. Neither party has any pre-established trust relationship with the other beyond what is declared in the Agent Card at the time of discovery [30]. The trust assumptions that govern A2A interactions are therefore symmetric rather than hierarchical. This section identifies the four trust assumptions most relevant to the privacy analysis in Chapter 5.

The first assumption is that **authentication is declared but not mandated**. The A2A specification expects a server to declare its authentication requirements in its Agent Card, but does not require any scheme to be present and does not mandate a specific one. A server may declare, among other standard security scheme types, API key authentication (a simple password-like token) or **OAuth 2.0** (an industry-standard authorisation protocol that lets a user grant a third party limited access to their account without sharing their password), or it may declare no authentication at all [30]. The protocol leaves the choice entirely to the server implementor, and the client must accept whatever scheme is declared in order to interact with that server. Leo et al. identify the absence of enforced identity and authentication at the communication layer as a structural weakness in agentic AI systems, enabling impersonation and related threats [31]. Because the A2A specification declares authentication schemes but does not mandate their enforcement, a deployment that omits authentication still conforms to the specification: from the client's perspective it is therefore indistinguishable from a deployment that enforces it [30]. A client delegating a task containing personal data to a server that declares no authentication has no protocol mechanism to prevent unauthorised third parties from accessing the same endpoint.

The second assumption is that **identity claims are self-asserted and unverified**. As established in Section 3.3.3, any party can publish an Agent Card at a well-known URL on a domain they control and populate it with arbitrary identity and capability claims. The A2A specification does not require Agent Card content to be cryptographically signed or verified against an authoritative registry [30]. Ferrag et al. identify the absence of cryptographic identity verification in LLM agent protocols as creating conditions for impersonation attacks in which a malicious agent presents itself as a legitimate service and thereby obtains personal data that was intended for a different recipient [21]. A spoofed Agent Card published

at a convincing domain could lead a client to delegate tasks containing sensitive personal data to an unintended recipient, with no protocol mechanism to detect the deception before data is transmitted.

The third assumption concerns the **absence of data handling obligations in the protocol**. When a client delegates a task to a server agent, the task message and all data it contains pass into the server's processing environment. The A2A specification defines no obligations for the server regarding how it handles that data: there are no protocol-level constraints on retention, secondary use, further disclosure, or deletion [30]. Such, Espinosa and García-Fornes establish that this is a structural property of multi-agent systems at large: once data crosses a communication boundary between agents, the sending agent has no architectural mechanism to enforce purpose limitation or data minimisation on the receiving side [26]. This pattern is especially consequential when the receiving agent operates under a different organisational regime, where the absence of shared data governance at the protocol level means that no common constraint binds the two parties. Applied to A2A, this means that the third trust assumption is a structural feature of how peer-to-peer agent communication is designed, not an incidental gap in the specification: the protocol deliberately leaves data governance to the deploying organisations rather than encoding it at the protocol level.

The fourth assumption is the **absence of a prior trust relationship in cross-organisational deployments**. When the A2A Client and the A2A Server belong to different organisations, neither party has any prior knowledge of the other beyond what is declared in the Agent Card. There is no shared governance framework, no contractual data processing agreement at the protocol level, and no mutual accountability mechanism defined by the specification [30]. When personal data is transmitted across this boundary, the sending organisation has no architectural guarantee about how it will be handled, retained, or further disclosed by the receiving party. When data crosses an organisational trust boundary enforced only by contractual rather than architectural means, the risk of unintended disclosure becomes a property of the architecture rather than of any individual interaction [44].

Table 3.2 summarises the four trust assumptions and their privacy implications.

Table 3.2: A2A trust assumptions and their privacy implications. Derived from the A2A specification [30].

Trust assumption	Basis in specification	Privacy implication
Authentication declared but not mandated	Server declares scheme in Agent Card; no scheme is required	Client cannot verify access control before transmitting data
Identity claims self-asserted and unverified	No mandatory Agent Card signing or registry verification	Spoofed Agent Card may lead to data disclosure to unintended recipient
No data handling obligations defined	Specification silent on retention, secondary use, deletion	Receiving agent's data governance practices are unknown to sender
No prior trust relationship	Agent Card is sole source of identity information	Personal data crosses organisational boundary without architectural safeguards

In summary, A2A governs every agent-to-agent interaction in the system. Its four structural trust assumptions, optional authentication, unverified identity claims, absent data handling obligations, and no prior trust relationship, produce a different category of risk from MCP. Where MCP concentrates control and risk in a central host, A2A distributes both: each agent independently decides what data to include in a task and what to do with data it receives, without any architectural mechanism to enforce consistency across those decisions. The result is that privacy risks in A2A deployments arise not from a single point of failure but from the accumulated effect of many individual design choices, none of which the protocol constrains. Section 3.4 examines what happens when MCP and A2A operate together in the same system.

3.4. MCP and A2A Combined

As established in Section 3.3.1, MCP and A2A are complementary rather than competing: MCP governs vertical agent-to-tool communication and A2A horizontal agent-to-agent communication [28]. Sapkota et al. ground this in a broader taxonomy of agentic AI systems, distinguishing tool access from peer coordination as fundamentally different interaction patterns addressed by different protocol prim-

itives [19]. The two protocols therefore act as complementary layers rather than alternatives, each handling a distinct class of interaction within the same system. Figure 3.5 shows the combined MCP and A2A stack.

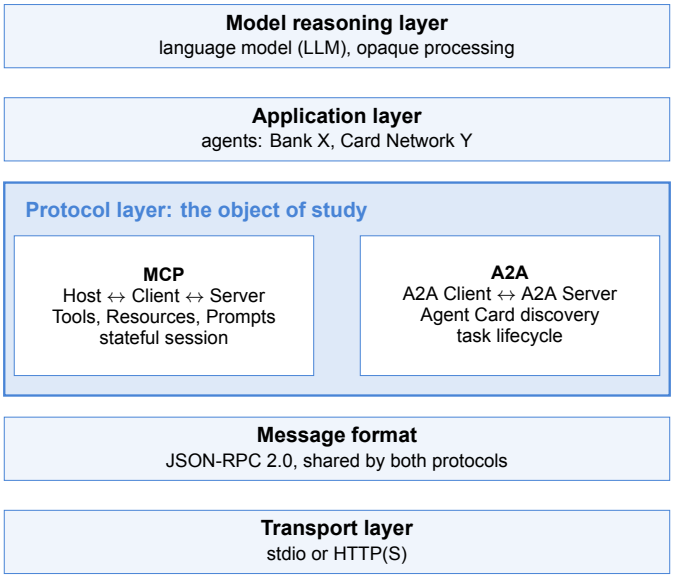

Figure 3.5: The combined MCP and A2A stack. The protocols differ only within the *protocol layer* and share the layers above and below, so a single agent can use both at once. (Author's own illustration.)

Both protocols were designed with the same overriding objective: to remove friction from agent communication. MCP was created to eliminate the $M \times N$ integration burden, and A2A to dissolve the silos that prevented agents from different frameworks from interoperating. In both cases the design goal is maximal interoperability and minimal friction, achieved through standardisation, permissive payloads, and automatic discovery. This same objective is what places the two protocols in tension with privacy. Privacy protection generally requires the opposite of frictionless exchange: it requires data minimisation, purpose limitation, and deliberate constraints on what is transmitted, to whom, and under what conditions. The properties that make MCP and A2A effective as interoperability standards, that any agent can connect, that payloads are unconstrained, and that capabilities are openly discoverable, are therefore the very properties that, left unaddressed, generate the privacy threats analysed in this thesis. The protocols are not privacy-hostile by intent; privacy was simply not among the design objectives, and the consequences of that omission are the subject of the chapters that follow.

This layered design has a direct consequence for privacy analysis. Because both protocols operate within the same system simultaneously, personal data retrieved through an MCP tool call may be incorporated into an A2A task message and transmitted to a remote agent operated by a different organisation. On the contextual-integrity view introduced in Section 3.1.2, the norms governing the original retrieval, which concern the relationship between the agent and its data source, do not necessarily apply to this subsequent cross-organisational transmission [9]. The combination of the two protocols can therefore create information flows that neither protocol individually anticipates or constrains. Figure 3.6 illustrates a host agent in one infrastructure discovering and delegating tasks to a remote infrastructure.

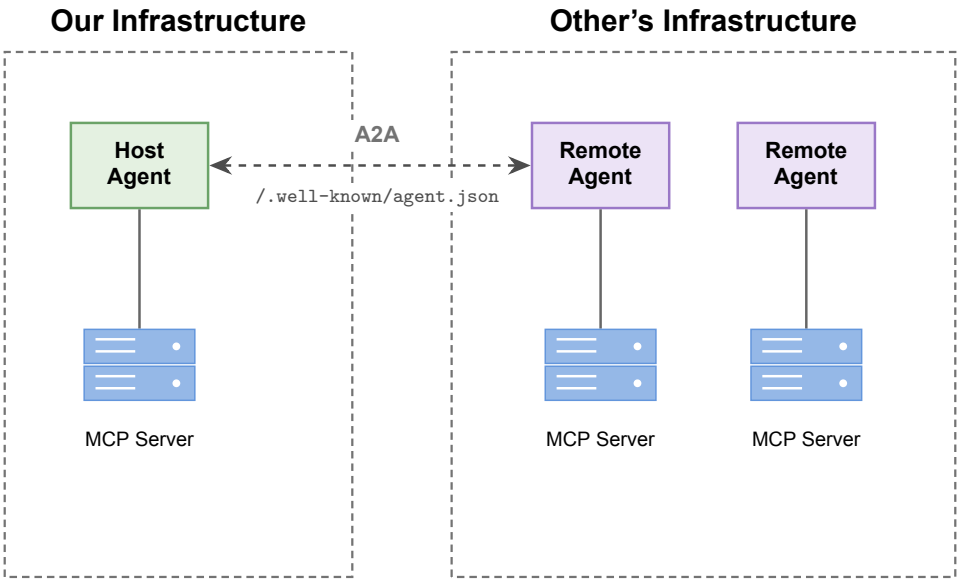

Figure 3.6: A host agent in one infrastructure discovers and delegates tasks to remote agents in another via A2A, retrieving each remote agent's capabilities from its Agent Card at the well-known path `/.well-known/agent.json`. Independently, every agent uses MCP to reach its own MCP servers (tools, resources and prompts). Author's own illustration based on the MCP and A2A specifications [55, 30].

Two structural properties shared by both protocols amplify this concern. The first is that neither protocol imposes semantic constraints on the personal data that may appear in its payload fields. Following Forouzan's account of protocol semantics, a protocol that treats message payload content as an opaque field places no protocol-level constraint on that content [56]; sensitive personal data can therefore be transmitted without any structural safeguard, regardless of what the application intends. Both MCP and A2A exhibit this property: MCP tool responses and A2A DataParts are both unconstrained at the protocol level with respect to content. The second is that both protocols support cross-organisational deployments without defining data handling obligations for the receiving party. Spiekermann and Cranor establish that protection mechanisms embedded in a system's architecture provide stronger privacy guarantees than contractual commitments between parties, because architectural protections make certain flows structurally impossible rather than merely contractually prohibited [58]. Neither MCP nor A2A provides architectural protection of this kind: both place the entire burden of cross-organisational data governance on the policy layer.

A further structural property, of a different kind, is common to both protocols and follows directly from their purpose. MCP and A2A are machine-to-machine protocols: in both, every participant is a software component: the host, client, and server in MCP, and the client and server agents in A2A. Neither protocol defines any interface through which a data subject, such as the cardholder whose data is processed, could observe the exchange, be notified that it is taking place, or intervene in it. The data subject is not a party to the protocol at any point [55, 30]. The implications of this property for cardholder awareness and intervention are analysed in Chapter 5.

Despite these shared properties, the two protocols create structurally different trust configurations that produce different privacy risk profiles. Reading the two specifications together, MCP's architecture can be characterised as a hub-and-spoke configuration: the host is the single central point through which all client connections and data flows are managed, and its compromise therefore endangers all connected servers and all data that passes through it. A2A's architecture, by contrast, creates a flat peer configuration: each agent independently discovers and interacts with other agents, and there is no central orchestrating component. This structural difference is not stated in these terms in either specification; it is an analytical characterisation derived from the architectural descriptions in the two specifications, the analysis of MCP by Hou et al., and the security analysis of A2A in Louck, Stulman and Dvir [7, 29]. In MCP, a compromise of the host endangers all connected servers and all data that passes through the host. In A2A, a spoofed or compromised server agent endangers only the data transmitted in the tasks delegated to that specific agent, but the absence of a central authority means there is no architectural mechanism to detect or contain the compromise before data is disclosed. These differences mean that each protocol creates a distinct category of privacy risk that

requires separate analysis. Chapter 5 therefore analyses each of the four protocol configurations, MCP within an organisation, MCP between organisations, A2A within an organisation, and A2A between organisations, individually through the LINDDUN framework.

3.5. Chapter Summary

This chapter has established the technical foundation required for the privacy threat analysis in Chapter 5. It introduced the concept of a communication protocol and explained why protocol design choices have direct privacy implications, before describing MCP as a vertical protocol governing agent-to-tool communication and A2A as a horizontal protocol governing agent-to-agent communication. The two protocols operate as complementary layers within the same system and share two structural properties that are analytically significant for what follows: the absence of semantic constraints on payload content, and the absence of data handling obligations for the receiving party in cross-organisational deployments. The protocol components described in this chapter map directly to the LINDDUN Data Flow Diagram elements used in Chapter 4: MCP servers and A2A agents correspond to processes, data sources to data stores, communication channels to data flows, organisational boundaries to trust boundaries, and human actors such as fraud analysts and cardholders to external entities.

An important analytical boundary applies throughout this chapter and the analysis that follows. The descriptions of MCP and A2A presented here are based on the protocol specifications: the documents that define what the protocols are and how they work by design. This means that the privacy properties identified in Chapter 5 are properties of the protocol design, not of any particular software implementation. A privacy risk that follows from MCP's session model, for example, is present in every system that uses MCP as specified, regardless of how carefully that system was built. This is a stronger claim than identifying an implementation flaw: it means the risk cannot be removed by fixing code, only by changing the architecture or adding controls that the protocol does not itself provide. This distinction between design-level and implementation-level risks is maintained throughout the analysis in Chapter 5 and is directly relevant to what types of remediation are possible.

4 Use Case and Data Flow Diagram

4.1. Introduction

This chapter translates the use case introduced in Chapter 2, as it is realised through the MCP and A2A protocols described in Chapter 3, into the formal inputs required for the LINDDUN privacy threat analysis in Chapter 5. The central output is a **Data Flow Diagram (DFD)**: a structured diagram that maps how personal data moves through the system, which components process it, where it is stored, and which organisational boundaries it crosses. In LINDDUN, the DFD is not merely an illustration; it is the analytical instrument through which privacy threats are systematically identified. Every privacy threat identified in Chapter 5 is traced to a specific element or boundary in the diagram produced here, making the accuracy and completeness of this chapter directly consequential for the rigour of the analysis that follows.

The chapter proceeds in five steps. Section 4.2 describes the use case in full. Section 4.3 enumerates all system components and trust boundaries that appear in the Data Flow Diagram. Section 4.4 defines the notation and conventions used in the diagram. Section 4.5 presents the full system Data Flow Diagram, which serves as the direct structural input to the LINDDUN mapping table in Chapter 5. Finally, Section 4.6 situates the system in its organisational context by analysing the actors whose roles, interests, and relationships shape its governance.

As established in Section 1.3, the use case is the analytical vehicle through which the privacy properties of MCP and A2A are examined, not the primary object of study. The findings produced in Chapters 5 and 6 are therefore primarily statements about MCP and A2A as protocols. Some threats are protocol-inherent, meaning they arise from specific design decisions in the protocol specifications and would appear in any use case where these protocols are deployed. Others are use-case-inherent, meaning they arise from the nature of automated fraud detection and would be present regardless of which communication protocol was used. This distinction is elaborated in Chapter 8.

4.2. Use Case Description

4.2.1. Selection Rationale

The use case was selected because it satisfies the three criteria set out in Section 2.3.2. It involves both MCP and A2A in active roles, realises all four analytically distinct protocol configurations simultaneously, and processes sensitive personal data across an organisational trust boundary. A full account of the selection reasoning is given in Chapter 2; this section focuses on the system description.

4.2.2. Scenario and Workflow

When a cardholder initiates a payment, Bank X processes the transaction and assesses its risk level. Bank X is the card-issuing bank: it holds the cardholder's account, maintains their transaction history and identity records, and is ultimately responsible for approving or declining the payment. Card Network Y is the card network operator: it routes payment messages between Bank X and the merchant's bank, and maintains a network-wide view of transaction patterns across many issuing banks. Because Bank X and Card Network Y each hold different but complementary data about the same cardholder and transaction, effective fraud detection requires both organisations to contribute to the risk assessment. Neither party alone has a complete picture.

When Bank X's systems detect signals associated with elevated fraud risk, such as an unusual transaction amount, an unfamiliar device, or a geographically inconsistent location, the automated fraud detection workflow is triggered. The cardholder is not notified at this point. From their perspective, the payment is simply pending. Meanwhile, a coordinated sequence of agent interactions runs automatically across both organisations to determine whether the transaction should be approved, declined, or **escalated** to a human analyst (meaning: forwarded to a person for a manual decision because the automated system cannot reach a conclusion with sufficient confidence).

Bank X's Fraud Detection Agent uses MCP to access data sources within Bank X's own organisation, querying internal systems including transaction history, card status, device fingerprint, and **KYC records**. KYC stands for Know Your Customer: the identity verification records that financial institutions are legally required to collect and maintain for every account holder, including name, address, date of birth, and identity document references. These records are gathered during **onboarding**, the process by which a new customer opens an account and their identity is verified. The agent simultaneously uses MCP to access a data source operated by a different organisation, querying an external

sanctions screening service, which checks whether the cardholder or their transaction counterparty appears on government-maintained AML lists.

The agent then initiates an A2A interaction with Card Network Y's Risk Intelligence Agent, a communication between agents of different organisations. In this exchange it transmits a **behavioural profile** (a structured summary of patterns derived from the cardholder's recent transaction activity, such as spending frequency, typical geographic locations, and device usage patterns) and receives a **risk score** in return. A risk score is a numerical or categorical value produced by the receiving agent that summarises its assessment of the likelihood that the transaction is fraudulent, based on the data available to that agent. Card Network Y's agent queries its own data sources through MCP within Card Network Y's organisation as part of this interaction.

Finally, if the combined risk score falls within a range that is too uncertain for an automated decision, Bank X's agent delegates the case to an internal review agent through A2A within Bank X's organisation, routing the case to a human analyst who makes the final decision. If the risk assessment is conclusive in either direction, the transaction is approved or declined automatically, without human involvement.

Throughout this workflow, the cardholder's personal data, including transaction records, device identifiers, behavioural patterns, and identity information, flows across multiple agent boundaries and between two independent organisations. The cardholder has no visibility into these flows and no means to influence them. This is the privacy context that motivates the analysis in this thesis.

All four protocol configurations are present in this single workflow, each introducing distinct privacy risks that will be analysed systematically in Chapter 5.

4.2.3. Modelling Simplifications

In a real payment system, the cardholder interacts not directly with the fraud detection agent but with a bank application layer, such as a mobile banking application, a web portal, or a point-of-sale terminal, which then forwards transaction data to the fraud detection system. This intermediary layer has been omitted from the DFD in this thesis. The reason is that the bank application layer does not involve MCP or A2A and therefore falls outside the protocol-level scope of this analysis. Including it would add a DFD element without analytical benefit. Data flow DF1 (Cardholder to Fraud Detection Agent) therefore represents the transaction data as it enters the fraud detection system, abstracting over the application layer through which it was collected.

4.2.4. Basis for Data Elements

The data elements used in this use case are not arbitrary: each is grounded either in the research literature on real-time payment fraud detection, in the operational data a card-issuing bank inherently maintains, or in a regulatory obligation. Transaction history and identity records are established standard inputs to production fraud detection; the device fingerprint is a recognised emerging signal; card status and authentication history are operational issuer data; and the external sanctions element corresponds to mandated anti-money-laundering screening. The full derivation of each element, with its sources and the data store in which it resides, is set out in Appendix G (Table G.1).

4.2.5. Validation Basis

The validation basis of the use case differs across its two layers. Its data foundations–the data elements and the flows between them–are grounded in the established literature on conventional payment fraud detection (Table G.1) and are therefore well supported independently of any agentic deployment. The agentic configuration that operates on those data is, by contrast, forward-looking: because production deployments of agentic AI for fraud detection are still emerging, this layer cannot be validated against existing systems, and its realism is instead assessed through expert validation (Chapter 7). The cross-organisational A2A configuration between Bank X and Card Network Y is the most forward-looking element: while A2A-based inter-agent collaboration between financial institutions is not yet in widespread production, the protocol is specifically designed for cross-organisational agent communication and the technical capability for such deployments exists [30]. The use case is therefore best understood as a near-term scenario reflecting the direction of current deployments rather than a description of a live system. In those interviews, the practitioners consulted, who have direct experience of agentic AI in the financial sector, recognised the scenario and its threats as plausible and realistic, and none judged the use case to be implausible.

The next section enumerates each system component, trust boundary, and data flow, producing the complete inventory that the LINDDUN analysis in Chapter 5 requires.

4.3. System Components and Trust Boundaries

This section enumerates the elements of the fraud detection system that appear in the Data Flow Diagram presented in Section 4.5. The enumeration follows the five standard DFD element types used in LINDDUN: processes, data stores, external entities, data flows, and trust boundaries [44]. Each element is identified by name and role so that the reader can trace every node and arrow in the diagram that follows. The section closes with a complete data flow inventory table that records the personal data types carried on each flow, which serves as the direct input for the LINDDUN mapping table in Chapter 5.

Processes

Processes are system components that receive, transform, or transmit data. Three processes are present in the use case. Each is an LLM-based AI agent with its own reasoning capabilities and tool access, connected to the rest of the system through MCP and A2A.

P1 **Bank X Fraud Detection Agent.** Receives transaction data, queries internal and external data sources via MCP, communicates with Card Network Y's agent via A2A, and produces a risk assessment. If the assessment is conclusive, it communicates the outcome directly to the merchant's bank. If the assessment is ambiguous, it delegates the case to the Internal Review Agent via A2A and subsequently forwards the analyst's final decision to the merchant's bank.

P2 **Bank X Internal Review Agent.** Receives escalated cases from the Fraud Detection Agent via A2A within Bank X, routes them to a human analyst for a final decision, and returns the analyst's decision to the Fraud Detection Agent via A2A.

P3 **Card Network Y Risk Intelligence Agent.** Receives a behavioural profile from Bank X's Fraud Detection Agent via A2A, queries Card Network Y's internal data sources via MCP, and returns a structured risk score.

Data Stores

Data stores are repositories where data is held, either persistently or temporarily. Five data stores are present in the use case.

DS1 **Bank X Transaction History Database.** Holds the cardholder's prior transaction records, including amounts, timestamps, merchant categories, and geographic locations.

DS2 **Bank X KYC and Identity Records.** Holds cardholder identity information collected during onboarding. KYC (Know Your Customer) records are the identity verification documents that financial institutions are legally required to collect and maintain for every account holder, including name, address, date of birth, and identity document references.

DS3 **Bank X Card Status and Device Fingerprint Store.** Holds real-time card state information and device-level identifiers, including device fingerprints and authentication history.

DS4 **Card Network Y Transaction Pattern Database.** Holds network-wide transaction data aggregated across many issuing banks, used to derive **behavioural baselines** (statistical reference points for what normal transaction activity looks like across the network, against which individual transactions are compared) and detect cross-bank fraud patterns. **Transaction velocity** refers to the rate at which transactions occur over a short period: an unusually high number of transactions in a brief window is a common fraud signal.

DS5 **External Sanctions Screening Database.** Holds AML sanctions lists and related identity risk indicators, operated by a third-party organisation that resides outside both TB1 and TB2.

External Entities

External entities are actors outside the system boundary that initiate or receive data but whose internal workings are outside the scope of the analysis. Four external entities are present in the use case.

EE1 **Cardholder.** Initiates the payment transaction and ultimately receives an approval or decline decision. The cardholder has no visibility into the agent interactions that take place during the risk assessment. The cardholder resides outside both TB1 and TB2.

EE2 **Merchant's Bank.** Receives the final transaction outcome from Bank X's Fraud Detection Agent and processes the payment accordingly. The merchant's bank resides outside both TB1 and TB2.

EE3 **Human Fraud Analyst (Bank X).** Receives escalated cases from the Internal Review Agent and issues the final decision in ambiguous cases where the automated risk assessment is inconclusive. Although the analyst operates within Bank X's organisational context, human actors are modelled as external entities in LINDDUN because they are treated as outside the automated system boundary.

EE4 **External Sanctions Screening Service.** Receives screening queries from Bank X's Fraud Detection Agent via MCP and returns results. This entity is also represented as DS5 above, reflecting its dual role as both an organisational actor and a data source. It resides outside both TB1 and TB2, as it is operated by a third party independent of both Bank X and Card Network Y.

Trust Boundaries

A **trust boundary** in a DFD marks the line between parts of a system that operate under different levels of organisational control and accountability. Within a trust boundary, all components are operated by the same organisation and subject to the same governance, security policies, and data protection obligations. A data flow that crosses a trust boundary moves from one organisational sphere of control into another, where different rules may apply and where the sending party cannot enforce how the data is handled once it has been received. In this thesis, trust boundaries correspond directly to organisational boundaries: each boundary encloses the systems of one organisation. Two trust boundaries are present in the use case.

TB1 **Bank X trust boundary.** Encloses all components operated and controlled by Bank X: P1, P2, DS1, DS2, and DS3. Data flows that cross TB1 either originate from outside the organisation, such as the payment data sent by the cardholder, or are directed to components outside Bank X's control, such as the sanctions screening service or Card Network Y's agent. The human fraud analyst (EE3) is placed outside TB1 in the DFD, consistent with the LINDDUN convention that human actors are modelled as external entities.

TB2 **Card Network Y trust boundary.** Encloses all components operated and controlled by Card Network Y: P3 and DS4. Data flows that cross TB2 involve the exchange of derived personal data between Bank X and Card Network Y. This boundary is the most privacy-sensitive in the use case, as it involves the transmission of a behavioural profile from Bank X and a risk score from Card Network Y, both carrying personal data processed under the separate governance structures of two independent financial institutions. The cardholder (EE1), the merchant's bank (EE2), and the external sanctions screening service (EE4/DS5) all reside outside both TB1 and TB2.

Data Flow Inventory

Table 4.1 provides a complete inventory of all data flows in the use case. Each flow is assigned a unique identifier, a source and destination drawn from the elements enumerated above, the protocol used where applicable, and the specific personal data types transmitted.

Basis for personal data types. The personal data listed for each flow was determined by combining three sources. First, the data store contents defined above establish what each store holds and therefore what a query to that store can return. Second, the protocol specifications for MCP and A2A (described in Chapter 3) establish the structural form of each flow: MCP tool call parameters carry whatever the invoking agent passes as arguments, and MCP results carry whatever the server returns from the underlying data source; A2A task messages carry whatever the sending agent includes in the task message or DataPart. Third, the use case description above establishes the purpose of each flow and therefore which data elements are necessary for that purpose. Together these three sources fully determine what personal data each flow carries. This derivation method follows the standard LINDDUN practice of reading data flows from the combination of the system model, the underlying data stores, and the protocol specifications [44]. Flows where the protocol column shows a dash (DF1, DF15, DF16, DF18) do not use MCP or A2A; they represent interactions at the boundary of the automated system, such as the cardholder's initial payment trigger or the analyst's decision, and are included to provide a complete picture of the system's data flows.

Table 4.1: Data flow inventory with personal data types. Derived from the use case description, the data store contents defined in Section 4.3, and the MCP and A2A protocol specifications [55, 30].

ID	Source	Destination	Protocol	Personal Data Transmitted
DF1	EE1 Cardholder	P1 FDA	(	Transaction amount, merchant details, device identifier, geographic location
DF2	P1 FDA	DS1 Tx History	MCP	Transaction history query parameters
DF3	DS1 Tx History	P1 FDA	MCP	Prior transaction records: amounts, timestamps, merchant categories, geographic locations
DF4	P1 FDA	DS2 KYC Records	MCP	Identity query parameters
DF5	DS2 KYC Records	P1 FDA	MCP	Name, address, date of birth, identity document references
DF6	P1 FDA	DS3 Card/Device	MCP	Card status and device fingerprint query parameters
DF7	DS3 Card/Device	P1 FDA	MCP	Card state, device fingerprint, authentication history
DF8	P1 FDA	EE4/DS5 Sanctions	MCP	Name, date of birth, identity attributes for sanctions screening
DF9	EE4/DS5 Sanctions	P1 FDA	MCP	Sanctions screening result and risk indicators
DF10	P1 FDA	P3 CNYA	A2A	Derived behavioural profile: transaction velocity, geographic patterns, device consistency indicators
DF11	P3 CNYA	P1 FDA	A2A	Structured risk score and supporting network-level indicators
DF12	P3 CNYA	DS4 CNY Patterns	MCP	Network pattern query parameters
DF13	DS4 CNY Patterns	P3 CNYA	MCP	Aggregated cross-bank transaction pattern data
DF14	P1 FDA	P2 IRA	A2A	Full risk assessment context, transaction details, and escalation reason
DF15	P2 IRA	EE3 Analyst	)	Escalated case details for human review
DF16	EE3 Analyst	P2 IRA	,	Final approve or decline decision
DF17	P2 IRA	P1 FDA	A2A	Final decision returned to Fraud Detection Agent for communication to merchant's bank
DF18	P1 FDA	EE2 Merchant Bank	,	Transaction outcome: approved or declined

MCP intra-org flows Identity data flows A2A flows Unshaded: boundary flows (no MCP/A2A).

FDA = Fraud Detection Agent; IRA = Internal Review Agent; CNYA = Card Network Y Risk Intelligence Agent; Tx = Transaction; CNY = Card Network Y.

The most privacy-sensitive flows are those that cross TB1 or TB2, because these carry personal data out of one organisation's control and into another's; they are the focus of the LINDDUN threat elicitation in Chapter 5. The next section defines the notation used to render these elements as a Data Flow Diagram.

4.4. DFD Notation and Conventions

This section describes the symbols, labelling conventions, and simplifications used in the Data Flow Diagram presented in Section 4.5. The notation follows the standard DFD symbology used in LINDDUN [44], which is directly derived from the original DFD notation introduced by DeMarco [62] and further formalised by Yourdon and Constantine [63].

Table 4.2 summarises the five DFD element types and their graphical representations used throughout the diagrams in this chapter. Figure 4.1 shows the symbols themselves.

Table 4.2: DFD element types and their symbols, following the notation of DeMarco [62]. Author's own illustration.

Element type	Symbol	Description
External entity	Rectangle	Actors outside the system boundary that initiate or receive data flows but whose internal workings are outside the scope of the analysis.
Process	Circle	Components that receive, transform, or transmit data. Labelled with a number and a descriptive name.
Data store	Double horizontal lines (open ends)	Repositories where data is held. Labelled with the letter D followed by a number and a descriptive name.
Data flow	Directed arrow	The connection along which data moves. Labelled with the flow identifier and the personal data transmitted. Bidirectional exchanges are shown as two separate arrows.
Trust boundary	Dashed rectangle	Encloses elements sharing the same organisational trust level. Labelled with the trust boundary identifier.

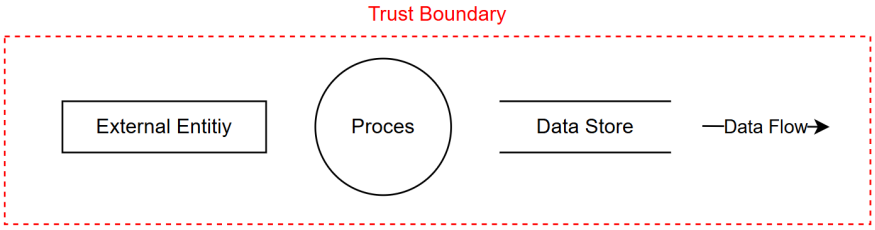

Figure 4.1: Visual representation of the DFD element type symbols used in this chapter, following the notation of DeMarco [62]. Author's own illustration.

Trust boundary representation. The two trust boundaries identified in Section 4.3 are represented consistently across all diagrams. TB1 encloses the internal systems of Bank X and TB2 encloses the internal systems of Card Network Y. The cardholder, the merchant's bank, and the external sanctions screening service reside outside both zones. Where a data flow crosses a trust boundary, this is made visually explicit by the arrow passing through the dashed boundary line.

Protocol labelling. Each data flow arrow is annotated with the protocol used to carry it, either MCP or A2A, in square brackets. This annotation is included because the protocol used on a given flow is central to the privacy analysis: the same data transmitted via MCP and via A2A carries different structural privacy implications, as established in Chapter 3. Where a flow does not use either protocol, for example the flows between the Internal Review Agent and the human analyst, no protocol annotation is included, consistent with the identifier assignments in Table 4.1.

Scope and simplifications. Section 4.5.1 presents a full system overview diagram showing all elements and flows simultaneously. This single diagram, combined with the data flow inventory in Table 4.1, provides the complete structural input for the LINDDUN threat elicitation in Chapter 5. The four protocol configurations present in the use case are distinguished analytically through the flow groupings introduced in Chapter 5 rather than through separate diagrams, which avoids visual redundancy without reducing analytical coverage.

4.5. Data Flow Diagrams

4.5.1. Full System Overview

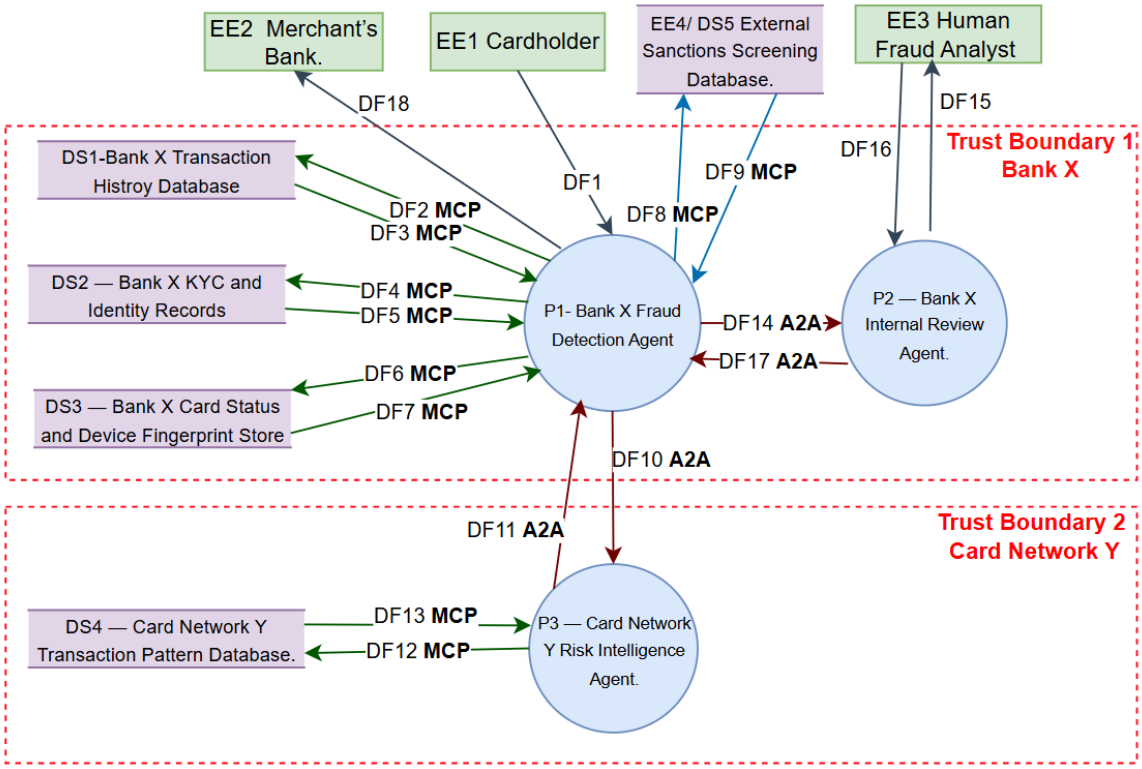

Figure 4.2: Full system Data Flow Diagram of the hypothetical financial fraud detection use case, showing all processes, data stores, external entities, and data flows across trust boundaries TB1 (Bank X) and TB2 (Card Network Y). Author's own illustration.

4.6. Actor Analysis

The system components are operated by, and affect, seven actors with distinct and partly conflicting interests: Bank X, Card Network Y, the cardholder, the human fraud analyst, the external sanctions service, the merchant's bank, and the financial regulator. From the CoSEM perspective, reading these actors alongside the Data Flow Diagram makes visible the governance structures and power asymmetries that the technical elements realise. The full actor mapping–role and position, primary interest, and key conflicts for each–is set out in Appendix G (Table G.2) and revisited in Chapter 8.

Structural observations. This actor configuration produces two structural asymmetries that are directly relevant to the privacy analysis that follows. First, the cardholder is the only actor with no technical representation in the system: every other actor has agents, data stores, access rights, or regulatory authority within the architecture, while the cardholder is an external entity with no interface through which to observe or contest the processing. Second, governance responsibility is distributed across Bank X and Card Network Y, which operate under separate regulatory frameworks and have no shared oversight mechanism. Privacy controls that are technically feasible within TB1 cannot be enforced at Card Network Y, and vice versa. The trust boundary TB2 is therefore not only a technical boundary but also an organisational and governance boundary whose crossing has direct consequences for the cardholder's practical recourse.

4.7. Chapter Summary

This chapter has produced the formal artefacts required for the LINDDUN threat elicitation in Chapter 5. The use case was described in full, and the system was enumerated into three processes (P1–P3), five data stores (DS1–DS5), four external entities (EE1–EE4), and two trust boundaries (TB1 for Bank X and TB2 for Card Network Y). All eighteen data flows were documented with their personal data types in Table 4.1. The full system Data Flow Diagram in Figure 4.2 brings all elements and flows together as the direct structural input to the LINDDUN mapping table applied in Chapter 5. There, each data flow is assessed against six LINDDUN threat categories across the four protocol configurations present in the

use case. The actor analysis in Section 4.6 then situated this technical architecture in its organisational and governance context, identifying seven actors: Bank X, Card Network Y, the cardholder, the human fraud analyst, the financial regulator, the external sanctions screening service, and the merchant's bank. Two structural asymmetries follow directly from this configuration: the cardholder is the only actor with no technical representation inside either trust boundary, and governance responsibility is distributed across two independent organisations with no shared oversight mechanism.

This chapter, together with Chapter 3, provides the answer to the first sub-question: *how do MCP and A2A function within the use case, and how can the system be represented as a DFD?* Chapter 3 established how MCP and A2A function technically, covering their architectures, session and task models, and trust assumptions. The present chapter translated that technical understanding into a formal DFD with all five LINDDUN element types, and situated the system in its organisational and actor context, providing the structural representation required for the threat elicitation in Phase 2.

5 Privacy Threat Identification

5.1. Introduction

This chapter tells the story of what happens to the cardholder's privacy when their payment is processed through the fraud detection system described in Chapter 4. It follows the data as it moves through the workflow, step by step, from the moment the payment is initiated to the moment the final decision is sent to the merchant's bank. At each step, the chapter identifies the privacy threats that arise from the way the system is designed and from the specific properties of MCP and A2A.

The analysis is based on the LINDDUN Pro privacy threat modelling methodology [64], applied systematically to all eighteen data flows in the use case. Six of the seven LINDDUN categories were assessed: Linkability, Identifiability, Non-repudiation, Detectability, Data Disclosure, and Unawareness and Unintervenability. Non-compliance is excluded because its analysis requires a legal methodology, not a technical one, as established in Section 2.3.1.

The full technical elicitation, including all threat node codes (the identifiers LINDDUN's threat trees assign to each specific threat condition), position-level tables, and the consolidated threat overview, is in Appendix D. This chapter presents the findings as a readable account, organised around the four protocol configurations that correspond directly to the research question: MCP within an organisation, MCP between organisations, A2A within an organisation, and A2A between organisations. Three additional groups cover flows that do not use MCP or A2A directly.

Each section ends with an explicit analytical conclusion about what that configuration reveals about the protocols. Section 5.8 identifies four patterns that emerge across all configurations. Section 5.9 provides the chapter summary.

Note on abbreviations. The analysis uses the system component labels introduced in Chapter 4 (P1, P2, P3 for the three agents; DS1–DS5 for the data stores; EE1–EE4 for the external actors; TB1 and TB2 for the two trust boundaries; DF1–DF18 for the data flows). These labels are defined in full in Table 4.1 and Figure 4.2. Readers unfamiliar with the system should consult Chapter 4 before reading this chapter.

5.2. Analytical Setup

How the eighteen flows were organised into nine groups

The use case contains eighteen data flows. Analysing each one strictly in isolation would produce repetition without additional insight, because many flows share the same structural properties: the same protocol, the same type of elements involved, and the same relationship to the organisational trust boundary. Flows that share all three of these structural dimensions are therefore grouped and analysed together. Where flows within a group carry meaningfully different personal data, those differences are highlighted within the group discussion.

Nine flow groups result from this process. Their assignment to the four protocol configurations and the corresponding chapter sections is given in Table 5.1. Three of the groups (H, G, and I) fall outside the protocol configurations entirely: the cardholder's payment initiation, the human analyst interaction, and the final payment response.

This grouping was discussed at a high level with a co-author of the LINDDUN framework, who identified no inconsistencies in the approach [48]. The full position-level elicitation for each group is in Appendix D.

Table 5.1 shows how the groups map to protocol configurations and chapter sections.

Table 5.1: Assignment of flow groups to protocol configurations and chapter sections. Author's own elaboration.

Protocol configuration	Groups	Section
MCP within an organisation	A, B	5.3
MCP between organisations	C, D	5.4
A2A within an organisation	F	5.5
A2A between organisations	E	5.6
Boundary and human actor flows	H, G, I	5.7

Scope note on the LINDDUN knowledge base

The elicitation throughout this chapter uses the standard, publicly established LINDDUN Pro threat trees. A recently proposed GenAI-specific extension of LINDDUN [65] could not be applied, because its threat trees have not been released publicly; this is a deliberate scope decision whose implications are discussed in Section 9.2.

5.3. MCP Within an Organisation

The fraud detection workflow begins the moment a payment triggers an alert inside Bank X. The Fraud Detection Agent (P1) immediately starts gathering everything it needs to assess the risk. Using MCP, it queries three of Bank X's internal databases in rapid succession: the transaction history database (DS1), the identity records database (DS2), and the card status and device fingerprint database (DS3). Across the boundary at Card Network Y, the Risk Intelligence Agent (P3) does the same with the network-wide pattern database (DS4). This entire retrieval phase, eight data flows in total, takes place inside the respective organisations' trust boundaries. No data leaves either organisation at this stage.

The queries: what goes out

When P1 queries DS1, DS2, and DS3, each query carries a small identifying reference: a card number, an account identifier, or an identity key. These are compact messages, not large data transfers. But all four queries from P1 share the same MCP session identifier. This is not a design choice that the developer makes or avoids: it is a structural property of the MCP protocol. Every tool call issued within a single MCP session carries the same identifier.

The consequence is that anyone with access to the session logs can see the full sequence: the same agent, in the same session, queried the transaction database, then the identity database, then the device database. Even without reading the contents of any query, this pattern reveals that a specific cardholder is under assessment. The session identifier makes separate, individually innocuous queries into a linked chain that is attributable to one person. This is what LINDDUN calls **Linkability**: the ability to connect separate interactions back to the same individual without needing to know their name.

One of the four queries, DF4 from P1 to DS2, carries a direct identity reference sufficient to retrieve the cardholder's full KYC record. The other three carry card or account references that could be combined with other data to identify the person. All four queries are fully automated and completely invisible to the cardholder. They have no way of knowing that their identity records, transaction history, and device fingerprint are being accessed simultaneously.

The responses: what comes back

The databases respond by sending back the full records. DS2 returns name, address, and date of birth. DS1 returns the complete transaction history. DS3 returns card status and device fingerprint. DS4 returns network-wide pattern data.

The shift from queries to responses is the most privacy-significant transition in this phase of the workflow. The queries carried small reference keys. The responses carry the actual personal data those keys unlock. More importantly, P1 now holds all three Bank X records simultaneously, within the same MCP session. A combined profile of the cardholder, linking their identity, their spending behaviour, and their device, has been assembled. This profile does not exist anywhere in the system as a single stored record. It was never created by a human decision. It was produced automatically, as an inevitable consequence of how MCP's session model works.

Returning the full KYC record for an identity verification also raises a separate concern: the response returns name, address, and date of birth in full, when the fraud assessment might only require a partial confirmation. The protocol provides no mechanism to request or enforce a minimised response. What the data store returns is what the agent receives.

What this reveals about MCP

The intra-organisational MCP configuration establishes the most important protocol-level finding in this analysis: MCP's session model makes cross-store linkability a structural consequence of using the protocol, not an incidental risk that careful implementation can avoid. An architect who implements MCP with full attention to security still produces this linkability, because it follows from the specification rather than from implementation quality. Reducing it requires architectural choices, such as scoping or segmenting sessions so that fewer interactions share an identifier, or replacing MCP with a stateless query pattern that removes the shared session entirely.

The return flows also reveal an asymmetry in MCP's request-response model: the response phase activates substantially more privacy risk than the query phase, because it carries the actual personal data rather than a pointer to it. Privacy analysis that focuses only on what an agent sends, and not on what it receives, will systematically underestimate this risk.

The full technical elicitation for Groups A and B, including all threat node codes and position-level tables, is in Appendix D, Sections D.2 and D.3.

5.4. MCP Between Organisations

Before escalating to Card Network Y, P1 runs one more check: it queries an external sanctions screening service (EE4/DS5) to verify whether the cardholder appears on any anti-money laundering list. This service is registered as an MCP-compatible tool, so P1 invokes it exactly as it would invoke an internal data store. From the protocol's perspective, querying an external third-party service is structurally identical to querying an internal database. The difference, which the protocol does not express, is that the data now crosses Bank X's organisational boundary (TB1).

Sending the cardholder's identity outside Bank X

To run the sanctions check, P1 sends the cardholder's name, date of birth, and nationality to EE4/DS5 as an MCP tool call (DF8). This is the first moment in the workflow when directly identifying personal data leaves Bank X's organisational control.

Once this data crosses TB1, Bank X loses the ability to enforce any rules about how it is used. The external service may retain it, log it, or share it further. Bank X cannot compel them to delete it. The MCP session identifier that linked the internal queries also travels with this external call, meaning an observer with access to the session log can see the entire chain: internal data retrieval followed by external disclosure of the same person's identity, in one attributable session. The cardholder has no idea that their name and date of birth have been transmitted to a third-party organisation.

A further risk arises if the transmission is unencrypted: the payload would then be visible to any party on the network path between Bank X and the external service. MCP recommends but does not mandate encryption for HTTP-based transport.

The result that comes back

The screening service returns its outcome to P1 (DF9). The response does not contain the cardholder's name. It contains only a result, for example a match indicator. But this result arrives back in the same MCP session that already holds the cardholder's full identity, transaction history, and device data. The session context makes the result fully attributable to a specific individual, even though the response itself carries no direct identifier.

A positive match result is particularly sensitive. It links the cardholder to a sanctions database entry. This is an association with serious legal and reputational consequences. The external service retains a log of the query and the response. Bank X cannot access or delete that log.

An observer can also infer the outcome without reading the content, simply by noting whether particular response fields are present or absent. The shape of a response can be as revealing as its contents.

What this reveals about MCP

The cross-boundary MCP configuration reveals that the organisational boundary amplifies every privacy threat identified in the previous section. The same session model that creates linkability inside Bank X now extends that linkability into a different governance regime, and neither party can fully control the evidence record that results. The privacy harm is not primarily about the format of the data: a sanctions screening result with no direct identifier is still fully attributable to a specific person because the session context ties it to everything assembled earlier. Reducing the content of a cross-boundary MCP flow does not eliminate its identifiability risk if the session travels with it.

The full technical elicitation for Groups C and D is in Appendix D, Sections D.4 and D.5.

5.5. A2A Within an Organisation

When the automated risk assessment produces an ambiguous result, P1 escalates the case to the Internal Review Agent (P2) inside Bank X. This escalation uses A2A. P1 packages everything it has assembled and sends it to P2 as an A2A task submission (DF14). P2 reviews the case and returns a decision (DF17). The entire interaction takes place inside TB1.

How A2A transmits a complete context

A2A is designed for communication between autonomous agents. Its task model is built on a simple principle: the sending agent packages everything the receiving agent might need to reason independently and reason well. This is what makes A2A effective for coordination. It is also what makes it privacy-relevant.

DF14 carries everything P1 has assembled during the current session: the transaction details, the name and address retrieved from DS2, the device fingerprint from DS3, and the outcome of the sanctions screening. This is the most data-rich single message in the entire workflow. It is also substantially more than P2 strictly needs. P2 needs to know whether the risk indicators are significant enough to warrant a decline decision. It does not need the underlying raw records from every database P1 consulted.

A2A provides no mechanism to prevent this over-transmission. Unlike a database query, where the developer defines exactly which fields to retrieve, an A2A task submission passes whatever the sending agent includes. There is no protocol-level prompt, check, or constraint that limits the payload to what the receiving agent actually requires. The developer must actively decide to minimise the task context, and nothing in the protocol encourages or enforces that decision.

What this reveals about A2A

The intra-organisational A2A configuration reveals that A2A introduces a fundamentally different type of privacy risk from MCP. MCP's risk arises from what the protocol structurally exposes: by default, a shared session identifier links the tool calls within a session, and that linkage is not removed by careful implementation alone. A2A's primary risk arises from what the protocol structurally permits: the full context payload is not required, but nothing prevents it. MCP creates a privacy problem that is structural by default. A2A creates a privacy problem that is avoidable in principle but not prevented in practice. Both require attention, but they require different types of response.

Privacy-aware deployment of A2A requires a design discipline at the payload level: an explicit decision to restrict what is included in the task context. MCP requires intervention at a different layer. Because the linkage follows from the session model rather than from payload content, mitigating it calls for architectural choices, such as scoping or segmenting sessions so that fewer interactions share an identifier, or replacing MCP with a stateless query pattern, rather than for payload-level restraint.

The full technical elicitation for Group F is in Appendix D, Section D.6.

5.6. A2A Between Organisations

Before escalating internally to P2, P1 first reaches out to Card Network Y. It sends a derived behavioural profile of the cardholder to P3 at Card Network Y via A2A (DF10) and receives a risk score in return (DF11). This is the cross-organisational A2A exchange, and it is structurally different from every other configuration in the analysis.

Sending a profile without a name

The profile that P1 sends does not contain the cardholder's name, account number, or date of birth. It contains derived behavioural attributes: how many transactions the cardholder made recently, in which geographic area, and whether the device used was consistent with their history. It looks, at first glance,

like anonymous data.

It is not. Card Network Y's database (DS4) holds transaction records from many issuing banks across the network. When P3 receives the profile and processes it against DS4, it is comparing a pattern against a library of millions of patterns from known accounts. A profile stating "three transactions in Rotterdam on a Tuesday afternoon, average value approximately 180, consistent device" may not contain a name, but if DS4 holds records from earlier interactions involving the same device and the same geographic pattern, P3 can match the profile to a specific individual.

The re-identification risk does not reside in what Bank X sends. It resides in the combination of what Bank X sends and what Card Network Y already knows. An organisation inspecting DF10 in isolation would not classify it as directly identifying. But that assessment is incomplete. The receiving agent's data holdings determine identifiability, not the format of the transmitted data. This is a finding with no equivalent in the MCP analysis.

Bilateral evidence of the exchange

Once both parties have exchanged A2A messages, both hold an independent record of the interaction. Bank X holds a log of the task it sent and the response it received. Card Network Y holds a log of the task it received and the response it produced. Neither organisation can remove the other's copy. If a dispute arises about what was shared, each party can produce their own unilateral record. The cardholder has no knowledge that a derived profile about them was sent to a separate financial institution.

What this reveals about A2A

The cross-organisational A2A configuration reveals a finding that challenges a common assumption in privacy engineering: that pseudonymisation, replacing direct identifiers with derived attributes, provides adequate protection in cross-organisational data sharing. In the context of A2A between large financial institutions, it does not. The receiving agent may hold exactly the data needed to reverse the pseudonymisation, and the A2A protocol provides no mechanism to verify or constrain this before the transmission occurs.

The implication is that privacy impact assessments for cross-organisational A2A flows must account for the knowledge state of the receiving agent. This requires organisations to understand what data their counterpart holds before sharing derived profiles, and to implement contractual or technical controls on what the receiving party does with the data after receipt, because the sending organisation cannot control the correlation process once the boundary has been crossed.

The full technical elicitation for Group E is in Appendix D, Section D.7.

5.7. Boundary and Human Actor Flows

Three flow groups fall outside the four protocol configurations. They do not use MCP or A2A, but they are not peripheral. Two of them, the payment initiation and the human analyst interaction, are the entry and exit points of the human element in an otherwise automated system. The third is the terminal payment response.

The cardholder initiates a payment

When the cardholder taps their card or confirms a payment online, they send a message that carries the transaction amount, merchant details, a device identifier, and their geographic location (DF1). This is the only moment in the entire workflow where the cardholder actively participates. From this point forward, everything happens without their knowledge or involvement.

The device identifier is a persistent technical fingerprint of the device used to make the payment. It can be cross-referenced against the device records in DS3 to link the transaction to the cardholder's identity without needing their name. When geographic location is sufficiently precise, it too can function as an identifying attribute. The automated assessment chain that follows this initial payment is completely invisible to the cardholder. They have no idea that their identity records, transaction history, and behavioural profile are about to be accessed, combined, and shared across two organisations.

The full technical elicitation for Group H is in Appendix D, Section D.8.

A human analyst reviews the full case

When the automated system cannot reach a confident decision, it escalates the case to EE3, the human

fraud analyst at Bank X. P2 sends the analyst a complete case dossier (DF15): a fully integrated record combining transaction history, identity attributes, device data, and the sanctions screening outcome. The analyst reads it, makes a decision, and sends the decision back (DF16).

This group produces the broadest set of privacy threats in the entire analysis. It also involves no automated protocol. The reason is the nature of the recipient. An automated agent processes data within its defined parameters and produces an output. A human analyst does something qualitatively different: they can retain the information mentally or in notes, discuss it with a colleague, reference it in a future informal conversation, or draw conclusions that go beyond the formal decision. None of these informal possibilities are captured by the system's data governance rules. The cardholder has no idea that a specific named human being is reading their complete financial and identity profile. There is no notification, no right to object, and no mechanism to request that the case be handled differently.

The analyst interaction also carries the highest Non-repudiation weight in the workflow, because a human action is legally attributable in ways that an automated protocol step is not. If the analyst makes a decision on the basis of the dossier, that decision and the underlying information are permanently associated with a named individual in a way that carries evidentiary significance.

The full technical elicitation for Group G is in Appendix D, Section D.9.

The final decision is sent to the merchant's bank

The last step is the simplest. P1 sends an approval or decline decision to the merchant's bank (EE2, DF18), with a transaction reference. This is the most privacy-minimised flow in the workflow: the payload contains only what is functionally required, with no personal attributes beyond the transaction reference.

The cardholder receives the decision outcome at their device or the merchant's terminal. They do not receive any explanation of the process that produced it, any indication that their data was shared with Card Network Y, or any notification that a human analyst may have reviewed their complete profile. The decision arrives as a fact, without context.

The full technical elicitation for Group I is in Appendix D, Section D.10.

5.8. Cross-configuration Patterns

Comparing the threat findings across all nine flow groups produces four structural patterns. Before presenting them, one observation about the comparison itself is important. MCP and A2A are not equivalent protocols doing the same job. MCP is an agent-to-tool protocol: it governs how an agent accesses a data source. A2A is an agent-to-agent protocol: it governs how one autonomous agent delegates work to another. Comparing their threat profiles therefore does not show which protocol is more dangerous. It shows that they introduce different types of threat that require different responses.

Pattern 1: The cardholder is unaware at every single step. In LINDDUN, this concern spans two distinct conditions: *Unawareness* is a gap in the cardholder's *knowledge* (they do not know what is happening to their data, which data is involved, or who can access it) while *Unintervenability* is a gap in their *agency*: even if aware, they have no mechanism to access, contest, correct, or halt the processing. There is one privacy threat that appears across every flow group, every protocol configuration, and every position in the analysis without a single exception: the cardholder does not know what is happening to their data. This is not a risk that careful implementation can eliminate. Neither MCP nor A2A contains any mechanism for notifying the person whose data is being processed, for requesting their consent, or for giving them any ability to see or influence what is happening. The cardholder is structurally excluded from the system that processes their most sensitive personal information.

This makes Unawareness the most significant finding in the analysis. All other threats describe specific harms: a profile being assembled, an identity being disclosed, a record becoming permanent. Unawareness describes the condition under which none of those harms can ever be detected or challenged by the person most affected. A cardholder who does not know that their behavioural profile was shared with Card Network Y cannot object. A cardholder who has no access or contestation mechanism cannot correct an erroneous sanctions match, cannot request deletion of a retained profile, and cannot seek redress if re-identification occurs.

Pattern 2: Crossing an organisational boundary amplifies every threat. Comparing the intra-organisational MCP groups with the cross-boundary MCP groups, and the intra-organisational A2A group with the cross-organisational A2A group, the same shift is visible in both protocols: as soon as data crosses the organisational boundary, additional threats are activated across every LINDDUN

category. The boundary, not the choice of protocol, is the primary amplifier of threat exposure. An organisation deploying MCP or A2A purely within its own systems faces a materially different risk profile from one that uses either protocol to communicate with external parties. This difference is structural, not a matter of configuration or careful implementation.

Pattern 3: MCP and A2A create different types of threat that need different responses. Within the same intra-organisational boundary context, MCP's Linkability threat is the structural default: it follows directly from the session model and is not removed by careful implementation alone. Reducing it requires deliberate architectural choices, such as scoping or segmenting MCP sessions, or replacing MCP with a stateless query pattern. A2A's Data Disclosure risk is different: it arises from what the protocol permits, not from what it enforces. A developer who actively restricts the task payload can reduce it, although nothing in the protocol guides or rewards this discipline. MCP therefore requires architectural-level remediation, including session scoping; A2A requires design-discipline remediation at the payload level. Treating them as equivalent problems leads to the wrong solutions.

Pattern 4: The human analyst is a privacy risk, not just a safeguard. Group G, the human analyst interaction, produces a threat profile as broad as the cross-boundary groups, yet involves no automated protocol. The human-in-the-loop is often assumed to improve privacy by introducing oversight and judgment. In this system, it does the opposite: it introduces the widest-scope access to the most sensitive aggregated data, through a channel that falls partially outside the system's automated governance structures. Calibrating the scope of human access is therefore not a secondary concern. It belongs in the same analytical tier as decisions about cross-boundary data transmission.

5.9. Chapter Summary

This chapter followed the cardholder's data through the entire fraud detection workflow and identified the privacy threats that arise at each step, applying LINDDUN Pro to all eighteen data flows across the four protocol configurations and the boundary and human actor flows. The full technical evidence behind these findings is collected in Appendix D. Chapter 6 takes the fifty-four threat entries produced by that analysis and selects the eight most significant for detailed description and expert validation.

Comparing the configurations surfaced the four cross-cutting patterns set out in Section 5.8: the cardholder's structural unawareness at every step, the amplification of every threat once data crosses an organisational boundary, the distinct and non-interchangeable risks introduced by MCP and A2A, and the human analyst as the broadest-scope access point in the workflow.

A fifth observation connects these findings across protocols, and is visible only because both protocols were analysed within the same system model. The cross-store profile that P1 assembles through MCP within Bank X is precisely the profile subsequently transmitted to Card Network Y via A2A: the MCP aggregation finding is the direct input for the A2A re-identification finding. An analysis of either protocol in isolation would miss this causal chain.

6 Threat Prioritisation and Descriptions

6.1. Introduction

Chapter 5 identified fifty-four privacy threat entries across the four protocol configurations and the boundary and human actor flows. Fifty-four entries is too many to describe in equal depth, and not all of them are equally significant. This chapter works through a structured reduction: from fifty-four entries to a prioritised register, from the register to a set of cross-configuration patterns, and from those patterns to eight threats selected for full description. The eight descriptions are the primary output of this chapter and the direct input to the expert interviews in Chapter 7.

An important framing note applies to all eight threats. They are privacy threats identified across the use case as a whole, meaning they arise from the complete system of data flows, components, and trust boundaries described in Chapter 4. They are not all equally specific to MCP or A2A as protocols. Some, most notably TH-01, TH-02, TH-04, and TH-06, arise directly from structural properties of the protocol specifications and would not exist if a different communication architecture were used. Others, most notably TH-07 and TH-08, are properties of the use case itself and would be present in any automated fraud detection system regardless of which protocol it used. TH-03 and TH-05 sit between these poles: MCP and A2A enable or amplify them by default, but do not uniquely cause them. All eight are valid findings of a LINDDUN analysis of this system; the distinction between protocol-inherent and use-case-inherent threats is discussed in depth in Section 8.1. Within this chapter, the selection criteria applied in Section 6.5 favour protocol-specific threats where possible, so that the descriptions are maximally relevant to the research question.

The chapter proceeds in four steps. Section 6.2 establishes the prioritisation methodology, defining what likelihood and impact mean in this context and how they are combined into four priority levels. Section 6.3 applies this methodology to all fifty-four entries, producing the full prioritised register. Section 6.4 reads across the register to identify structural patterns: which configurations produce the most severe threats, which LINDDUN categories are most consistently elevated, and what the register reveals about the relationship between protocol choice and privacy risk. Section 6.5 then uses these patterns to explain which threats are selected for full description and why. Section 6.6 presents the eight descriptions. Section 6.7 closes the chapter.

6.2. Prioritisation Methodology

Qualitative over quantitative

This thesis applies a qualitative prioritisation approach, consistent with the LINDDUN PRO tutorial's recommendation [64]. A quantitative approach was considered but rejected. The use case is a hypothetical scenario: precise numerical inputs such as occurrence frequencies cannot be reliably estimated from structural analysis alone. Stating that a specific threat will occur in 97% of deployments would create false precision. A qualitative ordinal scale accurately reflects the level of confidence that structural analysis provides without overstating it.

A methodological consequence of qualitative scoring is that the scores reflect the analytical judgement of the researcher rather than empirically measured values. This introduces a validity concern: the scores could be influenced by prior knowledge of which threats the analysis is expected to find significant. This concern is addressed in two ways. First, the scoring criteria for likelihood and impact are defined in advance and applied consistently across all fifty-four entries, so that the basis for each score is explicit and auditable. Second, the expert interviews in Chapter 7 serve specifically to validate these scores against the practical experience of domain experts. Confirmations, revisions, or rejections from those interviews are the primary mechanism for assessing whether the qualitative scores reflect a plausible real-world threat profile. The interviews therefore form an important part of assessing the prioritisation: they offer an external check on whether these qualitative scores are plausible.

What likelihood means here

Likelihood reflects how necessarily a threat follows from the protocol. A structural threat, present in every deployment, scores **High**; a configurational threat, dependent on an avoidable implementation choice, scores **Medium**; a contingent threat, requiring an unlikely combination of conditions, scores **Low** (Table 6.1). Cross-boundary flows raise likelihood by introducing parties outside the deploying organisation's control.

What impact means here

Impact reflects the severity for the cardholder, set by three factors: the data type exposed, the boundary scope (cross-boundary flows reduce the cardholder's practical recourse, since data that has left Bank X's boundary can no longer be addressed through its governance channels), and the cardholder's intervention capacity, that is, whether a materialised threat can be detected or contested (Table 6.2). The intervention factor applies to flows without a downstream transparency mechanism; it does not apply uniformly, as Detectability and Non-repudiation are assessed on their own merits.

Defining the scores

Table 6.1 and Table 6.2 define the three scores for each dimension. Table 6.3 combines them into four priority levels using a conservative matrix: High likelihood with Medium impact, or Medium likelihood with High impact, both yield High priority. Critical requires High on both dimensions; Low arises when at least one dimension is Low and neither is High; the remaining combinations yield Medium.

Table 6.1: Likelihood score definitions. Author's own definition.

Score	Definition
High	The threat arises from a structural property of the protocol and materialises in every instance of this flow type, regardless of implementation choices.
Medium	The threat depends on a specific but plausible implementation choice, such as the absence of encryption, broad logging retention policies, or the absence of data minimisation at the agent level.
Low	The threat requires an unlikely combination of conditions, such as a malicious insider with specific system access combined with a specific timing window, or an adversary with sustained access to internal network traffic.

Table 6.2: Impact score definitions. Author's own definition.

Score	Definition
High	The threat affects directly identifying data, crosses an organisational boundary, leaves the cardholder with no practical means of intervention, or enables further threats by establishing linkability or identification that other threats depend on.
Medium	The threat affects pseudonymous or derived data within an organisational boundary, or affects cardholder awareness without directly exposing identity. Consequences are meaningful but bounded by the deploying organisation's governance controls.
Low	The threat affects metadata or technical attributes requiring significant additional effort to link to an individual, or consequences are limited to conditions unlikely to affect the cardholder directly.

Table 6.3: Priority matrix combining likelihood and impact scores. **Critical** **High** **Medium** **Low**. Author's own definition.

	Impact: Low	Impact: Medium	Impact: High
Likelihood: High	**Medium**	**High**	**Critical**
Likelihood: Medium	**Low**	**Medium**	**High**
Likelihood: Low	**Low**	**Low**	**Medium**

6.3. Prioritised Threat Register

The full prioritised threat register (the complete table of every elicited threat with its likelihood, impact, and priority score) is provided in Appendix D (Table D.1), where each entry is listed together with its LINDDUN threat node codes. Each row corresponds to one LINDDUN category per flow group: the Gr and Cat. columns identify the entry, L and I give the likelihood and impact scores, and Priority gives

the resulting level. Rather than reproduce all fifty-four rows in the main text, this chapter works from the register directly and presents its structure visually in Table 6.4 (Section 6.4).

The register produces four priority levels. **Critical** means the threat is structural and has severe consequences: it will materialise in every deployment and the cardholder has no practical recourse. **High** means the threat is either structural with moderate consequences, or has severe consequences but depends on a plausible configuration choice. **Medium** means the threat is real but bounded in its consequences or scope. **Low** means specific conditions and limited consequences.

Of the fifty-four entries, eighteen receive Critical priority and nineteen receive High priority, together forming thirty-seven entries that exceed the Medium threshold. This number is still too large for individual description, and not all Critical or High entries are equally informative. Section 6.4 reads across the register to identify the structural patterns that explain why certain configurations produce more severe threats than others, and Section 6.5 uses those patterns to justify the final selection of eight threats for full description.

6.4. Priority Profiles Across Protocol Configurations

Visual summary

Table 6.4 provides a visual summary of the dominant priority level per LINDDUN category per protocol configuration. Each cell shows the highest priority achieved by any group within that configuration for that category. The table makes the two-dimensional structure of the analysis immediately visible: reading across a row shows how a category varies with boundary scope; reading down a column shows how a configuration varies across categories.

Table 6.4: Dominant priority level per LINDDUN category per protocol configuration. Each cell shows the highest priority achieved by any group within that configuration. Author's own analysis.

Configuration	L	I	N_R	D	D_D	U
MCP within org. (A, B)	C	C	H	L	H	C
MCP between orgs. (C, D)	C	C	H	M	C	C
A2A within org. (F)	H	H	L	L	H	H
A2A between orgs. (E)	C	C	C	M	H	C
Human actor (G)	C	C	H	M	C	C

C Critical H High M Medium L Low

Four structural observations are immediately visible in Table 6.4. First, the U column is Critical in every configuration except intra-organisational A2A, where it is High: Unawareness is universal, and its severity reaches the maximum wherever data leaves the organisation, reaches a human actor, or is returned in full within MCP, leaving High only where processing stays both intra-organisational and free of full-record returns. Second, the boundary is the dominant severity amplifier for the categories most directly tied to personal data content. In A2A, crossing the organisational boundary converts Linkability, Identifiability, and Unawareness from High to Critical; in MCP these categories already reach Critical within the organisation through the full-record return flows of Group B, so the boundary instead elevates Data Disclosure from High to Critical. Third, Non-repudiation differentiates the two cross-boundary protocols: it is Critical only in A2A cross-organisational flows, where the bilateral evidence structure prevents either party from unilaterally removing the record, whereas MCP cross-boundary flows produce one-sided server logs that receive High. Fourth, Detectability is consistently Low or Medium across all configurations: it is not a primary driver of severity in this system because the personal data flows are not visible to the cardholder regardless of configuration.

The rationale for each individual priority assignment, configuration by configuration, is provided in Appendix D (Section D.1).

From profiles to selection

The four structural patterns identified in Section 5.8, now reflected in the priority profiles above, define what a faithful selection of threats must capture: the universal Unawareness finding, the severity contrast that the organisational boundary produces within each protocol, the structurally distinct mechanisms of MCP and A2A, and the elevated exposure introduced by the human analyst. Section 6.5 turns these requirements into explicit selection criteria and applies them to the thirty-seven Critical and High entries.

6.5. From Register to Descriptions: Selection Rationale

What a threat description adds

A register entry records only a category, likelihood and impact scores, and the resulting priority. A threat description adds a self-contained account of a single threat: the DFD elements affected, the protocol mechanism that produces it, the underlying evidence and assumptions, and its relationship to other threats. Each is written to be evaluable without prior knowledge of the use case, a requirement for the expert interviews in Chapter 7, where interviewees assess the descriptions' plausibility and completeness against their practical experience.

The reduction from fifty-four to eight

The reduction from fifty-four entries to eight descriptions proceeds in two steps. The first step applies a priority threshold: only entries that receive Critical or High priority are candidates for description. This reduces the field from fifty-four to thirty-seven entries. Entries that receive Medium or Low priority are real threats but are either bounded in their consequences, dependent on unlikely conditions, or adequately captured by the patterns identified in the register analysis. They are documented in the register and in Appendix D but do not receive individual descriptions.

The first step is score-driven; the second is qualitative and does not reapply those scores, so severity is not weighed twice. It applies three criteria to the thirty-seven candidates. First, **protocol specificity**: preference is given to threats directly attributable to a structural property of MCP or A2A, since a threat that would arise from any inter-system data exchange is less informative for this thesis than one that follows specifically from MCP's session model or A2A's task context mechanism. Second, **configuration coverage**: the selection must span all four protocol configurations with at least one intra-organisational and one cross-boundary threat per protocol, so that the boundary contrast identified in the register analysis is visible in the descriptions, and it must also include the universal Unawareness finding, which no single configuration owns. Third, **analytical significance**: where several candidates in the same configuration satisfy the first two criteria, the one that best illustrates the underlying mechanism and is most useful for expert evaluation is selected. As a result, the eight selected threats are not simply the highest-scoring entries: TH-01 and TH-06 carry High rather than Critical priority but are selected ahead of higher-scoring candidates as the clearest examples of a structural MCP and A2A mechanism in their configuration.

The human analyst configuration is represented without a standalone description. Its most severe privacy consequence, the analyst's access to the complete cardholder dossier, is inherited from the upstream intra-organisational A2A over-disclosure rather than generated at the human flow itself, and is therefore documented as the direct consequence of that threat (TH-06); the analyst's awareness and intervention gaps are covered by the two universal Unawareness threats. Applying these criteria to the thirty-seven candidates yields the eight threats described below.

The eight selected threats

Table 6.5 lists the eight threats that result, grouped by the basis on which each was selected: those that follow from a structural property of MCP or A2A, those whose severity is driven by the organisational boundary crossing, and the two universal findings that no single configuration owns. The full descriptions follow in Section 6.6.

Table 6.5: The eight threats selected for full description, grouped by selection basis, with configuration, LINDDUN category, and priority. Full descriptions follow in Section 6.6. **Crit.** Critical **High**. Author's own analysis.

ID	Configuration	Category	Priority	Selection basis
Protocol mechanism: follows from a structural property of MCP or A2A				
TH-01	MCP within	Linkability	**High**	MCP session identifier links every query in an assessment; structural, not configurational.
TH-02	MCP within	Identifiability	**Crit.**	A single MCP session aggregates full identified records at the querying agent.
TH-04	A2A between	Non-repudiation	**Crit.**	A2A's task lifecycle creates bilateral, irremovable evidence of sharing.
TH-06	A2A within	Data Disclosure	**High**	A2A transmits the full task context by default; no built-in minimisation.
Boundary effect: severity driven by crossing the organisational boundary				
TH-03	MCP between	Data Disclosure	**Crit.**	Identified data leaves TB1; Bank X loses all governance control at the destination.
TH-05	A2A between	Identifiability	**Crit.**	A derived profile is re-identifiable using the recipient's own data holdings.
Universal: present across all groups, no protocol-level mitigation				
TH-07	All groups	Unawareness	**Crit.**	The cardholder is unaware of the processing at every stage; addressable only outside the protocol.
TH-08	All groups	Unintervenability	**Crit.**	No access or contestation pathway exists anywhere in the workflow.

6.6. Threat Descriptions

The eight threat descriptions below follow the LINDDUN PRO template [64], extended with a *Why it matters* paragraph. Each description is self-contained so that a reader without prior knowledge of the full analysis can evaluate the threat on its own merits. The coloured left bar and title background indicate priority: **red** = Critical, **orange** = High.

TH-01: MCP session identifier enables cross-store linkability

Category	Linkability	**Nodes**	`L.1.1`
Priority	**High** (L: High; I: Medium)	**Related**	TH-02, TH-03
DFD	P1 (source), DF2/DF4/DF6 (flows), DS1/DS2/DS3 (destinations)		
Assets	Transaction history, KYC records, device fingerprint, cardholder identity		

Summary. MCP maintains a stateful session context across all tool calls within a single fraud assessment. The session identifier on DF2, DF4, and DF6 links all queries to the same cardholder transaction. Anyone with access to session logs at P1 or at the destination data stores can reconstruct the full sequence of data accesses. This linkability is structural to MCP's session model, not a consequence of any particular implementation choice.

Why it matters. Bank X deliberately stores transaction history, KYC records, and device data in separate databases to limit the impact of any single breach. MCP removes this separation at the protocol level without any administrator making a mistake. An insider or attacker with access to session logs can reconstruct exactly which data about which cardholder was queried, when, and in what sequence, information that is more revealing than any individual record. The cardholder has no way to know this is happening and no way to prevent it.

Assumptions: MCP session identifiers are logged at both the source agent and the destination data store under standard operational logging practice.

TH-02: MCP return flows aggregate identified cardholder data at the querying agent

Category	Identifiability, Linkability	**Nodes**	`I.1.1, I.2.1.1, L.2.2.1`
Priority	**Critical** (L: High; I: High)	**Related**	TH-01, TH-05
DFD	DS1/DS2/DS3 (sources), DF3/DF5/DF7 (flows), P1 (destination)		
Assets	Name, address, date of birth, transaction records, device fingerprint		

Summary. DF3, DF5, and DF7 return records from DS1, DS2, and DS3 to P1. Each individual response may be justifiable in isolation, but P1 aggregates all three within the same session context, producing a cross-store profile combining name, address, and date of birth (DF5), behavioural transaction records (DF3), and device data (DF7). This aggregation is a direct consequence of MCP enabling an agent to query multiple sources within a single session without a built-in minimisation mechanism.
Why it matters. The combined profile reveals far more about a person than any single database holds. If P1 is compromised, if a developer stores the payload unnecessarily, or if the data is passed further than intended, the cardholder is fully exposed. This profile is also the direct input to TH-05: a derived version is transmitted to Card Network Y via A2A on DF10.
Assumptions: P1 retains the full returned payload without applying data minimisation before using it in the risk assessment.

TH-03: Identified personal data transmitted to external sanctions service without pseudonymisation

Category	Data Disclosure, Identifiability	**Nodes**	`DD.1.1, I.1.1`
Priority	**Critical** (L: High; I: High)	**Related**	TH-04
DFD	P1 (source), DF8 (flow), EE4/DS5 (destination)		
Assets	Name, date of birth, identity attributes		

Summary. DF8 transmits name, date of birth, and identity attributes from P1 to EE4/DS5, an external sanctions screening service outside both TB1 and TB2. Once transmitted, this data leaves Bank X's control entirely. The plaintext transmission constitutes a disclosure of identified personal data for which Bank X cannot enforce minimisation, retention limits, or access controls at the destination.
Why it matters. Once name and date of birth leave TB1, Bank X has no technical means to enforce how the external service uses, stores, or shares that data. A false positive, the cardholder incorrectly matched to a sanctions entry, can result in transaction refusal, account suspension, or reputational damage, while the cardholder does not know their data left the bank at all.
Assumptions: The sanctions screening service does not offer a privacy-preserving query interface such as hashed name matching or private set intersection.

TH-04: A2A task metadata creates non-repudiable bilateral evidence of data sharing

Category	Non-repudiation	**Nodes**	`Nr.1.1, Nr.1.3`
Priority	**Critical** (L: High; I: High)	**Related**	TH-05
DFD	P1/P3 (sources and destinations), DF10/DF11 (flows)		
Assets	Derived behavioural profile, A2A task metadata, agent identity records		

Summary. A2A assigns a persistent task identifier to the interaction on DF10. This identifier is embedded in all messages exchanged, including the profile on DF10 and the risk score on DF11. The result is a non-repudiable bilateral record: Bank X cannot deny transmitting a behavioural profile to Card Network Y, and Card Network Y cannot deny having received it. This is inherent to A2A's task lifecycle model and exists independently of application-level logging.
Why it matters. Non-repudiation is assessed here from the data subject's perspective, not the organisation's. In security terms, a permanent bilateral record is a desirable audit property. As a privacy threat it means that for every transaction assessed via this workflow, there exists irremovable evidence at Card Network Y that Bank X shared a cardholder profile with them, evidence the cardholder cannot contest, correct, or request deletion of, even if the underlying data sharing was later found to be erroneous or unlawful. Neither organisation can unilaterally delete the record

held by the other party.
Assumptions: Both P1 and P3 retain A2A task logs for audit purposes under standard operational practice.

TH-05: Derived behavioural profile enables re-identification at Card Network Y

Category	Identifiability	**Nodes**	`I.2.2, I.2.3`
Priority	**Critical** (L: High; I: High)	**Related**	TH-02, TH-04
DFD	P1 (source), DF10 (flow), P3 (destination)		
Assets	Transaction velocity, geographic patterns, device consistency indicators		

Summary. DF10 transmits a derived behavioural profile from P1 to P3. The profile contains no direct name or account number. However, Card Network Y holds network-wide transaction data in DS4. By correlating the received profile with DS4, P3 can re-identify the cardholder. This risk is elevated because the recipient operates under a separate governance structure and holds data that Bank X cannot inspect or control.
Why it matters. The cardholder and Bank X may believe the transmitted profile is anonymous because it contains no name. It is not. Card Network Y processes transactions from millions of cardholders and can match the received behavioural pattern against its own data. Once re-identified, Card Network Y can build a risk profile of the specific cardholder using data Bank X never shared and cannot control.
Assumptions: The behavioural profile on DF10 is at a granularity sufficient to distinguish the cardholder from others in DS4.

TH-06: Full risk assessment context on intra-organisational A2A exceeds data minimisation

Category	Data Disclosure	**Nodes**	`DD.2.1`
Priority	**High** (L: High; I: Medium)	**Related**	TH-07, Group G Data Disclosure
DFD	P1 (source), DF14 (flow), P2 (destination)		
Assets	Transaction details, KYC-derived attributes, device fingerprint, risk scores, escalation reason		

Summary. DF14 transmits the full risk assessment context from P1 to P2, including transaction details, KYC-derived attributes, device fingerprint, and the escalation reason. P2's function is to route the case to a human analyst and return the analyst's decision. This routing function does not require the full context. Transmitting the full context is a consequence of A2A's message-passing model, which sends a complete task object by default without a built-in attribute selection layer.
Why it matters. Although the flow is intra-organisational within TB1, every additional attribute transmitted to P2 is an attribute that the human analyst will subsequently see. This excess payload is the direct cause of the Critical Data Disclosure in Group G: the analyst views a complete personal dossier when only a risk summary was needed. Crucially, the excess is not the result of a misconfiguration. A2A by default passes the complete task context to the receiving agent because rich context is what makes agent reasoning effective. There is no built-in attribute selection mechanism in the protocol. Addressing TH-06 therefore requires a deliberate implementation choice (a context filtering layer, a system prompt restriction, or an explicit summary-first design) that must be built on top of A2A. Without that choice, the default behaviour of the protocol produces this threat every time intra- organisational task delegation is used.
Assumptions: P1 passes the full accumulated assessment context as the A2A task payload without applying attribute selection.

TH-07: Cardholder structurally unaware of automated multi-agent processing

Category	Unawareness	**Nodes**	`U.1.1`
Priority	**Critical** (L: High; I: High)	**Related**	TH-08, all Group U rows in Table D.1
DFD	EE1 (cardholder), all flows DF1 through DF18		
Assets	All personal data processed in the workflow		

Summary. The cardholder initiates a payment on DF1 and receives a decision via DF18. In between, their data is queried via MCP across three internal databases, transmitted to an external sanctions service, shared with Card Network Y via A2A, escalated internally via A2A, and presented to a human analyst. At no point is the cardholder informed. This satisfies U.1.1 at the source position in all nine flow groups. The unawareness is structural: MCP and A2A are machine-to-machine protocols with no data-subject-facing interface, and no configuration option within either protocol can address this.
Why it matters. Privacy rests on the idea that individuals can exercise meaningful control over information about themselves. That control begins with knowledge: you cannot make a choice, form an objection, or exercise any right over processing you do not know is happening. Unawareness removes this informational precondition entirely. In the three seconds between tapping a card and receiving a decision, three internal databases are queried, an external sanctions service is contacted, a profile is sent to a separate organisation, a second AI agent is invoked, and a human analyst may read the full dossier. None of this is visible to the cardholder. They cannot object to it, consent to it selectively, or ask for it to be done differently. These options may exist in principle, but the cardholder has no knowledge that the processing is occurring at all. If the decision is wrong, they do not know why. If their data has been shared with an external party, they will never find out through the system.
Assumptions: No out-of-band notification mechanism exists that would inform the cardholder of the fraud assessment process.

TH-08: No access or contestation mechanism exists for the cardholder

Category	Unintervenability	**Nodes**	`U.2.2`
Priority	**Critical** (L: High; I: High)	**Related**	TH-07
DFD	EE1 (cardholder), all flows DF1 through DF18		
Assets	All personal data processed in the workflow		

Summary. The cardholder has no means to discover which data was accessed, which databases were queried, which profile was shared with Card Network Y, or whether a human analyst reviewed their case. MCP and A2A are machine-to-machine protocols with no data-subject-facing interface. TH-07 establishes that the cardholder is uninformed; TH-08 establishes that even if they were informed, no pathway exists to access, rectify, or contest the data used. Together they mean the cardholder is both uninformed and unable to intervene.
Why it matters. TH-07 establishes that the cardholder does not know what is happening to their data. TH-08 establishes the second layer: even a cardholder who somehow became aware of the processing would have no technical pathway to do anything about it. There is no access interface, no correction mechanism, and no contestation pathway within this system. Unintervenability is therefore not a consequence of unawareness, it is an independent structural property of the architecture. The two threats compound each other: unawareness means the cardholder does not know control is absent; unintervenability means control is absent regardless. If any other threat in the register materialises (incorrect data, re-identification, or a profile shared beyond its intended purpose), the cardholder has no pathway to detect it, correct it, or seek redress. This applies equally to all nine flow groups and both protocols.
Assumptions: No system-level or application-level access mechanism is implemented that would allow the cardholder to query or contest data processing decisions.

6.7. Chapter Summary

This chapter applied a qualitative prioritisation framework to the fifty-four threat entries from Chapter 5. Likelihood and impact were each assessed on a three-point ordinal scale and combined using a conservative priority matrix producing four levels: Low, Medium, High, and Critical. The visual summary in Table 6.4 makes the structure of the register immediately readable: Unawareness is Critical in every configuration except intra-organisational A2A, the organisational boundary converts High to Critical across most categories, and Non-repudiation is the sharpest differentiator between intra-organisational and cross-boundary flows.

Of the thirty-seven Critical or High entries, eight were selected for full descriptions on the basis of protocol specificity, configuration coverage, and analytical significance. Each description is designed

to be self-contained and evaluable by an expert who has not followed the full analysis. The descriptions serve two purposes: they provide an auditable account of the most significant findings from this thesis, and they are the primary input for the expert interviews in Chapter 7. The interviews ask domain experts in MCP, A2A, and financial AI systems to assess the plausibility of each described threat on the basis of their practical experience. Confirmations, refinements, and additions from those interviews will determine how the findings are presented and qualified in the Discussion in Chapter 8.

This chapter, together with Chapter 5, provides the answer to the second sub-question: *what privacy threats are present in the system, and what is their relative severity?* Chapter 5 identified fifty-four threat entries across all eighteen data flows by applying the LINDDUN mapping table and threat trees to each flow group. The present chapter assigned a qualitative priority to each entry and produced detailed descriptions of the eight most significant threats, selected on the basis of protocol specificity, configuration coverage, and analytical significance.

7 Expert Interviews

7.1. Introduction

This chapter addresses the third sub-question: *to what extent are the identified threats recognised and validated by domain experts, and which additional threats do experts identify?* It presents the design, conduct, and results of the semi-structured expert interviews carried out to assess the plausibility and completeness of the threat inventory and descriptions produced in Chapters 5 and 6.

The chapter proceeds as follows. Section 7.2 describes the interview design and analytical approach. Section 7.3 presents the selection criteria and profiles of the participating experts. Section 7.4 describes how the interviews were conducted and how the data were analysed. Section 7.5 presents the validation results per threat description. Section 7.6 documents additional privacy threats raised by experts, expert observations on governance responsibility, and expert perspectives on the expected adoption of MCP and A2A in practice. Section 7.7 closes the chapter with the answer to the third sub-question.

7.2. Interview Design

The purpose of the expert interviews is twofold. First, to assess whether the eight threat descriptions produced in Chapter 6 are recognised as plausible and realistic by practitioners with direct experience of agentic AI systems, privacy by design, or financial sector AI deployments. Second, to surface any privacy threats arising from MCP or A2A that the LINDDUN-based analysis did not identify.

Semi-structured interviews were selected as the method for this phase. In a semi-structured interview, the interviewer follows a predefined set of topics and questions but allows the conversation to develop beyond those boundaries when the participant raises relevant considerations. This format allows systematic coverage of the eight predefined threat descriptions while leaving room for experts to raise considerations that the analysis did not anticipate [66]. A fully structured format such as a survey would have prevented experts from articulating nuance or raising additional threats. An unstructured format would have made the results difficult to compare across experts.

Each interview was structured around four phases. The first phase established the expert's background and their familiarity with MCP, A2A, and agentic AI systems. The second phase presented the research context, the use case, and the technical properties of the two protocols. The third phase presented the expert with the eight threat descriptions from Chapter 6 and asked them to assess each on two dimensions: plausibility (does this threat arise in practice?) and severity (does the assigned priority level reflect the real-world risk?). The fourth phase invited the expert to identify any privacy threats arising from MCP or A2A that they consider significant but that are not captured in the existing threat inventory. The full interview guide is provided in Appendix C.

7.3. Expert Selection and Profiles

Participants were selected through purposive sampling, a method in which participants are chosen deliberately because they possess the knowledge and experience most relevant to the research question [67]. Three expertise profiles were targeted: practical experience with MCP, A2A, or agentic AI deployment; experience with privacy by design or data protection in AI systems; and experience with AI in the financial sector. Participants were drawn from practitioners at Avanade, a technology consultancy with active engagements in agentic AI and digital security, and from professionals referred through the initial participant network. The profiles of the seven experts are summarised in Table 7.1. An eighth session was conducted with Kim Wuyts, a co-author of the LINDDUN framework, as an open consultation on the applied methodology rather than a structured threat validation.

Table 7.1: Expert profiles, based on interviews conducted in May 2026. Profile codes: (A) Agentic AI / MCP / A2A deployment, (B) Privacy / Data Protection / Security, (C) Financial sector AI. MCP/A2A familiarity: Y = familiar, P = partial, N = not familiar.

Expert	Area of expertise	Profile	MCP/A2A
E1	Information protection, data security, privacy compliance	B	N
E2	Data science, AI architecture, AI security, regulatory governance	A,B	Y
E3	Modern workplace engineering, Copilot practice lead	A,B	Y
E4	Modern workplace governance, Copilot adoption	A	P
E5	AI/agent development at a large Dutch bank	A,C	P
E6	Data security management, financial sector data governance	B,C	N
E7	Security architecture, Microsoft security solutions	B	N

Seven structured interviews were completed, covering all three targeted expertise profiles. Agentic AI deployment and engineering is represented by E2, E3, E4, and E5. Data protection, security, and governance is represented by E1, E2, E3, E6, and E7. Financial sector experience is represented by E5 and E6. E5 is the only participant currently building and operating agentic AI inside a regulated financial institution, providing the most operationally grounded perspective in the group. Several experts cover multiple profiles, offering perspectives that bridge technical AI implementation and organisational security governance.

7.4. Interview Conduct and Analysis Method

Conduct

Interviews were conducted in May 2026. Each interview lasted approximately 60 minutes and was conducted online via video call. All participants were based in the Netherlands but at different locations across the country. A tight interview schedule, with all seven sessions planned within a single two-week window, made online interviews the most practical format. The video call format also allowed the researcher to share the Data Flow Diagram and the threat descriptions on screen during the interview, which supported the discussion in Phase 2 and Phase 3 without requiring participants to navigate between documents independently.

At least 48 hours before each interview, participants received two documents: the participant information sheet including the informed consent form, and the eight threat descriptions from Chapter 6 as a standalone document. This advance distribution allowed experts to form an initial view of each threat before the session, reducing reading time during the interview and producing richer discussion. Each interview was audio recorded with the explicit written consent of the participant. Signed consent forms are retained by the researcher and are available upon request to the thesis committee.

Recordings were selectively transcribed after each session, meaning that the researcher identified and transcribed the expert's substantive assessments of each threat and the reasoning behind them, rather than producing a word-for-word transcript of the full session. This approach was chosen to focus the analytical record on the content most directly relevant to the research questions. Transcripts were anonymised before analysis: names and any identifying references were replaced with the corresponding expert identifier (E1 through E7). All audio recordings and identified transcripts are stored on TU Delft-approved storage accessible only to the researcher and supervisors, and will be deleted no later than one month after publication of the thesis. The use of AI-based tools in processing the interview material, and in the preparation of this thesis more generally, is documented in Appendix B.

Each interview followed a four-phase structure, set out in full in Appendix C. **Phase 1** (approximately 10 minutes) covered the expert's background and established the profile used in the analysis in Section 7.3. **Phase 2** (approximately 15 minutes) presented the technical and research context, using the full system Data Flow Diagram from Chapter 4 as a shared visual reference and briefly introducing the LINDDUN basis on which the threat descriptions had been produced. **Phase 3** (approximately 30 minutes) covered the validation itself: for each of the eight descriptions the expert was asked whether they recognised it as a realistic privacy threat and whether the assigned priority reflected the risk as experienced in practice, and was invited to elaborate, qualify, or challenge any aspect. At the end of Phase 3 the expert ranked the threats from most to least urgent (Q3.3); this ordering provides a validation dimension independent of the per-threat Confirmed/Refined/Challenged verdict and is analysed in Section 8.2. **Phase 4** (approximately 10 minutes) was an open reflection on threats potentially missing from the eight descriptions, on who should bear responsibility for cardholder privacy in a distributed multi-agent system, and on expected near-term deployment in the financial sector.

Analysis method

Responses were analysed by comparing each expert's assessment with the threat descriptions from Chapter 6. For each threat and each expert, the assessment was classified as one of three outcomes based on the transcribed portions of the interview: **Confirmed** (the expert recognises the threat as realistic and considers the priority appropriate), **Refined** (the expert recognises the threat but qualifies it, adjusts the priority, or adds conditions), or **Challenged** (the expert disputes the threat's plausibility or considers the priority substantially overstated). Where an expert's assessment contained both confirming and qualifying elements, the dominant position in the transcript was used. These classifications are summarised in the validation matrix in Table 7.2. Substantive comments and qualifications are reported per threat in Section 7.5. The interview transcripts were prepared with AI assistance: the enterprise version of Microsoft 365 Copilot, provided through Avanade, supported their transcription, summarisation, and anonymisation. The author then read and interpreted the transcripts directly, and the AI tools were not used to generate the assessments or the analysis. The full AI-usage disclosure, including the data-protection reasons for choosing an enterprise tool, is given in Appendix B.

7.5. Validation Results per Threat Description

Table 7.2 presents the validation matrix summarising expert assessments across the seven structured interviews. Substantive comments are discussed per threat in the subsections below; the full per-expert summaries, including each expert's profile, detailed priority assessments, and notable quotes, are provided in Appendix E.

Table 7.2: Validation matrix, based on interview data collected in May 2026. C = Confirmed, R = Refined, Ch = Challenged.

Threat	E1	E2	E3	E4	E5	E6	E7
TH-01	R	C	R	R	C	R	R
TH-02	C	R	C	C	R	C	C
TH-03	C	R	R	C	R	R	C
TH-04	R	R	R	R	R	R	R
TH-05	C	C	R	C	R	C	C
TH-06	R	R	R	R	C	C	R
TH-07	R	Ch	Ch	R	Ch	Ch	Ch
TH-08	C	Ch	Ch	R	Ch	Ch	Ch

TH-01: MCP session identifier enables cross-store linkability

No expert challenged TH-01; the dominant verdict was Refined, with experts accepting the threat but attaching conditions on auditability, access control, and deployment context. The most operationally grounded confirmation came from E5: their bank already issues single-use, per-query identifiers on the customer-facing side to prevent linkability, yet the server-side MCP session identifier still enables cross-database linkage by internal staff, direct evidence that the threat arises even where an organisation actively mitigates it.

The Refined verdicts converged on the session identifier's dual purpose. E3, E4, and E6 independently raised auditability: under PSD2 and anti-money-laundering rules banks must maintain full audit trails, and the session identifier is the very mechanism that makes those trails coherent across a fraud assessment. Removing or randomising it would serve privacy at the cost of the auditability that regulation demands. The threat, however, lies in what MCP's stateful session design enables by default, rather than in the audit log itself: the same identifier appears in every log entry across every store queried, so any administrator with logging access can reconstruct a cardholder's complete cross-store access profile without further authorisation.

Experts proposed two complementary mitigations, compartmentalising log access by team (E4, E6) and scoping the session identifier per database call rather than per task (E3), while E7 noted that governance of log access remains the decisive control regardless of identifier design. The assigned High priority was accepted by all who commented. The structural insight therefore concerns MCP's session design rather than logging itself: logging is required and correct, but MCP's default session model turns individually justified logs into a cross-database linkability instrument that exceeds what any single access event would produce, manageable only through deliberate governance and access-control choices that do not happen automatically on deployment.

TH-02: MCP return flows aggregate identified cardholder data at the querying agent

TH-02 received the strongest overall confirmation of any threat, with no expert challenging it or the Critical priority. E1, E4, E6, and E7 confirmed without qualification. E6 gave the most institutionally grounded reasoning: banks deliberately separate data stores by access role, and aggregating them at a single agent undermines the separation-of-duties principle that underpins financial data governance. E4 made the same point through a hospital analogy, where combining a patient's name, room, and diagnosis creates a sensitivity far greater than any single item.

E7 proposed destroying agent-assembled profiles immediately after each session. This is a partial control rather than a resolution: the aggregation is structural and occurs the moment the agent receives the return data, so the combined profile exists in working memory, and potentially in intermediate logs, throughout the session, before any deletion policy can act. A prompt-injection attack, system compromise, abnormal termination, or diagnostic logging could expose it within that window. TH-02 therefore remains valid at Critical priority regardless of post-session deletion, which limits persistence but neither prevents the aggregation event nor removes the in-session risk. Treating the profile as temporary also overrides MCP's default shared session context and must be deliberately built and verified by the implementing organisation.

The two Refined verdicts came from E2 and E5. E2 noted that responsible banks already apply anonymisation and data minimisation to sensitive identity data, making the risk less severe in practice. E5 proposed sequential processing where feasible, querying each source and retaining only its conclusion while discarding raw data before the next, though this may not be architecturally possible where cross-source comparison is required.

TH-03: Identified personal data transmitted to external sanctions service without pseudonymisation

TH-03 was confirmed or refined by all seven experts, with no Challenged verdicts. E1 noted that external data sharing is typically governed by contracts whose compliance is rarely monitored, and that the cardholder is often not informed their data left the bank. E4 raised a compliance concern: because MCP tends to transmit the maximum available context, it is hard to specify in advance exactly what personal data reaches an external party. E7 raised the most politically significant observation in the set: Mastercard and Visa are American companies subject to US law, so a US government order could compel disclosure in ways that override the contractual protections nominally in place with European banks. E7 and E5 independently proposed the same structural solution, an EU-supervised neutral intermediary that prevents direct bank-to-network data exchange.

The four Refined verdicts converged on limiting what is transmitted: E2 flagged interception of MCP traffic in transit as the primary risk; E3 questioned whether names and dates of birth need be sent at all rather than a citizen-service or account number; E5 proposed a shared pseudonymous identifier agreed between banks and screening services; and E6 considered outbound volume minimal in practice, with contractual safeguards and audit reports as mitigants. None of these, however, is a property of MCP itself: each is an implementation choice, governance arrangement, or multilateral standard built on top of the protocol. MCP transmits whatever the agent places in the payload and imposes no constraint on what crosses the boundary, so TH-03 is present by default in any cross-boundary deployment and resolved only when such mitigations are actively implemented.

TH-04: A2A task metadata creates non-repudiable bilateral evidence of cross-organisational data sharing

The non-repudiation category is the least intuitive of those assessed. LINDDUN evaluates it from the data subject's perspective: the question is whether the person whose data was processed can ever have that processing undone or contested, not whether the organisation can prove what happened. In TH-04 the persistent A2A task identifier leaves both Bank X and Card Network Y holding an attributable, undeletable record that the cardholder's profile was shared (Section 6.6); the cardholder was never informed and cannot request deletion or contest it. The threat lies in permanence rather than misuse, a data-subject framing that the consultation on the applied methodology supported.

TH-04 produced the most uniform result in the matrix: all seven experts gave a Refined verdict, none confirming Critical without qualification but none challenging it as implausible. The common thread was that bilateral logging is an inherent property of inter-organisational exchange, not unique to A2A,

so severity depends on governance rather than the protocol. E5 treated it as a risk to accept, the benefit of fraud detection outweighing the existence of a record that a check occurred; E6 noted that technical alternatives such as rights management are impractical where models must reason over plaintext. Partial mitigations were offered, namely NDAs and automatic retention limits (E4), encryption with dual-party certificates (E7), and retaining the task identifier while anonymising or discarding the personal content (E3), each depending on counterparty compliance. The unanimous Refined result reflects a pattern across E5, E6, and E7: threats whose mitigation requires governance and legal instruments rather than technical design are assessed as lower effective priority, consistent with the fraud-detection framing in Section 6.4, where a legal basis for processing moderates the weight of individual privacy threats.

TH-05: Derived behavioural profile enables re-identification at Card Network Y

No expert challenged TH-05 as unrealistic, and the Critical priority was accepted by all who addressed it. E1, E2, E6, and E7 confirmed without qualification. E2 framed re-identification as structurally easy today regardless of whether AI is involved, given the volume of data held about individuals; E6 described profiling for secondary purposes as largely uncontrollable once data crosses an organisational boundary; and E7 gave the decisive concrete point, that Mastercard already holds the mapping between card numbers and identities, so any card-linked attribute shared with it can be linked back without further effort. The two Refined verdicts came from E3, who noted that cardholder terms authorising data sharing provide a legal basis that may reduce practical severity, and E5, who accepted the threat as structurally unavoidable and again proposed the neutral-intermediary architecture raised under TH-03. The broad confirmation reinforces a key finding from Chapter 5: whether transmitted data can identify a person depends not only on its content but on what the recipient already holds, which cannot be assessed from the sending side alone and is what makes cross-organisational A2A exchanges structurally different from intra-organisational ones.

TH-06: Full risk assessment context on intra-organisational A2A exceeds data minimisation principle

TH-06 drew mostly Refined verdicts, and its assigned High priority was the most contested of any threat. E6 named it the most important threat from a banking-operations view: once personal data enters a cloud-hosted language model the organisation loses the verifiable oversight that is a hard regulatory requirement in the financial sector, a concern that reaches beyond A2A itself. E5 confirmed and proposed returning only aggregated, non-identifiable information to the human analyst, with the full profile available on explicit request. Among the Refined verdicts, E1 questioned unprompted why TH-06 was rated High rather than Critical, since forwarding a complete identity profile to an analyst is comparable in severity to the Critical threats, a practitioner-driven challenge to the assigned priority. E3 and E4 argued the threat is addressable at the implementation layer through system instructions and safety filters, with E3 drawing a parallel to Microsoft Copilot prompts that already prevent full transcript disclosure; E7 proposed a summary-first, details-on-request interface, noting a parallel with Security Operations Centre workflows. None of these responses is a property of A2A: the protocol passes the full task context by default, with no built-in mechanism to determine what subset the receiving party needs, so TH-06 arises whenever A2A is used for internal delegation without an explicit context-filtering step. This places it in the same structural position as TH-03, where mitigations exist in principle but must be deliberately built on a protocol that does not enforce them.

TH-07: Cardholder structurally unaware of automated multi-agent processing

TH-07 was, with TH-08, the most heavily challenged threat, and no expert confirmed the Critical priority. The five Challenged verdicts shared one argument: the terms accepted on account opening constitute a sufficient legal basis for automated fraud detection, making the cardholder's unawareness legally acceptable even if structurally real. E2 judged it not applicable on this basis; E3 saw unawareness as a structural feature of automated payment processing that the cardholder accepts implicitly, with regulation the more effective lever; and E5, E6, and E7 placed it among governance and legislative problems rather than technical threats, E7 adding that cardholders broadly prioritise payment speed over transparency. The two Refined verdicts accepted the threat but suggested a downgrade: E1 noted that terms and conditions provide at least a formal notification basis, and E4 proposed a cardholder-facing notification for automated screening, analogous to the existing travel-declaration feature, while

acknowledging its own privacy implications. The pattern reflects the contextual point in Section 6.4: in the fraud-detection setting, the existence of a legal basis for processing substantially reduces the contextual urgency of threats that are structurally present but legally authorised.

TH-08: No access or contestation mechanism exists for the cardholder

TH-08 was, with TH-07, among the most heavily challenged threats, and its Critical priority the most widely questioned. E1 confirmed it, noting that the only avenue under data-protection law is post-hoc deletion: the cardholder can request removal after the fact but cannot stop or adjust the process while it runs. E4 gave a Refined verdict, grouping it with TH-04, TH-05, and TH-07 as a fundamental trade-off of fraud detection whose resolution lies in governance and contractual rights, and judged some reduced contestability an acceptable cost. The five Challenged verdicts repeated the TH-07 reasoning that accepting account terms constitutes informed acceptance of the processing; E6 framed the contestability gap as a service-quality rather than a privacy problem, and E7 again pointed to regulation as the appropriate lever. TH-07 and TH-08 were challenged by the same five experts (E2, E3, E5, E6, E7) on the same legal-consent argument, a view most pronounced among those with direct financial-sector exposure (E5 and E6), suggesting a relatively settled practitioner position that consent-based frameworks make unawareness and unintervenability acceptable features of the fraud-detection context rather than violations requiring technical remediation. The notable exception is E1, who refined TH-07 but confirmed TH-08 at Critical, regarding the inability to intervene once data is released as a more clear-cut harm than the unawareness preceding it.

7.6. Additional Threats and Expert Observations

Additional privacy threats raised by experts

Five threats or privacy-relevant concerns not present in the existing eight descriptions emerged from the interviews and the consultation on the applied methodology.

Language model training data as a privacy vector (E2). E2 raised the risk that privacy-sensitive data used to fine-tune a locally deployed language model may surface in responses to unrelated users. This threat is intrinsic to the model's nature and independent of the communication protocol. It falls outside the flow-level analysis because the risk arises from the statistical properties of the model itself rather than from any individual data flow.

Inference-driven database enrichment (consultation on the applied methodology). During the consultation on the applied methodology, the possibility was raised that agents accumulating transaction and customer identity data could infer sensitive attributes over time, such as household composition, spending patterns, or health indicators, and enrich internal databases subsequently accessible to other processes. This represents a downstream data exposure risk linked to the reasoning capabilities of the agent, which was not captured in the flow-level analysis because the enrichment occurs as a consequence of repeated processing rather than as a property of any individual data flow.

Prompt injection as an entry-point risk (E3). E3 identified prompt injection via the cardholder-facing interface as a realistic attack vector, noting an ongoing contest between injection attacks and safety countermeasures. Prompt injection is the technique of embedding instructions in user input that redirect the model's behaviour in ways the system designer did not intend. While this is primarily a security threat rather than a privacy threat in the LINDDUN sense, it is relevant to privacy because a successful injection could redirect the agent's data access in ways that violate the access boundaries the system was designed to enforce.

The language model reasoning layer as an opaque processing environment (E6). E6 raised a privacy risk that is structurally different from all eight identified threats and from the other additional risks above. When the Fraud Detection Agent reasons over cardholder data, that reasoning happens inside a large language model. In a cloud-hosted deployment, the inputs sent to the model, the working context containing the cardholder's combined profile, and the intermediate reasoning steps are all processed on infrastructure the deploying organisation does not own, cannot inspect, and cannot audit in real time. The language model is, in this sense, a black box: input goes in, output comes out, and what happens in between is invisible to the organisation that deployed it.

This is a privacy risk because the cardholder's data, including their transaction history, identity attributes, and behavioural profile, enters an opaque computational process rather than travelling to an identifiable organisation via an auditable flow. MCP and A2A define what data moves between which agents, but

say nothing about what happens to it inside the model's working context. A system can therefore be fully compliant with both protocols while still exposing cardholder data to a processing environment the deploying organisation cannot oversee, a genuine structural risk in any agentic system that uses cloud-hosted language models to process personal data.

Geopolitical jurisdiction over cross-boundary data (E7). E7 raised a privacy risk that none of the other six experts identified: the geopolitical dimension of transmitting cardholder data to Mastercard or Visa. Both organisations are incorporated under United States law. Under US legislation such as the CLOUD Act, a US government order can compel these companies to disclose data held on their systems, including data received from European banks. This applies regardless of where the data is physically stored and regardless of the contractual protections in place between the sending bank and the card network.

This is a privacy risk because the cardholder has no knowledge that their profile may become subject to a foreign government's legal process, and contractual agreements between Bank X and Card Network Y cannot override a US compulsion order directed at the receiving organisation. It is an instance of what Nissenbaum's contextual integrity framework identifies as a norm violation [9]: data shared in a financial-services context flows to one where the governing norms differ entirely from what the cardholder could have anticipated. LINDDUN does not capture it, because it assesses risk within the system's defined trust boundaries, and US jurisdiction over the receiving organisation lies outside them.

Governance responsibility

All seven experts were asked who should bear primary responsibility for cardholder privacy in a distributed multi-agent system where no single organisation designed or controls the full workflow. The answers clustered around two positions. One position, expressed by E1 and E4, was that the deploying organisation, Bank X in this use case, bears primary responsibility because the trust relationship with the cardholder sits with that institution, regardless of how many third parties are involved downstream. A second position, articulated by E5 and E7, was that bilateral contractual responsibility between commercial entities is structurally inadequate for cross-jurisdictional data exchange. On this view, a government-controlled or EU-supervised intermediary body is the appropriate governance solution for inter-organisational exchange between banks and payment networks. E2 observed that in practice almost every organisation she has worked with has serious gaps in AI governance, noting that ungoverned grey zones between organisational roles are the primary attack surface for privacy and security incidents alike.

E1 raised a related distinction that cuts across all three positions: the difference between governance by contract and governance by architecture. Contractual protections, such as data processing agreements, non-disclosure agreements, and retention limits, are the dominant governance instrument for the threats involving cross-boundary data exchange. E1 argued that these are insufficient without verification, because a contract states what a counterparty should do but provides no technical guarantee that they will do it. Architecture-level measures, such as automatic data deletion after processing, session scoping, and context filtering, enforce constraints at the system level and do not depend on counterparty compliance. E1 raised the open question of whether such measures are technically feasible within MCP and A2A as currently specified. The answer, as the analysis in Sections 7.5 and 7.6 demonstrates, is that they are possible but must be deliberately built on top of the protocols, because neither MCP nor A2A enforces them by default. This makes the contract-versus-architecture distinction directly relevant to the platform governance argument in Section 8.6: as long as privacy controls remain contractual rather than architectural, their effectiveness depends entirely on the willingness and capacity of each deploying organisation to implement and verify them.

Adoption trajectory

Experts were asked whether they expected MCP and A2A to be widely deployed in financial fraud detection in the near term. The consensus was that deployment is likely but not imminent in its current form. E3 observed that MCP has been adopted across major platforms within approximately eighteen months, but that security and privacy controls have been added after the fact rather than built in from the start. E5, the only participant working inside a regulated financial institution, noted two dominant bottlenecks. The first is non-deterministic testing, since language models cannot be validated with a single test run the way conventional software can; the second is cross-functional governance approval, which takes substantially more time than the technical build. E2 observed that banks will likely adopt custom adaptations of these protocols with proprietary compliance layers rather than deploying them

unchanged. E6 argued that any agentic architecture must demonstrate auditability to meet existing financial sector regulatory requirements, and that current cloud-hosted language model deployments cannot guarantee the location of data processing, which is a blocker for data protection compliance in the Netherlands.

Two observations from E2 are relevant to how that adoption is likely to unfold in practice. First, E2 found no evidence of privacy or security by design in MCP when examining it for a financial services deployment: the protocol was built for interoperability and speed, and privacy was not a design consideration at the specification stage. This means that organisations adopting MCP cannot rely on the protocol itself to provide privacy properties; they must build those properties on top of it, and they must do so explicitly. Second, E2 separated the protocol specification from its implementation, noting that even a well-specified protocol provides no privacy guarantee if the developers connecting components do not understand how to do so in a privacy-respectful way. The gap between what a protocol permits and what a specific implementation actually does is therefore itself a privacy risk, one that is invisible in any protocol-level analysis, including this thesis, but that is likely to be the dominant source of privacy incidents in real-world deployments.

7.7. Key Findings and Synthesis

This chapter has addressed the third sub-question: *to what extent are the identified threats recognised and validated by domain experts, and which additional threats do experts identify?*

Seven structured expert interviews were conducted with practitioners spanning agentic AI deployment, data security, and financial sector AI. An eighth session was a high-level consultation on the applied methodology with a co-author of the LINDDUN framework. The validation results support the threat inventory substantially, with six of the eight threats confirmed or refined by a majority of experts and no threat entirely rejected.

TH-02 (data aggregation at the Fraud Detection Agent) and TH-05 (re-identification at the card network) received the broadest confirmation, with five Confirmed verdicts each. TH-01 (session identifier linkability) and TH-03 (cross-boundary transmission to the external sanctions service) were confirmed or refined by all seven experts, with the refinements converging on limiting what data is transmitted as the appropriate response rather than disputing the threat itself. Crucially, all proposed mitigations for TH-03 sit outside the protocol: limiting the payload, using pseudonymous identifiers, and contractual safeguards are implementation and governance choices that must be deliberately built on top of MCP, none of which the protocol enforces by default. TH-04 (bilateral record of data sharing between Bank X and Card Network Y) produced a uniform Refined result: all seven experts accepted the structural property but framed it as a trade-off inherent to inter-organisational data exchange rather than a correctable protocol flaw. TH-06 (A2A transmitting more data than the receiving agent needs) was confirmed by experts with financial sector experience and refined by those with implementation backgrounds, with E1 independently suggesting an upgrade to Critical. TH-07 (cardholder unaware of the processing) and TH-08 (cardholder unable to access or contest the processing) were the most widely challenged. Majorities argued that consent-based processing frameworks make these legally acceptable features of the fraud detection context, rather than genuine privacy violations requiring technical remediation.

Five additional privacy-relevant concerns emerged that fall outside the LINDDUN-based inventory, spanning model training data, inference-driven database enrichment, prompt injection, the model reasoning layer, and geopolitical jurisdiction, as set out in Section 7.6. These represent limitations of the flow-level analytical scope rather than failures of the threat identification method.

The three levels of the governance burden

Two cross-cutting observations from the interviews warrant emphasis. E1 drew a distinction between governance by contract and governance by architecture, arguing that contractual protections are insufficient without verification and that architecture-level measures are more robust because they do not depend on counterparty compliance. E2 identified a gap between protocol specification and implementation quality, noting that even a well-specified protocol provides no privacy guarantee if developers do not connect components in a privacy-respectful way. This observation aligns with a long-standing theme in the privacy engineering literature: that privacy is determined by how a system is actually engineered, not only by what its design specifies [58]. The same concern has recently been carried into the secure-by-design discussion of generative AI systems [68].

Together these observations suggest that the governance burden operates at three levels. E2's distinction separates the protocol as specified (the design level) from the way it is actually built (the implemen-

tation quality level). E1's emphasis on architecture-level measures identifies the level in between: the deployment architecture that a specific organisation constructs on top of the protocol, where controls such as session scoping and data deletion are added because the protocol does not enforce them by default. These are levels of governance responsibility, distinct from the technical communication stack introduced in Chapter 3 (Figure 3.1): that stack describes the layers through which data physically moves, whereas these three levels describe where in the governance chain a privacy control must be applied. A privacy failure can originate at any of the three levels, and a measure taken at one does not remedy a weakness at another, so each must be addressed independently. This framing is taken up again as a practical contribution in Section 9.3 of Chapter 9.

Summary

The expert interviews confirm that the privacy threats identified in this thesis are structurally real and recognised by practitioners. The most actionable findings from the practitioner perspective are TH-01, TH-02, TH-03, and TH-06, where technical design choices can substantially reduce risk. TH-04, TH-05, TH-07, and TH-08 are acknowledged as real but are considered by the majority of experts to require governance, contractual, and regulatory solutions rather than protocol-level remediation. The urgency ranking collected at the end of each interview (Q3.3) provides an independent check on this pattern by capturing how experts weigh the threats relative to one another; its results are analysed against the assigned priority levels in Section 8.2.

It should be noted that the validation results presented in this chapter are qualitative in nature. The assessments reflect the professional judgements of seven practitioners and one methodology expert, drawn from a specific network and organisational context. The findings cannot be generalised to all practitioners in the field, but they do provide evidence that the identified threats are recognised as relevant by professionals with direct experience in the areas most relevant to this thesis. These findings are discussed further in Chapter 8.

8 Discussion

This chapter interprets the findings of Chapters 5, 6, and 7 across six dimensions. Section 8.1 positions the results against the research gap and the literature set out in Chapter 1. Section 8.2 integrates the expert validation with the threat analysis. Section 8.3 discusses findings that were not anticipated. Section 8.4 analyses the governance implications, including four structural trade-offs, emergence, system-boundary effects, and the stakeholder conflicts that leave the data subject structurally excluded. Section 8.5 discusses implications for deploying organisations, including Privacy by Design. Section 8.6 addresses the platform governance of MCP and A2A. Section 8.7 summarises the chapter.

Throughout, the eight threats carried forward from Chapter 6 are referred to by name, not by their threat-register codes. For reference, these are: session linkability (the shared MCP session identifier that links the tool calls within a session), cross-store aggregation (the assembly of a combined profile at the Fraud Detection Agent), cross-boundary identity transmission (the sending of identified data across an organisational boundary without minimisation), the bilateral non-repudiable record (the matching task record A2A leaves on both sides of an exchange), profile re-identification at the card network, excessive task context (the over-broad message content A2A permits), cardholder unawareness, and cardholder unintervenability. The mapping to the register codes is given in Chapter 6.

8.1. Positioning Against the Research Gap and Literature

Chapter 1 defined the research gap as the absence of a structured, data-subject-centred privacy analysis of MCP or A2A at the protocol level within a concrete deployment (Section 1.2.6), sitting between two independent literatures (Section 1.2.5). The security literature on these protocols [29, 10, 21, 31] examines authentication, integrity, and access control, protecting the deploying organisation against external attackers. The MAS privacy literature supplies the relevant threat categories, linkability, disclosure, inference, and data minimisation [26, 35], but was built for classical agents exchanging formally typed messages, not the open, language-driven flows of MCP and A2A. More recent work by Patil, Stengel-Eskin and Bansal addresses compositional privacy risks in modern LLM-based multi-agent collaboration [12], yet does not provide a structured, per-protocol threat inventory of the kind developed here. The two closest privacy studies, Yao et al.'s intent-inversion attack on MCP and Wang et al.'s empirical mitigation-and-benchmarking work [33, 34], deliver neither a structured, data-subject-centred inventory spanning both protocols nor one grounded in a concrete deployment and prioritised by severity.

The analysis in this thesis fills that gap directly, and doing so clarifies how its findings relate to each of these literatures. The relationship to the security literature is the sharpest. The strongest protocol-level threats identified here, session linkability, cross-store aggregation, the bilateral non-repudiable record, and excessive task context, are invisible from a security standpoint because they are not security failures. A correctly functioning MCP session identifier that links three database queries is, to a security engineer, a feature working as intended; the A2A task identifier that leaves a permanent record on both sides is the same. They register as privacy threats only when the system is read from the data subject's side, which is the perspective the security literature does not adopt. This is consistent with the established distinction between security and privacy threat modelling that motivated the gap [11, 32]: a system can be fully secure while still producing serious privacy harm through legitimate, correctly authorised processing. The contribution of this thesis is best read as confirming that distinction concretely for MCP and A2A, rather than as overturning the security findings, which remain valid for the threats they address.

The relationship to the MAS privacy literature is one of inheritance and transposition. The threat categories that literature identified, linkability, disclosure, inference, and data minimisation, recur in the findings here, which supports the expectation in Chapter 1 that they would remain relevant. The present analysis shows how the design choices of MCP and A2A produce these long-recognised categories by default, in an open, cross-organisational setting the classical methods were never built for. The first of the three scientific contributions stated in Section 1.2.6, the first structured, data-subject-centred analysis of these protocols, is therefore located here. The applicability of LINDDUN to protocol-level analysis that this rests on had already been shown in an adjacent context by Mirzamohammadi et al. for the Solid protocol [40]; the present work extends that demonstrated suitability to MCP and A2A.

The relationship to the work of Liao et al. on extending LINDDUN to generative-AI systems is narrower [65]. Within the broadened framing they propose, cardholder unawareness is one threat category among several. The analysis here suggests something more specific for MCP and A2A: across every

configuration examined, unawareness recurred without exception, and no choice available at the protocol layer removed it. This claim should be read with care. That no awareness mechanism was found in the configurations studied does not establish that none could exist. The GenAI-specific threat trees from Liao et al. were also not publicly available, and so could not be applied here (Section 5.2). The recurrence of unawareness is therefore reported as a strong and consistent pattern within the scope analysed, not as a proof of impossibility.

A methodological counterpart deserves note: had this analysis surfaced no protocol-level privacy threats, that would not have established the protocols as privacy-safe. It would more likely have indicated that the elicitation was applied at too coarse a level of abstraction, since the design properties examined here, such as persistent session identifiers, full task-context objects, and unconstrained payloads, are observable in the specifications independently of the method used to surface them.

Threats the protocols cause and threats they merely inherit

A question runs through all eight threat descriptions: is a given threat caused by something specific to MCP or A2A, or is it a privacy problem that any automated fraud-detection system would have, whatever protocol carried the data? The distinction matters for the research question, which concerns threats that emerge from these protocols, not threats that merely co-occur with their use. The eight threats can be read as lying along a spectrum, and the strength of the protocol-specific claim varies across it.

Four threats are most plausibly attributed to protocol design. Session linkability follows from MCP's stateful session model, which assigns one persistent identifier to every tool call in a session, so that a correct deployment produces the linkage by default. Cross-store aggregation follows from MCP delivering full responses from every queried store to the orchestrating agent at once. The bilateral non-repudiable record follows from how A2A hands a task between agents, leaving each side an independent copy. Excessive task context follows from A2A's task object being designed to carry rich context, typically more than the receiving agent requires. For each, replacing the protocol with a stateless, context-minimal alternative would largely remove the threat, which is what grounds attributing it to the protocol's specific design rather than to the use case.

Two threats lie at the other end and are better described as inherited from the use case. Cardholder unawareness and unintervenability are real privacy problems that LINDDUN identifies correctly, but they are properties of any automated fraud-detection system. They would be present in a system built on conventional REST APIs, on a message queue, or on any other communication technology, because they follow from the system's purpose rather than from protocol architecture. The contribution for these two therefore concerns resolution rather than causation: even protocols with the lowest inherent privacy footprint cannot resolve them, since their cause lies outside the protocol layer.

Two further threats sit between the poles, and the protocol-level claim is weaker for them. Cross-boundary identity transmission is partly a protocol matter, since MCP transmits whatever the agent places in the message and enforces no minimisation; but any protocol used without minimisation would produce the same outcome, so what MCP contributes is permissiveness by default rather than an active design choice to disclose. Profile re-identification at the card network is the weakest protocol-level claim of the eight: re-identifying an individual from a behavioural profile is a data-science problem that would exist independently of A2A. A2A supplies the route by which the profile crosses the organisational boundary, but the risk derives from the data itself and from the receiving party's existing knowledge base.

That the fraud-detection setting makes it hard to separate what the protocol causes from what the use case demands was also raised in a high-level consultation on the applied methodology with one of the original LINDDUN authors. This lends some external support to treating it as a genuine feature of the problem, not an artefact of how the analysis was framed. It does not weaken the findings: all eight threats are genuinely present in the system and correctly identified. It does suggest where the protocol-specific contribution is strongest: the four design-caused threats, where the causal link between specification and privacy risk is direct. The two middle cases are best read as the protocols amplifying risks they did not originate. The two use-case threats are evidence of conditions the protocol layer cannot reach.

Design-inherent threats and implementation-level leaks

A second distinction locates the contribution more precisely, on a different axis from the spectrum above. That spectrum concerns *which* part of the system causes a threat; this distinction concerns the *level* at which the analysis operates. The present work is a qualitative privacy analysis of the protocol *design*

of MCP and A2A: it reasons from the specifications and the system architecture, not from any concrete software implementation. The threats it identifies are therefore design-inherent. They are present in a correct, faithful implementation and would remain even if that implementation were free of software defects. Session linkability, for example, does not depend on a coding error; it follows from what the MCP session model specifies.

This scoping has a consequence for how the findings should be read. Removing every design-inherent threat identified here would not make a deployed system free of privacy risk. A separate class of privacy leaks arises at the implementation level: overly broad logging, endpoints that expose more data than the specification requires, the absence of transport encryption where the specification permits but does not mandate it, insecure storage of session state, or ordinary software defects. These are real and can be severe, but they are out of scope here by construction, because a specification-level analysis cannot discover a flaw that exists only in a particular vendor's code. The expert interviews surfaced this boundary directly, with E2 observing that even a well-specified protocol offers no privacy guarantee if developers connect components in a privacy-disregarding way.

This boundary defines the durability of the contribution. Implementation leaks are, in principle, fixable: they can be patched, reconfigured, or caught in code review, which makes them ultimately a quality-assurance matter. Design-inherent threats cannot be patched away, because they follow from the specification itself; removing them requires changing the specification or adding a compensating layer outside the protocol. Identifying which privacy threats are located in the design therefore tells protocol designers and deploying organisations which risks they cannot engineer their way out of within the current specifications. This is the sense in which the analysis contributes at the conceptual-design level. It also marks the limit of what the thesis claims: within its scope it cannot characterise the privacy behaviour of any specific deployed implementation.

Reach beyond the financial use case

Because the strongest findings are design-inherent, their reach plausibly extends beyond the financial fraud-detection scenario in which they were derived: the four protocol-caused threats are properties of MCP and A2A as communication infrastructure, so on the design-level reasoning above they should appear in any faithful deployment irrespective of sector. The use case functioned as the analytical vehicle that made these properties observable and concrete, not as a boundary on where they apply, and the contribution is therefore best read as a general statement about the privacy properties of agentic AI communication protocols, illustrated through finance rather than confined to it. How far that transfer actually reaches (which threats generalise as a matter of protocol architecture and which, notably the severity weighting and legal basis, remain domain-specific and would need reassessment in each sector) is examined in full in the methodological limitations (Section 9.2).

8.2. Expert Validation: What the Interviews Confirm, Qualify, and Add

Chapter 7 presented the validation results per threat description. This section integrates those results with the threat analysis of Chapters 5 and 6, drawing three kinds of conclusion: what the validation confirms, what it qualifies, and what it adds that the LINDDUN analysis did not capture.

What the validation confirms

Six of the eight threats were confirmed or refined by a clear majority of experts, and none was rejected as implausible by all participants (Chapter 7). The breadth of confirmation matters for three reasons. First, the two most strongly confirmed threats, cross-store aggregation and profile re-identification at the card network, are those whose plausibility depends most on domain knowledge that structural analysis alone cannot supply: that banks deliberately separate data stores as a governance design, and that card networks hold enough network-wide data to make re-identification feasible. Both were confirmed by the practitioners best placed to judge those assumptions, which increases confidence in the structural analysis. Second, the unanimous acceptance of session linkability, cross-boundary identity transmission, and excessive task context supports treating these as structurally inherent to the protocol designs, since the proposed mitigations all added controls on top of the protocol rather than disputing the threat. Third, E1's was the only proposal to upgrade a threat, suggesting that excessive task context be rated Critical rather than High. It concerns the same mechanism behind the human-analyst finding in Section 8.3, and sits at the boundary between High and Critical despite the conservative placement at High in Chapter 6.

This breadth should be read with caution. The validation is confirmatory by design: the eight threats were constructed by the author and then put to the experts, an arrangement more likely to elicit agreement than independent rediscovery, so a high confirmation rate is partly a property of the method. The panel of seven was small, and assembled for relevant expertise rather than as a representative sample. And where experts pushed back on cardholder unawareness and unintervenability, this chapter re-reads the challenge as lower contextual urgency rather than over-attribution; that move is defensible, but it makes the findings harder to falsify, since it absorbs disagreement as a context effect. A sufficiently strong and consistent pattern of pushback should count as evidence of over-attribution, not only of lower urgency. The validation is therefore best read as support for the plausibility of the threats rather than independent corroboration, with the panel's limitations examined further in Chapter 9.

What the validation qualifies

Three qualifications emerged across multiple experts and carry weight beyond individual responses. The first concerns cardholder unawareness and unintervenability, and turns on the distinction between them. Unawareness is a property of the data subject's knowledge: the cardholder does not know what is happening to their data. Unintervenability is a property of their agency: even if aware, they have no mechanism to access, contest, correct, or halt the processing. Five of seven experts (E2, E3, E5, E6, E7) challenged these two threats on legal acceptability rather than plausibility, arguing that both are accepted features of fraud detection under a statutory legal basis.

This is best read alongside a point the experts raised: cardholders are already unaware of the conventional machine-learning fraud detection that banks run today, with no more visibility or control than in the agentic scenario. The unawareness is therefore not introduced by MCP and A2A; it is the prevailing condition of automated fraud detection. What the agentic configuration changes is the breadth of what the cardholder is unaware of, rather than whether they are unaware at all: a profile aggregated across internal stores and transmitted to a separate organisation, instead of a score computed within one institution. The structural condition is old; its scope is what these protocols extend. The experts' appeal to a legal basis touches the LINDDUN Non-compliance category, which this thesis deliberately excluded as requiring a legal rather than a technical methodology (Section 2.3.1); it is invoked here only as context for interpreting their judgements, not as a compliance finding. These two threats remain structurally real, and would be significant violations absent a clear legal basis, while their contextual urgency in this deployment is lower than their structural severity alone would suggest. This is why the findings are reported in two registers (Table 8.1), applied to deployment decisions in Section 8.5.

This resolution has to answer a sharper version of its own logic. Five of the seven experts challenged cardholder unawareness and unintervenability, and by the standard set out above, that a sufficiently strong and consistent pattern of pushback should count as evidence of over-attribution rather than be re-read as lower urgency, five out of seven is exactly such a pattern. Keeping the two threats at Critical structural severity therefore needs more than the two-register device; it needs a reason the pushback does not meet that test. The reason lies in the content of the disagreement, not its size. None of the five contested that the cardholder is in fact unaware, or that they in fact have no means to intervene; the LINDDUN identification was not disputed. What they contested was its standing: that this unawareness is legally accepted under a statutory fraud-prevention basis and is therefore not a practical concern. That is a judgement about acceptability and urgency, which falls in the Non-compliance register this thesis deliberately scoped out (Section 2.3.1), not a judgement that the threat was mis-identified. Over-attribution, in the sense that would warrant downgrading, would require experts to say the threat is not real, that the cardholder is aware, or can intervene, or that LINDDUN mis-categorised the flow. Had the pushback taken that form, the structural rating would have been lowered. Because it did not, the two threats are kept as structurally real and severe and reported as lower on contextual urgency, but this remains a contestable call, and the verdicts are tabulated in full (Chapter 7) precisely so a reader who weights the five challenges differently can reach the opposite conclusion from the same evidence.

The second qualification concerns the bilateral non-repudiable record. Its unanimous Refined verdict reflects a pattern visible in E5, E6, and E7: threats whose mitigation requires governance instruments rather than technical design are assessed as lower effective priority even when the structural concern is accepted. The record is therefore best read as a genuine governance challenge beyond what any single deploying organisation can resolve, rather than as a weakness in the analysis.

The third qualification concerns cross-boundary identity transmission. The LINDDUN analysis identified it as a data-governance risk: once data crosses the organisational boundary, Bank X cannot enforce what the external service does with it. E7 added a jurisdictional dimension: data transmitted to US-

incorporated card networks becomes subject to US government compulsion regardless of contractual protections. This amplification derives from the legal jurisdiction of the receiving organisation rather than any technical property of MCP, and reinforces that contractual safeguards alone cannot resolve cross-boundary identity transmission.

What the validation adds

The interviews surfaced five privacy-relevant concerns outside the LINDDUN-based inventory, listed and discussed as a structural limit of flow-level modelling in Section 8.3; two of them extend specific threats and warrant attention here. E6's observation about the language-model reasoning layer as an opaque processing environment is structurally significant: LINDDUN identifies what data flows between which components, but cannot identify what happens to that data once it enters a cloud-hosted model's working context, so the exposure identified here is a lower bound. E7's jurisdictional observation similarly extends cross-boundary identity transmission and profile re-identification: for US-incorporated card networks, even contractual governance may be overridden by a foreign legal process. Both bear on the platform governance discussion in Section 8.6, and both confirm that the inventory is the view LINDDUN provides rather than an exhaustive map.

Two orderings of the same threats

Experts were also asked to rank the eight threats from most to least urgent, and the pattern is consistent: the most urgent are the protocol-caused threats with technically addressable mitigations, decisions that can be made before deployment such as session scoping, payload minimisation, and context filtering. The technically oriented experts (E2, E3, E4) ranked cross-store aggregation and session linkability highest, while E5 and E6, with financial-sector experience, ranked cross-boundary identity transmission and profile re-identification higher. Cardholder unawareness and unintervenability ranked lowest, consistent with their Challenged verdicts, because resolving them requires cross-organisational agreement and possibly regulatory intervention rather than a design change. This produces two orderings of the same threats. The prioritisation in Chapter 6 ranks them by *structural severity*, combining likelihood and impact independently of deployment; the expert validation produces a second ordering by *contextual urgency*, weighing each threat in the legal and operational context of fraud detection. Table 8.1 sets the two side by side. The divergence is not a contradiction: a threat does not become less real because it ranks lower on urgency. The clearest case is unawareness and unintervenability, Critical on structural severity yet lowest on contextual urgency, which is why the conclusion in Chapter 10 answers the research question on both axes rather than collapsing them into one.

Table 8.1: The eight threats under two orderings: structural severity (Chapter 6) and contextual urgency after expert validation. The orderings answer different questions and do not coincide, most visibly for cardholder unawareness and unintervenability. Author's own analysis.

Structural severity (Chapter 6)	Contextual urgency (expert validation)
Critical: cross-store aggregation, cross-boundary identity transmission, bilateral non-repudiable record, profile re-identification, cardholder unawareness, cardholder unintervenability **High:** session linkability, excessive task context	**Highest urgency (protocol/architecture-addressable):** session linkability, cross-store aggregation, cross-boundary identity transmission, excessive task context
Ranked by likelihood and impact, independent of deployment context.	**Lower urgency (governance/regulation-dependent):** bilateral non-repudiable record, profile re-identification, cardholder unawareness, cardholder unintervenability

8.3. Unexpected Findings

Three findings emerged from the analysis that were not anticipated at the outset.

The human-analyst interaction produces as broad a threat set as a cross-organisational boundary crossing. The human-analyst flows (DF15 and DF16) were expected to produce a limited threat set, since they involve only human-mediated communication within a single organisation. Instead they produced the highest concentration of Critical-priority threats in the analysis, across all six assessed LINDDUN categories and at all three positions. The reason is structural: the case dossier passed to the analyst via DF15 aggregates the complete fraud-assessment record, including all data retrieved from

all internal stores and all intermediate inferences. A human analyst who receives this is exposed to a richer combination of personal data than any single agent handles, because the dossier is the output of the entire upstream chain. The finding bears directly on excessive task context and was reflected in E1's independent question of whether that threat should be rated Critical rather than High.

Intra-organisational A2A produces Critical threats despite operating within a single trust boundary. The initial expectation was that privacy risk would scale with the number of trust boundaries crossed, so that intra-organisational flows would produce lower-severity threats than cross-boundary flows. Group F (DF14 and DF17) falsifies this. The A2A task delegation between P1 and P2 inside Bank X produces Data Disclosure and Unawareness threats at Critical priority, because the task context object carries the full aggregated cardholder profile regardless of whether the receiving agent is inside or outside the boundary. The organisational boundary is not, in this architecture, a meaningful privacy boundary: the exposure is driven by the context model, not by the trust-boundary structure. The implication is that privacy assessments of agentic systems should not treat same-organisation flows as inherently lower risk. This finding also qualifies the second scientific contribution claimed in Section 1.2.6, that the organisational boundary is the primary determinant of severity. The boundary amplifies threats once crossed, as the cross-configuration analysis in Section 5.8 sets out, but the A2A context model can raise intra-organisational flows to Critical on its own. The boundary is therefore a primary amplifier, not the sole determinant.

The interviews surfaced five privacy concerns not captured by the LINDDUN analysis. The flow-level analysis identifies threats at the level of individual data flows between components. The five additional concerns raised by experts (Chapter 7), however, operate at a different level: the model's training data and reasoning layer, inference-driven enrichment over accumulated history, prompt injection against access-control boundaries, and the jurisdiction of receiving organisations. Together these represent a structural limit of flow-level privacy threat modelling for systems with inference capabilities: LINDDUN identifies what the flows carry and what the components do, not what the components can infer, accumulate, or be induced to do by adversarial input.

These three findings warrant a reflective caution that applies to the analysis as a whole. Because the work rests on a single, author-constructed use case, it is not always possible to separate findings that reflect properties of the protocols from findings that reflect modelling choices in the scenario. The breadth of the human-analyst threat set follows in part from the decision to route a complete case dossier to the analyst on DF15; a scenario in which the analyst received a filtered summary would likely have produced a narrower set. The intra-organisational A2A result similarly depends on how the task context object was modelled. These choices were made to reflect plausible default behaviour, but a different and equally defensible scenario could have yielded different results, so the unexpected findings are best understood as existence proofs, demonstrations that such risks can arise, and not as claims about how frequently they will. The protocol attribution that underpins the strongest threats rests on a counterfactual: that a stateless, context-minimal alternative would remove them. This is reasoned from the specifications, not demonstrated by comparing implemented systems. Testing it empirically against such an alternative is left to the future work set out in Section 10.3.

8.4. Governance Implications

The privacy threats identified here are best understood as the governance consequences of deliberate design trade-offs. MCP and A2A were built to be fast and frictionless because the problems they address, real-time fraud detection, cross-organisational risk assessment, automated screening, require speed that human-mediated processes cannot provide. The same session model that links three database queries in milliseconds also links them in the logs. The same task model that delegates a risk assessment in one call also creates a bilateral record of that delegation. These are features whose privacy implications were not addressed at the design stage. The governance question is therefore not whether MCP and A2A should be used; their functional value is clear. It is what governance structures need to be in place before they are deployed in systems that process personal data at scale.

Four structural trade-offs

The analysis reveals four trade-offs that governance frameworks must address explicitly.

Speed versus privacy. The efficiency advantage of MCP and A2A is inseparable from their privacy risk. The session model that assembles a fraud profile in seconds is the model that creates cross-store linkability. The task model that enables real-time cross-organisational collaboration is the model that produces non-repudiable bilateral evidence. At the protocol level, speed and privacy pull in opposite

directions.

Auditability versus privacy. Financial institutions are legally required to maintain complete audit trails. The MCP session identifier and the A2A task identifier are what make those trails possible; without them there is no audit trail, and with them there is session linkability and a non-repudiable bilateral record of cross-organisational sharing. Better protocol design cannot dissolve this tension, because it is a structural conflict between two legitimate requirements. Governance frameworks must decide explicitly which requirement takes precedence in which context, and under what conditions linkability evidence may be retained, accessed, and used.

Cross-organisational collaboration versus governance control. A2A lets Bank X and the card network collaborate on fraud detection in a way neither could achieve alone. Once data crosses the boundary on DF10, however, Bank X loses governance control over what the counterpart does with it. Cross-boundary identity transmission and profile re-identification are the direct consequences. No technical measure applied by Bank X reaches data after it has left its systems; the only available lever is contractual, and contractual instruments are not enforced at the protocol level. What such agreements must cover is set out in Section 8.5.

Agent autonomy versus data subject control. MCP and A2A are designed for autonomous execution without human intervention at each step, which is the source of their efficiency. Autonomous execution also means that consequential decisions about a data subject's personal data are taken without that subject ever being in the loop. Cardholder unawareness and unintervenability are the structural expression of this trade-off: more autonomy yields less transparency and less intervenability for the data subject. As Section 8.2 noted, this continues a condition already present in conventional automated fraud detection. Its distributional meaning, that the cardholder is the one stakeholder with no place in the system's design, is examined below, in the discussion of stakeholder conflicts.

Privacy risks as emergent system properties

Several of the identified threats are not attributable to any single component. They are emergent: they arise from the interaction of components that are each individually unobjectionable. This is a known characteristic of multi-agent privacy rather than a novel property of agentic systems, and it has a direct precedent in the literature. Patil, Stengel-Eskin and Bansal describe a compositional privacy risk: combining non-sensitive information across multiple agents can produce a violation that no single agent's output would have caused [12]. The same non-compositionality is well established for security, where the secure composition of individually secure components is a long-recognised problem [69]. The contribution here is therefore a concrete demonstration of how MCP and A2A exhibit this already-recognised property.

The clearest case is profile re-identification at the card network. The profile transmitted on DF10 contains no name or account number; the card network's dataset is a legitimate operational resource; the A2A protocol functions as designed. None of these, in isolation, constitutes a violation, yet their combination produces a re-identification risk that examining any single component could not have predicted. This matches the compositional risk of [12] and the general notion of emergence in complex-systems theory, where system-level properties cannot be predicted from the properties of the parts [42]. Cardholder unawareness and unintervenability are emergent in the same sense when considered together. Each results from specific design choices, but together they constitute a condition of complete informational exclusion more consequential than either alone. This condition arises from the combined protocol stack having no mechanism for inclusion, rather than from any actor's decision.

The governance implication is that privacy impact assessments evaluating individual flows or components in isolation will systematically miss emergent risks. This echoes the warning already present in the MAS privacy literature reviewed in Section 1.2.5 [12] and extends it to the MCP-plus-A2A setting: effective governance requires a whole-system perspective that asks what properties emerge from the combination of components, not only what each component does on its own.

System boundaries as governance boundaries

The two trust boundaries in the DFD, TB1 between Bank X and the external world and TB2 between Bank X and the card network, are not only technical. They are governance boundaries where different legal frameworks, accountability structures, and organisational interests meet. When data crosses TB2 on DF10, it moves from a context in which Bank X is the data controller under Dutch financial regulation to one in which the card network operates under its own regime and its own data interests. Sociotechnical systems theory has long treated the management of system boundaries as a central

concern: the interface between a system and its environment, and between coupled subsystems, is where many of the most consequential governance problems arise, not within the components themselves [43]. A2A creates a seamless technical interface across TB2 while doing nothing to address the governance discontinuity that boundary represents, so the technical ease of crossing masks the governance complexity of doing so. Organisations deploying A2A for cross-organisational communication should treat each boundary crossing as a governance event requiring explicit attention rather than a routine technical integration.

One objection to this finding deserves to be met head-on, because it touches the central claim that the organisational boundary is the primary amplifier of severity. LINDDUN is, by construction, a boundary-aware method: it directs the analyst to examine threats precisely where data crosses a trust boundary, so a sceptic could argue that finding the boundary decisive is partly an artefact of a method built to foreground boundary crossings rather than a property of the system. The objection has real force and cannot be dismissed; a method organised around a different unit of analysis might well have surfaced a different primary amplifier. Three considerations nonetheless place the finding beyond a mere method effect. First, LINDDUN flags threats at a crossing but does not assign them severity; the severity weighting here came from the impact scoring, which turned on the data's sensitivity and on what the receiving card network already holds (Chapter 6), so the amplification is driven by what sits on the far side of the boundary, not by the act of crossing it. Second, the method did not mechanically rank crossings highest: the intra-organisational A2A context model reached Critical without crossing any boundary at all (Section 8.3), which is direct evidence that boundary-dominance was not pre-ordained by the analysis. Third, the boundary's salience is corroborated outside LINDDUN, by the governance discontinuity set out above and by E7's jurisdictional observation, neither of which depends on the threat-modelling method. The honest formulation is therefore conditional: within a flow-level privacy analysis, the organisational boundary is the primary amplifier of severity, a claim about this class of method and this system, not an unconditional fact about agentic AI.

Stakeholder conflicts and the excluded data subject

The privacy threats arise in a system whose actors have partly aligned and partly conflicting interests. The full stakeholder analysis, mapping each stakeholder to its primary interest and the key conflicts, was presented as part of the use case in Section 4.6 of Chapter 4 (Table G.2). The purpose here is to read that mapping against the threats now identified, in order to characterise the kind of problem these conflicts represent.

Two observations follow. First, the cardholder is the only stakeholder with no representation in the system's design or governance. Bank X shapes it through its agent configuration, the card network through its agent specification, the external service through its API, and the regulator through compliance requirements; the cardholder shapes none of these. This absence follows from MCP and A2A being machine-to-machine protocols that were never intended to include the data subject as a participant. Second, the conflicts are primarily distributional rather than technical: they concern who bears the costs of the system's operation (the cardholder) through privacy exposure, and who receives the benefits (the institutions) through fraud prevention and efficiency. Governance frameworks that address only the technical layer will therefore be insufficient; effective governance requires mechanisms for representing and balancing the interests of all stakeholders, including those the current design structurally excludes.

8.5. Implications for Deploying Organisations

For organisations considering MCP and A2A in sensitive data workflows, the findings suggest several governance requirements that go beyond standard security hardening.

First, the cross-organisational A2A boundary requires explicit data-processing agreements that address not only what data is transmitted (DF10) but what the receiving organisation may do with it, including whether re-identification against network-wide datasets is permissible and how long the bilateral task evidence may be retained. These are requirements that cannot be met at the protocol level and must be negotiated between organisations before deployment.

Second, cardholder unawareness and unintervenability cannot be addressed by either organisation acting alone. Adding a notification or contestation mechanism requires a system-level decision and coordination between Bank X, the card network, and potentially the protocol layer itself. Organisations should treat these two as system-level governance requirements agreed across all parties before deployment, rather than as implementation tasks for a single development team.

Third, and cutting across the above, the mitigations for session linkability, cross-store aggregation, cross-boundary identity transmission, and excessive task context all sit outside the protocols. Session-identifier compartmentalisation, post-session profile destruction, cross-boundary payload minimisation, and context filtering before A2A task delegation must be deliberately engineered on top of MCP and A2A; none is enforced by default. A deploying organisation that installs either protocol and follows all published security guidance will still produce all four threats unless it additionally makes privacy-specific design choices the documentation does not require. This is a structural gap, not a configuration oversight. Whether that gap is best closed by each deploying organisation in turn or at the protocol layer itself is the subject of Section 8.6.

The structural response to all three requirements is Privacy by Design [70]: building these protections into the architecture from the outset, since none of the four trade-offs in Section 8.4 can be resolved after deployment by configuration alone.

Reading the findings in two registers

For a deploying organisation, the two orderings in Table 8.1 translate into a single instruction: treat the findings in both registers at once. The governance investment should scale with the sensitivity of the data: the more sensitive it is, the more the structural threats demand deliberate, privacy-specific controls instead of default protocol behaviour.

This bears on the research question of whether MCP and A2A can be used where privacy matters, and the answer is a qualified yes. The protocols were built for speed, efficiency, and rich context-sharing, not for privacy, so they carry the structural threats into every deployment; but those threats are addressable, at a cost, through controls layered on top. Whether such controls can be made strong enough for the most sensitive processing, and whether they introduce privacy weaknesses of their own, is a question this analysis raises but cannot settle, and it is taken up as future work in Section 10.3.

How much each threat matters is context-dependent

The two registers point to something more general: how much a given threat matters depends heavily on the deployment, and several of the eight threats show this especially clearly.

Session and cross-store linkability is the first. In LINDDUN terms it is unambiguously a privacy threat, since it lets data about a person be correlated across contexts. In banking, however, linkability is in part a regulatory requirement: anti-money-laundering and audit obligations compel institutions to keep linkable, traceable records, as the interviews repeatedly noted. That dual character is easy to over-read: it does not make the linkage less intrusive for the cardholder, whose contexts are joined either way, but limits the bank's room to act on it, since a capability the law mandates cannot simply be switched off. The obligation also justifies only purpose-bounded linkage, whereas the MCP session identifier links every query regardless of purpose, so the compliance rationale covers part of the threat, not all of it. The picture shifts without that obligation. An agent acting on behalf of an individual over their own data (reaching across that person's email, calendar, health, and shopping services over MCP) would, on the same mechanism, join previously separate contexts into a single cross-domain profile with no audit duty behind it. Whether that is worse than the banking case is uncertain (the contexts were never meant to meet, yet the user connected the services themselves), but the linkability is structurally identical in both; only the justification, and the freedom to act on it, differ.

The bilateral non-repudiable record is the second. A2A's task lifecycle leaves a permanent record of every inter-agent exchange on both sides, which neither party can unilaterally remove. For the two banks, mutually accountable, contractually bound, and operating under a regulator, such a record is low-risk and in places even useful as evidence. For the cardholder it is more ambivalent: a permanent, unilaterally unremovable trace of whose data was shared with whom, and when, remains a privacy exposure that the institutional trust mitigates but does not erase, and reading it as benign would quietly substitute the banks' perspective for the data subject's that this analysis otherwise keeps. Its weight turns less on trust between the parties than on who can reach the record later (a regulator, a court order, a breach), which the banking relationship constrains but does not control. In a lower-trust setting, between competing firms, or between an individual and an unfamiliar service, the same indelible record becomes a standing liability available for leverage, disclosure, or compulsion. The structural threat is identical throughout; what changes is how much the surrounding relationship absorbs it, which is precisely why the severity of non-repudiation cannot be read off the protocol alone but has to be weighed against the relational context of the deployment.

Cardholder unawareness and unintervenability are the third, and here the context-dependence runs

deepest. MCP and A2A neither cause these conditions nor offer any means to remedy them, and whether that absence is a serious failing or barely matters depends on who the data subject is. Where the data subject directs the system, an agent acting on their behalf over their own data, they are present and consenting, and the threat is much reduced, though not eliminated, since directing a task is not the same as being aware of which tools the agent invokes or where the data travels underneath. Where the data subject is a third party excluded from a chain that processes their data, as the cardholder is here, the same absence is a genuine structural exclusion. In the banking case this is often said to be absorbed at a different layer, through the terms and conditions a cardholder accepts. That reading should be treated with caution: a general card-acceptance agreement is a weak basis for informed consent to real-time, autonomous multi-agent processing of a behavioural profile, and leaning on it risks dressing up a structural gap as a solved problem. The high protocol-level severity and that thin contractual coverage are claims about different layers, and conflating them would understate the gap that remains.

This is why it is doubtful that a remedy should simply be baked into the protocol: a mandatory awareness-and-intervention mechanism would be essential where the data subject is an excluded third party and largely redundant where they are the user directing the system. If it belongs anywhere, it is as a capability the deployment can invoke for third-party data subjects, not as a fixed protocol feature, which places the responsibility on the deploying organisation and, ultimately, the regulator, rather than on MCP and A2A, as the governance discussion below develops. How that responsibility should be assigned across an open ecosystem of independently operated agents remains genuinely open.

Cross-boundary transmission without pseudonymisation is a fourth case, and what it turns on is the layer at which a safeguard lives, not the use case alone. Neither MCP nor A2A pseudonymises the personal data it carries across an organisational boundary; whether the data leaves in identified or pseudonymised form is decided above the protocol, in the implementation. The threat reported here is therefore a protocol-level observation: the protocol provides no default, no enforcement, and no guarantee of pseudonymisation. It would be easy to read this as reassuring, a careful deployer can add the safeguard, and in a sensitive cross-organisational setting such as this one a careful deployer probably would. That reassurance is misplaced. Leaving the safeguard to the implementation means every organisation that adopts these protocols must independently recognise the need and build it, with nothing in the protocol to prompt or require it, and some will not. The distinction between a safeguard missing from the protocol and one missing from a given system is therefore not merely a scoping boundary; it is itself the finding. This thesis characterises the protocol level, where the safeguard is structurally absent, and leaves to the unmodelled implementation the separate question of whether any particular deployment supplies it, but the absence of a default is what makes that question arise at all.

Profile re-identification at the card network is the fifth, and the case where the protocol's role is most partial. The re-identification risk derives from the card network's own data holdings, which are large enough to reverse a pseudonymised profile, and in that sense it would exist independently of A2A; it is use-case-inherent rather than design-inherent, in the terms developed in Chapter 9. What A2A contributes is not incidental, however. It makes the cross-boundary transfer of a behavioural profile a routine, automated, and repeatable operation, where outside an agentic setting it would be a deliberate and comparatively rare exception. The protocol does not create the re-identification capability, but it changes how readily and how often the profile is placed within reach of the data holdings that can exploit it, which is a genuine protocol-level contribution to the risk, not merely its transport.

That leaves a harder question than the others: if the risk is partly inherent to the use case, how should cross-organisational fraud detection be done instead? No clean answer is available. Withholding the profile entirely would defeat the purpose, while pseudonymising it helps only until it meets a dataset rich enough to reverse it, which is exactly what the receiving party holds. The more defensible conclusion is that some cross-organisational analytics of this kind cannot be made privacy-preserving by technical means alone, and that the operative control is governance (limiting what the receiver is permitted to do with data it can, in principle, re-identify) rather than any safeguard at the protocol or implementation layer. That is a claim this thesis can establish in outline but not fully resolve, and a candidate for future work.

The human overseer as safeguard and as risk

One point deserves to be made plainly, with a distinction the analysis itself forces. Autonomous processing of this kind should always have a specific actor who is accountable for it; that much is close to non-negotiable, because where a system acts on people's data with no one answerable for what it

does, no one is positioned to catch it when it goes wrong. What is *not* automatically warranted is the usual way of discharging that accountability, placing a human in the loop to review the system's output, because, as the human-analyst finding showed (Section 8.3), that reviewer is exactly where the most sensitive data converges. Accountability and a data-seeing overseer are routinely treated as the same safeguard, yet this case pulls them apart: the human introduced to make the system answerable becomes a privacy exposure in their own right. This complicates the common assumption that a human in the loop makes an automated system inherently safer; for privacy, it can do the opposite.

The implication is not to remove the human but to separate the two functions by design, so that oversight is calibrated to what a reviewer actually needs to verify behaviour (sampled cases, aggregated decision metrics, or access to flagged transactions only) rather than to the full identified profile the system happens to assemble. What confirming that the models behave correctly requires is rarely everyone's complete data, and a deployment that grants the latter to achieve the former has mistaken access for accountability.

Are these the right protocols, and where do they fit?

The qualified yes above concerns whether MCP and A2A *can* be used where privacy matters; it leaves open whether they are the *right* protocols and where they fit best. Both questions follow directly from the finding that severity is determined by the organisational trust boundary rather than by the protocol choice. Substituting MCP and A2A for other agent-communication protocols such as ANP or Agora would not materially change the privacy profile, because these are interoperability protocols rather than privacy protocols and share the same boundary-agnostic design [10]. What changes the profile is an architectural decision rather than a protocol one: data-sovereignty designs that keep personal data under the data subject's control [40], confidential-computing deployment, or contextual-integrity-enforcing middleware that evaluates each flow before it is transmitted [34]. A different protocol is therefore not the remedy, precisely because the protocol is not the lever.

By the same boundary-amplification logic, the cross-organisational processing of third parties' personal data in a regulated sector, the case examined here, sits at the demanding end of the spectrum, and was chosen deliberately as a stress test of the protocols' privacy properties rather than as a typical deployment. The low-sensitivity, intra-organisational, and user-directed settings where these protocols instead fit comfortably are taken up under adoption below.

Is revising the protocols enough?

The argument so far concerns substituting one interoperability protocol for another. It leaves open a sharper question: whether revising MCP and A2A *themselves* (adding the session scoping, context filtering, and notification or consent mechanisms whose absence drives the design-inherent threats) would be enough to make agentic processing privacy-preserving. On the evidence of this analysis it would be necessary but not sufficient, because three classes of privacy risk would persist even under perfectly privacy-aware protocols. First, the distinction drawn earlier between design-inherent threats and implementation-level leaks means that a sound specification can still be deployed in a leaky way; protocol-level guarantees do not bind the configurations assembled on top of them. Second, the privacy judgment-action gap (an agent correctly recognising a data item as sensitive yet transmitting it because the task implicitly demands it [34]) is a property of the model rather than of the protocol that carries the data, and a revised protocol cannot remove it. Third, cardholder unawareness and unintervenability, which neither protocol can remedy and which the discussion of context-dependence above examines in full.

More fundamentally, the analysis points to a tension that sits in agentic AI itself rather than in any single protocol. This might seem to undercut the thesis (if the problem lies a layer above MCP and A2A, why analyse the protocols at all?), but the opposite holds: it is precisely the protocol-level analysis that demonstrates the tension is structural rather than incidental. Without showing that the threats follow from what MCP and A2A specify, the claim that agentic AI is privacy-hostile by nature would remain a conjecture; the protocol analysis is what grounds it. The tension has three faces, all visible in the findings. Rich context is forwarded through the chain, maximising what flows and minimising what any one trust boundary can contain; the agent acts autonomously, so the data processing is open-ended rather than fixed in advance; and the outcome of a task is non-deterministic, so what the system will do with the data cannot be fully audited beforehand. These are not defects to be patched but the very properties that make the systems useful.

That raises a question the protocol framing alone does not: for some sensitive settings the right answer

may be a deliberate decision not to use agentic AI, rather than a safer agentic protocol, in favour of more bounded systems whose data flows are constrained and whose outcomes are deterministic and auditable. The findings suggest where that trade-off tips: towards bounded systems when the data subject is an excluded third party, without consent or visibility, and the processing crosses an organisational boundary, the conditions under which the unawareness and boundary-amplification threats are most severe. Where exactly the line falls cannot be settled here, but the choice is a real one, and treating "which agentic protocol?" as the only question understates it. This trade-off, too, is a direction for future work.

Adoption in practice and the regulatory frame

Looking further ahead, and now beyond what the analysis strictly establishes, a reasoned expectation is that MCP and A2A will see their widest adoption where the data is not especially sensitive: developer tooling, internal enterprise automation, and agents acting on behalf of an individual over their own data. This is not a finding of the analysis but an inference from it, and the reasoning is specific: these are exactly the settings in which the unawareness and boundary-crossing threats identified here largely fall away, because the data subject is the user and no organisational boundary is crossed. For privacy-sensitive, cross-organisational settings it is correspondingly less clear that these protocols are the right instrument at all. What such settings arguably call for is a different class of protocol: one that keeps the advantages of agentic AI (autonomy, tool orchestration, cross-system context) while being privacy-safe by design rather than by added controls. Whether those two goals can be reconciled is genuinely open, and designing and evaluating such a protocol is perhaps the most valuable of the future-work directions this thesis points to (Section 10.3).

A prior question is whether an organisation such as a bank would deploy these protocols in this form at all. The analysis treats the agentic configuration as given, yet banking data in practice is highly compartmentalised, sitting in access-controlled silos each governed by its own authorisation and consent requirements, with processing that is comparatively static and tightly audited. Granting an autonomous agent the broad, cross-silo access the use case assumes is not a small step, and how and when institutions would extend such access, if at all, remains open. This does not weaken the analysis, because its claims are conditional: the structural threats describe what such a deployment would carry if assembled, not a prediction that banks will assemble it soon. The use case functions as an existence proof of the risks, not as a forecast of adoption.

That said, the interviews made clear that banks are actively exploring agentic AI, and that a tension between privacy and efficiency is already live in the sector (Section 8.4). On the reading of these interviews taken here, an interpretation of a small expert sample rather than an established sector-wide fact, user privacy appears to be weighted heavily chiefly because of regulation and the penalties non-compliance carries, rather than as an independent design goal. If that reading is right, adoption will be gated less by what the technology can do than by the point at which institutions are confident a deployment is legally approved and will not expose them to sanction, making the decisive variable regulatory as much as technical.

The regulatory environment is therefore part of the picture, and it differs sharply across jurisdictions in both substance and timing. It seems likely, though this goes beyond what the analysis can establish, that agentic AI will be deployed in less heavily regulated jurisdictions before it is approved for sensitive sectors in the European Union, so where these systems first appear is partly a regulatory question rather than a purely technical one. A comparative analysis of how different regulators treat autonomous agentic processing is a direction for future work this thesis can only flag.

Finally, the form a deployment takes matters as much as the protocols it uses, since the same MCP and A2A mechanisms carry very different exposure depending on configuration. Processing that stays fully automated and entirely within one organisation is the least exposed; routing the output to a human recipient concentrates risk at that point, as the human-analyst finding showed (Section 8.3); and transmitting personal data outside the organisation is the most exposed of all, in line with the boundary-amplification finding. A realistic privacy assessment therefore has to ask not only which protocols are used, but in which of these forms, since the same stack can be relatively safe or distinctly hazardous depending on whether a human and an organisational boundary sit in the path.

Pulling these threads together, the bottom line, stated as a considered judgement rather than a result the analysis proves, is that these protocols can be used in a high-stakes financial fraud setting under one of two conditions. Either the implementation layer develops privacy-preserving techniques strong enough to contain the structural threats and can demonstrably show that it has, making deployment

defensible; or the sector moves to the privacy-safe successor protocols discussed above. This is deliberately a sharp dichotomy, and real deployments will sit between its poles, lower data sensitivity, partial mitigation, or strong governance compensating for what technique cannot. But absent one of the two conditions, or a credible position between them, the honest answer to whether MCP and A2A belong in this use case today is: not yet.

8.6. Platform Governance of MCP and A2A

The argument of this section is that the privacy threats identified here expose a separation between the party that determines the privacy properties of MCP and A2A, the protocol designers, and the party that bears the consequences, the deploying organisations. Because the design choices that drive the strongest threats are fixed at the protocol level and cannot be changed by any individual deployer, effective privacy governance of agentic AI cannot rest on deploying organisations alone. It must also engage the protocol layer, where privacy protections could be built into the specifications and apply by default. This argument is normative, and one defensible reading of the findings rather than a conclusion they compel. The threat analysis stands independently of it: a reader could accept every threat identified here and still hold that responsibility should track legal accountability, which rests with the deploying organisation. The argument's force depends on premises about where the burden of preventing foreseeable harm ought to sit, which lie outside what a LINDDUN-based analysis can establish. The remainder of this section establishes the separation, addresses the objection that designers should not be answerable for it, distinguishes governance responsibility from legal liability, and sets out what the protocol layer would need to do.

MCP and A2A are open technical standards, but they did not emerge from neutral infrastructure. Both began as vendor-controlled specifications: MCP was introduced by Anthropic and A2A by Google. Both have since been placed under the vendor-neutral governance of the Linux Foundation, A2A through the Agent2Agent project in June 2025 and MCP through the Agentic AI Foundation in December 2025. That transfer is recent, and the design decisions that determine the privacy properties analysed here, the session model, the context-passing behaviour, and the task-record structure, were made by the originating vendors beforehand and are inherited by the specifications unchanged. Both donations were framed as a way to secure vendor-neutrality and broad adoption as durable infrastructure standards, on the reasoning that a single-vendor standard is harder for competitors to adopt with confidence [50, 49]. The transfer therefore formalised the existing specifications rather than revising them. Neutral governance does not by itself make a protocol privacy-respecting: it changes who stewards the standard, not what the standard does with personal data by default. This structure is characteristic of digital platforms, which Gawer characterises as modular architectures of a stable core and a periphery that third parties build on [71]. The privacy consequence here is asymmetric: the platform owner captures the value of broad adoption, while the risks of specific deployments fall to the organisations that build on the platform. This asymmetry follows from the economics of standards competition: because the value of a communication protocol grows with the number of agents and tools that adopt it, the overriding incentive is to become the default before rivals do, and releasing the specification as an open standard serves that end by lowering the barrier to adoption and entrenching the resulting network effects. Open-sourcing is therefore better read as a platform-capture strategy than as a concession, and the design properties analysed here were fixed before the standard was opened; speed to adoption, not privacy by default, is what these incentives reward.

A defender of the originating vendors could object that this asymmetry is not theirs to answer for. MCP and A2A were designed for interoperability and speed, not for privacy-sensitive processing. On this view, the responsibility for assessing a protocol's fitness for such a context rests with the organisation that chooses to deploy it there, much as the designers of a general-purpose transport protocol are not answerable for every application built on it. This objection has force, which is why the argument is framed in terms of governance responsibility rather than fault. The counter-position is narrower. These protocols are being adopted at scale in privacy-sensitive domains such as fraud detection, as the experts in Chapter 7 confirm is already underway. A design choice that determines the privacy properties of thousands of downstream deployments is therefore governance-relevant, whatever the purpose for which it was originally made.

The structural finding of this thesis creates a governance problem its own framing must address: the privacy properties of these systems follow from protocol-level design choices (the persistent session identifier, the full task-context object, the bilateral task record), yet the parties who made those choices are not the parties the law holds accountable for the resulting processing. The party determining the privacy properties is not the party answerable for them. This mismatch, more than any individual threat,

is the governance contribution of the analysis.

Two things must be kept separate. Legal accountability is the narrower: under the GDPR it turns on the controller and processor roles, which fall on the deploying organisation, not on the protocol designer, who does not process the data at all. How liability *should* be allocated when the design originates elsewhere needs a legal methodology and is left to future work (Section 10.3). Governance responsibility, where the burden of *preventing* privacy harm ought to lie, is the broader question the findings bear on. Here a distinction matters: between harms from genuinely unforeseeable misuse of a general-purpose protocol, for which blaming the designer would be unreasonable, and harms from foreseeable, structural properties of the design itself. The threats identified here fall on the second side. Forwarding rich identified context between autonomous agents is not an unanticipated misuse of MCP and A2A but the very thing they are built to do. The claim that designers bear no responsibility is therefore too strong; the defensible position is that they should not carry open-ended legal *liability* for every downstream deployment, yet do carry a governance responsibility for the privacy properties their design makes structural.

What follows is not vendor liability but stewardship paired with concrete design obligations. Once a protocol is adopted widely enough that its design causes harm at scale, its direction should pass to bodies able to govern it in the public interest, regulators, or open governance organisations such as the Linux Foundation, under which both protocols now sit. That transition is welcome but is not itself a remedy, since it inherits the design as already specified.

The substantive work is in the design, and the findings make concrete what privacy-by-default would mean at the protocol level. A session-scoping limit in MCP, so the session identifier does not persist across separate database calls within a task, addresses session linkability; a payload-minimisation guideline or attribute-selection layer in MCP reduces the default exposure in cross-boundary identity transmission; and a context-filtering mechanism in A2A's task object, requiring the sending agent to specify which attributes the receiver needs, addresses excessive task context. Each targets a specific design choice that produces a specific harm, and none requires action by deploying organisations; adding them to the specifications moves the burden from every individual deployer to the protocol layer, where it can be resolved once. Whether they would fully eliminate the threats, or introduce new trade-offs of their own, cannot be settled here and is left to future work.

Read in economic terms, this structure is a familiar one. The privacy cost of the protocols' default behaviour falls on the cardholder, a party to neither the protocol's design nor its deployment, while the benefits of broad, low-friction adoption accrue to the designers and the deploying organisations; the cost is, in the language of economics, a negative externality [72]. Because no actor in the chain internalises it, none has an incentive to absorb the expense of preventing it: protocol designers optimise for interoperability and adoption, deploying organisations for the fraud-loss reductions the system delivers, and privacy-specific controls add cost and friction without returning revenue to either. The result is a market failure in which privacy is structurally under-provided, not through negligence but through the ordinary operation of misaligned incentives. The same lens sharpens the remediation argument made above. Where a harm can be prevented by several parties, economic efficiency favours assigning the duty to the one that can avoid it at least cost, the least-cost avoider; here that is the protocol layer, where a single change to the specification removes the harm for every downstream deployment at once, rather than thousands of deployers independently rediscovering and re-engineering the same control. The regulatory implication that follows is therefore not incidental: correcting a market failure of this kind, in which private incentives diverge from the social cost, is the classic rationale for governance intervention.

The second implication is for regulators. Governing the privacy consequences of agentic AI by engaging only with deploying organisations, while leaving the protocol standards untouched, would be structurally incomplete: it regulates the party that processes the data but not the party that determines how exposed that data is by default. The findings provide a concrete basis for engaging the protocol layer directly, session linkability, cross-boundary identity transmission, and excessive task context name specific design properties that produce privacy harm by default, and can ground requirements directed at protocol designers and standards bodies, not only at the organisations that build on their work.

8.7. Chapter Summary

This chapter interpreted the findings of Chapters 5, 6, and 7 across six dimensions; its central conclusions can be drawn together briefly.

The contribution is distinct from the existing security literature on MCP and A2A, which analyses threats to the deploying organisation: the present analysis addresses threats to the data subject. Its most significant finding is that the strongest protocol-level threats, session linkability, cross-store aggregation, the bilateral non-repudiable record, and excessive task context, arise because the protocols function correctly, not because they are misconfigured or attacked. This distinction between security correctness and privacy risk, anticipated by the literature in Section 1.2.5 and here made concrete, is the analytical core of the work. The qualified verdicts on the bilateral record, unawareness, and unintervenability reflect the fraud-detection context's legal framework, not a weakness in the structural findings.

Where that contribution is strongest is clarified by the protocol-inherent versus use-case-inherent spectrum. The four threats above follow from specific design choices in MCP and A2A; cardholder unawareness and unintervenability would arise in any automated fraud-detection system, including the conventional machine-learning systems banks already run; cross-boundary identity transmission and profile re-identification sit between the poles, amplified by the protocols by default without being uniquely caused by them.

The governance implications are best understood, in the CoSEM framing, as a distributed problem. The four structural trade-offs require coordinated decisions across protocol designers, deploying organisations, and regulators, yet the current distribution of responsibility leaves the burden on deployers while the decisive design choices remain at the protocol layer. Two unexpected findings reinforce a whole-system reading: intra-organisational A2A reaches Critical severity without crossing an external boundary, and the human-analyst interaction carries the highest threat concentration in the system.

These dimensions return to the main research question. The threats are structural, not incidental: they follow from the protocols' design, are recognised as plausible by practitioners, can be separated from threats inherent to the use case, and form a distributed governance problem no single actor can resolve. They should be read in two registers, the general protocol-level claim and the context-specific weight under a fraud-prevention mandate, and at the level of conceptual design rather than implementation quality. Chapter 9 then reflects on the limitations and contributions of the study, and Chapter 10 gives the consolidated answer.

9 Reflection and Contribution

9.1. Introduction

Following the interpretation of the findings in Chapter 8, this chapter steps back to reflect on the research process and to evaluate the quality and limitations of the study. Section 9.2 discusses the methodological limitations that bound the validity and generalisability of the findings. Section 9.3 describes the scientific and practical contributions of the thesis.

9.2. Methodological Limitations

Structural scope of the LINDDUN framework

LINDDUN was designed for systems with stable, specifiable data flows that can be represented as a static DFD. Agentic AI systems deviate from this assumption in a structurally important way: LLM-based agents can autonomously initiate additional tool calls or task delegations at runtime that were not specified in the original design and therefore cannot be represented in the DFD. The analysis in this thesis is bounded by the flows explicitly modelled in the DFD. Threats arising from autonomous agent behaviour outside these flows, including inference risks where agents derive sensitive attributes from individually innocuous data, and model memory effects where prior session content influences future responses, are not captured by the LINDDUN analysis. This is a structural limitation of applying a static, flow-level threat modelling methodology to a class of systems that is inherently dynamic, and it means that a category of privacy-relevant risks associated with MCP and A2A is outside LINDDUN's analytical domain regardless of how completely the DFD is modelled or how thoroughly the threat trees are applied.

This deserves to be put more pointedly, but as a distinction between two questions rather than as doubt about the method as used. LINDDUN was well suited to the question this thesis actually posed (a structural, protocol-level account of what MCP and A2A expose by design) and the limitation noted here concerns a different question it was never asked to answer. Much of what makes agentic systems privacy-relevant is dynamic and emergent rather than statically specifiable: risk arises from how autonomous agents behave at runtime, not only from the flows fixed in a DFD. The findings themselves point this way, the human-analyst concentration and the intra-organisational A2A result were both characterised as emergent system properties (Section 8.4), not as features of any single flow.

Recent work on MCP- and A2A-based systems names further risks of the same kind: token-lifetime and over-broad access scopes that belong to the authorisation architecture rather than the data flows [29]; AgentCard linkability, arising in the discovery phase before any modelled flow begins; the privacy judgment-action gap, where an agent recognises a data item as sensitive yet shares it because the task demands it, a property of model inference rather than of the carrying flow [34]; and consent fatigue in long agentic chains, which operates at the interaction-design level [29]. The common thread is decisive: none of these lives at the data-flow level, so no flow-level method can see them. This is not a gap in how LINDDUN was applied but a structural boundary of the approach itself, so the fair question is not whether LINDDUN suited the question posed, but whether the privacy risks of agentic AI as a whole can be mapped by a flow-level method alone. They cannot, which is what motivates the complementary, dynamic-layer analysis taken up next.

A substantial set of risks at the intersection of classic multi-agent-systems threats and privacy also plausibly went unnamed here, precisely because LINDDUN reasons about privacy in isolation and does not reach the agent-coordination, delegation, and trust dynamics that multi-agent-systems security studies. A concrete example is delegation trust: when one agent hands a task to another, it cannot verify how the receiving agent will subsequently handle the personal data in the task context, a risk that is neither a property of the delegating flow nor of any single component, and so falls between the two analyses. Surfacing that overlap is arguably one of the clearest gaps a single-framework analysis leaves open.

As established in the framework comparison in Chapter 2, MAESTRO and LINDDUN are positioned as complementary: MAESTRO analyses threats across the architectural layers of an agentic stack, including authorisation architecture and agent discovery, while LINDDUN analyses privacy threats in individual data flows and trust-boundary crossings. This division is consistent with the MAESTRO authors' own observation that LINDDUN "must be paired with other frameworks to create a comprehensive threat model" [45], and this thesis supplies the privacy-specific layer that pairing calls for. A

combined analysis applying both frameworks to the same system would therefore reach considerably more of the privacy-relevant risk than either alone (those at the data-flow level identified here, those at the authorisation, discovery, and agent-inference levels by MAESTRO) without, on the argument of the previous paragraph, capturing the dynamic, runtime risks that neither static method represents. What this division leaves out is worth stating precisely, because the gap is not one a more detailed flow model could close. A DFD, however fine-grained, captures a system as a fixed set of flows between fixed components; it is a snapshot. The risks that matter most in agentic systems are not in the snapshot but in what unfolds over time: what a model does with data inside its own reasoning context, how agents discover and authenticate one another, and how autonomous delegation produces consequences no single flow reveals. Adding flows does not help, because the object being represented is behavioural and temporal, not structural, which is why the limitation is a property of the method's form, not of how carefully it was applied. MAESTRO is built around exactly this dynamic, layered character, and the Liao et al. generative-AI extensions to LINDDUN address the adjacent model-inference layer, where data is exposed through what a model infers rather than through what a flow carries [65]. The conclusion that follows is modest and consistent with the previous paragraphs: LINDDUN and MAESTRO are views of different layers rather than competing accounts, and a study combining a flow-level method with a dynamic, multi-agent one would map substantially more of the privacy picture than either alone, though, as noted, not the whole of it. Both the combined application and the GenAI extension are natural directions for future research.

LINDDUN generative AI extensions not applied

During the final stage of this research, the LINDDUN team published a domain-specific extension of the framework for generative AI systems [65]. This extension affects three of the seven LINDDUN threat types, with its most significant additions in the Data Disclosure and Unawareness & Unintervenability categories, and contributes a set of GenAI-specific threat characteristics and examples to the knowledge base. The fully elaborated threat trees are provided only as supplementary material to that publication, which has not been released at a publicly accessible location. Correspondence with the LINDDUN research group during the analysis phase confirmed that the threat trees were not yet available, and the supplementary link remains unavailable at the time of writing.

Because the GenAI-specific threat trees were not (and are still not) available as part of the LINDDUN instrument, they could not be incorporated into the analysis. The threat identification in Chapter 5, the prioritisation in Chapter 6, and the expert validation in Chapter 7 are all based on the standard, publicly established LINDDUN threat tree patterns [73]. This is a deliberate scope decision rather than an omission: a privacy threat analysis must be reproducible against a defined and available knowledge base, and applying a threat tree set that is not publicly available would have undermined that reproducibility. As discussed in Chapter 8, the two categories the extension refines most extensively, Data Disclosure and Unawareness & Unintervenability, are precisely those this thesis found most consequential at the protocol layer, so the extension is expected to enrich rather than overturn the findings. Re-running the elicitation against the GenAI-extended threat trees, once they are fully integrated into the published framework, is a direction for future research.

Single, hypothetical use case

The use case is a hypothetical scenario constructed from the protocol specifications and domain knowledge rather than an empirical observation of a live system. This means that the DFD is a modelling artefact whose completeness cannot be verified against a real deployment. If a real system were to contain flows or components not represented in the DFD, the threat inventory would need to be extended. The use case was constructed to satisfy the analytical criteria established in Chapter 2, and its representativeness is supported by the literature on financial multi-agent AI deployments cited in Chapter 4. Nevertheless, the hypothetical nature of the case introduces a gap between the modelled system and the range of real-world deployments that the findings are intended to describe.

The hypothetical design is, however, a deliberate methodological choice with a specific analytical advantage. Because the use case was constructed to realise all four protocol configurations simultaneously, the analysis is able to isolate the contribution of the protocol specifications themselves from the implementation decisions of any particular organisation. A study of a live system would conflate what the protocol causes by design with what a specific implementation team chose to do or omit. The findings of this thesis are therefore not a description of what is wrong with one specific system; they are a characterisation of what is structurally present in any system built on these protocols with this architecture. That claim is stronger in one sense and weaker in another: stronger because it is not con-

tingent on the decisions of a single deploying organisation, and weaker because it cannot be verified against operational data. The expert interviews in Chapter 7 partially address this second limitation by providing practitioner assessments of whether the identified threats are recognised as realistic in actual deployment contexts. Whether the findings generalise must be addressed carefully, because they rest on a single, author-constructed use case, and a scenario built by the analyst can embed the conclusions it is meant to test. The use case is also modelled at a high level: several implementation layers, where concrete mitigations such as context-filtering would sit, are deliberately absent, so the analysis characterises what is structurally present at the protocol level and stays silent on what implementation measures could counter. What carries the generalisation despite the single case is the distinction developed in Chapter 8 between design-inherent and use-case-inherent threats, though that distinction is itself an analytical claim drawn from one case, and would need several to be confirmed.

The force of the distinction is that the design-inherent threats do not depend on the modelling choices that make the single-case basis fragile. Session linkability, cross-store aggregation at the querying agent, the bilateral non-repudiable record, and the over-broad task context follow from properties of the MCP and A2A specifications, not from how the fraud scenario was drawn. There is therefore good reason to expect them in any faithful deployment of these protocols, regardless of sector: a healthcare agent querying patient-record systems over MCP and delegating a referral over A2A would, on the same mechanisms, be subject to the same linkability, aggregation, and bilateral-record patterns. This is a reasoned expectation rather than a demonstrated result, healthcare was not analysed, but it is grounded in the specifications rather than in the use case, which is what makes it more than a guess.

What does not carry across is the severity weighting. The impact scores in Chapter 6 depend on the data processed and on the holdings of the parties: profile re-identification at the receiving agent is severe because of what the card network already knows, which is specific to that setting and would differ in healthcare or human-resources contexts. The legal basis is domain-specific too, as the two-register reading in Section 8.5 sets out: the general register, in which the protocols are privacy-problematic by design, transfers; the fraud-detection register, resting on a statutory fraud-prevention basis, does not. A deployment under a different legal basis would re-weight the same structural threats.

The findings are therefore most directly applicable, at full specificity, to financial-sector deployments; the design-inherent threats and the four patterns are expected to generalise as a matter of protocol architecture, while the severity ranking and legal weighting require reassessment against each domain. A multi-sector comparative study to test both the transfer and the design/use-case distinction itself is identified as future research.

Qualitative prioritisation

The likelihood and impact scores assigned in Chapter 6 reflect the analytical judgement of the researcher rather than empirically measured values. The criteria for each score level were defined in advance and applied consistently, but the possibility of systematic bias in the direction of findings the analysis was expected to produce cannot be entirely excluded. The expert interviews in Chapter 7 mitigate this limitation by providing a practitioner assessment of the plausibility of the scores, although, as Section 8.2 notes, that validation is confirmatory in design rather than fully independent, so it reduces this risk without removing it.

Expert sample and generalisability of validation

Seven structured expert interviews were conducted, with participants drawn primarily from Avanade, a technology consultancy jointly owned by Microsoft and Accenture. This organisational context is relevant for interpreting the validation results. Consultants at a firm of this type work across many different client organisations, sectors, and technology environments simultaneously, which means their practical experience is not limited to a single institutional context. Several participants held expertise across agentic AI deployment, data security governance, and financial sector AI systems, often gained through multiple client engagements rather than a single employer. This breadth of exposure provides a different kind of generalisability from, for example, interviewing practitioners at a single bank: it reflects cross-sector patterns rather than the specifics of one organisation's deployment.

That said, the sample remains a sample. All participants operate within a professional network that has a particular relationship with Microsoft technologies, including Copilot and Azure AI services, which are among the primary platforms on which MCP is currently deployed commercially. This may introduce a degree of familiarity with MCP in particular that is not representative of the broader practitioner population. The validation findings are therefore most directly applicable to the type of enterprise technol-

ogy context in which these practitioners work, and their transferability to, for example, open-source or research-oriented deployments of MCP and A2A requires separate assessment. The qualitative nature of the validation means the findings indicate plausibility as judged by a specific group of informed practitioners, not empirical prevalence across all deployments.

Protocol version specificity and rapid evolution

The analysis is based on the MCP specification revision 2025-03-26 and the A2A specification version 0.2.5. Both protocols were actively developed during the period of this research, and this pace is not incidental. MCP reached production adoption across major platforms within approximately eighteen months of its introduction in late 2024, and both protocols were moved under vendor-neutral foundation governance during the research period, as set out in Section 8.6. Their development is therefore accelerating, not stabilising. Structural findings that are tied to specific protocol mechanisms, such as MCP's session model or A2A's task lifecycle, reflect the behaviour of the analysed versions and may not apply to future versions if those mechanisms are modified.

Two observations qualify this limitation. First, the threats that this thesis identifies as structurally protocol-inherent, primarily session linkability, cross-store aggregation, the bilateral non-repudiable record, and the over-broad task context, arise from design choices that are central to what makes these protocols useful: MCP's stateful session, A2A's bilateral task evidence, and A2A's rich context model. These are not incidental features that are likely to be removed in a routine update; they are the architectural foundations of the protocols' core functionality. Removing them would require a fundamental redesign rather than a version increment. The structural findings are therefore more stable than the version numbers suggest. Second, the rapid evolution of both protocols means that any future analysis should begin by comparing the specification versions. If MCP introduces session scoping or A2A introduces context filtering, the corresponding threats would be structurally resolved and would no longer appear in a fresh analysis. Longitudinal tracking of how the privacy properties of MCP and A2A change across versions is a natural direction for future research.

The data-subject perspective

A consequence of the framing must be made explicit, because it bears on how the threat inventory itself should be read. By taking the data subject's perspective throughout, the analysis sees the system through the eyes of the person whose data is processed, and does not represent the bank or the other organisations, for whom the same flows serve legitimate business and regulatory purposes. This is deliberately one-sided, and the one-sidedness is the contribution rather than a flaw in it: the deployer-centred view, in which these flows are efficiency and fraud-prevention gains, is already well supplied by the security and confidentiality literature this thesis positions itself against (Section 1.2.6); the data-subject view is the side that was missing. The cost of the choice is that the inventory records the exposures borne by the cardholder and not the benefits accruing to the institutions, so it should be read as a privacy account, not a balanced cost-benefit one, a full appraisal of these deployments would have to weigh both.

That the inventory depends on this choice is itself a methodological finding. LINDDUN is by construction a data-subject-centred framework, its categories defined around a person whose data is processed, so the choice of *which* person is not a neutral setting but partly constitutes which threats become visible. Re-centring on a different actor in the chain would leave the flow-level threats largely intact, since session linkability, cross-store aggregation, and the bilateral record are properties of how identifiable data moves, not of whose data it is; but the use-case-bound threats would change, with cardholder unawareness, unintervenability, and re-identification replaced by that actor's equivalents, and some would lose their force entirely, as cardholder unawareness does once the data subject is the user directing the system. Dropping the data-subject frame altogether, for the deploying organisation's view, would not re-weight the analysis but change its kind, turning it into the confidentiality-and-security account the existing literature already provides. The threats identified here are visible precisely because of the data-subject lens, and a differently centred analysis would have produced a different inventory, which is why the choice of perspective is a methodological commitment to be stated, not a default to be left implicit.

Non-compliance scoped out

The frequency with which the experts returned to legal acceptability is itself worth registering: in the settings where these systems would be built, the question of whether a deployment is permitted weighs

at least as heavily as the question of what it exposes, which is consistent with the reading in Section 8.4 that adoption is gated more by regulatory approval than by technical privacy properties. Assessing that acceptability directly was nonetheless kept outside the scope of this work, and the reason is methodological rather than practical. Determining whether a deployment complies with the AI Act, DORA, NIS2, or PSD2 calls for legal interpretation of statute and case law, not the technical threat modelling applied here; it answers a different question with a different method, and folding it in would have changed the kind of study this is rather than deepened it. The threats identified here are the input such an assessment would reason about, not a substitute for it. A dedicated legal-technical analysis of agentic-AI fraud detection under those instruments is, for that reason, one of the more valuable directions for future work, and arguably the one the experts themselves most often pointed toward.

9.3. Scientific and Practical Contributions

Scientific contributions

This thesis makes three scientific contributions. Each is positioned in detail against the security and multi-agent-systems privacy literature, and against the gap identified in Section 1.2.6, in the discussion in Section 8.1; they are stated here as consolidated claims. First, it provides the first structured, data-subject-centred privacy threat analysis of MCP and A2A, taking the perspective of the individuals whose personal data is processed rather than that of the deploying organisation addressed by the existing security literature. Second, it demonstrates that the organisational boundary, rather than the protocol choice, is the primary amplifier of threat severity in multi-agent AI systems, which challenges the assumption that protocol selection is the primary lever for privacy risk management. Third, it identifies a cross-protocol threat chain, running from MCP data aggregation through A2A cross-organisational transmission to re-identification at the receiving agent. This chain is visible only when both protocols are analysed within the same system model, and the closest prior studies, confined to a single protocol or to runtime benchmarking, could not surface it.

Four transferable structural patterns

The central transferable output of this thesis is a set of four structural patterns, derived in Section 5.8 and recurring throughout the analysis, that characterise how privacy risk arises in MCP- and A2A-based systems. They are stated here as a consolidated, reusable set because they, rather than the eight use-case-specific threat descriptions, are the findings that transfer to other agentic AI deployments. They are analytical findings, not mitigations: this thesis ends at the identification and characterisation of privacy threats and does not propose privacy-enhancing technologies to resolve them, which is left to future research.

While designing and evaluating privacy-enhancing measures falls outside the scope of this thesis, the analysis points clearly toward where the most valuable directions lie, because the most severe threats found here share a single root. Cross-store aggregation at the querying agent, excessive task context, and unpseudonymised cross-boundary transmission all stem from the same default: MCP and A2A forward more identified data through the chain than any individual step requires. The direction with the widest reach is therefore data minimisation at the protocol level, context-filtering and pseudonymisation by default, since one principle, constraining what enters the chain, bears on three of the design-inherent threats at once, on exactly the layer where the analysis locates their cause. This is the protocol-level counterpart to the concrete measures sketched in Section 8.6.

A second direction addresses what minimisation cannot: the universal unawareness and unintervenability threats. These concern visibility and agency rather than data volume, which no reduction in what flows can resolve, and they are structurally the most significant threats found, present in every configuration, addressable by no protocol mechanism, and the precondition under which the other harms go unnoticed. Their *contextual* urgency is the most context-dependent of any threat here: in fraud detection, a statutory basis and the fact that cardholder intervention is neither expected nor desired hold it low, whereas in a setting without that basis it rises sharply. The fitting response is therefore not built into the protocol but invoked by the deployment when the data subject is an excluded third party: a notification-and-intervention capability, the bounded form of data-subject control, rather than full data sovereignty, which a system in which the cardholder is structurally not a party cannot realistically provide.

A third class of threat, re-identification at the receiving party, is only partly technical, and for it the operative direction is governance: constraining what the receiver may do with data it can in principle re-

identify, as argued in Section 8.5. Identifying these directions is as far as this thesis takes the question; designing and evaluating the measures themselves is left to future work (Section 10.3). The patterns set out below are offered in the same spirit, as a reusable analytical lens rather than as solutions.

1. **Structural unawareness.** The data subject is unaware at every step. Neither protocol provides any mechanism for notifying, obtaining consent from, or giving visibility to the person whose data is processed. This makes Unawareness the only threat present at every position in every configuration, and the precondition under which all other harms go undetected.
2. **Boundary amplification.** The second scientific contribution above generalises into a reusable pattern: crossing an organisational trust boundary is the primary amplifier of threat exposure, and the same protocol mechanism produces a materially more severe threat profile once data leaves the originating organisation. Stated as a pattern, it provides a fast diagnostic, namely that the trust-boundary structure of a deployment, not its protocol stack, is the first thing to examine when estimating privacy risk.
3. **Protocol-divergent threat types.** MCP and A2A produce different *kinds* of privacy threat that require different kinds of response. MCP threats such as session linkability follow from what the protocol enforces and are architectural; A2A threats such as broad task payloads follow from what the protocol permits and are matters of design discipline. Treating the two protocols as interchangeable misdiagnoses the risk.
4. **The processing output as a risk concentrator.** The point at which an automated processing chain delivers its result to a human actor concentrates more privacy risk than any individual automated step. This shows that privacy assessment of agentic systems must take a whole-system view rather than evaluating each component in isolation.

These four patterns are the form in which the thesis contribution is most readily reused: they apply to any MCP- or A2A-based system and provide a compact analytical lens for reasoning about its privacy properties before a full LINDDUN analysis is carried out.

These patterns are deliberately broader than the protocol-level answer to the main research question. Of the four, only protocol-divergent threat types is purely protocol-inherent; boundary amplification turns on the organisational boundary rather than the protocol, while structural unawareness and the human risk concentrator reflect the use case and the system architecture. The protocol-level answer therefore remains the narrower set of protocol-inherent threats identified in Section 10.2; these patterns extend that answer into a reusable, system-level lens rather than replacing it.

Practical contributions

The eight threat descriptions produced in Chapter 6 provide practitioners deploying MCP and A2A in sensitive sectors with a structured account of the most significant privacy risks, together with the protocol mechanisms that produce them. The finding that Unawareness is structurally universal in agentic AI protocols identifies a category of risk that cannot be addressed through protocol-level measures alone, directing attention to architectural and organisational transparency mechanisms as necessary complements to technical privacy controls.

A further practical contribution, drawn from the expert interviews in Chapter 7, is the observation that the governance burden for these protocols is distributed across three distinct levels, each requiring a different kind of intervention and a different responsible party. The *protocol design level* concerns what MCP and A2A enforce or permit by specification; privacy weaknesses at this level can only be removed by the protocol maintainers. The *deployment architecture level* concerns the privacy controls, such as session scoping, payload minimisation, and post-processing data deletion, that an adopting organisation must build on top of the protocol because it is not enforced by default. The *implementation quality level* concerns whether the developers connecting the components do so in a privacy-respecting way, since even a sound protocol and a sound architecture provide no guarantee if the implementation is careless.

The practical value of this distinction is that it tells each actor where their responsibility lies. A privacy failure can originate at any of the three levels, and a measure taken at one level does not compensate for a weakness at another. For an organisation adopting MCP or A2A, this provides a structured way to allocate privacy responsibility rather than assuming that the choice of a reputable protocol is by itself sufficient.

The distinction between what a protocol enforces by design and what its implementation actually does reflects a long-standing theme in the privacy engineering literature: privacy outcomes depend on how a

system is engineered, not only on its design specification [58]. This concern has recently been carried into the secure-by-design discussion of generative AI systems [68]. That literature, however, is oriented towards engineering practice and does not partition responsibility across the three levels identified here. The contribution of this thesis is the synthesis of the expert observations into a three-level division of the privacy governance burden, rather than the underlying principle itself. The derivation of these three levels from the expert observations is set out in Section 7.7 of Chapter 7.

Applying this level-based view to the eight threats makes the allocation of responsibility concrete. Table 9.1 maps each threat to the governance level at which it can be addressed and to the actor that bears primary responsibility for addressing it. The table assigns *responsibility*, not remediation: it identifies who is positioned to act on each threat, consistent with the protocol-inherent versus use-case-inherent distinction in Chapter 8, and deliberately stops short of prescribing specific privacy-enhancing techniques, which remain outside the scope of this thesis and a matter for future work. The allocation follows directly from where each threat originates. Threats caused by a protocol specification can only be removed by the protocol maintainers; threats arising from how a system is assembled on top of the protocol fall to the adopting organisation; and threats inherent to the fraud detection purpose, rather than to either protocol, require coordinated action across organisations and, ultimately, the regulator. The protocol maintainers here are the standardisation bodies that now govern the specifications rather than the originating companies alone, as the reflection that follows sets out.

The governance argument of Chapter 8 concerned where responsibility for these threats *ought* to lie; this section turns to where, realistically, it *will* fall, and the two do not coincide. The case made there is that the most efficient locus is the protocol layer, yet several factors make a near-term protocol-level remedy unlikely, so in practice much of the burden will rest on deploying organisations and regulators instead. Both MCP and A2A are by now market-standard protocols that an individual adopter cannot rewrite, and both now sit under vendor-neutral Linux Foundation governance with consensus-based change processes, MCP through its Specification Enhancement Proposals and steering group, A2A through the Linux Foundation project. That stewardship is the appropriate home for the protocols, as argued earlier, but it also means no single maintainer can amend either specification quickly, so the protocol-level fixes that would resolve these threats at source will arrive slowly if at all.

The two protocols differ in how this plays out. The MCP threats are architectural, following from what the specification enforces, so removing them at source requires precisely that slow, collective governance route. The A2A threats follow from what the specification permits, and a deploying organisation can mitigate them without any change to the standard. This makes the A2A threats the easier to address technically, but, for that very reason, the more likely to go unaddressed, because they depend on each deployer independently choosing to act rather than on a property of the protocol. That asymmetry is itself an argument for the protocol-level direction advocated earlier: what is left to the discipline of every individual deployer is, predictably, done by only some. The allocation of responsibility that follows, summarised in Table 9.1, should therefore be read as a description of where the burden will land under current governance, not as an endorsement that this is where it ideally belongs.

Table 9.1: Allocation of responsibility for addressing each threat, by governance level and primary responsible actor, with a brief description of each threat. The table assigns responsibility, not specific mitigations, which are outside the scope of this thesis. Author's own analysis.

Threat	Governance level	Primary responsible actor
MCP session identifier enables cross-store linkability	Protocol design; deployment architecture	Protocol maintainers (MCP), who define the session model; the adopting organisation where session scoping is configurable
MCP return flows aggregate identified cardholder data at the querying agent	Deployment architecture	The adopting organisation (Bank X), which controls how query responses are aggregated at the orchestrating agent
Identified data transmitted to the external sanctions service without pseudonymisation	Deployment architecture	The adopting organisation, which controls what is placed in the payload before cross-boundary transmission
A2A task metadata creates a permanent bilateral record of data sharing	Protocol design; cross-organisational governance	Protocol maintainers (A2A), who define the task-record model; the two organisations jointly, through a data processing agreement
Derived behavioural profile enables re-identification at Card Network Y	Cross-organisational governance	The receiving organisation (Card Network Y), bound by an agreement with Bank X governing re-use of the transmitted profile
Full risk-assessment context on intra-organisational A2A exceeds data minimisation	Deployment architecture	The adopting organisation, which controls the context passed in A2A task objects
Cardholder structurally unaware of the automated multi-agent processing	Use case; system-level governance	All parties jointly, and ultimately the regulator: no single actor can introduce data-subject notification alone
No access or contestation mechanism exists for the cardholder	Use case; system-level governance	All parties jointly, and ultimately the regulator: providing access or contestation is a system-level and legal decision

Together, the limitations and contributions set out in this chapter frame how far the findings can be carried; the consolidated answer to the research question follows in Chapter 10.

10 Conclusion

This chapter draws the thesis to a close. It does not introduce new analysis. Section 10.1 answers the three sub-questions in turn, Section 10.2 gives a direct and consolidated answer to the main research question, Section 10.3 states the principal limitations and directions for future research, and Section 10.4 reflects on the broader significance of the findings. Before those answers, Table 10.1 summarises what each analytical chapter produced and how it feeds the answer to the main research question, including the attribution of the eight threats on which the protocol-level answer rests.

Table 10.1: Outputs of Chapters 5–10 and their role in answering the main research question. The lower panel gives the attribution of the eight threats: the protocol-level answer is the protocol-inherent subset, the partial pair is qualified, and the use-case-inherent threats are reported but excluded from protocol-level claims. Author's own elaboration.

Chapter	Output (what it produces)	Role toward the MRQ answer
Ch. 5 Threat identification	54 threat entries across nine flow groups; four cross-cutting structural patterns	Surfaces the full threat space and begins separating protocol- from use-case-driven risk
Ch. 6 Prioritisation	Eighteen Critical and nineteen High; eight threats described and classified by attribution	Performs the protocol-inherent versus use-case-inherent attribution on which the answer rests
Ch. 7 Expert validation	Seven interviews: all eight realistic, six confirmed by a majority; five out-of-scope concerns	External check; the protocol-inherent threats received the broadest confirmation
Ch. 8 Discussion	Two registers (structural severity and contextual urgency); boundary over protocol; privacy as a governance problem	Interprets severity and weight; shows the protocol-level finding is structural, not incidental
Ch. 9 Reflection	Limitations; three scientific contributions; four transferable structural patterns	Marks the protocol-level answer apart from the broader, system-level lens
Ch. 10 Conclusion	Answers SQ1–SQ3 and the main research question; seven future-research directions	States the answer: eight threats, of which four protocol-inherent, two partial, two use-case-inherent

Protocol-inherent (4)	Partial (2)	Use-case-inherent (2)
TH-01 session linkability	TH-03 cross-boundary transmission	TH-07 unawareness
TH-02 cross-store aggregation	TH-05 re-identification at the card network	TH-08 unintervenability
TH-04 bilateral non-repudiation		
TH-06 excessive task context		

10.1. Answer to the Sub-questions

SQ1: How do MCP and A2A function within the use case, and how can the system be represented as a DFD?

MCP and A2A serve two distinct roles in the fraud detection use case. MCP governs the vertical interaction between an agent and the data sources within an organisation, operating through a stateful session in which an agent issues tool calls to internal stores. A2A governs the horizontal interaction between agents, operating through a task lifecycle in which one agent delegates work to another and receives a result. The use case realises all four protocol configurations simultaneously, MCP within and across organisations and A2A within and across organisations, which is what makes it analytically suitable for isolating the privacy properties of each. The system was represented as a single Data Flow Diagram comprising three agent processes, five data stores, four external entities, eighteen data flows, and two organisational trust boundaries. This DFD, together with the data flow inventory, provided the complete structural input to the LINDDUN analysis, in which every identified threat is traced to a specific element or boundary in the diagram.

SQ2: What privacy threats are present in the system, and what is their relative severity?

Fifty-four privacy threat entries were identified across the eighteen data flows and nine flow groups by applying the LINDDUN mapping table and threat trees to each group. A qualitative prioritisation combining likelihood and impact placed eighteen entries at Critical priority and nineteen at High. From these, eight threats were selected for full description on the basis of protocol specificity, configuration coverage, and analytical significance. Of the eight, six are Critical priority (cross-store aggregation, cross-boundary transmission without pseudonymisation, the bilateral non-repudiable record, profile re-identification at the card network, cardholder unawareness, and cardholder unintervenability) and two

are High priority: session linkability and excessive intra-organisational task context. These priority levels reflect structural severity. They are not the same ordering as the degree to which each threat is caused by the protocols. Because the protocols are the object of study, the answer to the main research question rests on which threats are protocol-inherent rather than on which carry the highest severity, and the two orderings do not coincide. The protocol-inherent threats are identified and separated from the use-case-inherent ones in Section 10.2. Three patterns dominate the analysis. First, Unawareness is the only truly universal threat, present at every position in every flow group. Second, the organisational boundary is a stronger amplifier of threat severity than the choice of protocol: the same protocol mechanism produces a more severe threat once data crosses an organisational boundary. Third, MCP and A2A introduce different *kinds* of threat that call for different remedies: MCP's linkability is a structural default of its session model and is reducible only through architectural measures such as session scoping, whereas A2A's disclosure risk follows from what the protocol permits rather than what it enforces and is reducible through payload-level design discipline, so treating the two as the same problem leads to the wrong mitigations.

SQ3: To what extent are the identified threats recognised and validated by domain experts, and which additional threats do experts identify?

Seven structured expert interviews confirmed that all eight threats are recognised as realistic; no threat was rejected as implausible by all participants. The protocol-inherent threats received the broadest confirmation, with cross-store aggregation and profile re-identification most strongly endorsed by the financially experienced experts. The qualifications that experts raised concerned the contextual urgency of the threats rather than their structural presence; no qualification disputed that a threat was present. A majority argued that the cardholder-facing threats, cardholder unawareness and unintervenability, are reduced in contextual urgency by the statutory legal basis under which fraud detection operates, and the unanimous refinement of the bilateral non-repudiable record reflected the view that governance-dependent threats rank lower in effective priority than technically addressable ones. Beyond the inventory, experts surfaced five privacy-relevant concerns that fall outside the flow-level scope of LINDDUN, most notably the language model reasoning layer as an opaque processing environment and the geopolitical jurisdiction risk of transmitting data to US-incorporated card networks. The effect of these qualifications is that the eight threats can be ordered in two distinct ways, by structural severity and by contextual urgency, which do not coincide; the two orderings are set side by side in Table 8.1 in Chapter 8.

10.2. Answer to the Main Research Question

This thesis set out to answer the following question:

> **"What privacy threats emerge from the use of MCP and A2A communication protocols, as identified through a LINDDUN-based analysis of a financial fraud detection scenario?"**

The analysis conducted in this thesis identifies eight privacy threats that arise from the use of MCP and A2A communication protocols in a financial multi-agent AI system. These eight threats span five of the six assessed LINDDUN categories, Linkability, Identifiability, Non-repudiation, Data Disclosure, and Unawareness and Unintervenability, together with the four protocol configurations present in the use case. Detectability was assessed throughout the elicitation in Chapter 5, and Detectability threats were identified, but none of them was among the eight threats selected for full description. The eight were chosen for their protocol specificity, configuration coverage, and analytical significance, and the Detectability findings scored lower on these criteria than the threats that were selected.

Four of the eight threats are protocol-inherent: they arise directly from specific design decisions in the MCP and A2A specifications and would appear in any system where these protocols are deployed, regardless of the application domain. Session linkability arises because MCP's session architecture persistently links all tool calls within a session to the same data subject in the logs. Cross-store aggregation arises because MCP delivers full responses from all queried data stores to the agent simultaneously, producing a combined profile that exceeds the sensitivity of any individual store. The bilateral non-repudiable record arises because A2A's task lifecycle model creates permanent bilateral records on both sides of every inter-agent interaction that neither party can unilaterally remove. Excessive task context arises because A2A's task context object carries the full assessment record to the receiving agent by default, with no built-in mechanism to restrict what is sent to what the recipient strictly needs. These four threats are findings about MCP and A2A as protocols rather than about the financial domain. As properties of the protocol designs, they are present in any faithful deployment and therefore

transfer to other sectors, although the severity weighting of each threat and the legal basis that governs it remain domain-specific, as set out in Chapter 9.

Two threats are use-case-inherent: they arise from the nature of automated fraud detection rather than from the protocols specifically and would be present in any comparable system regardless of which communication technology was used. Cardholder unawareness and unintervenability reflect the structural exclusion of the data subject from a processing chain that concerns them directly, as is already the case in the conventional machine-learning fraud-detection systems banks operate today. MCP and A2A do not cause these conditions; they provide no mechanism to remedy them either.

Two threats sit between the poles. Cross-boundary transmission without pseudonymisation is partly protocol-inherent: MCP transmits whatever data the agent includes in the message and imposes no minimisation constraint by default. Profile re-identification at the card network is partly use-case-inherent: the re-identification risk derives from the data holdings of the card network and would exist independently of A2A. A2A's contribution is not merely to transport the profile but to make its cross-boundary transfer a routine, automated operation, bringing it within reach of those holdings far more readily than would otherwise be the case. Read together, three of these threats form a cross-protocol chain rather than standing as isolated findings: MCP first aggregates a behavioural profile across Bank X's internal stores, A2A then carries that profile across the organisational boundary, and the receiving card network can re-identify it against its own data holdings. Each link is comparatively benign in isolation; the privacy risk emerges only from the two protocols operating in sequence within one system, which is why a single-protocol analysis could not have surfaced it.

Expert validation confirmed that all eight threats are recognised as realistic by practitioners. The four protocol-inherent threats received the broadest confirmation, with no expert challenging their plausibility. The interviews additionally surfaced five privacy-relevant concerns that fall outside the LINDDUN analytical scope, set out in the answer to SQ3 above and discussed in Chapter 8.

10.3. Limitations and Future Research

The findings are bounded by four principal limitations, discussed in full in Chapter 9. The analysis operates on a static Data Flow Diagram of a hypothetical system, and therefore does not capture threats arising from the autonomous, runtime behaviour of LLM-based agents or from flows not represented in the model. LINDDUN, as a flow-level framework, cannot see privacy risks located in the inference behaviour of the language models themselves, in the authorisation architecture, or in the discovery phase that precedes the modelled flows. The prioritisation is qualitative and reflects the analytical judgement of the researcher, mitigated but not eliminated by the expert validation. Finally, the structural findings are tied to the specific protocol versions analysed (MCP revision 2025-03-26 and A2A version 0.2.5), which were under active development throughout the research. The findings are also grounded in a single financial use case: the design-inherent threats transfer to other sectors as a matter of protocol architecture, but their severity weighting and legal context require reassessment in each new domain.

These limitations define seven directions for future research. First, the privacy-enhancing measures suggested for the eight threats should be evaluated empirically to determine which most effectively reduce risk without compromising the functional value of the protocols. Second, the LINDDUN approach could be extended, or combined with the GenAI threat-tree extensions of Liao et al. [65] and with an agentic-AI framework such as MAESTRO, to capture the autonomous flows and inference-level risks that fall outside the present scope. Third, a multi-sector comparative analysis, applying the same method to healthcare, human-resources, or public-sector deployments, would test the expected transferability of the design-inherent findings. Fourth, longitudinal tracking of how the privacy properties of MCP and A2A change across specification versions would establish whether the structural threats identified here are resolved, amplified, or carried forward as the protocols evolve. Fifth, the re-identification risk identified at the card network warrants empirical validation against real card-network data holdings, which could not be examined within the scope of this thesis. Sixth, a dedicated legal analysis should assess how MCP- and A2A-based multi-agent systems fall under the emerging regulatory frameworks for AI, most notably the EU Artificial Intelligence Act [53]. The Act entered into force during this research, and its provisions and accompanying standards were still being put into practice throughout it. Such an analysis requires a legal rather than a technical methodology, and was therefore outside the present scope. Seventh, this thesis assesses the privacy threats of MCP and A2A as given and does not evaluate whether alternative protocol designs would avoid the protocol-inherent threats identified here. Because four of the eight threats follow from the protocols' specified session and task models

rather than from the use case, a natural question is whether a different design, for instance one incorporating selective disclosure, data-subject participation, or capability-scoped context objects, would mitigate them at the protocol level. A comparative privacy evaluation across agent communication protocols, including the more specification-immature options such as ANP and Agora once they mature, would establish whether the threats documented here are specific to MCP and A2A or general to the machine-to-machine agentic paradigm. A more fundamental version of the same question, raised in Chapter 8, is whether the right answer for the most privacy-sensitive settings is a safer agentic protocol at all, or a deliberate decision not to use agentic AI in those settings, in favour of more bounded systems whose data flows are constrained and whose task outcomes are deterministic and auditable. Establishing where that trade-off tips is a direction for future work that the present analysis can frame but not resolve.

10.4. Broader Significance

The findings of this thesis carry a significance that extends beyond the specific protocols and use case analysed. MCP and A2A are the first generation of open, interoperable standards for agentic AI communication. The privacy properties they exhibit, including the structural exclusion of the data subject, the absence of built-in minimisation mechanisms, and the production of non-repudiable cross-organisational records, will be reproduced in every system built on these standards unless they are addressed at the governance level. A technical fix applied by a single deploying organisation cannot correct a structural property of the protocol layer.

This is the central lesson for Complex Systems Engineering and Management. The privacy risks identified in this thesis are not engineering failures that can be resolved by better code. They are emergent properties of a sociotechnical system in which multiple organisations, multiple protocols, and multiple governance frameworks interact without a common accountability structure for the data subject. Addressing them requires coordination between protocol designers, deploying organisations, and regulators, precisely the kind of multi-actor, multi-level governance challenge that the CoSEM discipline is designed to analyse and address. Reaching that analysis required integrating the four kinds of knowledge that define the CoSEM perspective: the technical, in the protocol-level design choices and the data-flow modelling; the institutional, in the allocation of governance responsibility across designers, deploying organisations, and regulators; the economic, in the externality and incentive structure that leaves privacy under-provided; and the social, in the position of the cardholder as a data subject excluded from a process that concerns them. It is the combination, rather than any one of these alone, that brings the privacy problem into view and locates it as a design challenge no single discipline could frame on its own. The contribution of this thesis is therefore not only a catalogue of privacy threats but a demonstration that privacy in agentic AI systems is fundamentally a governance problem, and that the tools of sociotechnical systems analysis are necessary to make that problem visible.

Full Reference List

[1] Doug Bonderud. *AI Decision-Making: Where Do Businesses Draw the Line?* IBM Think Insights. Quotes the 1979 IBM training statement “A computer can never be held accountable…” IBM. 2024. URL: https://www.ibm.com/think/insights/ai-decision-making-where-do-businesses-draw-the-line (visited on 05/27/2026).

[2] Mustafa Suleyman and Michael Bhaskar. *The Coming Wave: Technology, Power, and the Twenty-First Century's Greatest Dilemma*. Quoted passage on p. 1: “Soon you will live surrounded by AIs. They will organize your life, operate your business, and run core government services.” New York: Crown, 2023. ISBN: 9780593593950.

[3] Xinpeng Hong, Changgang Zheng, and Noa Zilberman. “In-Network Machine Learning for Real-Time Transaction Fraud Detection”. In: *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI 2024)*. Vol. 392. Frontiers in Artificial Intelligence and Applications. Open access; documents that low-latency real-time fraud detection is a core requirement for financial institutions, with systems operating at microsecond-scale latency to match the speed of modern payment fraud. IOS Press, 2024. DOI: 10.3233/faia240828.

[4] O. Martínez, Patricia Sánchez, and Elena Mira Alcaraz. “A Review of Machine Learning and Deep Learning Approaches for Fraud Detection Across Financial and Supply Chain Domains”. In: *Research Square* (2025). Preprint under review; systematic review of 80+ studies (2015–2025) identifying real-time processing constraints and the integration of anomalous patterns across data as core requirements for fraud detection. DOI: 10.21203/rs.3.rs-7893661/v1. URL: https://doi.org/10.21203/rs.3.rs-7893661/v1.

[5] Jianfeng Li and Dexiang Yang. “Research on Financial Fraud Detection Models Integrating Multiple Relational Graphs”. In: *Systems* 11.11 (2023). Builds fraud detection on a heterogeneous information network combining user–device, user–merchant, and user–address signals, evidencing that effective detection depends on combining identity, device, and behavioural data rather than any single source, p. 539. DOI: 10.3390/systems11110539. URL: https://doi.org/10.3390/systems11110539.

[6] Yang Liu et al. “Pick and Choose: A GNN-based Imbalanced Learning Approach for Fraud Detection”. In: *Proceedings of the Web Conference 2021 (WWW '21)*. Establishes that fraud detection relies on rich, multifaceted behavioural interactions represented across heterogeneous account–device graphs. ACM, 2021, pp. 3168–3177. DOI: 10.1145/3442381.3449989. URL: https://doi.org/10.1145/3442381.3449989.

[7] Xinyi Hou et al. “Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions”. In: *ACM Transactions on Software Engineering and Methodology* (2026). Published 2026; earlier circulated as arXiv preprint 2503.23278 (2025). DOI: 10.1145/3796519. URL: https://doi.org/10.1145/3796519.

[8] Rajesh Kumar Singh et al. “Applications of Artificial Intelligence for Optimizing Banking and Financial Operations: A Systematic Literature Review and Future Research Agenda”. In: *Journal of Economic Surveys* 39.5 (2025), pp. 2284–2302. DOI: 10.1111/joes.12693.

[9] Helen Nissenbaum. “Privacy as Contextual Integrity”. In: *Washington Law Review* 79.1 (2004). https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10. Introduces contextual integrity: a privacy violation occurs when personal data flows in ways that violate the norms of the context in which it was originally shared; directly applicable to cross-organisational agent data flows, pp. 119–157.

[10] Zeynab Anbiaee et al. “Security Threat Modeling for Emerging AI-Agent Protocols: A Comparative Analysis of MCP, A2A, Agora, and ANP”. In: *arXiv* (2026). doi: 10.48550/arXiv.2602.11327, https://doi.org/10.48550/arXiv.2602.11327.

[11] K. Wuyts, R. Scandariato, and W. Joosen. “Empirical evaluation of a privacy-focused threat modeling methodology”. In: *Journal of Systems and Software* 96 (2014). doi: 10.1016/j.jss.2014.05.075, https://doi.org/10.1016/j.jss.2014.05.075, pp. 122–138.

[12] V. Patil, E. Stengel-Eskin, and M. Bansal. “The Sum Leaks More Than Its Parts: Compositional Privacy Risks and Mitigations in Multi-Agent Collaboration”. In: *arXiv* (2025). doi: 10.48550/arXiv.2509.14284, https://doi.org/10.48550/arXiv.2509.14284.

[13] M. Wooldridge and N. R. Jennings. “Intelligent agents: Theory and practice”. In: *The Knowledge Engineering Review* 10.2 (1995). doi: 10.1017/S0269888900008122, https://doi.org/10.1017/S0269888900008122, pp. 115–152.

[14] Michael Wooldridge. *An Introduction to MultiAgent Systems*. 2nd. Wiley Publishing, 2009. ISBN: 0470519460.

[15] M. Luck, P. McBurney, and C. Preist. *Agent Technology: Enabling Next Generation Computing (A Roadmap for Agent Based Computing)*. https://eprints.soton.ac.uk/257309/. AgentLink, 2003.

[16] A. Vaswani et al. “Attention is all you need”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 30. doi: 10.48550/arXiv.1706.03762, https://doi.org/10.48550/arXiv.1706.03762. NeurIPS proceedings: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. 2017, pp. 5998–6008.

[17] J. Devlin et al. “BERT: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. doi: 10.18653/v1/N19-1423, https://doi.org/10.18653/v1/N19-1423. 2019, pp. 4171–4186.

[18] Wayne Xin Zhao et al. “A Survey of Large Language Models”. In: *arXiv preprint* arXiv:2303.18223 (2023). Widely cited survey covering LLM capabilities including summarisation, question answering, code generation, and multi-step reasoning. DOI: 10.48550/arXiv.2303.18223. URL: https://arxiv.org/abs/2303.18223.

[19] R. Sapkota, K. I. Roumeliotis, and M. Karkee. “AI agents vs. agentic AI: A conceptual taxonomy, applications, and challenges”. In: *Information Fusion* (2026). doi: 10.1016/j.inffus.2025.103599, https://doi.org/10.1016/j.inffus.2025.103599.

[20] Ziwei Ji et al. “Survey of Hallucination in Natural Language Generation”. In: *ACM Computing Surveys* 55.12 (2023). Canonical peer-reviewed survey of hallucination in LLMs; 2,701 citing publications; defines hallucination as the generation of content that is plausible in form but factually incorrect or unsupported by the source, pp. 1–38. DOI: 10.1145/3571730.

[21] M. A. Ferrag et al. “From prompt injections to protocol exploits: Threats in LLM-powered AI agents workflows”. In: *ICT Express* (2025). doi: 10.1016/j.icte.2025.12.001, https://doi.org/10.1016/j.icte.2025.12.001.

[22] Lei Wang et al. “A survey on large language model based autonomous agents”. In: *Frontiers of Computer Science* 18.6 (2024). doi: 10.1007/s11704-024-40231-1, https://doi.org/10.1007/s11704-024-40231-1, p. 186345.

[23] Lewis Hammond et al. *Multi-Agent Risks from Advanced AI*. Preprint; not yet peer-reviewed. Identifies risks in multi-agent LLM systems including trust breakdown, emergent collective behaviour, and collusion between agents from different organisations. 2025. DOI: 10.48550/arxiv.2502.14143. arXiv: 2502.14143 [cs.AI].

[24] Alan Chan et al. *Infrastructure for AI Agents*. Preprint; not yet peer-reviewed. Analyses the infrastructure requirements for cross-organisational AI agent deployments, including open ecosystems where agent identities and data flows may not be fully visible to any single party. 2025. DOI: 10.48550/arxiv.2501.10114. arXiv: 2501.10114 [cs.AI].

[25] Allan Dafoe et al. “Cooperative AI: machines must learn to find common ground”. In: *Nature* 593.7857 (2021), pp. 33–36. DOI: 10.1038/d41586-021-01170-0.

[26] J. M. Such, A. Espinosa, and A. García-Fornes. “A survey of privacy in multi-agent systems”. In: *The Knowledge Engineering Review* 29.3 (2014). doi: 10.1017/S0269888913000180, https://doi.org/10.1017/S0269888913000180, pp. 314–344.

[27] Nicholas R. Jennings. “On agent-based software engineering”. In: *Artificial Intelligence* 117.2 (2000), pp. 277–296. DOI: 10.1016/s0004-3702(99)00107-1.

[28] Yingxuan Yang et al. *A Survey of AI Agent Protocols*. Preprint; not yet peer-reviewed. Surveys the current landscape of AI agent protocols including MCP, A2A and competing proposals, analysing trade-offs between expressiveness, security and interoperability. 2025. DOI: 10.48550/arxiv.2504.16736. arXiv: 2504.16736 [cs.AI].

[29] Yedidel Louck, Ariel Stulman, and Amit Dvir. *Security Analysis of Agentic AI Communication Protocols: A Comparative Evaluation*. Preprint; not yet peer-reviewed. First empirical comparative security analysis of agentic AI communication protocols; evaluates A2A, CORAL and ACP across 14-point vulnerability taxonomy. 2025. DOI: 10.48550/arxiv.2511.03841. arXiv: 2511.03841 [cs.CR].

[30] A2A Project. *Agent2Agent (A2A) Protocol Specification, Version 0.2.5*. Accessed: 2026-04-13. 2025. URL: https://a2a-protocol.org/v0.2.5/specification/.

[31] Martin Leo et al. "From threat to trust: Assessing security risks of agentic AI systems". In: *International Journal of Information Security* 25.1 (2026). DOI: 10.1007/s10207-025-01185-y.

[32] Laurens Sion, Dimitri Van Landuyt, and Kim Wuyts. "Privacy Risk Assessment for Data Subject-Aware Threat Modeling". In: *Proceedings of the IEEE Security and Privacy Workshops (SPW)*. Establishes that security threat modelling is insufficient for privacy: privacy risk must be assessed from the data subject perspective, not only from the organisation's perspective. IEEE, 2019, pp. 64–71. DOI: 10.1109/spw.2019.00023.

[33] Yunhao Yao, Zhiqiang Wang, and Haoran Cheng. "IntentMiner: Intent Inversion Attack via Tool Call Analysis in the Model Context Protocol". In: *arXiv preprint* arXiv:2512.14166 (2025). doi: 10.48550/arXiv.2512.14166, https://doi.org/10.48550/arXiv.2512.14166.

[34] Shouju Wang et al. *Privacy in Action: Towards Realistic Privacy Mitigation and Evaluation for LLM-Powered Agents*. Preprint; not yet peer-reviewed. Introduces PrivacyChecker (a contextual-integrity-based mitigation) and PrivacyLens-Live, a dynamic benchmark recreating MCP and A2A environments to measure privacy leakage by LLM agents at runtime. https://doi.org/10.48550/arXiv.2509.17488. 2025. arXiv: 2509.17488 [cs.CR].

[35] E. Yu and L. M. Cysneiros. "Designing for Privacy in a Multi-agent World". In: *Workshop on Deception, Fraud and Trust in Agent Societies*. doi: 10.1007/3-540-36609-1_16, https://doi.org/10.1007/3-540-36609-1_16. Springer, 2003, pp. 209–223.

[36] G. Piolle, Y. Demazeau, and J. Caelen. "Privacy management in user-centred multi-agent systems". In: *International Workshop on Engineering Societies in the Agents World*. doi: 10.1007/978-3-540-75524-1_20, https://doi.org/10.1007/978-3-540-75524-1_20. Springer, 2007, pp. 354–367.

[37] Martijn Warnier and Frances M. T. Brazier. "Anonymity services for multi-agent systems". In: *Web Intelligence and Agent Systems* 8.2 (2010). Presents a pseudonym- and handle-based anonymity service for agent platforms (implemented for AgentScape), directly addressing the linkability of agent interactions. Extended journal version of the earlier IAS 2007 conference paper "Organized Anonymous Agents"., pp. 219–232. DOI: 10.3233/wia-2010-0188.

[38] S. Maliah, R. Brafman, and G. Shani. "Increased privacy with reduced communication in multi-agent planning". In: *Proceedings of the International Conference on Automated Planning and Scheduling*. Vol. 27. doi: 10.1609/icaps.v27i1.13821, https://doi.org/10.1609/icaps.v27i1.13821. 2017, pp. 209–217.

[39] M. Kossek and M. Stefanovic. "Survey of recent results in privacy-preserving mechanisms for multi-agent systems". In: *Journal of Intelligent & Robotic Systems* 110.3 (2024). doi: 10.1007/s10846-024-02161-9, https://doi.org/10.1007/s10846-024-02161-9, p. 129.

[40] Omid Mirzamohammadi et al. "Security and Privacy Threat Analysis for Solid". In: *Proceedings of the IEEE Secure Development Conference (SecDev)*. Applies both STRIDE and LINDDUN to the Solid communication protocol in a financial analytics use case; confirms that privacy threats in protocol specifications are systematically missed by security-only analyses. IEEE, 2023, pp. 196–206. DOI: 10.1109/secdev56634.2023.00033.

[41] Xin Jie Chua et al. "Banking Done Right: Redefining Retail Banking with Language-Centric AI". In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Documents Ryt Bank, the first regulator-approved deployment of conversational AI as a primary banking interface, orchestrated by four LLM-powered agents; evidences active deployment of multi-agent AI in the financial sector. Also available as arXiv:2510.07645. Suzhou, China: Association for Computational Linguistics, 2025, pp. 646–658.

[42] John H. Holland. *Emergence: From Chaos to Order*. Oxford: Oxford University Press, 1998.

[43] Eric Trist. “The Evolution of Socio-technical Systems”. In: *Perspectives on Organization Design and Behavior*. Ed. by Andrew H. Van de Ven and William F. Joyce. New York: Wiley, 1981, pp. 19–75.

[44] M. Deng et al. “A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements”. In: *Requirements Engineering* 16.1 (2011). doi: 10.1007/s00766-010-0115-7, https://doi.org/10.1007/s00766-010-0115-7, pp. 3–32.

[45] Cloud Security Alliance. *Agentic AI Threat Modeling Framework: MAESTRO*. Accessed: 2026-04-07. 2025. URL: https://cloudsecurityalliance.org/blog/2025/02/06/agentic-ai-threat-modeling-framework-maestro.

[46] NIST. *Privacy Framework | NIST*. 2025. URL: https://www.nist.gov/privacy-framework (visited on 03/10/2026).

[47] Kim Wuyts and Wouter Joosen. *LINDDUN privacy threat modeling: a tutorial*. CW Reports CW685. LIRIAS1652885. Available at https://lirias.kuleuven.be/1652885. Department of Computer Science, KU Leuven, July 2015.

[48] Kim Wuyts. *Personal communication on LINDDUN applicability to agentic AI protocols and on LINDDUN PRO reporting format*. Personal communication. Co-author of the original LINDDUN framework and a leading contributor to its development; formerly at the DistriNet Research Group, KU Leuven. Communicated by email, March 2026. Mar. 2026.

[49] Linux Foundation. *Linux Foundation Launches the Agent2Agent Protocol Project*. Accessed: 2026-04-13. 2025. URL: https://www.linuxfoundation.org/press/linux-foundation-launches-the-agent2agent-protocol-project-to-enable-secure-intelligent-communication-between-ai-agents.

[50] Anthropic. *Donating the Model Context Protocol and Establishing the Agentic AI Foundation*. Accessed: 2026-05-28. 2025. URL: https://www.anthropic.com/news/donating-the-model-context-protocol-and-establishing-of-the-agentic-ai-foundation.

[51] European Parliament and Council of the European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation)*. Official Journal of the European Union, L 119, 4 May 2016, pp. 1–88. CELEX 32016R0679. Accessed: 2026-05-30. 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[52] European Parliament and Council of the European Union. *Directive (EU) 2015/2366 of the European Parliament and of the Council of 25 November 2015 on Payment Services in the Internal Market (Payment Services Directive 2)*. Official Journal of the European Union, L 337, 23 December 2015, pp. 35–127. CELEX 32015L2366. Accessed: 2026-05-30. 2015. URL: https://eur-lex.europa.eu/eli/dir/2015/2366/oj.

[53] European Parliament and Council of the European Union. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*. Official Journal of the European Union, L series, 2024/1689, 12 July 2024. CELEX 32024R1689. Accessed: 2026-05-30. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

[54] Shaosheng Cao et al. “TitAnt: Online real-time transaction fraud detection in Ant Financial”. In: *Proceedings of the VLDB Endowment* 12.12 (2019), pp. 2082–2093. DOI: 10.14778/3352063.3352126.

[55] Anthropic. *Model Context Protocol Specification*. Accessed: 2026-04-07. 2025. URL: https://modelcontextprotocol.io/specification/2025-03-26.

[56] B. A. Forouzan. *Data Communications and Networking*. 5th. ISBN: 978-0-07-337622-6, New York, NY: McGraw-Hill, 2012.

[57] A. S. Tanenbaum and D. J. Wetherall. *Computer Networks*. 5th. ISBN: 978-0-13-212695-3, https://www.pearson.com/en-us/subject-catalog/p/computer-networks/P200000003188. Upper Saddle River, NJ: Pearson Prentice Hall, 2011.

[58] S. Spiekermann and L. F. Cranor. “Engineering Privacy”. In: *IEEE Transactions on Software Engineering* 35.1 (2009). doi: 10.1109/TSE.2008.88, https://doi.org/10.1109/TSE.2008.88, pp. 67–82.

[59] Anthropic. *Introducing the Model Context Protocol*. Accessed: 2026-04-07. 2024. URL: https://www.anthropic.com/news/model-context-protocol.

[60] Google. *Announcing the Agent2Agent Protocol (A2A)*. Accessed: 2026-04-13. 2025. URL: https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/.

[61] Ali Dorri, Salil S. Kanhere, and Raja Jurdak. "Multi-Agent Systems: A Survey". In: *IEEE Access* 6 (2018), pp. 28573–28593. DOI: 10.1109/ACCESS.2018.2831228.

[62] Tom DeMarco. *Structured Analysis and System Specification*. Yourdon Press, 1979.

[63] Edward Yourdon and Larry L. Constantine. *Structured Design: Fundamentals of a Discipline of Computer Program and Systems Design*. Prentice Hall, 1979.

[64] Laurens Sion, Kim Wuyts, and Wouter Joosen. *LINDDUN PRO: Interaction-Based Privacy Threat Elicitation Technique*. Online operational description of LINDDUN PRO. Canonical peer-reviewed reference is Sion, Wuyts & Yskout (2018), "Interaction-Based Privacy Threat Elicitation," IEEE EuroS&PW, doi: 10.1109/eurospw.2018.00017. 2023. URL: https://linddun.org/pro/ (visited on 05/30/2026).

[65] Qianying Liao et al. "A LINDDUN-based Privacy Threat Modeling Framework for GenAI". In: *arXiv preprint arXiv:2603.06051* (2026). doi: 10.48550/arXiv.2603.06051, https://doi.org/10.48550/arXiv.2603.06051.

[66] Barbara DiCicco-Bloom and Benjamin F. Crabtree. "The Qualitative Research Interview". In: *Medical Education* 40.4 (2006). doi: 10.1111/j.1365-2929.2006.02418.x, https://doi.org/10.1111/j.1365-2929.2006.02418.x, pp. 314–321.

[67] Steve Campbell et al. "Purposive Sampling: Complex or Simple? Research Case Examples". In: *Journal of Research in Nursing* 25.8 (2020). doi: 10.1177/1744987120927206, https://doi.org/10.1177/1744987120927206, pp. 652–661.

[68] Dalal Alharthi and Ivan Roberto Kawaminami Garcia. "A Call to Action for a Secure-by-Design Generative AI Paradigm". In: *arXiv preprint* arXiv:2510.00451 (2025). doi: 10.48550/arXiv.2510.00451, https://doi.org/10.48550/arXiv.2510.00451.

[69] Daryl McCullough. "Noninterference and the Composability of Security Properties". In: *Proceedings of the 1988 IEEE Symposium on Security and Privacy*. IEEE Computer Society Press, 1988, pp. 177–186.

[70] Ann Cavoukian. "Privacy by Design: The 7 Foundational Principles". In: *Information and Privacy Commissioner of Ontario* (2010). URL: https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf.

[71] Annabelle Gawer. "Bridging Differing Perspectives on Technological Platforms: Toward an Integrative Framework". In: *Research Policy* 43.7 (2014), pp. 1239–1249. DOI: 10.1016/j.respol.2014.03.006.

[72] A. Acquisti, C. Taylor, and L. Wagman. "The Economics of Privacy". In: *Journal of Economic Literature* 54.2 (2016). doi: 10.1257/jel.54.2.442, https://doi.org/10.1257/jel.54.2.442, pp. 442–492.

[73] DistriNet, KU Leuven. *LINDDUN Threat Trees*. 2024. URL: https://linddun.org/threat-trees/ (visited on 04/17/2026).

[74] Ibrahim Adabara, Bashir Olaniyi Sadiq, and Aliyu Nuhu Shuaibu. "A Review of Agentic AI in Cybersecurity: Cognitive Autonomy, Ethical Governance, and Quantum-Resilient Defense". In: *F1000Research* 14 (2025). Peer-reviewed (F1000 open peer review), gold OA. References the MAESTRO framework for Agent-to-Agent protocols as a guiding approach for resilient agentic AI systems., p. 843. DOI: 10.12688/f1000research.169337.1.

[75] Sourya Joyee De and Daniel Le Métayer. "PRIAM: A Privacy Risk Analysis Methodology". In: *Data Privacy Management and Security Assurance (DPM 2016), Lecture Notes in Computer Science*. Peer-reviewed Springer LNCS paper proposing PRIAM, a privacy risk analysis methodology organised around data catalogues, feared events, and harm trees; 71 Smart Citations. Springer International Publishing, 2016, pp. 221–229. DOI: 10.1007/978-3-319-47072-6_15.

[76] Samuel Wairimu, Leonardo Horn Iwaya, and Lothar Fritsch. “On the Evaluation of Privacy Impact Assessment and Privacy Risk Assessment Methodologies: A Systematic Literature Review”. In: *IEEE Access* 12 (2024). Peer-reviewed (IEEE Access), gold OA. Systematic literature review of PIA and PRA methodologies; classifies LINDDUN as a privacy threat model (PTM) and PRIAM as a privacy risk assessment (PRA), and positions them as complementary., pp. 19625–19650. DOI: 10.1109/access.2024.3360864.

[77] Alex Sangers et al. “Secure Multiparty PageRank Algorithm for Collaborative Fraud Detection”. In: *Financial Cryptography and Data Security*. Lecture Notes in Computer Science. Documents cross-institutional transaction graph analysis where each institution holds only a partial view of cardholder activity; directly supports the use case architecture in Chapter 4. Springer, 2019, pp. 605–623. DOI: 10.1007/978-3-030-32101-7_35.

[78] Financial Action Task Force. *International Standards on Combating Money Laundering and the Financing of Terrorism & Proliferation — The FATF Recommendations*. FATF, Paris. Adopted February 2012, updated October 2025. Recommendation 10 (customer due diligence) and Recommendation 6 (targeted financial sanctions) establish the screening obligations referenced in Chapter 4. 2025. URL: https://www.fatf-gafi.org/en/publications/Fatfrecommendations/Fatf-recommendations.html.

[79] Tim Finin et al. “KQML as an Agent Communication Language”. In: *Proceedings of the Third International Conference on Information and Knowledge Management*. ACM, 1994, pp. 456–463. DOI: 10.1145/191246.191322.

A Privacy Threat Trees

This appendix reproduces the seven LINDDUN threat trees used in the elicitation, one for each threat category: Linkability (Figure A.1), Identifiability (Figure A.2), Non-repudiation (Figure A.3), Detectability (Figure A.4), Data Disclosure (Figure A.5), Unawareness and Unintervenability (Figure A.6), and Non-compliance (Figure A.7).

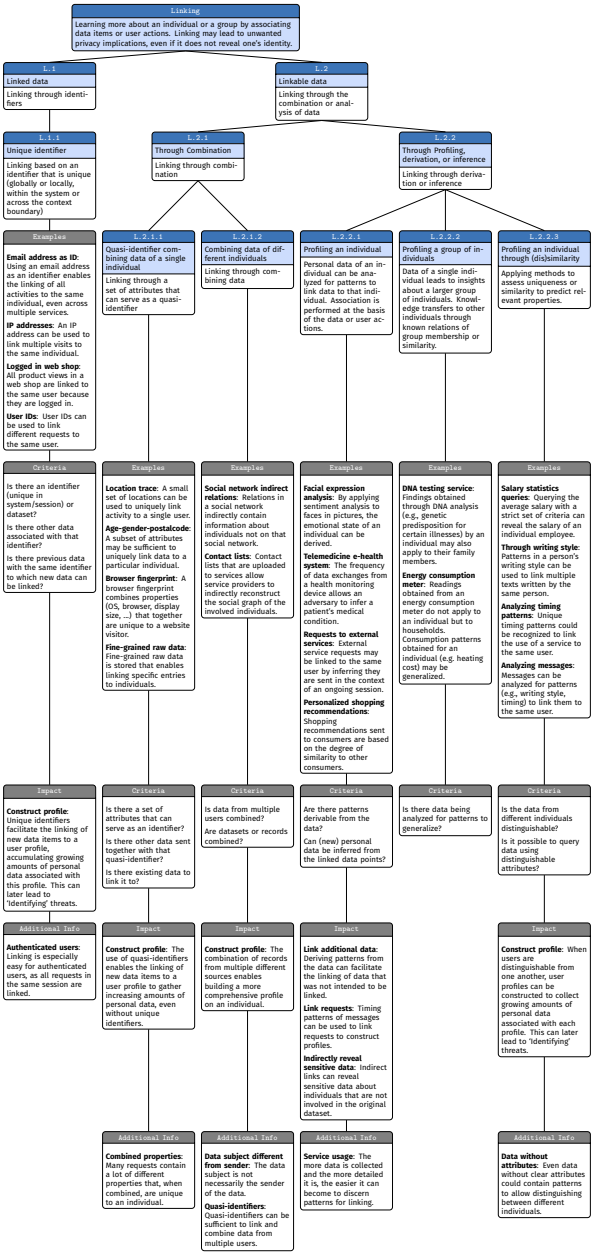

Figure A.1: LINDDUN Linking threat tree, detailing the sub-characteristics of the Linkability threat category [73].

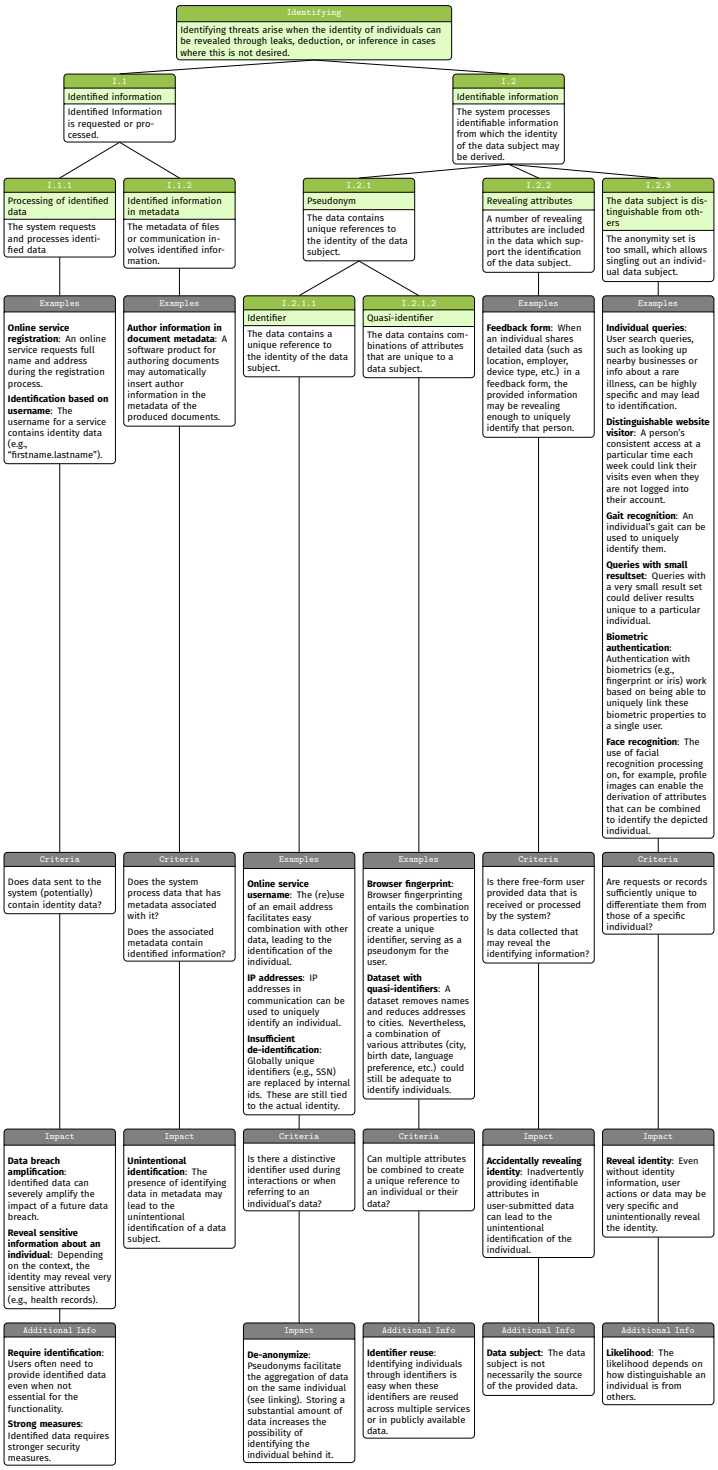

Figure A.2: LINDDUN Identifying threat tree, detailing the sub-characteristics of the Identifiability threat category [73].

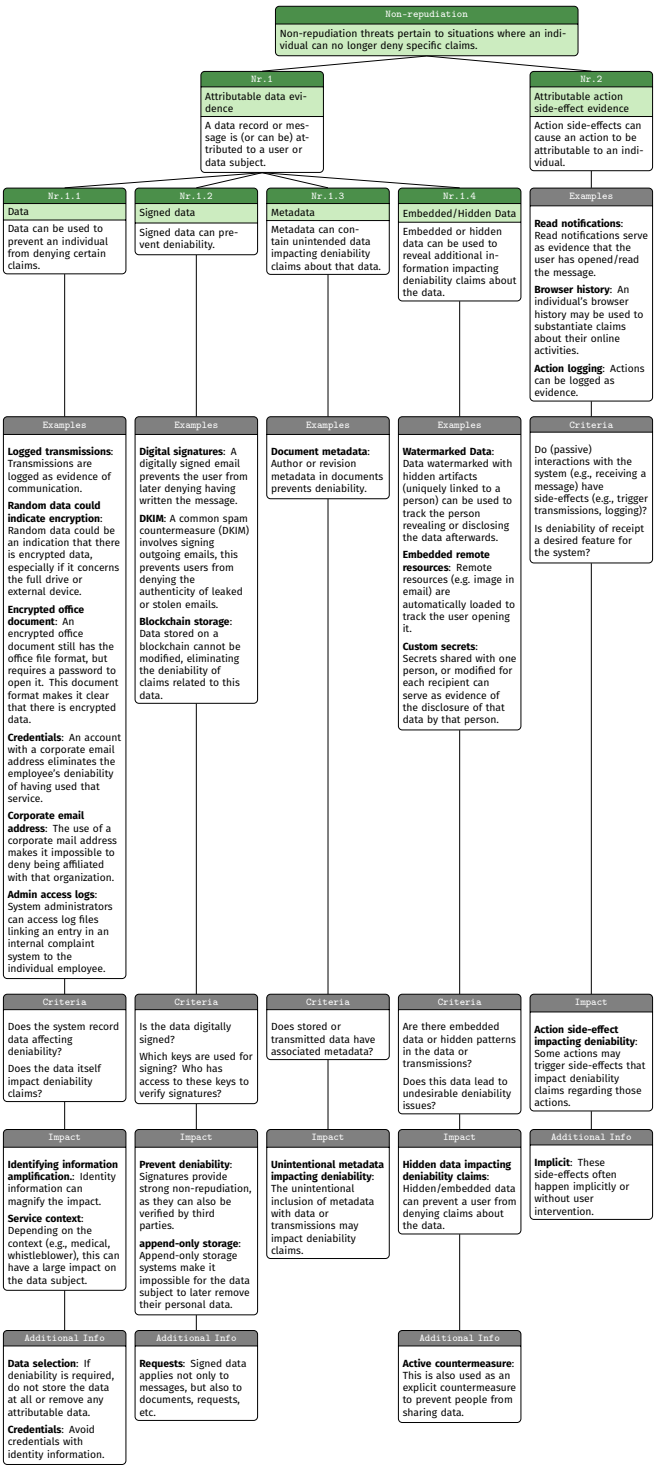

Figure A.3: LINDDUN Non-repudiation threat tree, detailing the sub-characteristics of the Non-repudiation threat category [73].

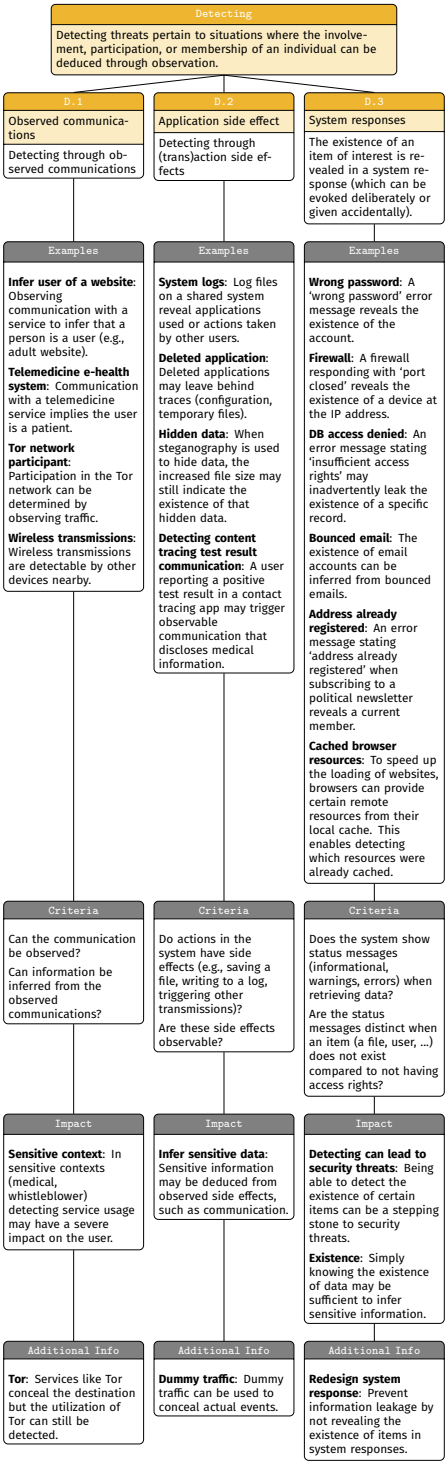

Figure A.4: LINDDUN Detecting threat tree, detailing the sub-characteristics of the Detectability threat category [73].

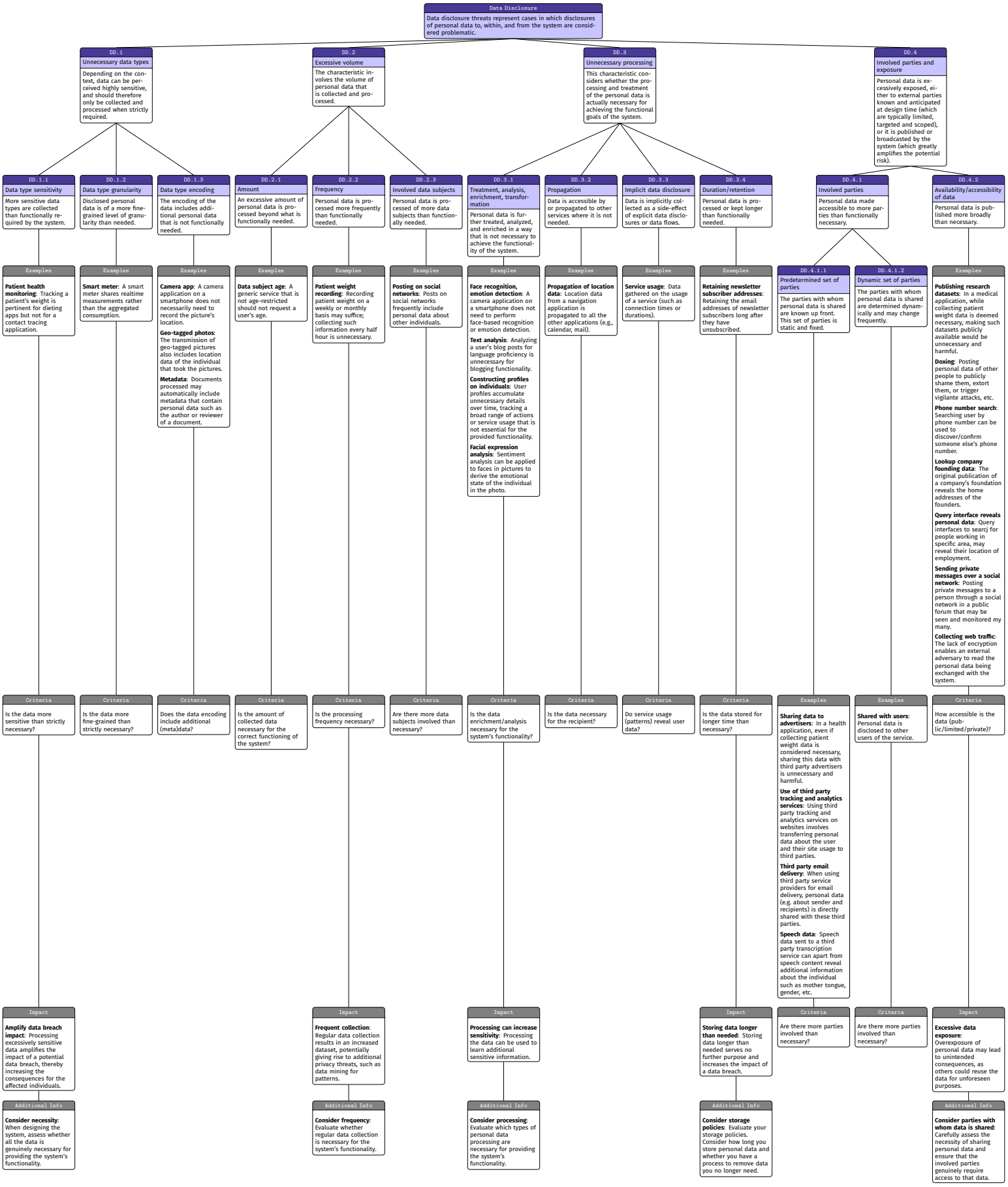

Figure A.5: LINDDUN Data Disclosure threat tree, detailing the sub-characteristics of the Data Disclosure threat category [73].

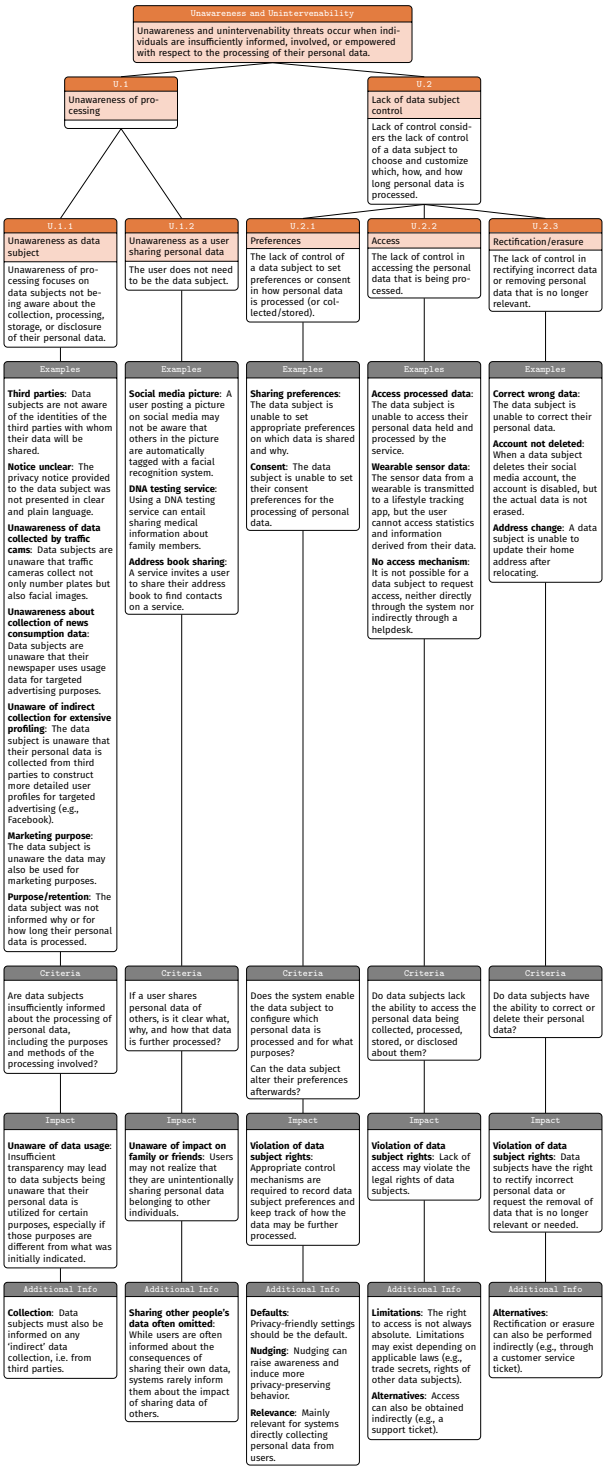

Figure A.6: LINDDUN Unawareness and Unintervenability threat tree, detailing the sub-characteristics of the Unawareness and Unintervenability threat category [73].

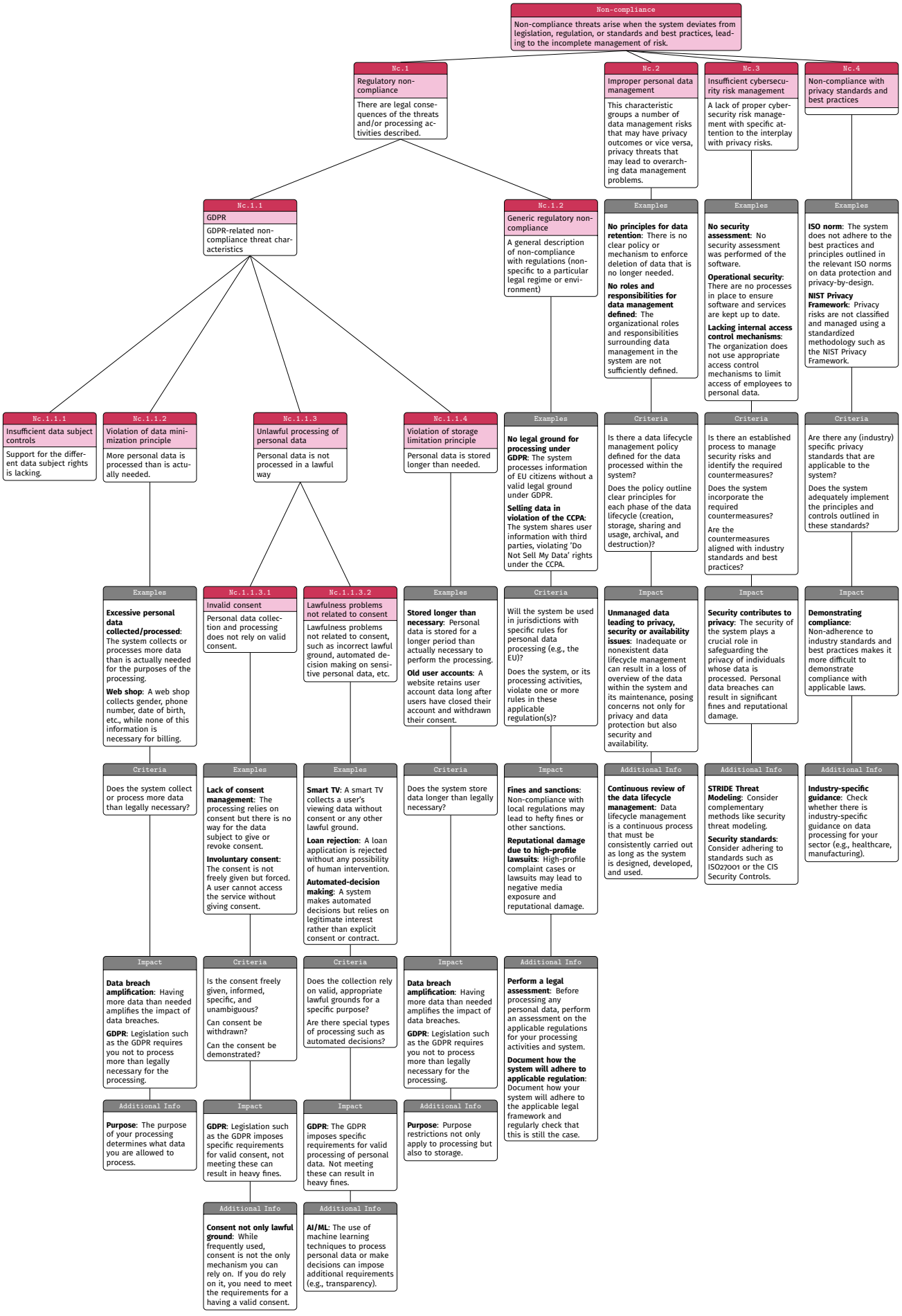

Figure A.7: LINDDUN Non-compliance threat tree, detailing the sub-characteristics of the Non-compliance threat category [73].

B AI Disclosure

This thesis was developed with the support of generative artificial intelligence (AI) tools, used during the period of this research. The general-purpose and research-support tools used were OpenAI's ChatGPT, Google's NotebookLM and Gemini, Anthropic's Claude, Microsoft Copilot, and the literature-search assistants Scite and Elicit. Some of these assistants were accessed through Model Context Protocol (MCP) connectors: Scite, for instance, was queried over an MCP connection to retrieve and cross-check sources, which gave the author first-hand familiarity with one of the two protocols that this thesis studies. For the processing of the expert-interview material specifically, the enterprise version of Microsoft 365 Copilot provided through Avanade was used, as described below. AI support was used only as a writing, editing, and research-support aid. It helped improve language, structure, clarity, and consistency throughout the thesis, and supported literature search and the handling of interview transcripts. AI was not used to generate independent research findings, fabricate data, or replace the author's own analysis and judgment.

More specifically, AI was used for the following purposes.

Language and clarity improvement
AI was used to improve grammar, sentence structure, readability, and academic tone. Drafts written by the author were refined to make the text clearer and more coherent, while the original meaning and argument remained the author's own.

Structure and formatting support
AI was used to help organize sections of the thesis according to academic requirements. This included improving the structure of chapters, aligning subsections, and supporting formats such as the methodology section.

Rewriting and editing
AI was used to rewrite text in a clearer way, shorten or expand sections, and convert notes or bullet points into fluent academic text.

Conceptual clarification
AI was occasionally used to clarify concepts and definitions and to reformulate explanations in a more precise way. These clarifications were always based on literature selected, interpreted, and evaluated by the author.

Consistency checking
AI was used to check whether the research question, sub-questions, methodology, and chapter structure were aligned and internally consistent.

Processing of expert-interview material
For the expert interviews, the enterprise version of Microsoft 365 Copilot provided through Avanade was used to support the summarisation, paraphrasing, and anonymisation of the interview transcripts, which at that stage still contained personally identifiable information about the participants. This tool was chosen deliberately with privacy in mind: unlike consumer AI tools, it operates under enterprise data protection, meaning that the data is processed within the organisational tenant and is not used to train foundation models. The summaries and paraphrases produced were reviewed, corrected, and interpreted by the author, and the anonymisation was verified by the author before any interview material was used in the thesis. The reading, interpretation, and analysis of the interviews remained the author's own work.

Examples of prompts and outputs
The following pairs are representative of the prompts used and the kind of output they produced. They illustrate the supporting role the tools played; in every case the underlying content, judgement, and final wording remained the author's own. The outputs shown are paraphrased and abbreviated.

- *Language and clarity.*
  Prompt: "Improve the grammar, readability, and academic tone of the following paragraph without changing its meaning: [paragraph]."
  Representative output: a cleaner version of the same paragraph, with the argument and technical content unchanged.
- *Logical order of the thesis.*
  Prompt: "Is my thesis now in a logical order? Here are my chapter titles with a one-line summary of each: [list]."

Representative output: confirmation that the progression from problem and research gap, through methodology, the protocols, the case, the elicitation and prioritisation, and expert validation, to discussion and conclusion is coherent, with a reminder to introduce each key term before its first use.

- *Comparison with other theses.*
  Prompt: "Compare my thesis against other TU Delft CoSEM master's theses in terms of structure and scope. Based on them, is anything missing from mine?"
  Representative output: the thesis contains the elements common to comparable work (a problem and gap, an explicit methodology, a worked case, results, a discussion, and a conclusion); the two aspects worth making more visible are the explicit link between the contribution and the prior literature, and the societal and CoSEM framing in the discussion.
- *Rewriting notes into prose.*
  Prompt: "Rewrite these notes into fluent academic prose, adding nothing new: MCP = client–server; A2A = agent-to-agent; both emerged 2024–2025; neither privacy-first."
  Representative output: "MCP follows a client–server model and A2A an agent-to-agent model. Both protocols emerged in 2024–2025, and neither was designed with privacy as a primary goal."
- *Conceptual clarification.*
  Prompt: "Based on the LINDDUN documentation, explain in one or two sentences the difference between Unawareness and Unintervenability."
  Representative output: Unawareness concerns knowledge (the data subject does not know that, or how, their data is processed; Unintervenability concerns agency) they cannot access, correct, or contest that processing.
- *Consistency checking.*
  Prompt: "Here are my main research question, three sub-questions, and chapter titles. Check whether they are aligned and flag any inconsistencies: [list]."
  Representative output: the sub-questions map onto the main question and the chapter sequence, with a note to introduce one key term earlier than its current first use.

AI was not used to:

- replace the author's own selection, reading, and critical assessment of the literature (search assistants supported finding sources, but all sources were read, selected, and evaluated by the author);
- make methodological decisions on behalf of the author;
- generate original research findings or conclusions;
- fabricate data, sources, or references.

All final decisions regarding the research design, literature selection, interpretation, methodology, analysis, and argumentation were made by the author. The author takes full responsibility for the final content of this thesis.

C Interview Guide

This appendix contains the interview guide used in the expert interviews described in Chapter 7. The guide was shared with participants in advance of the interview, together with the eight threat descriptions from Chapter 6 and the participant information sheet. The interview consists of four phases.

Phase 1: Introduction (10 minutes)

The opening phase establishes the expert's background and their familiarity with the topics covered in the thesis. This information is used to contextualise their assessments in the analysis.

Q1.1: Personal background

- Could you briefly introduce yourself?
- What is your educational background?
- What is your current role, and what does it involve day-to-day?
- Could you walk me through the roles you have held previously that are most relevant to this research?

Q1.2: Familiarity with AI and agentic AI

- Do you have experience working with AI systems? If so, in what capacity?
- Are you familiar with the concept of agentic AI, systems in which agents autonomously execute multi-step tasks, use tools, and communicate with other agents?
- Have you worked with or encountered agentic AI systems in a professional context?

Q1.3: Familiarity with MCP and A2A

- Are you familiar with MCP (Model Context Protocol) or A2A (Agent-to-Agent protocol)? If so, in what context?
- Have you used, deployed, or reviewed systems that use either of these protocols?

Q1.4: Security and privacy background

- Do you have experience in security, privacy, or data protection? If so, could you describe the nature of that experience?
- Have you been involved in privacy or security assessments of AI systems or communication protocols?

Phase 2: Thesis Context and Technical Background (15 minutes)

This phase explains the research context, the two protocols, the use case with the Data Flow Diagram, and the analytical framework. The goal is to ensure the expert has sufficient understanding of the technical and privacy context before assessing the threats.

Security versus privacy: why the distinction matters. Before explaining the protocols, it is important to clarify the distinction between security and privacy, because it determines what this thesis is and is not about.
Security is concerned with protecting a system and its data assets on behalf of the *deploying organisation*. A security analysis asks: can an external attacker gain unauthorised access? Are authentication mechanisms functioning correctly? Can the system be manipulated by an adversary? Security threats include prompt injection, token theft, supply chain attacks, and authentication weaknesses.
Privacy is concerned with protecting personal data on behalf of the *data subject*, the individual whose data is being processed. A privacy analysis asks: is more data collected than necessary? Can interactions be linked back to a specific person? Does the data subject know what is happening to their data? Can they do anything about it? Privacy threats include excessive aggregation, re-identification, unintended disclosure, and structural exclusion of the data subject from the process.

The critical point is that a system can be *fully secure* and still produce serious privacy violations. A correctly functioning MCP session identifier that links three database queries to the same cardholder is not a security vulnerability, it is working exactly as designed. But from the data subject's perspective, it collapses the separation between databases that the organisation deliberately created as a privacy safeguard. This thesis focuses exclusively on this second kind of risk: threats to the data subject that arise from the normal, correct operation of MCP and A2A, not from attacks against them.

Thesis overview. This master's thesis applies the LINDDUN privacy threat modelling framework to two agentic AI communication protocols: MCP and A2A. LINDDUN is a structured method that identifies privacy threats at the level of individual data flows between system components. The thesis asks which privacy threats arise from these protocols in a financial fraud detection system, and to what extent practitioners recognise those threats as realistic.

What MCP is: and why it matters for privacy. MCP (Model Context Protocol) is an open protocol introduced by Anthropic in 2024. It defines how an AI agent communicates with external tools and data sources. An MCP *host* application runs one or more MCP *clients*, each of which maintains a stateful session with an MCP *server* that exposes data or tools. The agent calls tools on the server, for example, querying a customer database or invoking a screening service, and the server returns structured results.
The key privacy-relevant property of MCP is the **session identifier**. Every tool call within a single session carries the same session ID. This means that if an agent queries three different databases in one session, say, transaction history, identity records, and device data, all three queries are linked by the session identifier. Anyone with access to the session logs can reconstruct exactly which data about which customer was queried, when, and in what sequence, even if the individual databases are separately governed. The session model also means that all results return to the same agent within the same session context, allowing the agent to combine data from multiple sources into a single profile that no individual database would produce on its own.
A second privacy-relevant property is that MCP has **no built-in data minimisation mechanism**. The protocol does not ask whether a query is proportionate; it simply executes it. An agent can request and receive more data than the task requires, and the protocol places no constraint on this.

What A2A is: and why it matters for privacy. A2A (Agent-to-Agent protocol) is an open protocol introduced by Google in 2025. It defines how two autonomous AI agents communicate across organisational or system boundaries. An A2A *client agent* delegates a task to a *server agent*, which executes it autonomously and returns a result as an *artifact*. The server agent's internal reasoning, tools, and data access are fully opaque to the client.
The key privacy-relevant property of A2A is the **persistent task identifier**. Every A2A interaction is assigned a task ID that both the sending and receiving agent retain independently. This means that if Agent A at Organisation 1 sends a profile to Agent B at Organisation 2, there is a permanent record of that interaction at both sides that neither party can unilaterally delete. This creates non-repudiable bilateral evidence of data sharing for every task that passes through the protocol.
A second privacy-relevant property is that A2A transmits **a full task context object** by default. When Agent A delegates a task to Agent B, it sends everything it knows about the case. There is no built-in mechanism that asks what Agent B actually needs to complete its task versus what Agent A simply happens to have available.

The use case: financial fraud detection at Bank X and Card Network Y. The thesis analyses a hypothetical but realistic financial fraud detection workflow. When a cardholder taps their card, Bank X's fraud detection agent automatically assesses the transaction. If fraud signals are detected, the following sequence runs, all invisible to the cardholder:

1. Bank X's agent uses **MCP** to query three internal databases: transaction history (DF2/DF3), KYC identity records (DF4/DF5), and device fingerprint data (DF6/DF7). All three queries share the same MCP session identifier.
2. Bank X's agent uses **MCP** to query an external sanctions screening service (DF8/DF9), sending the cardholder's name and date of birth in plaintext and receiving a sanctions match result.
3. Bank X's agent uses **A2A** to share a derived behavioural profile with Card Network Y's Risk Intelligence Agent (DF10). Card Network Y queries its own network-wide database (DF12/DF13) and returns a risk score (DF11).

4. If the risk score falls in an ambiguous range, Bank X's agent uses **A2A** internally to escalate the case (DF14) to a human analyst (DF15/DF16), who makes the final decision (DF17).
5. The cardholder receives a payment decision (DF18).

The Data Flow Diagram from Chapter 4 is presented here as a visual reference. The diagram shows all 18 data flows, the two trust boundaries (TB1 between Bank X and the external world, TB2 between Bank X and Card Network Y), and the five system components (P1 = Bank X's Fraud Detection Agent, P2 = Bank X's Internal Review Agent, P3 = Card Network Y's Risk Intelligence Agent, EE1 = Cardholder, EE2 = Payment Terminal, EE3 = Human Analyst, EE4 = External Sanctions Service, DS1–DS5 = data stores).

The analytical approach. LINDDUN identifies six privacy threat categories. For each of the 18 data flows in the diagram, the analysis asks which of the six categories applies. The result is a threat register of 54 entries, prioritised by likelihood and impact. Eight threats were selected for detailed description and form the basis of Phase 3.

Phase 3: Threat Validation (30 minutes)

The eight threat descriptions were shared in advance. For each description, the expert is asked two questions after a brief verbal summary. The goal is not to test knowledge but to assess whether the threat is recognised as realistic and whether the priority reflects real-world risk.

Q3.1: Plausibility (asked per threat)

Do you recognise this as a realistic privacy threat in agentic AI systems using MCP or A2A? Please explain your reasoning.

- **Yes, I recognise this in practice.**
- **Yes, but only under specific conditions** (please describe).
- **Uncertain** (please describe what would help you assess this).
- **No, I do not consider this realistic** (please explain why).

Q3.2: Priority (asked per threat)

The threat has been assigned a priority of Critical or High. Does this reflect the risk as you experience it in practice?

- **The priority is appropriate.**
- **The priority is too high** (please explain).
- **The priority is too low** (please explain).
- **I cannot assess the priority** (please explain).

Q3.3: Urgency ranking (asked at end of Phase 3)

Now that we have discussed all eight threats, could you rank them from most to least urgent, that is, which threats do you consider most important to address first, and why? There is no right or wrong answer; I am interested in your personal assessment based on your professional experience.

Please rank the following threats from 1 (most urgent) to 8 (least urgent):

Rank	Threat
_____	TH-01: MCP session identifier enables cross-store linkability
_____	TH-02: MCP return flows aggregate identified cardholder data
_____	TH-03: Identified personal data to external sanctions service
_____	TH-04: A2A task metadata creates non-repudiable bilateral evidence
_____	TH-05: Derived behavioural profile enables re-identification
_____	TH-06: Full risk assessment context exceeds data minimisation
_____	TH-07: Cardholder structurally unaware of automated processing
_____	TH-08: No access or contestation mechanism for the cardholder

The eight threats are summarised below with a brief verbal introduction for each. The full descriptions were sent in advance.

TH-01: MCP session identifier enables cross-store linkability. *Priority: High.*
In one sentence: *The session identifier that MCP uses to link database queries also destroys the separation between databases that the organisation deliberately built.*
Bank X stores transaction history, KYC records, and device data in separate databases to limit exposure. MCP queries all three within a single session, and all three queries carry the same session identifier. An insider or attacker with log access can reconstruct exactly what data was accessed about which cardholder, when, and in what order, information that no individual database would reveal. *Privacy sensitivity: the session model structurally removes the separation that organisational data governance deliberately created.*

TH-02: MCP return flows aggregate identified cardholder data at the querying agent. *Priority: Critical.*
In one sentence: *MCP makes it trivial for an agent to combine data from three separate databases into a complete personal profile that no individual database was ever meant to hold.*
P1 receives name, address, date of birth, transaction history, and device fingerprint in a single session. None of the individual databases holds this complete profile. MCP makes combining them trivial. If P1 is compromised or the data is logged unnecessarily, the cardholder's complete personal profile is exposed. *Privacy sensitivity: the aggregated profile is the direct input to TH-05, it is what gets sent to Card Network Y.*

TH-03: Identified personal data transmitted to external sanctions service without pseudonymisation. *Priority: Critical.*
In one sentence: *The moment name and date of birth leave Bank X's systems via MCP, the bank loses all governance over what happens to that data.*
DF8 sends the cardholder's name and date of birth to an external service outside Bank X's control. Once the data crosses the boundary, Bank X cannot enforce how it is stored, shared, or retained. A false positive, the cardholder incorrectly matched to a sanctions entry, can cause transaction refusal or account suspension, and the cardholder never knows their data left the bank. *Privacy sensitivity: loss of governance over identified personal data is irreversible once the data is transmitted.*

TH-04: A2A task metadata creates non-repudiable bilateral evidence of data sharing. *Priority: Critical.*

In one sentence: *Every A2A interaction creates a permanent record at both sides that neither organisation can ever unilaterally delete.*
Every A2A interaction between Bank X and Card Network Y is assigned a task ID that both parties retain independently. Neither can delete the other's copy. For every transaction assessed this way, there is a permanent record at Card Network Y that Bank X shared a cardholder profile with them, even if the sharing was later found to be erroneous or unlawful. *Privacy sensitivity: this is unique to A2A; MCP cross-boundary flows do not produce bilateral evidence.*

TH-05: Derived behavioural profile enables re-identification at Card Network Y. *Priority: Critical.*
In one sentence: *A profile with no name is not anonymous if the recipient already knows enough to match it to a specific person.*
The profile on DF10 contains no name, only transaction velocity, geographic patterns, and device consistency. Bank X considers this anonymous. Card Network Y processes transactions from millions of cardholders and can match the received pattern against its own data. Once re-identified, Card Network Y can build a risk profile of the specific cardholder using data Bank X never provided and cannot control. *Privacy sensitivity: anonymity depends on what the recipient already knows, not on what was sent.*

TH-06: Full risk assessment context on intra-organisational A2A exceeds data minimisation. *Priority: High.*
In one sentence: *A2A sends everything by default, so the human analyst ends up with a complete personal dossier when all they needed was a risk summary.*
When P1 delegates the case to P2 via A2A, it sends the full context: name, KYC data, device fingerprint, sanctions result, risk score. P2's only function is to route the case to a human analyst. It does not need all of this. A2A sends the full task object by default. The result is that the human analyst receives a complete personal dossier when they only needed a risk summary. *Privacy sensitivity: excess disclosure at P2 is the direct cause of the Critical threat at the human analyst stage.*

TH-07: Cardholder structurally unaware of automated multi-agent processing. *Priority: Critical.*
In one sentence: *The cardholder taps their card and receives a decision, and everything that happened in between is permanently invisible to them by design.*
The cardholder taps their card. Three seconds later they receive a yes or no. In between: three internal databases queried, an external sanctions service contacted, a profile sent to another company, a second AI agent invoked, possibly a human analyst reading their full dossier. The cardholder knows none of this. Neither MCP nor A2A has a notification mechanism for the data subject. This is not a configuration choice, it is structural. *Privacy sensitivity: applies to every cardholder in every transaction assessed by this system.*

TH-08: No access or contestation mechanism exists for the cardholder. *Priority: Critical.*
In one sentence: *Even if the cardholder knew everything that happened to their data, there is no pathway in the system for them to request, correct, or contest anything.*
Even if the cardholder knew what had happened, they could do nothing about it. They cannot request which data was used, see the profile sent to Card Network Y, find out if an analyst reviewed their case, or appeal the decision through the system. MCP and A2A are machine-to-machine protocols with no data-subject-facing interface. *Privacy sensitivity: TH-07 means the cardholder is uninformed; TH-08 means that even if they were informed, they are unable to intervene. Together they are the most consequential finding of the thesis.*

Phase 4: Open Reflection (10 minutes)

This phase invites the expert to reflect freely on the analysis and share any perspectives that have not yet come up.

Q4.1: Missing threats

Based on your experience, are there privacy threats from MCP or A2A that you consider significant but that are not in the eight descriptions? If so: what is the threat, how does it arise from the protocol, what data or data subjects are affected, and how would you prioritise it?

Q4.2: Gaps in the analysis

The analysis covers only the flows explicitly modelled in the system diagram. It does not cover threats from autonomous agent behaviour at runtime or inference risks. Do you consider these omissions significant for the completeness of the analysis?

Q4.3: General observations

Is there anything else related to privacy in agentic AI, MCP, or A2A that you consider relevant and that has not come up?

Q4.4: Governance and responsibility

In this system, no single organisation designed or controls the full multi-agent workflow. Who do you think should be responsible for the privacy of the cardholder in a distributed multi-agent system like this, the deploying organisations, the protocol designers, the regulator, or some combination?

Q4.5: Adoption trajectory

Do you expect MCP and A2A to be widely deployed in the financial sector in the near term? Does that expectation affect how urgent you consider the identified privacy risks to be?

Closing: Thank you for your time. Your responses will be fully anonymised. You will appear in the thesis only as an expert identifier with a brief role description. Please let me know if you wish to review the relevant section before publication.

D Threat Elicitation: Full Technical Evidence

Chapter 5 presents the privacy threat findings as a narrative account, following the cardholder's data through the fraud detection workflow step by step. This appendix contains two things. First, the full prioritised threat register from Chapter 6 (Table D.1), listing all fifty-four entries with their likelihood, impact, and priority scores. Second, the complete technical evidence underlying the Chapter 5 findings: the LINDDUN mapping table, the flow grouping rationale, and, for each of the nine flow groups, a consolidated overview table (one row per LINDDUN category) and a position-level node table showing exactly which threat tree nodes were identified at which positions. This appendix is the primary reference for methodological reviewers and supervisors who wish to verify the analytical basis of the findings in Chapters 5 and 6.

Non-compliance (N_C) is excluded from all tables throughout this appendix because its analysis requires a legal methodology rather than a technical threat modelling approach, as established in Section 2.3.1.

Full prioritised threat register

Table D.1 presents the complete prioritised threat register. One row corresponds to one LINDDUN category per flow group. The Gr and Cat. columns identify the entry; L and I show the likelihood and impact scores; Priority shows the resulting level. Hi = High, Me = Medium, Lo = Low. The scoring criteria are defined in Section 6.2.

Table D.1: Prioritised threat register. One row per LINDDUN category per flow group. L = Likelihood, I = Impact. **Crit.** Critical **High** **Med.** **Low**. Author's own analysis.

Gr	Cat.	L	I	Priority
H, DF1: Payment Initiation				
H	L	Hi	Me	**High**
H	I	Hi	Me	**High**
H	N_R	Me	Lo	**Low**
H	D	Me	Lo	**Low**
H	D_D	Me	Me	**Med.**
H	U	Hi	Me	**High**
A, DF2,4,6,12: MCP Query Flows				
A	L	Hi	Me	**High**
A	I	Hi	Me	**High**
A	N_R	Me	Lo	**Low**
A	D	Me	Lo	**Low**
A	D_D	Hi	Me	**High**
A	U	Hi	Me	**High**
B, DF3,5,7,13: MCP Return Flows				
B	L	Hi	Hi	**Crit.**
B	I	Hi	Hi	**Crit.**
B	N_R	Hi	Me	**High**
B	D	Me	Lo	**Low**
B	D_D	Hi	Me	**High**
B	U	Hi	Hi	**Crit.**
C, DF8: MCP Cross-boundary Out				
C	L	Hi	Hi	**Crit.**
C	I	Hi	Hi	**Crit.**
C	N_R	Hi	Me	**High**
C	D	Me	Me	**Med.**
C	D_D	Hi	Hi	**Crit.**
C	U	Hi	Hi	**Crit.**
D, DF9: MCP Cross-boundary In				
D	L	Hi	Hi	**Crit.**
D	I	Hi	Hi	**Crit.**
D	N_R	Hi	Me	**High**
D	D	Me	Me	**Med.**
D	D_D	Hi	Me	**High**
D	U	Hi	Hi	**Crit.**

Gr	Cat.	L	I	Priority
E, DF10,11: A2A Cross-org.				
E	L	Hi	Hi	**Crit.**
E	I	Hi	Hi	**Crit.**
E	N_R	Hi	Hi	**Crit.**
E	D	Me	Me	**Med.**
E	D_D	Hi	Me	**High**
E	U	Hi	Hi	**Crit.**
F, DF14,17: A2A Intra-org.				
F	L	Hi	Me	**High**
F	I	Hi	Me	**High**
F	N_R	Me	Lo	**Low**
F	D	Me	Lo	**Low**
F	D_D	Hi	Me	**High**
F	U	Hi	Me	**High**
G, DF15,16: Human Analyst				
G	L	Hi	Hi	**Crit.**
G	I	Hi	Hi	**Crit.**
G	N_R	Hi	Me	**High**
G	D	Me	Me	**Med.**
G	D_D	Hi	Hi	**Crit.**
G	U	Hi	Hi	**Crit.**
I, DF18: Terminal Response				
I	L	Me	Lo	**Low**
I	I	Me	Me	**Med.**
I	N_R	Me	Lo	**Low**
I	D	Me	Lo	**Low**
I	D_D	Me	Lo	**Low**
I	U	Hi	Me	**High**

LINDDUN mapping table

The LINDDUN mapping table determines which threat categories are applicable to each element-type combination, and at which of the three **positions**: the **source** (S, the component sending the data), the **data flow** (FL, the data in transit), and the **destination** (Dst, the component receiving the data). A greyed-out position indicates the category is structurally inapplicable there. Two exclusions apply throughout: Detectability does not apply at the destination, and Unawareness does not apply at the data flow. The resulting mapping is shown in Table D.2.

Table D.2: LINDDUN mapping table restricted to the five element-type combinations present in the use case [73]. Greyed positions are structurally inapplicable. Non-compliance excluded.

Source	Dest.	L	I	N_R	D	D_D	U
Process	Process	S-FL-Dst	S-FL-Dst	S-FL-Dst	S-FL-Dst	S-FL-Dst	S-FL-Dst
Process	DataStore	S-FL-Dst	S-FL-Dst	S-FL-Dst	S-FL-Dst	S-FL-Dst	S-FL-Dst
Process	Ext. Entity	S-FL-Dst	S-FL-Dst	S-FL-Dst	S-FL-Dst	S-FL-Dst	S-FL-Dst
DataStore	Process	S-FL-Dst	S-FL-Dst	S-FL-Dst	S-FL-Dst	S-FL-Dst	S-FL-Dst
Ext. Entity	Process	S-FL-Dst	S-FL-Dst	S-FL-Dst	S-FL-Dst	S-FL-Dst	S-FL-Dst

Flow grouping rationale

The eighteen data flows in the use case were grouped into nine groups for analysis. Flows sharing the same element-type combination, protocol, and trust boundary context were analysed together. This approach was reviewed at a high level by a co-author of the original LINDDUN framework, who identified no inconsistencies in it, in documented personal correspondence [48]. The resulting groups are listed in Table D.3.

Table D.3: Data flow groups, constituent flows, element-type combinations, and structural rationale for grouping. Author's own elaboration.

Group	Flows	Element combination	Rationale
H	DF1	Ext. Entity → Process (no protocol)	Payment initiation from the cardholder; entry point; richest inbound payload; no protocol.
A	DF2, DF4, DF6, DF12	Process → DataStore (MCP)	Outgoing MCP query flows; all share the session identifier as a structural linkability mechanism.
B	DF3, DF5, DF7, DF13	DataStore → Process (MCP)	Incoming MCP record flows; return counterparts of Group A; carry full records, substantially raising privacy exposure.
C	DF8	Process → Ext. Entity (MCP, cross-boundary)	Only outgoing cross-boundary MCP flow; transmits identified personal data outside TB1.
D	DF9	Ext. Entity → Process (MCP, cross-boundary)	Incoming cross-boundary MCP result; attributable via session context.
E	DF10, DF11	Process → Process (A2A, cross-org.)	Cross-organisational A2A pair; derived profile to Card Network Y; risk score returned.
F	DF14, DF17	Process → Process (A2A, intra-org.)	Intra-organisational A2A pair; DF14 is the most data-rich payload in the workflow; no built-in minimisation.
G	DF15, DF16	Process/Ext. Entity → Ext. Entity/Process (no protocol)	Human analyst interaction; DF15 delivers the full aggregated assessment to an analyst exercising independent judgment.
I	DF18	Process → Ext. Entity (no protocol, cross-boundary)	Terminal payment response; minimal payload.

MCP groups (A–D) A2A groups (E, F) Unshaded: boundary/human actor flows (H, G, I).

Group index

Table D.4 maps each flow group to the section in which it is analysed.

Table D.4: Flow groups and their correspondence to Chapter 5 sections and appendix sections. Author's own elaboration.

Group	Description	Chapter 5 section	Appendix section
A	Outgoing MCP queries (intra-org.)	5.3	D.2
B	Incoming MCP return flows (intra-org.)	5.3	D.3
C	Outgoing MCP flow (cross-boundary)	5.4	D.4
D	Incoming MCP result (cross-boundary)	5.4	D.5
F	A2A pair (intra-org.)	5.5	D.6
E	A2A pair (cross-org.)	5.6	D.7
H	Cardholder payment initiation	5.7	D.8
G	Human analyst interaction	5.7	D.9
I	Terminal payment response	5.7	D.10

D.1. Per-configuration Scoring Rationale

This section gives the rationale for each priority assignment summarised in the heatmap of Section 6.4 (Table 6.4), configuration by configuration.

MCP within an organisation (Groups A and B). Group A receives High across Linkability, Identifiability, Data Disclosure, and Unawareness. The session identifier is structural to MCP and creates cross-store linkability in every deployment, but the impact remains Medium because the flows carry query keys rather than full records and all processing stays within TB1. Group B elevates Linkability, Identifiability, and Unawareness to Critical: the return flows carry name, date of birth, and transaction history, raising the data type factor substantially. Non-repudiation remains High rather than Critical because the evidence of processing is held internally and Bank X retains full governance control over it. Data Disclosure rises to High in Group B because full records are returned where a minimised subset would suffice.

MCP between organisations (Groups C and D). Linkability, Identifiability, Data Disclosure, and Unawareness are Critical in both groups: the loss of governance control at TB1 substantially raises the impact for the categories most directly tied to personal data content. Non-repudiation receives High rather than Critical because the evidence of cross-boundary disclosure, while held by an external party, does not by itself enable further harm to the cardholder beyond the disclosure already captured under Data Disclosure. Detectability receives Medium in both groups: the transmission is observable in principle but requires access to network logs that are not routinely accessible to the cardholder or to Bank X's governance function.

A2A within an organisation (Group F). All categories receive High or below. The absence of a built-in minimisation mechanism creates a structural Data Disclosure risk, but because the data stays within TB1 the impact is Medium rather than High. Unawareness receives High: the cardholder is unaware not only of data access but of autonomous decision-making about their transaction, yet the processing stays within the deploying organisation, which retains the ability to implement post-hoc transparency measures. Non-repudiation and Detectability receive Low, consistent with intra-organisational flows where internal evidence is under governance control and the relevant actors are identifiable.

A2A between organisations (Group E). Linkability, Identifiability, and Unawareness are Critical. Non-repudiation is the only category that is Critical here but not in the MCP cross-boundary groups: the bilateral evidence structure of A2A means that neither party can unilaterally remove the record of the exchange, which is a qualitatively stronger governance constraint than MCP's one-sided server log. Data Disclosure receives High rather than Critical because the transmitted payload on DF10 is a derived profile rather than directly identifying attributes, and the minimisation risk is configurational rather than structural.

Human actor flows (Group G). Linkability, Identifiability, Data Disclosure, and Unawareness are Critical. The analyst receives the most comprehensive cardholder profile in the entire workflow and can act on it through informal channels outside the system's governance boundary. Non-repudiation receives

High rather than Critical: human actions are legally attributable but the evidence is internal and under Bank X's governance control. Detectability receives Medium: unlike automated protocol flows, human analyst access is in principle auditable through access logs, though the cardholder still has no visibility.

How to read the elicitation tables

Each group below is presented with two tables. The **consolidated overview table** has one row per LINDDUN category. Its columns are: **Flow** (the data flows in this group), **Cat.** (the LINDDUN category), **Loc.** (the positions S/FL/Dst where this category applies), **Chars.** (the specific threat tree node codes, e.g. L.1.1 = Linkability, branch 1, sub-branch 1), and **Description** (plain-language explanation of why each node applies). The **position-level node table** presents the same information in a compact matrix. *n/a* indicates a structurally excluded position per the mapping table above.

D.2. Group A: Outgoing MCP Query Flows

Group A consists of four Process-to-DataStore MCP flows: DF2 (P1 → DS1), DF4 (P1 → DS2), DF6 (P1 → DS3), and DF12 (P3 → DS4). In each flow, an agent process issues a structured MCP query carrying parameters derived from cardholder data to a data store. DF2, DF4, and DF6 are intra-organisational within TB1; DF12 is intra-organisational within TB2. Per the mapping table, Detectability does not apply at the destination and Unawareness and Unintervenability does not apply at the data flow position; these exclusions apply throughout this group.

The per-position LINDDUN assessment for this group is set out in Table D.5.

Table D.5: Group A: Outgoing MCP Query Flows (P → DS), flows DF2, DF4, DF6, DF12. S = source, FL = data flow, Dst = destination. Author's own analysis.

Flow	Cat.	Loc.	Chars.	Description
P1/P3 → DF2/4/6/12 → DS1–4	L	S,FL,Dst	L.1.1, L.2.1.1, L.2.2.1	Session ID links sequential queries to same transaction (S). Query parameters are quasi-identifiers across flows (FL). Query logs accumulate access profile (Dst).
	I	S,FL,Dst	I.1.1, I.2.1.1	KYC query directly references named individual (DF4). Card/account reference is unique identifier (other flows).
	N_R	S,FL,Dst	Nr.1.1, Nr.1.3	MCP session context creates attributable evidence of each access (S,Dst). Request metadata forms persistent record (FL).
	D	S,FL	D.1, D.2	Consecutive tool call pattern reveals assessment in progress (S). Query flows reveal check type on internal network (FL).
	D_D	S,FL,Dst	DD.1.2, DD.1.3, DD.3.4	Excess query granularity discloses more than required (S). Encoding carries non-functional metadata (FL). Query logs retained beyond operational need (Dst).
	U	S,Dst	U.1.1, U.2.2	Cardholder unaware of coordinated multi-store access (S). No mechanism to inspect queried stores (Dst).

Position-level threat nodes:

Pos.	L	I	N_R	D	D_D	U
S	L.1.1	I.1.1 (DF4); I.2.1.1	Nr.1.1	D.2	DD.1.2	U.1.1
FL	L.2.1.1	I.1.1 (DF4); I.2.1.1	Nr.1.3	D.1	DD.1.3	n/a
Dst	L.2.2.1	I.1.1 (DS2); I.2.1.1	Nr.1.1	n/a	DD.3.4	U.2.2

D.3. Group B: Incoming MCP Return Flows

Group B consists of four DataStore-to-Process MCP flows: DF3 (DS1 → P1), DF5 (DS2 → P1), DF7 (DS3 → P1), and DF13 (DS4 → P3). In each flow, a data store returns the requested records to the querying agent as a structured MCP response. Group B is the return counterpart of Group A: where Group A carried query parameters derived from cardholder data, Group B carries the full records retrieved in response to those queries. This structural difference makes the privacy threats in Group B more severe in several categories, as the payload now contains full personal data records rather than compact query keys. Per the mapping table, Detectability does not apply at the destination and Unawareness and Unintervenability does not apply at the data flow position; these exclusions apply throughout this group.

The per-position LINDDUN assessment for this group is set out in Table D.6.

Table D.6: Group B: Incoming MCP Return Flows (DS → P), flows DF3, DF5, DF7, DF13. S = source, FL = data flow, Dst = destination. Author's own analysis.

Flow	Cat.	Loc.	Chars.	Description
DS1–4 → DF3/5/7/13 → P1/P3	L	S,FL,Dst	L.1.1, L.2.1.1, L.2.2.1	Session ID links returned records to cardholder (S). Returned attributes form rich quasi-identifier across flows (FL). P1 accumulates cross-store profile (Dst).
	I	S,FL,Dst	I.1.1, I.2.1.1, I.2.2, I.2.3	DS2 returns directly identifying KYC attributes via DF5 (all). DS1/DS3 return card-linked records. Combined data singles out subject (Dst).
	N_R	S,FL,Dst	Nr.1.1, Nr.1.3	Response logs create attributable evidence of disclosure (S,Dst). Response metadata forms persistent record (FL).
	D	S,FL	D.1, D.2, D.3	Response generation is detectable side-effect (S). Flows observable on internal network; response structure may reveal record existence (FL).
	D_D	S,FL,Dst	DD.1.1, DD.1.3, DD.3.4	Full KYC record returned when only subset needed (S). Encoding carries non-functional metadata (FL). Records retained beyond immediate processing (Dst).
	U	S,Dst	U.1.1, U.2.2	Cardholder unaware records are retrieved and disclosed autonomously (S). No access mechanism at P1 (Dst).

Position-level threat nodes:

Pos.	L	I	N_R	D	D_D	U
S	L.1.1	I.1.1 (DS2); I.2.1.1; I.2.2	Nr.1.1	D.2	DD.1.1; DD.1.2	U.1.1
FL	L.2.1.1	I.1.1 (DF5); I.2.1.1; I.2.2	Nr.1.3	D.1, D.3	DD.1.1; DD.1.3	DD.1.2; n/a
Dst	L.2.2.1	I.1.1; I.2.1.1; I.2.3	Nr.1.1	n/a	DD.2.1; DD.3.4	DD.3.1; U.2.2

D.4. Group C: Outgoing Cross-boundary MCP Flow

Group C contains a single flow: DF8, from P1 (Fraud Detection Agent) to EE4/DS5 (External Sanctions Screening Service), carrying identity attributes selected for sanctions screening, including name, date of birth, and nationality. This is a Process-to-ExternalEntity MCP flow that crosses the organisational trust boundary TB1, making it structurally distinct from the intra-organisational flows in Groups A and B. EE4/DS5 is registered as an MCP-compatible external tool within Bank X's MCP server configuration, meaning that DF8 is structured as a typed MCP tool call rather than a generic API request. The MCP-specific characteristics identified in Groups A and B therefore apply here, with the additional consequence that MCP protocol metadata, including the session identifier, tool name,

and parameter schema, crosses the organisational boundary and becomes visible to an external party outside Bank X's data governance framework. Per the mapping table, all threat categories apply at the source, data flow, and destination for this flow type.

The per-position LINDDUN assessment for this group is set out in Table D.7.

Table D.7: Group C: Outgoing Cross-boundary MCP Flow (P → EE), flow DF8. S = source, FL = data flow, Dst = destination. Author's own analysis.

Flow	Cat.	Loc.	Chars.	Description
P1 → DF8 → EE4/DS5	L	S,FL,Dst	L.1.1, L.2.1.1, L.2.2.1	Session ID links disclosure to same cardholder transaction (S). Identity attributes form quasi-identifier in transit (FL). External service accumulates screening profile (Dst).
	I	S,FL,Dst	I.1.1	Name and date of birth directly identify cardholder at all positions.
	N_R	S,FL,Dst	Nr.1.1, Nr.1.3	Bank X logs create evidence of external disclosure (S,Dst). Request metadata forms persistent record (FL).
	D	S,FL,Dst	D.1, D.2, D.3	Outgoing MCP tool call to external service is detectable (S). DF8 crosses TB1 and is observable on external network (FL). Response structure may reveal screening outcome (Dst).
	D_D	S,FL,Dst	DD.1.1, DD.1.3, DD.3.3, DD.3.4, DD.4.1.1, DD.4.1.2	Directly identifying data transmitted outside TB1 (S). Encoding carries metadata; unencrypted flow constitutes implicit disclosure (FL). Data retained and potentially shared by EE4/DS5 (Dst).
	U	S,Dst	U.1.1, U.2.2	Cardholder unaware identifying attributes left Bank X's boundary (S). No access or contestation mechanism at EE4/DS5 (Dst).

Position-level threat nodes:

Pos.	L	I	N_R	D	D_D	U
S	L.1.1	I.1.1	Nr.1.1	D.2	DD.1.1; DD.4.1.1	U.1.1
FL	L.2.1.1	I.1.1	Nr.1.3	D.1	DD.1.1; DD.1.3; DD.3.3	n/a
Dst	L.2.2.1	I.1.1	Nr.1.1	D.3	DD.3.4; DD.4.1.2	U.2.2

D.5. Group D: Incoming Cross-boundary MCP Result

Group D contains a single flow: DF9, from EE4/DS5 (External Sanctions Screening Service) to P1 (Fraud Detection Agent), carrying a sanctions screening result and risk indicators. This is an External Entity-to-Process MCP flow crossing into TB1 from outside both trust boundaries. The payload is a derived result rather than raw personal data, but the result is directly attributable to the cardholder whose identity was submitted on DF8.

The per-position LINDDUN assessment for this group is set out in Table D.8.

Table D.8: Group D: Incoming Cross-boundary MCP Result (EE → P), flow DF9. S = source, FL = data flow, Dst = destination. Author's own analysis.

Flow	Cat.	Loc.	Chars.	Description
EE4/DS5 → DF9 → P1	L	S,FL,Dst	L.1.1, L.2.1.1, L.2.2.1	Session ID links screening result to same cardholder (S). Result attributes form quasi-identifier in transit (FL). P1 accumulates screening history (Dst).
	I	S,FL,Dst	I.1.1, I.2.1.1, I.2.3	Positive match result is directly identifying (S,FL). Combined with session data, subject is singled out (Dst).
	N_R	S,FL,Dst	Nr.1.1, Nr.1.3	EE4/DS5 logs create third-party evidence of result release (S,Dst). Response metadata forms persistent record (FL).
	D	S,FL	D.1, D.2	Response is detectable side-effect (S). DF9 crosses TB1 and is observable on external network (FL).
	D_D	S,FL,Dst	DD.1.1, DD.1.3, DD.3.4	Result may include more risk indicators than required (S). Encoding carries non-functional metadata (FL). Result retained beyond immediate decision (Dst).
	U	S,Dst	U.1.1, U.2.2	Cardholder unaware of external screening result (S). No access or contestation mechanism at P1 (Dst).

Position-level threat nodes:

Pos.	L	I	N_R	D	D_D	U
S	L.1.1	I.2.1.1	Nr.1.1	D.2	DD.1.1	U.1.1
FL	L.2.1.1	I.2.1.1	Nr.1.3	D.1	DD.1.3	n/a
Dst	L.2.2.1	I.2.1.1	Nr.1.1	n/a	DD.3.4	U.2.2

D.6. Group F: Intra-organisational A2A Pair

Group F consists of two Agent-to-Agent flows: DF14 (P1 → P2) and DF17 (P2 → P1). DF14 carries the fraud assessment request issued by P1 to P2, including aggregated findings derived from the transaction data, risk attributes, and sanctions screening outcome assembled during the current assessment. DF17 carries the fraud decision returned by P2 to P1, including a risk score, a binary fraud decision, and a motivating rationale. Together, DF14 and DF17 form a single request-response pair within one A2A task. Group F is structurally distinct from the MCP flows in Groups A through E in several respects. A2A is an agent-to-agent protocol in which both parties are autonomous reasoning entities that maintain their own context, rather than a tool-invocation protocol in which an agent calls a passive data store or external service. A2A organises communication around task identifiers and conversation threads rather than MCP session identifiers. A2A payloads are richer than MCP tool call parameters: they carry aggregated assessment results, reasoning context, and decision rationales rather than compact query keys or raw records. Both flows are intra-organisational within TB1. Per the mapping table, Detectability does not apply at the destination and Unawareness and Unintervenability does not apply at the data flow position; these exclusions apply throughout this group.

The per-position LINDDUN assessment for this group is set out in Table D.9.

Table D.9: Group F: Intra-organisational A2A Pair (P → P), flows DF14, DF17. S = source, FL = data flow, Dst = destination. Author's own analysis.

Flow	Cat.	Loc.	Chars.	Description
P1/P2 → DF14/17 → P2/P1	L	S,FL,Dst	L.1.1, L.2.1.1, L.2.2.1	A2A task ID links escalation to same cardholder (S). Full assessment context is highly specific quasi-identifier (FL). Escalated cases accumulate at P2 (Dst).
	I	S,FL,Dst	I.1.1, I.2.1.1, I.2.3	Full assessment context on DF14 directly identifies cardholder at all positions. Combined data singles out subject at P2 (Dst).
	N_R	S,FL,Dst	Nr.1.1, Nr.1.3	A2A task logging creates attributable evidence of escalation (S,Dst). Message metadata forms persistent record (FL).
	D	S,FL	D.1, D.2	Escalation task initiation is detectable within TB1 (S). Internal A2A flows observable on internal network (FL).
	D_D	S,FL,Dst	DD.2.1, DD.1.3, DD.3.4	Full assessment context may exceed what is needed for decision (S). Encoding carries non-functional metadata (FL). Case retained beyond decision lifecycle (Dst).
	U	S,Dst	U.1.1, U.2.2	Cardholder unaware of human escalation (S). No access or contestation mechanism at P2 (Dst).

Position-level threat nodes:

Pos.	L	I	N_R	D	D_D	U
S	L.1.1	I.1.1 (if match); I.2.1.1	Nr.1.1	D.2	DD.2.1	U.1.1
FL	L.2.1.1	I.1.1 (if match); I.2.1.1	Nr.1.3	D.1	DD.1.2; DD.1.3	n/a
Dst	L.2.2.1	I.2.1.1; I.2.3	Nr.1.1	n/a	DD.3.1; DD.3.4	U.2.2

D.7. Group E: Cross-organisational A2A Pair

Group E contains two Process-to-Process A2A flows across the inter-organisational boundary between TB1 and TB2: DF10 (P1 → P3), carrying a derived behavioural profile including transaction velocity, geographic patterns, and device consistency indicators; and DF11 (P3 → P1), returning a structured risk score and network-level indicators. DF10 is the privacy-critical flow of the pair: it transmits a derived profile constructed from raw personal data held by Bank X to an agent operated by a separate organisation. DF11 returns a result but does not carry raw personal data about the cardholder.

The per-position LINDDUN assessment for this group is set out in Table D.10.

Table D.10: Group E: Cross-organisational A2A Pair (P → P), flows DF10, DF11. S = source, FL = data flow, Dst = destination. Author's own analysis.

Flow	Cat.	Loc.	Chars.	Description
P1/P3 → DF10/11 → P3/P1	L	S,FL,Dst	L.1.1, L.2.1.1, L.2.2.1	A2A task ID links profile to same cardholder transaction (S). Behavioural profile forms quasi-identifier (FL). P3 accumulates profiles from Bank X (Dst).

Flow	Cat.	Loc.	Chars.	Description
	I	S,FL,Dst	I.2.2, I.2.3	Behavioural profile contains revealing attributes (S,FL). Re-identification risk elevated by P3's access to network-wide data (Dst).
	N_R	S,FL,Dst	Nr.1.1, Nr.1.3	A2A task logging creates attributable evidence of cross-organisational exchange (S,Dst). Message metadata forms persistent record (FL).
	D	S,FL	D.1, D.2	A2A task initiation is detectable (S). Flows cross TB1/TB2 and are observable on inter-organisational network (FL).
	D_D	S,FL,Dst	DD.1.2, DD.1.3, DD.3.4	Profile may be more granular than required for risk score (S). A2A encoding carries non-functional metadata (FL). Profile retained at Card Network Y beyond Bank X's control (Dst).
	U	S,Dst	U.1.1, U.2.2	Cardholder unaware derived profile is shared with separate organisation (S). No access or contestation mechanism at P3 (Dst).

Position-level threat nodes:

Pos.	L	I	N_R	D	D_D	U
S	L.1.1	I.2.2	Nr.1.1	D.2	DD.1.2	U.1.1
FL	L.2.1.1	I.2.3	Nr.1.3	D.1	DD.1.3	n/a
Dst	L.2.2.1	I.2.3	Nr.1.1	n/a	DD.3.4	U.2.2

D.8. Group H: Cardholder Payment Initiation

Group H contains a single flow: DF1, from EE1 (Cardholder) to P1 (Fraud Detection Agent), carrying transaction amount, merchant details, device identifier, and geographic location. This is an External Entity-to-Process flow with no automated protocol. It is the entry point of the entire fraud detection workflow and carries the richest inbound payload, combining directly identifying data with behavioural and device-level attributes. Per the mapping table, Detectability does not apply at the destination and Unawareness and Unintervenability does not apply at the data flow position; these exclusions apply throughout this group.

The per-position LINDDUN assessment for this group is set out in Table D.11.

Table D.11: Group H: Cardholder Payment Initiation (EE → P), flow DF1. S = source, FL = data flow, Dst = destination. Author's own analysis.

Flow	Cat.	Loc.	Chars.	Description
EE1 → DF1 → P1	L	S,FL,Dst	L.1.1, L.2.1.1, L.2.2.1	Device identifier links transaction to cardholder (S). Attribute combination is quasi-identifier in transit (FL). P1 accumulates behavioural profile (Dst).
	I	S,FL,Dst	I.2.1.1, I.2.2, I.2.3	Device identifier and location are unique references (S,FL). Combined with P1 data, subject is singled out (Dst).
	N_R	S,FL,Dst	Nr.1.1, Nr.1.3	Authentication credential attributes initiation to cardholder (S,Dst). Metadata forms persistent record (FL).
	D	S,FL	D.1, D.2	Payment initiation is a detectable transmission event (S,FL).

Flow	Cat.	Loc.	Chars.	Description
	D_D	S,FL,Dst	DD.1.2, DD.1.3, DD.3.4	Location and device identifier more granular than required (S). Encoding carries non-functional metadata (FL). Full payload retained beyond immediate need (Dst).
	U	S,Dst	U.1.1, U.2.2	Cardholder unaware payment triggers multi-agent assessment (S). No control mechanism at P1 (Dst).

Position-level threat nodes:

Pos.	L	I	N_R	D	D_D	U
S	L.1.1	I.2.1.1, I.2.2	Nr.1.1	D.2	DD.1.2	U.1.1
FL	L.2.1.1	I.2.1.1	Nr.1.3	D.1	DD.1.3	n/a
Dst	L.2.2.1	I.2.1.1, I.2.3	Nr.1.1	n/a	DD.3.4	U.2.2

D.9. Group G: Human Analyst Interaction Pair

Group G consists of two flows: DF15 (P2 → EE3) and DF16 (EE3 → P2). DF15 carries the full case dossier escalated by P2, the Internal Review Agent, to the human fraud analyst, including transaction details, identity attributes, device data, and the sanctions screening outcome. DF16 carries the analyst's final approve-or-decline decision returned to P2. Together, DF15 and DF16 form a single escalation-and-response pair in which P2 refers a case to a human analyst, who reviews the full assessment and returns a decision. Group G is structurally distinct from all preceding groups in that one party is a human actor rather than an autonomous system component. The analyst exercises discretionary, independent judgment and has direct, unmediated access to the full set of personal data assembled during the fraud assessment workflow. These flows use no automated protocol; the analyst interacts with P2 through an internal review interface rather than through MCP or A2A. Both flows are intra-organisational within TB1. Per the mapping table, Detectability does not apply at the destination and Unawareness and Unintervenability does not apply at the data flow position; these exclusions apply throughout this group.

The per-position LINDDUN assessment for this group is set out in Table D.12.

Table D.12: Group G: Human Analyst Interaction Pair (P → EE), flows DF15, DF16. S = source, FL = data flow, Dst = destination. Author's own analysis.

Flow	Cat.	Loc.	Chars.	Description
P2/EE3 → DF15/16 → EE3/P2	L	S,FL,Dst	L.1.1, L.2.1.1, L.2.2.1	Case ID links presentation to same cardholder (S). Case details form highly specific quasi-identifier (FL). Analyst access creates human-side linkage (Dst).
	I	S,FL,Dst	I.1.1, I.2.1.1, I.2.3	Case details directly identify cardholder to analyst at all positions. Highest identifiability risk in the workflow.
	N_R	S,FL,Dst	Nr.1.1, Nr.1.3	P2 logs case presentation as evidence of disclosure to human actor (S,Dst). Interface logs presented data and decision (FL).
	D	S,FL	D.1, D.2	Case routing to analyst is detectable within TB1 (S). Presentation and decision flows observable on analyst network (FL).
	D_D	S,FL,Dst	DD.2.1, DD.1.3, DD.3.4, DD.4.1.2	Full case details may exceed what is needed for decision (S). Interface may log or cache beyond interaction (FL). Case logs retained at analyst system; risk of informal onward sharing (Dst).
	U	S,Dst	U.1.1, U.2.2	Cardholder unaware named individual is reviewing their case (S). No participation or contestation mechanism at analyst (Dst).

Position-level threat nodes:

Pos.	L	I	N_R	D	D_D	U
S	L.1.1	I.1.1 (DF15); I.2.1.1	Nr.1.1	D.2	DD.2.1	U.1.1
FL	L.2.1.1	I.1.1 (DF15); I.2.1.1	Nr.1.3	D.1	DD.1.1; DD.1.2; DD.1.3	n/a
Dst	L.2.2.1	I.1.1; I.2.1.1; I.2.3	Nr.1.1	n/a	DD.3.1; DD.3.4; DD.4.1.2	U.2.2

D.10. Group I: Terminal Payment Response

Group I contains a single flow: DF18, from P1 (Fraud Detection Agent) to EE2 (Merchant's Bank), carrying the transaction outcome: approved or declined. This is a Process-to-External Entity flow with no automated protocol, crossing the TB1 boundary. The payload is a decision outcome rather than identified personal data, but it is directly attributable to the cardholder whose transaction was assessed.

The per-position LINDDUN assessment for this group is set out in Table D.13.

Table D.13: Group I: Terminal Payment Response (P → EE), flow DF18. S = source, FL = data flow, Dst = destination. Author's own analysis.

Flow	Cat.	Loc.	Chars.	Description
P1 → DF18 → EE2	L	S,FL,Dst	L.1.1, L.2.1.1, L.2.2.1	Transaction reference links outcome to cardholder (S,FL). Accumulated outcomes at EE2 build behavioural profile (Dst).
	I	S,FL,Dst	I.2.1.1	Transaction reference is unique identifier linking outcome to specific cardholder at all positions.
	N_R	S,FL,Dst	Nr.1.1, Nr.1.3	Bank X logs create attributable evidence that decision was transmitted (S,Dst). Metadata forms persistent record (FL).
	D	S,FL	D.1, D.2	Outcome transmission is detectable (S). DF18 crosses TB1 and is observable on external network (FL).
	D_D	FL,Dst	DD.1.3, DD.3.4	No excess data at source (binary outcome only). Encoding carries non-functional metadata (FL). Outcome retained long-term at EE2 (Dst).
	U	S,Dst	U.1.1, U.2.2	Cardholder unaware of automated inter-bank outcome transmission (S). No contestation mechanism at EE2 (Dst).

Position-level threat nodes:

Pos.	L	I	N_R	D	D_D	U
S	L.1.1	I.2.1.1	Nr.1.1	D.2	,	U.1.1
FL	L.2.1.1	I.2.1.1	Nr.1.3	D.1	DD.1.3	n/a
Dst	L.2.2.1	I.2.1.1	Nr.1.1	n/a	DD.3.4	U.2.2

E Expert Validation Interview Summaries

This appendix presents structured summaries of the eight expert validation interviews conducted as part of the research. Interviews were held in May 2026 and followed a semi-structured guide covering four phases: participant introduction, thesis and protocol background, threat-by-threat validation, and open reflection. All interviews were recorded with the participant's consent; the substantive assessments and their supporting reasoning were selectively transcribed for analysis, and the quotations reproduced below are verbatim extracts from those passages. Seven participants are Avanade practitioners whose identities are anonymised in line with the informed consent agreement; they are referred to by a coded identifier (E1–E7). The eighth session, with Kim Wuyts, a co-author of the original LINDDUN framework and a leading contributor to its development, took the form of an open consultation on the applied methodology rather than a structured threat validation, and is summarised separately at the end of this appendix (Section E). Quotes are reproduced in the original language of the interview (Dutch or English); Dutch quotes are followed by an English translation in square brackets.

Reading the validation tables. Each interview summary contains a *Threat validation* table with one row per threat. The **Priority alignment** column records the expert's own assessment of the threat's priority. The cell **colour** corresponds to the LINDDUN priority level at which the expert would place the threat (**Critical**, **High**, **Medium**, **Low**, following the convention used in Chapter 6), and the **label** (e.g. *Aligned*, *Conditionally confirmed*, *Suggests downgrade*) describes the nature of the assessment relative to the priority assigned in Chapter 6. The consolidated Confirmed/Refined/Challenged coding used in the narrative analysis of Section 7.5 of Chapter 7 reduces this nuance to three categories for cross-expert comparison; the present appendix retains the richer per-expert wording.

E1 Security and Information Protection Specialist

Participant profile

E1 has several years of experience in information protection and data security, with a focus on privacy compliance and data governance tooling. Their work includes Privacy Impact Assessments and regulatory frameworks. They had no prior hands-on experience with MCP or A2A but were familiar with agentic AI concepts through internal training.

Threat validation

ID	Threat (short)	Priority alignment	Key observation
TH-01	MCP session linkability	Aligned	Recognises the threat. Collecting data from multiple databases makes protective measures essential, not only for the user but for the organisation itself.
TH-02	MCP data aggregation	Aligned	Confirms the threat. Without filtering measures, aggregation at a single agent represents a high personal and organisational risk.
TH-03	Cross-boundary transmission	Aligned	Recognises the threat; notes contracts exist but are rarely verified in practice, and that the cardholder is rarely informed their data left the bank.
TH-04	A2A bilateral logging	Partially aligned	Retention controls and configurable deletion timers could partially mitigate this; priority depends on implementation choices.
TH-05	Re-identification	Aligned	Confirms the threat as well known within organisations.
TH-06	A2A data minimisation	Questions upgrade	Aligns with High but questions why not Critical, given that an entire identity profile is forwarded to the human analyst.
TH-07	Cardholder unawareness	Suggests downgrade	Terms and conditions partially address unawareness, even if rarely read.

ID	Threat (short)	Priority alignment	Key observation
TH-08	Cardholder unintervenability	Aligned	Once data is released through a card tap, no in-process intervention is possible.

Additional observations

Governance by contract versus governance by architecture. E1 drew a distinction between contractual protections and technical controls. They consider contractual protections insufficient without verification, and raised the question of whether architecture-level measures such as automatic data deletion after processing are feasible within MCP and A2A.

Organisational accountability. E1 argued that Bank X bears ultimate privacy responsibility for the cardholder, because the trust relationship exists between them. Other parties in the system share secondary responsibility when incidents can be traced to their systems.

Notable quotes

"Als er geen maatregelen zijn om alleen de data te krijgen dat je nodig hebt, dan krijg je meer dan dat. En dat is niet alleen voor de user, maar ook voor het bedrijf zelf."

["If there are no measures to retrieve only the data you need, then you end up getting more than that. And that is not only a risk for the user, but also for the organisation itself."]

"Bedrijven horen dat wel met de users te delen wat er precies gedaan wordt met informatie, maar dat wordt niet altijd gecontroleerd."

["Organisations are supposed to inform users about exactly what is done with their information, but that is not always verified."]

E2 Senior Data and AI Architect

Participant profile

E2 has over two decades of experience in data science, AI architecture, and AI security. They have been involved in AI governance at a regulatory level and have hands-on deployment experience with an MCP-adjacent agentic system in a financial services context. They have a background in red-teaming AI systems.

Threat validation

ID	Threat (short)	Priority alignment	Key observation
TH-01	MCP session linkability	Confirmed	Immediately confirmed the threat as definitively present.
TH-02	MCP data aggregation	Partially confirmed	Responsible banks already apply anonymisation and data minimisation for sensitive data; the threat is less naive in practice than the model suggests, but not absent.
TH-03	Cross-boundary transmission	Confirmed with caveats	The communication channel itself is the primary risk: interception of MCP traffic in transit may be more dangerous than the destination organisation's data handling.
TH-04	A2A bilateral logging	Lower priority	Logging on both sides is a compliance requirement; the real risk is in how easily the logged identifier links to personal data within each system.

ID	Threat (short)	Priority alignment	Key observation
TH-05	Re-identification	Confirmed	A natural consequence of the volume of data currently gathered about people; extremely easy to identify individuals even from partially anonymised data.
TH-06	A2A data minimisation	Confirmed	MCP is more suitable than unmodified A2A for the use case, because A2A would require significant custom amendments for compliance.
TH-07	Cardholder unawareness	Not applicable	Terms and conditions accepted on account opening constitute informed consent.
TH-08	Cardholder unintervenability	Not applicable	Same reasoning as TH-07.

Additional observations

MCP was not built with privacy at its core. E2 found no evidence of built-in privacy or security by design in MCP, and recommended that any deployment include an explicit compliance roadmap with checks at each stage.

Implementation quality as a distinct vulnerability. E2 separated the protocol specification from its implementation. Even a well-specified protocol provides no guarantee of safety if developers do not connect components in a privacy-respectful way.

LLM training data as an additional privacy vector. E2 raised a threat outside the eight identified: privacy-sensitive data used to fine-tune a locally deployed LLM may surface in responses to unrelated users. This risk is intrinsic to the model's nature and independent of the communication protocol.

Governance vacuum. Based on broad consulting experience, E2 observed that almost every organisation has serious gaps in AI governance: no one knows who owns what. Ungoverned grey zones are the primary attack surface for both privacy and security incidents.

Notable quotes

"There is no such thing as zero trust in a non-deterministic model. So we have to find the best way that we can to inform the client about risks and about what is possible and what is not."

"You can have all of the checklists of the world on the governance side, but if on the implementation side the developers do not know how to connect things in a secure and privacy-respectful way, the framework provides no guarantee of safety."

"The way how they communicate this [data cross-boundary],[meaning the protocol and channel used to send data between organisations], that is where the real risk lies. It is not necessarily the organisations per se [both parties may be fully trustworthy and GDPR-compliant]. It is just the way how they communicate this [the protocol itself, MCP in this case, was not built with privacy at its core, and the data travelling between the two parties is briefly outside the control of either one]."

E3 Modern Workplace and Copilot Practice Lead

Participant profile

E3 is a senior practice lead with fourteen years of consulting experience. They hold Microsoft Most Valuable Professional status and have hands-on experience with MCP through personal projects and conceptual knowledge of A2A from published technical papers. Their background spans security, data protection, and identity access management.

Threat validation

ID	Threat (short)	Priority alignment	Key observation
TH-01	MCP session linkability	Conditionally confirmed	Accepts the threat where cross-store linkage is undesirable. Raises auditability as a competing interest; suggests scoping the identifier per individual database call rather than per task.
TH-02	MCP data aggregation	Confirmed	Cites recent Dutch data breaches as evidence that a single aggregation point is a realistic attack vector.
TH-03	Cross-boundary transmission	Confirmed	Questions whether names are necessary at all: a BSN number or IBAN should suffice for sanctions screening, reducing outbound exposure.
TH-04	A2A bilateral logging	Confirmed	Recording *that* a person was screened is itself sensitive information.
TH-05	Re-identification	Partially confirmed	Accepts the privacy concern; notes that customers sign contracts authorising the data sharing.
TH-06	A2A data minimisation	Confirmed with mitigation	Guard rails applied at the internal review agent via system prompts can restrict what is passed to the human analyst.
TH-07	Cardholder unawareness	Low priority	The cardholder just wants to pay; privacy awareness among end users is too low to make this a meaningful operational priority.
TH-08	Cardholder unintervenability	Low priority	Same reasoning as TH-07; suggests the EU AI Act may eventually address this structurally.

Additional observations

Rapid protocol adoption outpaces privacy maturity. E3 observed that MCP was adopted across major platforms within approximately eighteen months, with security and privacy controls added incrementally after the fact. This is a pattern of privacy-by-retrofit rather than privacy-by-design.

Prompt injection as a connected risk. E3 identified prompt injection as a realistic attack vector via the cardholder-facing entry point, and noted an ongoing cat-and-mouse game between injection attacks and guard-rail countermeasures.

Speed-to-market pressure. Business pressure and C-level enthusiasm for AI frequently lead to security and privacy being deprioritised during development. The biggest risk is that things are missed because they are not properly thought through or tested.

Notable quotes

"Ieder luikje is een risico. Er hoeft maar een ethernet outlet niet afgeschermd te zijn en iemand komt binnen. Dat gebeurt nog steeds."

["Every hatch is a risk. All it takes is one unshielded ethernet outlet and someone is inside. That still happens."]

"Je ziet dat MCP over tijd steeds beter wordt vanuit security en privacy oogpunt, maar in het begin was er echt kritiek op."

["You see MCP improving over time from a security and privacy perspective, but at the start there was real criticism."]

E4 Modern Workplace and Governance Specialist

Participant profile

E4 has nearly seven years of consulting experience, beginning in infrastructure and security before specialising in Microsoft 365, governance, and Copilot adoption. They hold Microsoft MVP status and in 2025 moved to a full-time role leading Copilot adoption trajectories for client organisations.

Threat validation

ID	Threat (short)	Priority alignment	Key observation
TH-01	MCP session linkability	Aligned	The session identifier is an immutable object: once present, it allows tracing everything associated with a query. Raises whether data is deleted after session close.
TH-02	MCP data aggregation	Confirmed	Uses a hospital analogy: patient name and room number are innocuous individually; combined with ward diagnoses they create a highly sensitive aggregate record.
TH-03	Cross-boundary transmission	Confirmed	Full MCP context may make it impossible to demonstrate GDPR compliance, which requires knowing exactly what was transmitted to an external party.
TH-04	A2A bilateral logging	Confirmed with contractual mitigation	NDAs and automatic retention limits are partial mitigants, but depend entirely on the counterparty's compliance.
TH-05	Re-identification	Confirmed	Differences between current rule-based and future agentic systems may be smaller than expected; the threat is not new in kind, only in degree.
TH-06	A2A data minimisation	Confirmed	Guard rails can restrict what is passed to the human analyst, but prompt injection vulnerabilities mean guard rails are never fully reliable.
TH-07	Cardholder unawareness	Accepted	More concern about the automation of what was previously a human-mediated process than about unawareness itself.
TH-08	Cardholder unintervenability	Accepted	Confirmed; the key question is whether log access is properly compartmentalised by team.

Additional observations

Compartmentalisation of log access. E4 proposed that access to log files should be divided by team, so that the team maintaining the fraud detection agent cannot simultaneously access logs from all downstream interactions. A system permitting only per-dataflow log queries would significantly reduce linkability risk in practice.

The automation of trust. People accept that human analysts see sensitive data because those analysts have signed NDAs and are individually accountable. Automated agents perform the same operations without equivalent accountability mechanisms.

Notable quotes

"Als je een ziekenhuis hebt en je hebt een patiëntnaam en een kamer, die twee samen doen niet zoveel. Op het moment dat je ook een kamernummer hebt en de ziektes op die kamers, zijn die twee samen opeens een stuk groter dan los van elkaar."

["A patient name and room number together are not that problematic. But add the diagnoses associated with those rooms, and the two combined are suddenly far more sensitive than either one in isolation."]

"We zijn iets koelanter als het om mensen gaat. Als het door een agent wordt gedaan, dan moet het 100% kloppen en afgedicht zijn."

["We are somewhat more lenient when it comes to humans. When it is done by an agent, everything must be 100% correct and fully sealed."]

E5 AI/Agent Developer at a Large Dutch Bank

Participant profile

E5 has sixteen years of consulting and technical architecture experience. At the time of the interview they were working as an AI/agent developer for the production chatbot of a large Dutch bank, responsible for all integrations with internal banking systems. They are the only participant currently building and operating agentic AI inside a regulated financial institution.

Threat validation

ID	Threat (short)	Priority alignment	Key observation
TH-01	MCP session linkability	Confirmed	The bank already generates single-use session IDs to protect the frontend, yet the MCP session ID on the server side still enables cross-dataflow linkage by internal staff. Confirms TH-01 in a live deployment.
TH-02	MCP data aggregation	Conditionally confirmed	If sequential per-source processing is feasible for the use case, the aggregation risk is significantly reduced.
TH-03	Cross-boundary transmission	Confirmed with solution proposal	Proposes a shared pseudonymous identifier agreed between banks and Interpol/Europol, not derived from BSN or birth date, to decouple screening queries from personal identity.
TH-04	A2A bilateral logging	Accepted as a trade-off	The benefit of safer money outweighs the small risk that someone knows a check was performed; this is a risk to accept rather than mitigate.
TH-05	Re-identification	Confirmed but unavoidable	Structurally unsolvable in the use case; proposes a government-controlled neutral intermediary between banks and payment providers.
TH-06	A2A data minimisation	Confirmed	The internal review agent could provide only aggregated, non-identifiable answers to the human analyst, with full context available on explicit request.
TH-07	Cardholder unawareness	Low priority	No easy solution beyond a checkbox; proactive disclosure by the cardholder could be a partial transparency mechanism.
TH-08	Cardholder unintervenability	Low priority	Classifies TH-01, TH-02, TH-03, and TH-06 as actionable by design; TH-04, TH-05, TH-07, and TH-08 as governance and legislative problems.

Additional observations

Real-world confirmation of TH-01. E5 described how their bank generates single-use card identifiers to protect the frontend but acknowledged that the server-side MCP session ID still allows internal log linkability. This is the most operationally grounded confirmation of TH-01 in the interview set.

A clear split between actionable and accepted risks. E5 explicitly divided the eight threats into two groups: those where technical design choices can make a substantial difference (TH-01, TH-02, TH-03, TH-06) and those representing accepted systemic trade-offs best addressed through law and contracts

(TH-04, TH-05, TH-07, TH-08).

Governance and non-deterministic testing as bottlenecks. E5 identified two dominant obstacles to responsible agentic AI deployment in regulated institutions: non-deterministic models cannot be validated with a single test run, and cross-functional governance sign-off requires substantially more time than the technical build.

Notable quotes

"Zelfs als dat ophalen van die card status geanonimiseerde data teruggeeft, doordat je vervolgens kan kijken naar die session identifier die eerder de naam terugkrijgt, kan je al die data alsnog aan elkaar linken."

["Even if the card status retrieval returns anonymised data, the fact that you can look at the session identifier, which earlier did retrieve the name, still allows you to link all of that data together."]

"Ik denk dat 1, 2, 3 en 6 degene zijn waarvan ik denk: hier kunnen wij echt een verschil maken door design en technische implementatie. De andere vier, dat moet meer aan de kant van governance en wetten worden opgelost."

["I think threats 1, 2, 3 and 6 are the ones where we can really make a difference through design and technical implementation. The other four need to be resolved through governance and law."]

E6 Data Security Manager

Participant profile

E6 is a data security manager with nine years of consulting experience, focusing on structured data protection, data governance, and compliance. They hold Microsoft Most Valuable Professional status and have extensive experience working with and inside large Dutch financial institutions.

Threat validation

ID	Threat (short)	Priority alignment	Key observation
TH-01	MCP session linkability	Partially confirmed	Extensive logging already occurs in financial institutions for legal audit. However, consolidating all logs in a single database would create a single person with omniscient access, a serious security and privacy risk.
TH-02	MCP data aggregation	Confirmed	Banks explicitly separate data stores by access role; aggregating these at a single agent directly undermines the separation of duties principle underpinning financial data governance.
TH-03	Cross-boundary transmission	Confirmed, moderate priority	In practice, outbound data to sanctions databases is likely minimal. Contractual safeguards and independent audit reports are applicable mitigants.
TH-04	A2A bilateral logging	Accepted as inherent	Bilateral logging is an inherent property of API-to-API communication; rights management analogies exist but are impractical where models must reason over plaintext data.
TH-05	Re-identification	Confirmed	Behavioural profiling for secondary purposes is a known and troubling practice; largely uncontrollable at the sending organisation's level.
TH-06	A2A data minimisation	Most critical for banks	When data leaves the bank's span of control into a cloud LLM, auditability collapses, and auditability is a hard regulatory requirement.

ID	Threat (short)	Priority alignment	Key observation
TH-07	Cardholder unawareness	Low priority	Customers do not mind not knowing how the decision was made, as long as payment proceeds; terms and conditions are accepted without being read.
TH-08	Cardholder unintervenability	User experience problem	The contestability gap is real but experienced primarily as a service quality issue rather than a privacy violation per se.

Additional observations

Auditability as the dominant bank concern. E6 argued that for regulated financial institutions, demonstrating what data was processed, when, under what conditions, and where, is a non-negotiable compliance requirement. Any agentic architecture eroding this auditability is a blocker for adoption. Cloud LLMs processing data across jurisdictions introduce GDPR compliance risk because the location of processing cannot be guaranteed.

The LLM as a black-box processing layer. The model reasoning layer in a cloud-hosted deployment is opaque: prompts, context, and reasoning are all processed in a location the bank cannot audit. This creates the same compliance gap as the communication protocol threats, through a different mechanism.

The consent-coercion paradox. E6 articulated the structural consent problem clearly: users are forced to accept terms and conditions to receive the service. This is an unavoidable fact, not a fair one, and substantially undermines the adequacy of consent as a legal basis for the processing described in this thesis.

Notable quotes

"Er is een reden waarom data in aparte databases wordt gehuis. Door dat bij elkaar te brengen in een agent, ook al heeft die maar heel even memory, ondermijn je dat principe."

["There is a reason why data is housed in separate databases. By bringing it together in an agent, even if the agent only holds it briefly, you undermine that principle."]

"De verwerking is een black box en je kan niet zien waar dat gebeurt. De auditability is een ijzer. Als je dat niet kan aantonen, kan je niet waarborgen dat je hebt voldaan aan de rechten van de mensen met wiens data je werkt."

["The processing is a black box and you cannot see where it happens. Auditability is a hard requirement. If you cannot demonstrate it, you cannot guarantee you have fulfilled the rights of the people whose data you process."]

E7 Security Manager

Participant profile

E7 is a security manager with eight years of consulting experience, beginning as a system engineer before specialising in security. Their work covers Microsoft security solutions, client consultations, design workshops, and implementation oversight. They had no prior familiarity with the LINDDUN framework but are experienced with the component concepts from security practice.

Threat validation

ID	Threat (short)	Priority alignment	Key observation
TH-01	MCP session linkability	Conditionally confirmed	Session IDs could be randomised to reduce inherent linkability, but logging always allows reassembly by someone with access rights.

ID	Threat (short)	Priority alignment	Key observation
TH-02	MCP data aggregation	Confirmed	Agrees fully; proposes that agent-assembled profiles be treated as ephemeral and destroyed immediately after each session concludes.
TH-03	Cross-boundary transmission	Confirmed	Notes the geopolitical dimension: Mastercard and Visa are US entities subject to US law; a government order could compel data disclosure in ways incompatible with GDPR.
TH-04	A2A bilateral logging	Partially mitigable	Encryption with dual-party certificates could reduce the sensitivity of logged identifiers, though with performance costs.
TH-05	Re-identification	Confirmed, unavoidable	Mastercard already knows which name belongs to which card number and can trivially re-identify pseudonymised profiles.
TH-06	A2A data minimisation	Confirmed	Draws a parallel to Security Operations Centre workflows, where analysts similarly receive excessive contextual data; proposes a summary-first, details-on-request interface.
TH-07	Cardholder unawareness	Low priority	Better disclosure of data-sharing partners could improve the situation without blocking fraud detection.
TH-08	Cardholder unintervenability	Low priority	Users prioritise payment speed over control; regulatory requirements are the appropriate lever.

Additional observations

Geopolitical dimension of cross-boundary data sharing. E7 raised a consideration not surfaced by other participants: Mastercard and Visa are American companies subject to US law. A US government order compelling data disclosure would override the GDPR-based contractual protections nominally in place. E7 proposed a neutral EU- or UN-supervised intermediary as a structural solution, independently echoing the same proposal made by E5.

Post-session data destruction. E7 proposed a concrete technical control: agent-assembled profiles should be treated as ephemeral and destroyed when the fraud assessment session concludes. Only the active session should hold the aggregated profile at any point in time.

Agentic AI as an intensification of existing risks. E7 noted that fraud detection already constructs behavioural profiles and logs decisions via ML-based systems. The transition to agentic AI changes the speed, scope, and degree of inference but not the fundamental architecture; the privacy threats are intensifications of existing risks rather than entirely new categories.

Notable quotes

"Als zo'n query klaar is, moet die data gewoon vernietigd worden. Het profiel moet worden afgebouwd. Alleen de actieve sessie mag dat geaggregeerde profiel op dat moment bevatten."

["Once a query is complete, that data must be destroyed. The profile must be dismantled. Only the active session should contain the aggregated profile at any given moment."]

"Mastercard en Visa zijn Amerikaanse bedrijven. Als de Amerikaanse president zegt: 'je moet deze data vrijgeven', dan zijn die bedrijven eventueel verplicht om dat te doen, terwijl je als eindgebruiker misschien helemaal geen idee hebt dat dit kan gebeuren."

["Mastercard and Visa are American companies. If the US president says 'you must release this data', those companies may be obliged to do so, while as an end user you may have no idea at all that this can happen."]

Kim Wuyts: Co-author of the original LINDDUN Framework

About this session

Kim Wuyts is a co-author of the original LINDDUN privacy threat modelling framework and a leading contributor to its development, which began during her PhD research at KU Leuven. Unlike the seven practitioner interviews (E1–E7), this session was an open consultation on the applied methodology rather than a structured threat validation. The discussion covered the LINDDUN PRO methodology as applied in this thesis, the analytical choices made during the elicitation, and broader suggestions for framing and extending the research. The session lasted approximately 30 minutes.

Summary of the discussion

The session opened with a walkthrough of the DFD and the grouping of data flows by protocol configuration. Wuyts indicated that grouping flows with the same protocol characteristics is an acceptable pragmatic interpretation that avoids an explosion of duplicate threats, provided that differences in data sensitivity across grouped flows are acknowledged where relevant. She also indicated that maintaining the LINDDUN position-level separation (source, flow, destination) as an analytical frame during elicitation is sufficient, even if the final output does not report every position separately, though, depending on the context and the organisation's requirements for the analysis, more detailed documentation of each position may still be advisable where a rigorous demonstration of the process is required.

The discussion then moved to the purpose limitation principle. She clarified that data minimisation and linkability must be assessed per process, not per organisational boundary. An agent performing fraud detection may lawfully access data that an agent performing customer optimisation may not, even within the same bank. This is consistent with the per-dataflow analytical approach used in this thesis and with the grouping structure in Chapters 5 and 6.

On the specific threats, Wuyts briefly engaged with TH-01 (MCP session identifier linkability) and TH-06 (A2A context exceeding data minimisation). For TH-01 she noted that the question of whether linkability constitutes a problem depends on the organisation's intent: if separation between data stores is deliberate, the session identifier that undermines it is a genuine threat. For TH-06 she considered the A2A data minimisation violation structurally important, while noting that in a fraud detection context pushback on grounds of legal necessity is likely. She suggested framing the threats in the discussion chapter against a less regulated context to make the structural severity more apparent to readers who might otherwise dismiss the findings as necessary trade-offs.

She also offered a clarification on TH-04 (A2A bilateral non-repudiation): non-repudiation in LINDDUN concerns the data *subject's* right to deny that their data was processed, not the organisation's operational interest. The threat is therefore valid if the individual would prefer that their screening by a card network not be permanently recorded.

Methodological suggestions

Framing threats in a non-fraud context. Wuyts suggested that the discussion chapter explicitly compare the threat priorities found in this thesis against what they would look like in a less regulated application of the same protocols. In a fraud detection context, legal necessity and regulatory mandate reduce the effective severity of several threats. In a general-purpose agentic AI system, the same structural properties of MCP and A2A would be far more problematic. This comparison would make the protocol-level findings more transferable beyond the specific use case.

Inference-driven database enrichment as an additional threat. She raised a threat not captured in the eight identified: agents that accumulate transaction and KYC data over time could infer sensitive attributes (such as household composition or spending patterns) and enrich internal databases that are subsequently accessible to other processes. This represents a downstream data disclosure risk linked to the inference capabilities of the agent itself, independent of the communication protocol.

Notable quotes

> *“MCP en agent-to-agent hebben by default deze problemen. Dat kan een interessante takeaway zijn: algemeen is dit slecht voor de privacy. Voor ons fraudedetectiesysteem, ik weet het niet, want in die context mag je de data al en moet je ze gebruiken.”*
>
> *[“MCP and agent-to-agent have these problems by default. That could be an interesting takeaway: in general, this is bad for privacy. For our fraud detection system, I am not sure, because in that context you are already permitted and required to use the data.”]*

F Supplementary Methodology Detail

This appendix collects supporting detail for the methodological choices made in Chapter 2: the full descriptions of the candidate frameworks that were evaluated but not selected, the further privacy methodologies that were considered, and the complete definitions of the seven LINDDUN threat categories.

F.1. Evaluation of the Candidate Frameworks

This section gives the full evaluation of the three candidate frameworks summarised in Table 2.1 of Chapter 2.

MAESTRO

MAESTRO (Multi-Agent Environment, Security, Threat, Risk, and Outcome) is a threat modelling framework developed by the Cloud Security Alliance specifically for agentic AI systems [45]. It organises threats across seven architectural layers: foundation models, data operations, agent frameworks, deployment infrastructure, evaluation and observability, a vertical security-and-compliance layer, and the agent ecosystem. Each layer is accompanied by a threat landscape that enumerates the threat types relevant to that layer, addressing risks such as adversarial manipulation, prompt injection, unauthorised tool access, and supply chain vulnerabilities [45]. Although MAESTRO is a Cloud Security Alliance white paper rather than a peer-reviewed academic article, it has begun to be cited as a reference framework for agentic AI risk assessment in peer-reviewed work [74].

Although MAESTRO provides useful coverage of security risks in agentic AI, it is primarily a security framework. The MAESTRO authors themselves describe LINDDUN as "essential for addressing privacy concerns in Agentic AI," but note that it "must be paired with other frameworks to create a comprehensive threat model" [45]; by the same logic, MAESTRO's security focus needs to be complemented by a privacy-specific framework to cover the data-subject dimension this thesis addresses. Two further considerations point in the same direction. First, LINDDUN rests on a decade-long academic foundation, beginning with its original peer-reviewed publication in *Requirements Engineering* [44] and further developed through subsequent extensions [64] and empirical evaluation [11]; MAESTRO, by contrast, was published in February 2025 and has not yet undergone academic peer review. Second, LINDDUN provides a reproducible threat elicitation methodology with publicly maintained threat trees, whereas MAESTRO's contribution is a layer-by-layer threat catalogue without an equivalent systematic elicitation procedure. These properties make LINDDUN better suited as the primary analytical instrument; MAESTRO is retained as a supplementary cross-check during the threat identification (Chapter 5) for risks at architectural levels that fall outside LINDDUN's data-flow scope. Applying LINDDUN to a communication protocol in this way also has methodological precedent: Mirzamohammadi et al. model the Solid protocol as a data flow diagram and run the standard per-flow LINDDUN elicitation against it, showing that the framework's reproducible procedure transfers directly to a protocol-level setting of the kind analysed here [40].

NIST Privacy Framework version 1.1

The NIST Privacy Framework version 1.1 is an organisational governance framework that structures privacy management around five core functions: identify, govern, control, communicate, and protect [46]. It is designed to help organisations assess and improve their privacy programmes at a strategic level, providing high-level guidance on organisational roles, responsibilities, and processes for managing privacy risk [46]. While the NIST Privacy Framework is well suited for establishing and evaluating an organisation's overall privacy posture, it operates at a level of abstraction that is too high for the analysis carried out in this thesis. The framework does not provide a method for identifying privacy threats in individual data flows or system components, nor does it offer tools for analysing the privacy implications of specific protocol interactions. Its unit of analysis is the organisation as a whole, whereas this thesis requires an analysis at the level of individual agent communications and trust boundary crossings. The NIST Privacy Framework is therefore not appropriate for the protocol-level privacy threat analysis carried out in this thesis.

LINDDUN

LINDDUN (Linkability, Identifiability, Non-repudiation, Detectability, Data Disclosure, Unawareness & Unintervenability, and Non-compliance) is a privacy-specific threat modelling framework developed by Deng et al. and published in the peer-reviewed journal *Requirements Engineering* [44]. Unlike

MAESTRO and the NIST Privacy Framework, LINDDUN operates at the level of individual data flows and trust boundaries. This is precisely the level at which MCP and A2A function: both protocols define how data moves between agents, tools, and data sources across organisational boundaries, and it is at these data flows and boundary crossings that privacy threats arise. LINDDUN is therefore structurally aligned with the analysis this thesis requires in a way that neither MAESTRO nor the NIST Privacy Framework can match. Its compatibility with AI agent systems is supported by direct correspondence with a principal developer of the framework indicating that the seven LINDDUN threat categories apply to agentic AI architectures [48]. The security analyses of MCP and A2A reviewed in Chapter 1 further confirm that a separate, privacy-specific analysis is necessary: security-oriented frameworks address different threat categories and cannot substitute for a framework focused on the protection of personal data with respect to the data subject [11].

F.2. Other Privacy Methodologies Considered

Other privacy methodologies were considered but did not match the analytical needs of this thesis. PRIAM [75] is a peer-reviewed privacy risk analysis methodology that uses data catalogues, feared events, and harm trees, but operates at the level of personal data items rather than data flows, which makes it less suited to the protocol-level analysis of agent communications carried out here. Data Protection Impact Assessments under GDPR Article 35 are a legal-compliance instrument rather than a technical threat modelling methodology; rather than competing with such a methodology, a DPIA is the legal vehicle that a method like LINDDUN can help populate [76]. Security frameworks such as STRIDE, PASTA, OCTAVE, Trike, and VAST, as catalogued in the MAESTRO blog post itself [45], are security-oriented and do not provide privacy-specific threat categories. A recent systematic literature review of privacy impact and privacy risk assessment methodologies confirms that LINDDUN and PRIAM are the two most established privacy-specific methodologies, with LINDDUN positioned as a privacy threat model and PRIAM as a privacy risk assessment [76].

F.3. LINDDUN Threat Categories

Table F.1 gives the full definition of each of the seven LINDDUN privacy threat categories, with examples drawn from the fraud-detection setting analysed in this thesis.

Table F.1: The seven LINDDUN privacy threat categories [44], with examples drawn from the fraud-detection use case analysed in this thesis.

Letter	Category	Description
L	**Linkability**	The ability to link two or more items of interest without knowing the identity of the data subject, such as linking two agent interactions to infer a behavioural pattern.
I	**Identifiability**	The ability to identify a data subject from information available in the system, such as identifying a customer from a tokenised card identifier combined with transaction metadata.
N	**Non-repudiation**	The inability of a data subject to deny an action or communication, such as an agent interaction being permanently logged in a way that can be attributed to an individual. In security contexts, non-repudiation is a desirable property: a signed record that proves a transaction occurred. In LINDDUN it is assessed as a privacy threat because it is evaluated from the data subject's perspective: an irremovable record that the individual cannot contest, correct, or have deleted even if the underlying processing was later found to be erroneous or unlawful.
D	**Detectability**	The ability to distinguish whether an item of interest exists, even without access to its content, such as detecting that a specific customer is under fraud investigation from the pattern of agent queries.
D	**Data Disclosure**	The exposure of information to parties who should not have access to it, such as personal data transmitted in an A2A message being visible to unintended recipients.
U	**Unawareness & Unintervenability**	Two related but distinct threats that together determine whether a data subject can exercise any meaningful control over their personal data. **Unawareness** is the absence of knowledge: the data subject does not know that processing is occurring, which data is involved, or which parties have access to it. Without this knowledge, no choice or objection is possible. **Unintervenability** is the absence of a mechanism to act: even a data subject who is informed cannot access, correct, or contest the data being processed about them. Unawareness removes the informational precondition for control; unintervenability removes the practical precondition. Together they establish a condition of complete exclusion from the processing that concerns the individual.
N	**Non-compliance**	The failure of a system to comply with applicable privacy obligations, policies, or standards, such as retaining personal data beyond the purpose for which it was collected.

G Use Case Supplementary Detail

This appendix collects supporting detail for the use case of Chapter 4: the full basis for each data element (Section G.1) and the complete actor mapping (Section G.2).

G.1. Data-Element Basis

The data foundations of the use case are grounded in the research literature on real-time payment fraud detection, in the operational data a card-issuing bank inherently maintains, or in a regulatory obligation; the summary claim is made in Chapter 4, and the supporting derivation is given here.

The use case is hypothetical but grounded in current practice. AI-driven systems for financial fraud detection and risk assessment are actively being deployed across the financial sector [8], and multi-agent architectures built on LLM-based agents are beginning to appear in production banking settings, such as Ryt Bank's regulator-approved retail-banking interface orchestrated by four LLM-powered agents [41]. The Model Context Protocol is increasingly used to connect such agents to internal tools and data sources across organisational systems [7]. These developments reflect a broader shift towards agentic, multi-agent AI architectures [19]; the use case applies that pattern to the real-time aggregation of transaction history, device attributes, and identity records for fraud detection.

The data elements used in this use case are not arbitrary design choices: each is grounded either in the research literature on real-time payment fraud detection, in the operational data that a card-issuing bank inherently maintains, or in a regulatory obligation. They fall into four groups.

Established standard inputs. Transaction history and identity records are the core, well-established inputs to production fraud detection. Cao et al. illustrate this in TitAnt, a fraud detection system deployed at Ant Financial, one of the largest payment processors in the world: TitAnt scores transactions in real time by combining engineered transaction- and identity-level features (such as account profile attributes and transfer context) with aggregated features derived from the transaction network [54]. Sangers et al. further show that collaborative fraud detection between financial institutions requires combining transaction-level features from multiple organisations that each hold complementary but partial views of the same cardholder's activity [77], which is precisely the cross-organisational structure modelled here.

Emerging signal. A **device fingerprint** (a set of technical attributes collected from a user's device, such as device model, operating system, screen resolution, and browser settings, used to distinguish one device from another without requiring a login) is included as a recognised but still emerging signal: TitAnt itself does not yet use it, but explicitly identifies device and IP information as a complementary direction for future extension [54].

Operational issuer data. Card status and authentication history are operational data that a card-issuing bank inherently maintains in order to authorise a payment at all. They are not drawn from the research literature but are present in any real issuing environment, and are included because the workflow cannot reach an authorisation decision without them.

Regulatory-obligation data. The external sanctions screening element corresponds to **anti-money laundering (AML)** checks: a legal obligation requiring financial institutions to screen transactions and counterparties against government-maintained lists of individuals and organisations prohibited from participating in financial transactions. The FATF Recommendations, the international AML/CFT standard, make both customer due diligence (Recommendation 10) and screening against targeted financial sanctions (Recommendation 6) binding obligations for financial institutions [78], so these checks are standard practice across the sector.

The cross-organisational A2A configuration between a card-issuing bank and a card network operator reflects the real structure of the payment card industry, in which these two types of institution each hold complementary data about the same transaction but are operationally independent. Taken together, these elements form the minimal realistic data set for a fraud detection workflow of this type. Table G.1 makes the derivation explicit: for each data element it states the basis and the data store in which it resides.

Table G.1: Basis and location of each data element in the use case. Each element is grounded in the research literature, in operational issuer data, or in a regulatory obligation, and is mapped to the data stores (DS) defined in Section 4.3. The data flows that carry each element are enumerated separately in Table 4.1. Author's own compilation.

Data element	Basis	Data stores
Transaction history	Established standard input: TitAnt [54]; cross-organisational combination of complementary partial views, Sangers et al. [77]	DS1, DS4
Identity / KYC records	Established standard input: TitAnt profile and identity features [54]; legally mandated customer due diligence, FATF Rec. 10 [78]	DS2
Device fingerprint	Emerging signal: identified by TitAnt as a complementary direction for future extension [54]	DS3
Card status	Operational issuer data inherent to payment authorisation	DS3
Authentication history	Operational issuer data inherent to payment authorisation	DS3
Network-wide transaction patterns	Cross-organisational complementary view, Sangers et al. [77]	DS4
Sanctions / AML screening	Regulatory obligation: FATF Recommendations 6 and 10 [78]	DS5

G.2. Actor Analysis

This section gives the full actor mapping for the use case, summarised in Section 4.6 of Chapter 4. Table G.2 maps each actor to its role and position in the system, its primary interest, and the key conflicts that arise between actors. The mapping is read again in Chapter 8, where the distributional character of these conflicts is examined in relation to the identified privacy threats.

Table G.2: Actors in the fraud detection system: role and position, primary interest, and key conflicts. Author's own elaboration.

Actor (role / position)	Primary interest	Key conflict
Bank X Card-issuing bank; operates P1–P2 and holds DS1–DS3 within TB1; sole contractual link to the cardholder	Fraud prevention, regulatory compliance, operational efficiency, liability minimisation	Conflicts with the cardholder on data minimisation and transparency; with Card Network Y on data governance once data crosses TB1 or TB2
Card Network Y Card network operator; independent organisation; operates P3 and DS4 within TB2; no contractual link to the cardholder	Network integrity, fraud reduction across issuing banks, protection of its own data assets	Conflicts with Bank X on re-use of the received profile; with the cardholder on transparency and data retention
Cardholder Data subject; initiates DF1 and receives DF18; account with Bank X only (its terms may include a general authorisation for automated screening); no visibility into the agent interactions	Privacy, accuracy of the fraud decision, right to information and contestation	Conflicts with all other actors: excluded from the system by design and bears the consequences of errors without recourse
Human fraud analyst Bank X employee; reviews escalated cases (DF15); receives the most complete cardholder profile assembled in the workflow	An accurate case decision with sufficient context, within Bank X's governance policies	Holds the fullest profile in the workflow; informal retention or sharing outside the workflow conflicts with data minimisation
External sanctions service Third-party data provider (EE4/DS5) outside both trust boundaries; receives identified personal data	Commercial service provision and completeness of sanctions data	Conflicts with the cardholder on false positives and retention; with Bank X on governance of the transmitted data
Merchant's bank Acquiring institution (EE2) outside both boundaries; receives the outcome on DF18; no part in the assessment	A timely, reliable authorisation outcome so the transaction can be settled	Receives an outcome of extensive processing without visibility; has no relationship with the cardholder governing the data it handles
Financial regulator Outside all trust boundaries; not a DFD element; sets the legal framework for both institutions	Consumer protection, financial stability, and compliance with data protection and financial regulation	Conflicts with both institutions on auditability versus privacy; currently lacks a framework for multi-agent distributed processing

H Communication Protocol Supplementary Detail

This appendix collects supplementary material supporting Chapter 3: the historical context of agent communication protocols, the full taxonomy of current protocols, the MCP and A2A message-format examples, and the protocol reference tables. The privacy-relevant architecture, payload, trust, and data-retention properties of MCP and A2A are presented in the main text of Chapter 3.

H.1. A Brief History of Communication Protocols

The challenge of enabling autonomous agents to communicate in a structured and interoperable way has been studied since the early 1990s, and three generations of agent communication protocols can be distinguished.

The first generation, comprising the Knowledge Query and Manipulation Language (KQML, 1993) [79] and the FIPA Agent Communication Language (2000) [14] (whose speech-act-based approach is discussed in Section 1.2.4 of Chapter 1), was designed for closed, single-organisation environments with known, trusted agents; privacy was not a design concern because data flows were fully controlled.

The second generation was defined by service-oriented architectures and web-based APIs, which replaced formal semantic constraints with lightweight, schema-based message formats. Cross-organisational data exchange became possible through this simplification, but without mechanisms for controlling what personal data was transmitted or how it was used once shared.

The third and current generation is defined by the emergence of LLM-based agent protocols. MCP and A2A operate in open ecosystems where agent capabilities are discovered dynamically, message payloads are unstructured, and sensitive personal data flows across organisational boundaries with minimal constraints [7, 28]. This shift from closed, trust-assumed environments to open, cross-organisational deployments is what makes the privacy properties of MCP and A2A a matter of significant concern, and what motivates the analysis in this thesis.

H.2. A Taxonomy of Current Agent Communication Protocols

With the rise of LLM-based agentic AI systems, several application-layer communication protocols have emerged to address the need for standardised agent interoperability. Anbiaee identifies four protocols that currently represent the state of the art: MCP, A2A, the Agent Network Protocol (ANP), and Agora [10]. Two dimensions are used to organise these protocols for the purposes of this thesis. These dimensions are chosen because they directly determine which privacy risks are present and which protocols can be analysed. The first dimension, communication type, determines which privacy boundary structures apply: a protocol that governs communication between an agent and a tool creates a different trust configuration from one that governs communication between two peer agents, and the two configurations produce structurally different data flows. The second dimension, adoption status, determines which protocols are mature enough to support a structured privacy analysis: only protocols with stable, complete, and publicly available specifications can be mapped to a Data Flow Diagram and systematically analysed. Together these two dimensions identify which protocols are analytically relevant and why.

The first dimension concerns communication type. A distinction can be drawn between **vertical protocols**, which govern communication between an agent and an external tool or data source, and **horizontal protocols**, which govern communication between two agents. MCP is a vertical protocol: it defines how an agent, acting as a client, accesses tools and resources exposed by a server. A2A, ANP, and Agora are horizontal protocols: they define how agents communicate as peers, delegating tasks and exchanging results. This distinction is analytically significant for this thesis because vertical and horizontal communication involve different trust boundary structures and therefore different privacy risk profiles.

The second dimension concerns adoption status. MCP was introduced by Anthropic in 2024 and has since been adopted across major development platforms, with official integrations in tools such as VS Code, Claude, and a growing ecosystem of third-party servers. A2A was introduced by Google in 2025 and is supported by a broad industry coalition. ANP, developed by Alibaba, and Agora, a research proposal, remain largely experimental and lack the production-grade documentation and ecosystem support available for MCP and A2A [29, 10].

Table H.1 summarises the four protocols across the dimensions most relevant to this thesis. Because MCP and A2A are the only two protocols with sufficient technical documentation, active community adoption, and real-world deployments to support a structured privacy analysis, the analysis in this thesis focuses exclusively on these two protocols. The full selection rationale, including the three criteria applied and the exclusion of ANP and Agora, is set out in the scope delimitation in Chapter 2.

Table H.1: Taxonomy of current agent communication protocols. Based on [10, 29, 7, 30, 55].

Protocol	Type	Initiator	Cross-org	Adoption status
MCP	Vertical (agent to tool)	Client-driven	Yes	Production; wide adoption
A2A	Horizontal (agent to agent)	Peer-to-peer	Yes	Production; broad coalition
ANP	Horizontal (agent to agent)	Peer-to-peer	Yes	Experimental; limited ecosystems
Agora	Horizontal (agent to agent)	Peer-to-peer	Yes	Research proposal only

Protocols analysed in this thesis.

H.3. Message Format Examples

Listing H.1 shows an MCP tool call in which personal data appears in both the request and the response, with no protocol-level field recording purpose, accessor identity, or retention period (discussed in Section 3.2.4). Listing H.2 shows an A2A task request carrying structured personal data in a `DataPart` with no schema enforcement at the protocol level (discussed in Section 3.3.4).

```
// Request: agent asks for transaction history
{
  "method": "tools/call",
  "params": {
    "name": "get_transactions",
    "arguments": {
      "account_id": "NL91ABNA0417164300",
      "from_date": "2025-01-01",
      "to_date":   "2025-03-31"
    }
  }
}

// Response: server returns personal transaction data
{
  "result": {
    "transactions": [
      { "date": "2025-01-14", "amount": -249.99,
        "counterparty": "Merchant␣X" },
      { "date": "2025-02-03", "amount": -89.00,
        "counterparty": "Merchant␣Y" }
    ]
  }
}
```

Listing H.1: An MCP tool call requesting transaction history. Personal data appears in both the request (account identifier) and the response (transaction records), with no protocol-level field recording purpose, accessor identity, or retention period.

```
{
  "jsonrpc": "2.0",
  "id": "req-001",
  "method": "message/send",
  "params": {
    "message": {
      "role": "user",
      "messageId": "msg-abc-123",
      "kind": "message",
```

```
      "parts": [
        {
          "kind": "text",
          "text": "Analyse␣the␣behavioural␣profile␣below."
        },
        {
          "kind": "data",
          "data": {
            "account_id": "NL91ABNA0417164300",
            "transaction_count_30d": 47,
            "avg_transaction_value": 312.50,
            "unusual_hours_flag": true
          }
        }
      ]
    }
  }
}
```

Listing H.2: Example A2A task request containing a DataPart. Structured personal data is transmitted in the `data` field with no schema enforcement at the protocol level.

H.4. Protocol Reference Tables

This section collects the reference tables for the protocol mechanics described in Chapter 3: the three constituent elements of a communication protocol (Table H.2), the comparison of MCP Tools and Resources (Table H.3), the privacy-relevant Agent Card fields (Table H.4), and the A2A task lifecycle states (Table H.5).

Table H.2: The three constituent elements of a communication protocol [56].

Element	Definition	Consequence if absent
Syntax	The structure and format of data: how bits and bytes are ordered, which fields appear in a message, and how field boundaries are delimited.	Messages cannot be parsed or read by the receiving party.
Semantics	The meaning assigned to each field and the actions that each message type requires from its recipient.	Messages can be parsed but not understood; no action can be derived from the content.
Timing	The rules governing when data may be sent, at what rate, and how the communicating parties synchronise their interaction.	Interactions that are individually understood may nevertheless fail under concurrent or asynchronous conditions.

Table H.3: Comparison of MCP Tools and Resources on dimensions relevant to privacy analysis. Derived from the MCP specification [55] and Hou et al. [7].

Dimension	Tools	Resources
Discovery method	`tools/list`	`resources/list`
Invocation method	`tools/call` with arguments	`resources/read` with URI
Side effects	Yes; may write, trigger APIs, modify state	No; read-only
Personal data location	`params` and `result` fields	URI (addressing) and response body
Schema enforcement	Input schema per tool (a structured description of the accepted inputs)	None beyond MIME type
Privacy exposure type	Data disclosed in arguments and results	Identity encoded in URI; bulk data in response

Table H.4: Agent Card fields and their privacy relevance [30].

Field	Content	Privacy exposure
`name` and `description`	Identity and purpose of the agent	Reveals the existence and function of the agent to any observer
`url`	Task endpoint address	Discloses operational infrastructure to any observer
`skills`	Capabilities with input and output descriptions	Reveals what data types the agent processes
`securitySchemes`	Required authentication scheme	Discloses security posture before any task is delegated
`defaultInputModes`	Accepted media (MIME) types, e.g. `text/plain`, `application/json`	Reveals the categories of personal data the agent accepts

Table H.5: A2A Task lifecycle states [30].

State	Meaning
`submitted`	Task received by the server and acknowledged; processing not yet started
`working`	Server is actively processing the task
`input-required`	Server requires additional input from the client before proceeding (paused)
`auth-required`	Server requires additional authentication before proceeding (paused)
`completed`	Task finished successfully; results are available
`failed`	Task terminated due to an error during processing
`canceled`	Task was canceled before completion
`rejected`	Task was rejected by the server and never started
`unknown`	Task state cannot be determined (e.g. invalid or expired ID)

Successful terminal state Unsuccessful terminal states.