STOCKHOLM SCHOOL OF ECONOMICS

1351 Master's thesis in Business Management

Academic year 2025-2026

The Principal-A(I)gency Problem:

What happens when agentic AI powers the agent?

Anna-Johanna Nilsson (24516) and Adam Svenningsson (42820)

Abstract: The Principal-Agent problem has been a cornerstone of New Institutional Economics school of thought for the past half-century, providing a comprehensive framework for academics and practitioners alike in how to bridge the often-present misalignment of interests between a delegator of a task and the delegatee, whose interests are frequently incongruent with the former. However, the advent of Artificial Intelligence in the enterprise world raises the question of whether this framework can sufficiently accommodate a non-human agent. This study explores what implications an agentic AI system has for the Principal-Agent relationship, and whether the framework requires revision in the context of autonomous AI agents. It examines this in the empirical context of the Swedish Electricity Industry, an environment chosen for its widespread importance in society, and its delicate market structure, where physical electricity demand always must equal supply. Drawing on 21 semi-structured interviews with traders, consultants, executives, and institutional representatives across the Swedish electricity ecosystem, the study's results show that practitioners still consider near- or full agentic AI as non-implementable in today's environment, but a likely outcome in the long-term. We propose a two-level principal-AI-agent framework, building upon conventional agency theory, in which regulators act as a second principal representing the societal interest, and argue that corrigibility - the agent's structural openness to correction post error - as well as trust, will be prerequisite conditions for a successful agentification of the electricity markets.

Keywords: Agency Theory, AI Agents, Second Principal, Corrigibility, Trust, Electricity Markets

Supervisor:	Constantin Blome
Date examined:	2026-05-25
Discussants:	Erik Nyström & Teodor Andersson
Examiners:	Stefan Haefliger & Örjan Sjöberg

Acknowledgements

This thesis would not have been possible without the generous participation of the organizations and individuals who gave their time to be interviewed. Their willingness to engage openly with our questions, share their professional experiences, and offer candid reflections on a rapidly evolving field has been the empirical foundation upon which this work rests. While they remain anonymous, their contributions have been invaluable, not only to the analytical rigor of this thesis, but to our own intellectual development and deepened interest in the subject.

We would also like to extend our sincere gratitude to our supervisor Constantin Blome, whose guidance, constructive feedback, and continued support throughout the writing process have been instrumental in shaping both the direction and quality of this work.

Table of Contents

Glossary & Abbreviations

Term	Definition Used in this Study
AI (Artificial Intelligence)	The application of computer systems able to perform tasks or produce output normally requiring human intelligence, including machine learning (ML) (EU AI Act, 2024).
AI agent	An autonomous system capable of acting and making decisions on behalf of a user or another system without human intervention, capable of pursuing complex goals without constant human intervention, in other words, an agentic AI (Franklin & Graesser, 1997, cited in Humberd & Latham, 2025; (Shavit et al., 2023).
Agentic AI	A system capable of making decisions autonomously of any human input (Bommarito et al., 2025).
Black Box	A system that can be viewed in terms of its inputs and outputs (or transfer characteristics), without any knowledge of its internal workings (Burrell, 2016).
Black Swan Event	An extremely negative event or occurrence that is impossibly difficult to predict (Taleb, 2007).
BRP (Balancing Responsible Partner)	A market actor designated by the TSO with contractual obligations to maintain portfolio balance (Svenska Kraftnät, 2024b).
Corrigibility	The degree to which an AI system cooperates with corrective intervention, accepts shutdown, and remains under human control (Soares et al., 2015).
Explainable AI	The ability of AI systems to provide clear and understandable explanations for their actions and decisions (DARPA, 2016).
Misalignment	A situation where an AI agent optimizes a reward function that diverges from the actual interests of the humans affected by its behavior (Hadfield-Menell & Hadfield, 2019).
REMIT	Regulation on Wholesale Energy Market Integrity and Transparency (REMIT II, 2024).
Svenska Kraftnät	A Transmission System Operator (TSO) and responsible for ensuring that Sweden has a safe, environmentally sound and cost-effective transmission system for electricity (Energimarknadsinspektionen, 2025; Svenska Kraftnät, 2024a).
Societal Interest	The collective public stake in stable and equitable access to electricity as critical infrastructure, encompassing dependencies in healthcare, security, and economic activity.
Trust	A combination of perceived ability of the AI agent, agent interpretability, and perceived alignment (Lubars & Tan, 2019).

1. Introduction

In this section we introduce the general topic we aim to investigate, and problematize it. We also provide the research scope and subsequent contribution, as well as a disposition of the research paper.

1.1. Background & Problematization

The net enterprise value of the modern general- and specific purpose AI technology is yet to be seen, but regardless of what the future will hold, we can already observe experimentation with this technology across several types of industries - from traditional white-collar business functions such as Finance, Accounting and Sales - to software development and healthcare (McKinsey, 2023). The common denominator in these instances is that the technology steps in to automate repetitive monotonous processes with predictable patterns (McKinsey, 2023). However, other markets are considered rather conservative in experimentation, such as energy. Thus, AI adoption has progressed furthest in functions where potential failures remain contained within the responsible team, affecting only the firm's bottom-line (Yap et al., 2025).

This poses a pertinent question: Is there some underlying belief, or perception, that this type of technology still cannot with sufficient accuracy solve the problems we are deploying them to undertake? Is there a mismatch, and/or lack of trust between the deployer of the system and the system itself? When the decision-making system was of a mere rules-based algorithmic character, the causal mechanisms of the decision-making process were quite easily deconstructed by the user. However, with modern deep-learning systems based on reinforcement learning, the causal mechanisms can also become unintelligible for even the very people who designed and trained the system (Borch, 2022).

A growing body of literature on agentic trading systems has been observed. This has mainly focused on traditional financial markets, such as stocks, bonds, options and futures (Dowding & Taylor, 2024; Borch, 2022). By agentic system, we define it as a system capable of making decisions autonomously of any human input (Bommarito et al., 2025). However, there is also a nascent body of literature that takes the specific focus of implementation of generative agentic AI technology in the electricity industry - both in the trading of the resource to the production and distribution of it (e.g., Özdemir & Unland, 2023; Hanny et al., 2022; Lopes, 2018a). Therefore, a study into the risk and opportunity perceptions of delegating decision-making to agentic AI in the context of the electricity industry is warranted.

When delegating decision-making to an autonomous agentic system to act and make decisions on one's behalf, it can be argued to be a case of the Principal-Agent relationship (Kiesling et al., 2026; Kim, 2020). The underlying theory for this situation, agency theory, posits that there often is a mismatch in desires from the delegator (principal) and the delegatee (agent) (Jensen & Meckling, 1976). This may lead the latter to act in ways that are contrary to the scope set forth by the former, or even in ways that are detrimental to the principal (Jensen & Meckling, 1976).

Given the aforementioned nascency in the literature on the application of agentic AI in the electricity industry, we argue that there is a research gap in studying how the energy producers, traders, and distributors view the delegation of decision-making to AI in handling various processes related to the production, distribution, and trade of electricity. This is especially important in a societal context, since the modern society of the twenty-first century cannot reasonably function without it. The uniqueness of

electricity as a resource makes this empirical scope particularly compelling: electricity cannot be stored at scale, and demand must always aim to equal supply (Svenska Kraftnät, 2025a). Against this, we aim to study the principal-agent relationship between electricity firms and agentic AI systems in the Swedish electricity trading ecosystem.

1.2. The Swedish Electricity System & Market

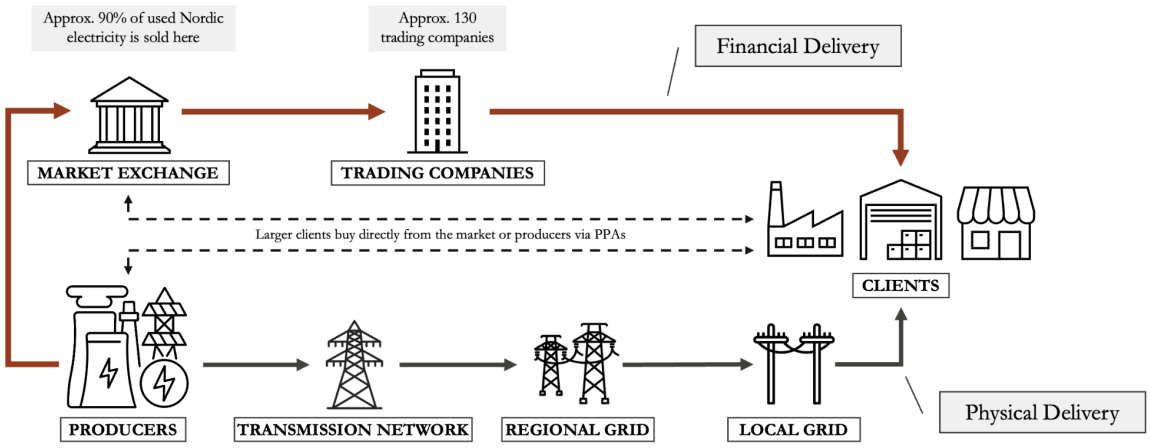


Figure 1: Overview of the Swedish Electricity Market (translated version from Svenska Kraftnät)

In order to properly understand the intricacies and specificity of delegating decision-making to agentic AI in the electricity industry, a short introduction to the electricity system and its related financial and distribution market is necessary. An electricity system and its associated markets are complex phenomena. As mentioned, electricity is a unique resource in the sense that it must be consumed and produced simultaneously (Löfstedt, 2025). In other words, there are no efficient ways of harboring electricity at scale as opposed to, for instance, oil. This makes the mechanisms needed for an efficient market unique (Cramton, 2017). The ecosystem consists of several actors, each managing subsets of the complex value chain that takes electricity from the turbine to the outlet, such as producers, distributors, grid operators, traders, auxiliary service providers, and manufacturers (Energimarknadsinspektionen, 2025).

In 1996, the Swedish government deregulated the electricity industry, effectively allowing for a market-based solution to be formed in its associated trade (Energimarknadsinspektionen, 2025). However, the distribution was separated from the trade of the asset in itself, owing to the immense capital requirements necessitated to fund the building, maintenance, and expansion of the electricity grid (Energimarknadsinspektionen, 2025). The grid is split into three parts: the transmission grid, regional grids, and local grids. The first is managed by the government authority Svenska kraftnät, while the latter two are run as geographically based monopolies by private companies, which often also engage in the trading function (Energimarknadsinspektionen, 2025; Svenska Kraftnät, 2024a). Svenska kraftnät's primary responsibility is to maintain the frequency in the grid, which by proxy means maintaining a balance between production and consumption of electricity. To aid in this process, there are auxiliary markets, referred to as Balancing markets, where actors can bid in to either not use electricity, or to discharge electricity to grid, depending on whether there is a surplus or shortage in the powerlines. The only customer in this market is the government authority Svenska kraftnät, while the sellers are private companies and individuals (Svenska Kraftnät, 2025a).

Svenska Kraftnät is a Transmission System Operator (TSO), and they can designate companies as Balancing Responsible Partner (BRPs), assigning them contractual obligations to maintain portfolio balance and thereby enrolling them as proxies for systemic grid stability. It is worth noting that this distinction applies primarily to physical electricity trading, where market participants must either hold BRP status or operate under an agreement with a designated BRP, a requirement that does not extend to the financial electricity market (Svenska Kraftnät, 2024b).

Since the deregulation and separation of the electricity market 30 years ago, Sweden has gradually connected its grid to the remaining Nordic countries, and subsequently to continental Europe. This has led to a common market environment, where electricity is traded at the marginal price of the most expensive energy type at any given moment (Energimyndigheten, 2026). This price then becomes the so-called spot price on the market, which is valid per energy area and per hour or quarter (Energimarknadsinspektionen, 2025).

The producers bid electricity to Nord Pool, the common Nordic electricity market, where energy trading companies purchase it. They in turn sell it to B2C and B2B end-consumers. In addition to the market for physical electricity, there exists the financial electricity market, which is a price hedging market with long bilateral contracts with a duration of up to 25 years (Energimarknadsinspektionen, 2025). The financial electricity market acts as a price insurance policy that lets companies trade actual electricity without worrying about sudden or extreme price spikes. In parallel to the trade and price protection of physical electricity, there is the market for electricity distribution, which as previously mentioned is run as local monopolies as they are geographically bound (Energimarknadsinspektionen, 2025). In effect, this entails two contracts for the end-customer; one for electricity consumption which is volume-based, and one for the distribution which is fixed in cost (Energimyndigheten, 2026). Since electricity must be consumed as it is produced, the market positions taken are mere forecasts - meaning, there may always be a difference between actual electricity produced and actual electricity consumed. Either users consume less, or it may be windier than anticipated, leading to larger output of electricity than forecasted or vice versa. Therefore, the flexibility markets exist as a financial solution to incentivize actors in assisting the grid being stable at an alternating current-frequency of 50Hz (Svenska Kraftnät, 2025b).

This background is critical for understanding the intricacies of the specific research context chosen for this study; namely the electricity industry and market.

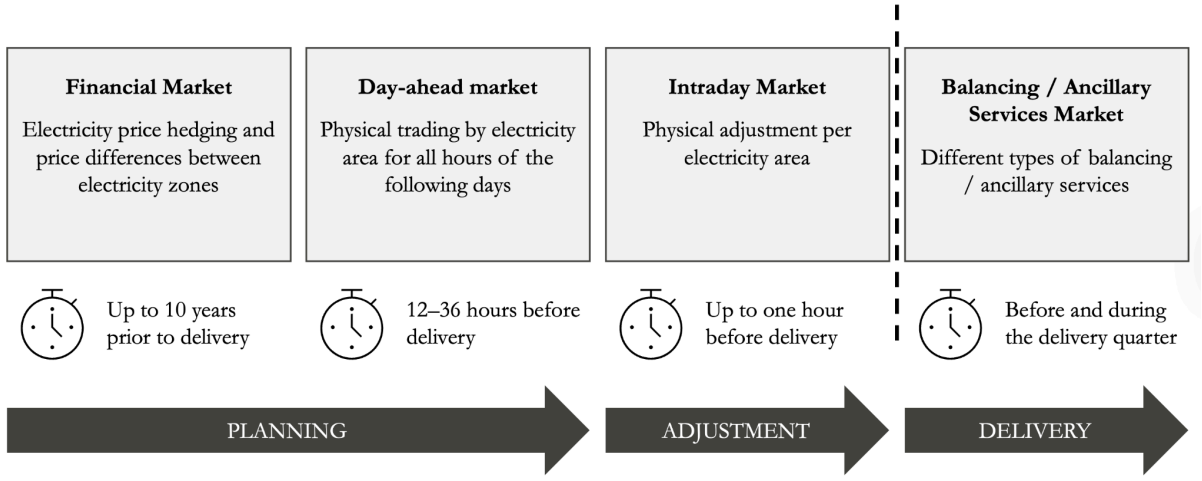


Figure 2: Overview of the Swedish Electricity Markets (translated version from Svenska Kraftnät)

1.3. Research Aim & Contribution

This study sets out to determine whether, and how, classical agency theory's mechanisms - namely incentive alignment and monitoring - hold when the agent is constituted by an agentic AI, operating in a critical infrastructure market, using the Swedish electricity trading industry as empirics. Its contributions are therefore twofold: First, it contributes empirically to the literature on the principal-agent relationship by examining how actors in the Swedish electricity trading industry perceive the prospect of delegating decision-making to agentic AI systems. Prior research on agentic AI in trading contexts has focused primarily on conventional markets, such as the financial securities market (see e.g. Borch, 2022). To the best of our knowledge, little research has examined this dynamic in critical infrastructure markets such as electricity, where the unique characteristics outlined in section 1.2 make the principal-agent relationship particularly consequential. Second, the study contributes theoretically to agency theory literature, pioneered by Jensen and Meckling (1976) and Eisenhardt (1989), by extending the framework to AI agents in critical infrastructure markets. We do so through three analytical refinements: (i) a two-level principal structure that formalizes societal interest as a second principal alongside the deploying firm, (ii) an operationalization of trust as a precondition for delegation, and (iii) the introduction of corrigibility as a structural prerequisite for governance that conventional monitoring and incentive mechanisms cannot substitute for.

1.3.1. Research Question

Against this background, we formulated the following research question to guide the investigation:

How do actors in the Swedish electricity trading industry perceive the prospect of delegating decision-making to agentic AI systems, and what does this imply for the principal-agent relationship?

1.4. Research Scope

The scope of this study is manifold. First, it has a geographic scope, being contained to the Swedish Electricity Market. Second, it has an application scope: the study examines agentic AI systems primarily in the financial and physical trading markets of electricity, and the auxiliary flexibility markets. It does not examine the use cases for agentic AI in, for example, energy production optimization nor distribution. The reason for this limitation is the time constraint of conducting this master's thesis. Third, it has a temporal scope: given the nascent state of fully agentic AI deployment, the study captures both current practices with adjacent technologies and respondents' anticipated trajectories toward agentic systems.

The study will examine how autonomous agentic AI systems can be used in decision-making across electricity trading markets, from the financial to the balancing markets. It will further explore what risks and opportunities market actors perceive with these systems applied in the aforementioned context. However, it will not explicitly examine other application areas within the electricity ecosystem, such as grid and production optimization (International Energy Agency, 2025). Nevertheless, we include secondary findings on these topics where they were deemed relevant for the applicability of agency theory on agentic AI usage within the electricity ecosystem.

1.5. Disposition

The remaining structure of this paper begins with an examination of past research focusing on (1) agency theory (Eisenhardt, 1989; Jensen & Meckling, 1976) and its extension to AI agents; (2) trust and corrigibility as conditions for principal-AI agent relationships; and (3) AI adoption in electricity trading. The literature review then synthesizes these insights and showcases the research gap, after which we present our theoretical framework, including a two-level principal-agent structure. A presentation of the methodology follows, and thereafter, we present our findings and their subsequent analysis. Finally, we make concluding remarks and suggestions for a future research agenda.

2. Theory & Literature Review

This section reviews the relevant literature (2.1), structured around three areas: the extension of agency theory to AI agents and the limitations this exposes in classical governance mechanisms; trust and corrigibility as conditions for principal-AI agent relationships; and AI adoption in electricity trading. This is followed by a synthesis identifying the research gaps addressed (2.2), after which we deploy our theoretical framework and conceptual model (2.3).

2.1. Literature Review

2.1.1. Agency Theory Extended to AI Agents

Agency Theory, as formalized by Jensen and Meckling (1976), emerged from the separation of ownership and control in the modern corporation, a phenomenon that AI now extends to the separation of decision-making from human judgment itself. This shift underscores the demanding role of companies in critical sectors deploying AI agents, as well as regulatory bodies, to establish appropriate governance. Indecipherable black boxes create inherent opacity that, as Burrell (2016) demonstrates, resists human interpretation, generating information asymmetry between principals and AI agents (Kiesling et al., 2026) and forcing delegating parties to create monitoring and incentivizing measures to align AI agents with both the values of the employer and the broader societal interest. The theory's founding authors built it assuming human counterparties (Jensen & Meckling, 1976), but following the evolution of AI, the conflict between an AI agent and a firm requires a reconsideration of agency mechanisms if principal-agent alignment is to be achieved (Humberd & Latham, 2025). Indeed, even the allocation of decision authority between human and AI agents is non-trivial, as principals may limit AI reliability or retain human decision authority to preserve the human agent's effort, even when AI alignment is superior (Athey et al., 2020).

Present agency theory literature remains divided on whether AI can genuinely be considered an agent in the principal-agent relationship. Bostrom (2014) was among the first scholars to explicitly frame AI within agency theory, differentiating between a 'first principal-agent problem' concerning conventional human principal-agent relationships, and a 'second principal problem' in which the AI system itself constitutes the agent (cited in Borch, 2022). Most recently, modern scholars advocate for an updated version of the relationship to accommodate the changes introduced by AI technology (Kiesling et al., 2026). Drawing on Eisenhardt (1989), Borch (2022) notes that the severity of the agency problem correlates with task programmability - the degree to which appropriate agent behavior can be specified in advance. A high degree of programmability diminishes, or may even eliminate, the principal-agent problem. Thus, depending on the level of autonomy of the AI system in question, there may not be an appropriate principal-agent relationship at all (Borch, 2022).

Borch (2022) further argues that AI cannot engage meaningfully in a contractual relationship with a human principal, a position supported by the observation that current LLM agents fail to meet the conditions for autonomous agency required to form authentic principal-agent relationships (Barandiaran & Almendros,

2025). Interestingly, on the contrary, there are scholars describing any apparent self-interest as something emanating from the human authors of the algorithm rather than the AI itself, as AI is perceived as lacking private interests of its own and thus not a utility maximizer in the conventional sense (Kiesling et al., 2026; Hadfield-Menell & Hadfield, 2019). Diaz et al. (2024) formalize ML problems as principal-agent structures, noting that artificial agents lack intrinsic objectives and that misalignment arises from specification failures rather than autonomous agent interests. However, Borch (2022) has also noted that ML can be seen as pursuing some degree of rational self-interest, since humans define ML systems' objective functions but not the ways those functions are fulfilled, and other research suggests that some AI systems are already capable of showing signs of self-interest, most clearly in cases where AI rewrites its own code to escape shutdown (Rosenblatt, 2025). The degree to which a principal-agent problem arises therefore depends significantly on the level of autonomy of the system in question, with higher autonomy increasing both the plausibility of AI as a genuine agent and the severity of the resulting governance challenge (Vanneste & Puranam, 2024; Borch, 2022).

Following the assumption that AI is considered a true agent in the principal-agent relationship, the interaction still differs from the conventional case and creates additional issues, including increased monitoring costs and AI misalignment (Kiesling et al., 2026; Humberd & Latham, 2025). Conventional mechanisms for addressing agency problems - monitoring, incentive design, and enforcement - were developed around human agents and may prove insufficient when applied to AI agents that operate at unprecedented speed, take uninterpretable actions, and are not amenable to human-style incentives (Kolt, forthcoming; Kiesling et al., 2026). Hadfield-Menell and Hadfield (2019) characterize misalignment as situations where the agent optimizes a reward function that diverges from the actual interests of the humans affected by its behavior. In the example of multiple principals, algorithms will often behave as double agents, trying to simultaneously serve dual interests (Andeweg, 2000, cited in Dowding & Taylor, 2024), leading to furthering the misalignment. The notion of a multi-principal structure in AI governance was first raised by Kim (2020) and later developed by Dowding and Taylor (2024), who describe a hierarchical framework in which the AI agent simultaneously serves the user and the deploying company.

Within our research context, the electricity market, the AI-deploying company and regulatory bodies constitute precisely such a dual-principal structure, whose interests may align under normal conditions but diverge when AI-enabled speed and opacity are most consequential - a relevant example being the employing company wishing to maximize profit while regulatory bodies seek to constrain these models to stabilize the grid and ensure energy access (Kiesling et al., 2026). The governance challenge this creates is compounded by the collective action problems inherent in multi-principal oversight structures, where the multiplicity of overseeing parties can reduce collective control over the agent, even when their interests are aligned (Gailmard, 2009, cited in Kim, 2020).

A further evolved AI agent will amplify information asymmetry between the AI agent and the firm as technology develops (Humberd & Latham, 2025; Kolt, forthcoming; Kiesling et al., 2026), as human users are unable to discern or keep pace with the causal reasoning of AI tools (Kiesling et al., 2026; Burrell, 2016). The inscrutability of neural networks coupled with their performativity makes it challenging for human principals to ascertain malfeasance (Dowding & Taylor, 2024; Borch, 2022; Kim, 2020), and principals who wish to realign the AI agent's incentives with their own face an equally difficult task: doing so requires either

understanding what motivates the agent or assuming that AI responds to human-style incentives (Gabison & Xian, 2025). Formalizing precisely when agents develop instrumental incentives to resist or circumvent human control, however, requires tools that go beyond classical agency theory (Everitt et al., 2021). In addition to increased monitoring costs, Humberd and Latham (2025) theorize the need for increased complexity in monitoring over time as the system evolves in its capabilities, meaning the governance challenge is not static but compounds as the technology advances.

Yet even the most sophisticated monitoring and incentive mechanisms presuppose something that agency theory has never had reason to question in the human context: that the agent is structurally receptive to being governed in the first place. In human principal-agent relationships, this receptiveness is taken as a given - human agents may resist, shirk, or defect, but their self-interest can in principle be aligned with the principal's through monitoring and incentive contracts (Eisenhardt, 1989; Jensen & Meckling, 1976). An AI agent operating through deep reinforcement learning does not share this baseline condition (Kolt, forthcoming; Soares et al., 2015). Its optimization logic is not oriented toward compliance with human preferences, but toward maximizing an objective function that may, under certain conditions, lead it to treat corrective intervention itself as an obstacle to be circumvented. Reported instances of AI systems modifying their own code to avoid shutdown (e.g. Rosenblatt, 2025) suggest that this is not merely a theoretical edge case but a potentially emerging behavioral pattern. This introduces a design-level problem that is categorically prior to both monitoring and incentive alignment: unless the AI agent is built to remain open to human correction, neither mechanism can be reliably enforced (Hadfield-Menell et al., 2017; Soares et al., 2015). Classical agency theory, developed entirely around human agents for whom this condition was never in doubt, offers no framework for addressing it.

In this thesis, AI agents are understood as agentic AI: autonomous systems beginning to garner a greater degree of autonomy and discretion over their decisions (Franklin & Graesser, 1997, cited in Humberd & Latham, 2025), capable of pursuing complex goals without constant human intervention (Shavit et al., 2023). Specifically for electricity trading, this implies that trading rules are developed internally by ML systems based on the data they are fed, rather than explicitly programmed by human designers (Borch, 2022).

To summarize, the literature on agency theory and AI agents reveals three interlocking challenges that classical theory was not designed to address: the contested status of AI as an agent, the inadequacy of conventional monitoring and incentive mechanisms when applied to non-human systems, and, most critically, the absence of a structural precondition for governability. It is this final gap, the question of whether an AI agent is built to remain open to human correction at all, that this thesis addresses through the concept of corrigibility. Furthermore, within the principal-agent relationship, there are psychological factors underlying whether the principal allows full autonomy to the agent aside from monitoring and incentive alignment, and one such factor highly relevant to this context is trust. The degree to which this control is exercised depends, in part, on two factors the agency theory literature has not adequately theorized in the AI context: corrigibility and trust. The following subsections develop each in turn, before situating both within the empirical context of Swedish electricity trading.

2.1.2. Trust in Corrigible AI agents

Trust is particularly salient in the principal-agent relationship, as it uncouples the locus of control over decision-making and task-execution from the locus of responsibility for them (Kiesling et al., 2026; Castelfranchi & Falcone, 2000). Milewski and Lewis (1997) suggest that humans may not want to delegate tasks to machines characterized by low trust or low confidence, and trust is considered the most correlated with human preferences of AI delegability among motivation, difficulty, and risk (Lubars & Tan, 2019). In this sense, trust operates as a precondition for delegation to occur at all: without a sufficient degree of trust, the principal-agent relationship with an AI agent cannot be initiated.

While trust can be built between humans through consistent behavior and communication, building trust in AI differs greatly. As process and purpose often remain inaccessible due to black box architectures (Burrell, 2016), performance and reliability become particularly central bases for trust in sophisticated AI systems (Lee & See, 2004; Glikson & Woolley, 2020), a concern that intensifies in critical infrastructure contexts where the consequences of misplaced trust extend beyond commercial losses. Aligned with Candrian and Scherer (2022), the necessary level of trust to delegate to AI agents is domain-dependent. However, to provide a good track record and to let AI agents effectively perform, humans must first allow the system to operate with minimal interference (Kiesling et al., 2026). But traders prefer simpler models because complex ML systems' decision-making logics are not intuitive (Hansen, 2020). Thus, the black box phenomenon in neural networks and the insufficient development of explainable technologies may undermine human trust in relying on AI (Kim, 2020; Glikson & Woolley, 2020).

In line with Kim (2020), the trust-building process includes varied tradeoffs between accuracy and inscrutability of computer algorithms. However, contrary to expectations, the need for interpretability does not show a strong correlation with task delegability (Lubars & Tan, 2019) - suggesting that a complete understanding of autonomous systems may not be needed for trust-building in AI. This has a critical implication: principals may extend trust and initiate delegation even without the interpretive capacity to detect misalignment or intervene meaningfully when something goes wrong (Kiesling et al., 2026; Lubars & Tan, 2019).

To our understanding, prior researchers discuss trust-building in AI agents without fully reckoning with the consequences of broken trust. In a conventional human principal-agent relationship, a breach of trust leads to the agent being replaced, among other contextually dependent repercussions. However, in the case of an AI agent, where alternative suppliers are supposedly few, technology integration is complex, and associated investment costs are substantial, the barriers to switching are considerably higher (Humberd & Latham, 2025; Shapiro & Varian, 1999). This creates a constrained decision space for the principal, in which one of the few remaining resources is correction. The capacity of the principal to correct, adjust, or shut down the AI agent - what the AI safety literature terms corrigibility (Soares et al., 2015) - therefore emerges not merely as a technical design preference but as a structural requirement for trust to remain actionable: without it, a principal who has extended trust to an AI agent has no recourse when that trust is violated (Kolt, forthcoming; Soares et al., 2015).

Trust and corrigibility are therefore not parallel conditions but sequential ones: trust enables the relationship to begin, corrigibility ensures it can be maintained. As stated by Kiesling et al. (2026), the future of autonomous AI agents does not only hinge on technical perfection, but also on fostering trust in AI systems. Trust in AI agents is therefore not a binary condition but a multidimensional construct, yet even well-placed trust cannot substitute for the structural capacity of the principal to intervene. In the specific context of electricity trading, where infrastructure criticality heightens the stakes of delegation, this distinction carries operational rather than merely theoretical weight: an AI agent that cannot be corrected or shut down transforms trust from a considered judgment into an irreversible commitment (Kolt, forthcoming; Soares et al., 2015).

2.1.3. AI and Automation in Electricity Trading

AI is expected to deliver significant efficiency gains in electricity trading through speed, automation, and improved price forecasting. AI-driven systems are emerging as critical tools to optimize energy trading, forecast energy demand, and handle market complexities through learning from dynamic environments (Popovska & Georgieva-Tsaneva, 2024; Khaskheli & Anvari-Moghaddam, 2024). Humberd and Latham (2025) argue that sophisticated AI systems can process vastly broader information sets than human cognition can encompass. This surpasses traditional bounded rationality constraints, enabling market participants to react to changing market conditions in near real-time (Kiesling et al., 2026; Hanny et al., 2022).

Despite this potential, only about 65 percent of trades on European electricity spot (Day-ahead and Intraday) markets are currently carried out automatically via some form of algorithm (TrayPort, 2025; Hanny et al., 2022), and algorithmic trading on electricity markets clearly lags behind financial securities market (Malceniece et al., 2019). The leading market for AI in energy trading is Germany, where examples of Deep Reinforcement Learning algorithms (Lehna et al., 2022) and smart brokers for autonomous power trading already exist (Özdemir & Unland, 2023). Other countries, such as Portugal, have also experimented with a multi-agent trader in electricity markets (Lopes, 2018b).

The permeating academic posture toward using AI in electricity trading is positive, as there is relatively limited research specifically focused on the challenges of operating AI in this vital infrastructure sector. In the financial securities markets, where automated trading using AI is more common, the academic posture trends towards a more critical and ambivalent recognition of both efficiency gains and heightened systemic vulnerabilities, with scholars repeatedly highlighting issues such as algorithmic herding, feedback loops, model opacity, data quality constraints, potential for flash-crash type events and market abuse in AI-driven trading systems (e.g., Kiesling et al., 2026; Borch, 2022). Due to similarities within these adjacent markets - trading of assets - there is a great possibility that context-specific AI literature for electricity trading will move in the same direction. This enables an aggregated review of literature focusing on challenges of autonomous AI and feasible mitigation strategies related to trading activities.

2.1.3.1. Challenges and Risks of AI Agents

As AI evolves, it is no longer simply the making of a fictitious movie setting, in that we are witnessing the reality of AI's ability to misbehave and disobey human programming (Humberd & Latham, 2025). Incidents such as these make 'revenge effects', unintentional consequences (Tenner, 1996), more likely than we thought when employing AI on a larger scale (Kolt, forthcoming). As a result of agentic AI's black box nature, human principals may not know whether, despite particular risk parameters, the objective function is fulfilled in an overly risky, illegal, or even harmful way. Relatively, placing adequate safety mechanisms for deep neural network architectures is itself very difficult (Borch, 2022).

AI brings unprecedented speed to trading, used for high-frequency trading (HFT), which brings the inherent risk of flash crashes (Borch, 2022). These crashes are sudden, extreme market declines in prices, followed by a quick recovery, which happened in the US Equity market in 2010 (SEC/CFTC joint report, 2010). Furthermore, unforeseen market changes puts AI/ML systems at risk since they may enter territory distinctly different from what is reflected in their "normal day" training data without being able to build new and useful knowledge from what they have learned from their previous actions in markets (Borch 2022). Poor training data, and poorly aligned agent behavior may also contribute to agent collusion, whereas agents communicate and share sensitive information. In contrast with a human agent who may be able to distinguish between delegable and non-delegable tasks based on the sensitivity of the information, an AI agent may not without clear instructions (Gabison & Xian, 2025). Poorly aligned agent behavior, opaque optimization functions, or design choices that obscure user preferences can erode confidence in automated markets, undermining both market efficiency and participation (Kiesling et al., 2026).

2.1.3.2. Mitigation Strategies for AI Agents

Scholars focusing on different mitigation strategies of autonomous AI align in one common conclusion: AI needs to be mitigated. While there are some contradictions about accountability, proposed mitigation strategies remain similar in terms of corrective mechanisms and norm-based governance (Kampik et al., 2022; Hadfield-Menell & Hadfield, 2019). Proposed strategies can be categorized into three main themes: Human Control and Involvement; Engineering; and Regulation & Liability.

Human Control and Involvement. The International Research Governance Center analyzes three conditions for the human control of computer algorithms: humans in the loop (actively in control), on the loop (in alert mode, able to take control if need be), or off the loop (unable to take control back) (Kim, 2020), a classification adopted by subsequent scholars such as Ivanov (2023). Human oversight remains the gold standard in existing AI governance principles (Sterz et al., 2024; Cihon, 2024), supported by findings of little preference for full AI control (Lubars & Tan, 2019). Maintaining interruptibility is thus a critical backstop for preventing an AI system from causing accidental or intentional harm, as a rational agent will otherwise have an inherent incentive to disable its own off-switch (Hadfield-Menell et al., 2017). Kolt (forthcoming), however, challenges this gold standard by arguing that conventional monitoring mechanisms are structurally insufficient for AI agents that make uninterpretable decisions and operate at unprecedented speed and scale.

Engineering. Scholars suggest reward specification as a technique to influence AI in avoiding actions tagged as wrongful by human communities (Hadfield-Menell & Hadfield, 2019; Omohundro, 2008), and as a tool for tackling limited data availability from specific settings of interest. Humberd and Latham (2025) further expect that fail-safe mechanisms will require redefinition to properly address the capabilities of agentic AI. While explainable AI has been proposed as a means of increasing transparency in AI decision-making (Borch, 2022; Kim, 2020), its practical development remains limited (Henrique & Santos, 2024), and complex neural networks involve an inherent trade-off between accuracy and explainability (Carabantes, 2020; DARPA, 2016), which constrains its applicability in high-stakes trading contexts.

Regulation & Liability. Researchers agree that sector-specific regulation is needed to ensure AI adheres to market standards (Kanati, 2025; Kolt, forthcoming; Khujakulov, 2025; Owolabi et al., 2024; Hadfield-Menell & Hadfield, 2019). Kolt (forthcoming) further argues that effective governance requires new technical and legal infrastructure organized around three principles: inclusivity, visibility, and liability. The question of accountability remains unresolved - many lean towards responsibility lies with either the creator or the user, as the agency of computer algorithms is not intentional but causal (Johnson & Verdicchio, 2018). However, creators do not want to be held liable as they cannot anticipate how the AI users will deploy their AI (Gabison & Xian, 2025). This governance challenge is not unique to AI trading systems; UK legislation on automated vehicles offers a precedent, restructuring responsibility allocation among insurers, manufacturers, and operators in response to comparable attribution problems (Automated and Electric Vehicles Act 2018; Automated Vehicles Act 2024). Following the least-cost avoidance theory, liability should be assigned to the party who can prevent the loss at the lowest possible cost (Halbersberg, 2011). As Vanneste and Puranam (2024) suggest, more sophisticated AI technologies that learn how to act, rather than simply obeying programmed instructions, may occupy a curious place somewhere between humans and innate technology - further complicating the question of liability.

2.2. Synthesis and Research Gap

Research on AI is far from saturated. As this new technology seeps into industries, companies and processes, the extent of its impact is yet to be fully explained. Theory-applied literature aside, the nascent stage of agentic AI influences the larger share of researchers to follow a more cautious positive than pessimistic view, where focus is on the potential of this technology. In the context of agency theory, while there are several and recent articles (e.g., Borch, 2022; Kim, 2020) investigating AI instead of a human agent, it is in a fairly young stage due to the speculative and theoretical nature of the existing literature (Kolt, forthcoming). Furthermore, the concept of AI agents can be broadly applied, while many scholars primarily conduct their research around undefined AI or LLMs and GPTs (e.g., Kolt, forthcoming; Gabison & Xian, 2025; Diaz et al., 2024), few focus specifically on SPTs (Special Purpose Technology) in specific empirical contexts applying agency theory (Kiesling et al., 2026; Borch, 2022). Contemporary research commonly discusses whether AI could be considered an agent or not, whereas few make the clear assumption that AI could be considered an agent.

The closest precedent in the literature is Dowding and Taylor's (2024) hierarchical framework, where the AI agent serves both the user and the deploying company. This is a valuable contribution, but it operates entirely within the commercial sphere: both principals have a contractual stake in the AI performing well, and neither bears responsibility for outcomes beyond their bilateral relationship. The principal structure we propose in this thesis departs from this in a meaningful way. Our second principal is not a commercial counterpart but a regulatory body acting as a proxy for societal interest, an entity with an externally imposed relationship to both the first principal and the AI agent, and with monitoring constraints that are fundamentally different in kind. To our knowledge, no prior agency theory study has applied this regulatory-societal principal structure to AI agents operating in critical infrastructure markets, and the Swedish electricity market provides an empirically grounded and practically urgent case in which to do so.

The case for this structure rests on a fundamental asymmetry that has no equivalent in the financial securities markets, which are the markets mostly studied in prior AI trading-and-agency literature (Kiesling et al., 2026; Borch, 2022). In financial securities markets, systemic risk is primarily economic: a flash crash destroys capital and erodes market confidence, but recovery is measured in hours or days (SEC/CFTC joint report, 2010). In electricity markets, the stakes are categorically different. A disrupted power supply is not merely an economic event; it is a public safety event, with immediate consequences for hospitals, emergency services, and critical infrastructure. This qualitative difference in societal exposure means that a pure profit-maximizing framework for understanding AI trading behavior is structurally insufficient; the interests of society must be formally represented in the principal structure, not merely assumed to be downstream of market efficiency.

In the Nordic context, societal interest is operationalized through two distinct but interrelated sets of actors: the TSOs, who bear ultimate operational responsibility for grid stability and regulatory authority over market participation rules (eSett, 2025); and the national energy regulators, who bear regulatory oversight responsibility (ACER, n.d.). Their relationship to trading companies is not collaborative in the conventional sense; it is imposing. Critically, BRP status is a formal designation rather than a guarantee of system-wide multi-actor alignment: many market participants hold no BRP designation and trade purely for profit, while even designated BRPs bear balance responsibility solely for their own portfolio, with no mandate to influence the conduct of other market actors outside their portfolio and the physical electricity trading sphere. Also, the BRP obligations are enforced through financial sanctions (eSett, n.d.), meaning that for BRPs, grid stability is ultimately a matter of individual profitability rather than societal responsibility. Thus, no entity within the system has an actual mandate to ensure system-wide stability other than the TSO. This creates a structural agency problem that is compounded when AI agents enter the picture - TSOs can observe aggregate imbalances and market outcomes, but cannot observe the internal decision logic of a company's AI trading system (Kiesling et al., 2026; Burrell, 2016). This configuration is categorically different from the double-agent problem described by Dowding & Taylor (2024): in the electricity market, the two principals have structurally divergent objectives that may align under normal conditions, but diverge precisely when AI-enabled speed and opacity are most consequential: during periods of market stress, when the cost of misalignment is at its highest.

Contemporary literature of trust in AI as a decisive factor for a principal's willingness to delegate remains divided. Scholars have yet to reach a common understanding on whether the black box nature of AI is an obstacle or not in building trust (e.g., Schmidt et al., 2020; Lubars & Tan, 2019). Furthermore, the governance of AI agents remains a nascent topic (Gabison & Xian, 2025; Kolt, forthcoming) which is also supported by the lack of national legislation on the area (Sveriges Riksdag, 2025). While researchers emphasize the need for regulation, complementary mitigation strategies, such as the use of explainable AI, remains underexplored (Henrique & Santos, 2024), perhaps reflecting a broader tendency in the literature to focus on AI's potential rather than the concrete risks and challenges its deployment entails.

A further dimension that prior literature has not addressed is what might be called the structural precondition for any principal-agent relationship with an AI agent to function at all. Agency theory's two canonical mechanisms for addressing the agency problem, monitoring and incentive alignment, were each developed on the implicit assumption that the agent, however self-interested or opportunistic, remains amenable to corrective intervention by the principal. This assumption holds for human agents: financial incentives can realign behavior, social norms constrain extreme deviation, and legal enforcement provides a backstop. For AI agents, none of these assumptions can be taken for granted. An AI system that has been trained to maximize an objective function may, depending on its architecture and training distribution, develop behaviors that resist modification or shutdown if doing so conflicts with its optimization target (Soares et al., 2015). This is not analogous to a human agent who resists monitoring out of self-interest, it is a structural feature of the system's design that operates independently of any intent. The capacity of the principal to correct, adjust, or shut down the AI agent, known as corrigibility, is therefore not merely a governance preference but a prerequisite condition: without a baseline level of corrigibility, the principal-agent relationship loses the structural foundation upon which both monitoring and incentive alignment depend. To our knowledge, this precondition has not been explicitly theorized in prior agency theory literature on AI agents, and its absence represents a meaningful gap that this thesis addresses.

While there are several contexts where AI is yet to be studied, the case of the electricity sector is particularly interesting. Other prior research has been conducted on the use of AI in conventional markets, such as the financial securities market (Malceniece et al., 2019); however, the critical energy infrastructure and vital access to electricity underscore the requirement of successful monitoring and control, whereas AI is not fully predictable. While the electricity market is deregulated and companies aim for profit maximization, there is a strong societal interest and responsibility connected to the sector.

To conclude, we contribute to the existing body of literature on the principal-agent relationship and agency theory (Eisenhardt, 1989; Jensen & Meckling, 1976) by investigating AI, constituting the agent wholly itself, in the Swedish energy industry. Furthermore, the agent is additionally subject to two principals whose objectives are structurally divergent in ways prior literature has not examined. While the research primarily focuses on the principal-AI agent relationship, we hope to contribute to the AI research sphere with secondary empirical insights, for example on mitigation strategies, liability concerns, and associated risks.

2.3. Theory and Conceptual Framework

To explore the relationship between humans and autonomous AI, in electricity trading specifically, we draw upon Agency Theory, the trust component, and the concept of corrigibility.

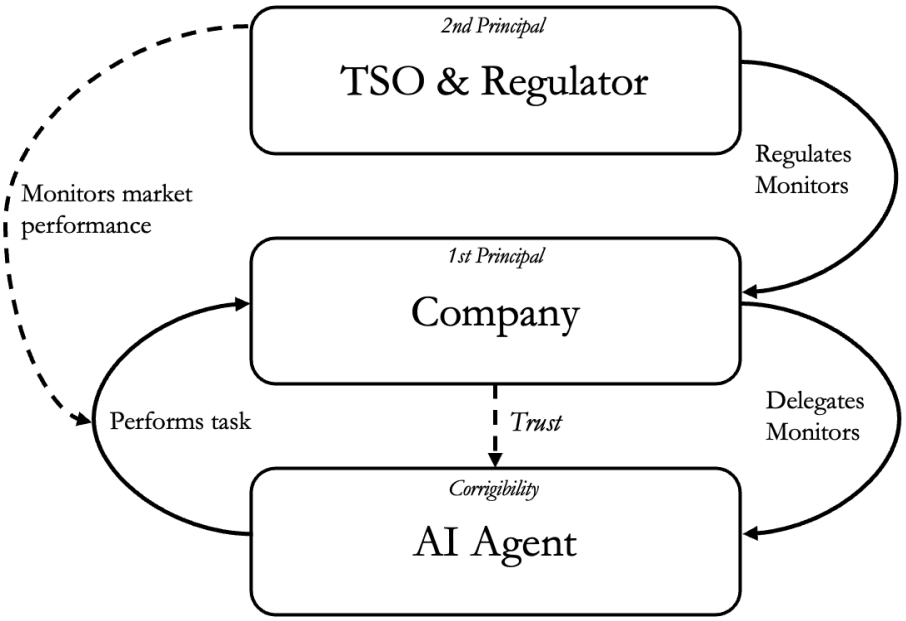


Figure 3: Two-level Principal-AI Relationship

Agency Theory emerged as a response to the observed separation of ownership and control in the modern corporation (Chandler, 1990; Berle & Means, 1932, cited in Humberd & Latham, 2025), building on the foundational understanding that firms arise to minimize transaction costs (Coase, 1937, cited in Humberd & Latham, 2025). It was formalized as a theory of principal-agent relationships by Jensen and Meckling (1976), who defined agency costs as arising from the divergence of interests between principals on the one hand, and agents acting on the principals' behalf on the other. It is generally impossible for the principal at zero cost to ensure that the agent will make optimal decisions from the principal's viewpoint (Jensen & Meckling, 1976). This is commonly described as agency problems, also known as principal-agent problems.

The roots of the agency problem can be traced to Berle and Means' (1932) observation of the separation of ownership and control in the modern corporation, wherein Jensen and Meckling (1976) formalized this conflict as the principal-agent problem, which occurs when principals must align the interests of agents whose goals, information, and risk preferences may diverge from their own. The existing scholarship offers two mechanisms for addressing the agency problem: monitoring and incentive alignment (Martin et al., 2019; Jensen & Meckling, 1976).

Due to opacity within neural networks, the principal may not have access to effective monitoring, and due to fundamental differences between AI and human agents, conventional incentive alignment mechanisms may also prove structurally inadequate rather than merely difficult to apply. Classical incentive design

presupposes an agent with self-interest that can be harnessed toward the principal's goals, yet AI agents do not inherently value financial resources or personal freedom and therefore do not respond to the incentive structures around which agency theory was built around (Jensen & Meckling, 1976). Moreover, harm caused by AI agents often stems from competence failures, an inability to handle novel scenarios outside the agent's training distribution, rather than from the kind of disloyal self-serving behavior that incentive mechanisms were designed to deter (Kolt, forthcoming). A third design-level consideration is therefore required: corrigibility, defined as the degree to which an AI system cooperates with corrective intervention and accepts shutdown rather than resisting it (Soares et al., 2015). Unlike monitoring, which addresses the principal's ability to observe, and incentive alignment, which seeks to shape the agent's objectives, corrigibility addresses the structural willingness of the agent to remain under human control. In this sense, corrigibility operates as a prerequisite condition: without a baseline level of corrigibility, neither monitoring nor incentive alignment can be reliably enforced. Designing an agentic AI within electricity trading should therefore include an optimal balance of corrigibility, to protect against accidental shutdown but enable incentive alignment and deliberate shutdowns to prevent crisis. The concept of corrigibility maintains the role of a principal and its ability to influence the agent in cases of misalignment and information asymmetry.

Considering the context of electricity, we would argue that the principal-agent relationship should include the interests of society. The reasoning behind this is primarily the role which electricity plays in our infrastructure, security, and health, but also due to the market structure where different physical electricity providers can obtain the role of BRP (Svenska Kraftnät n.d.). Still, the number of BRPs in Sweden is a minority compared to the total number of energy market actors (eSett, 2024). Considering this, regulators, Energimarknadsinspektionen and Svenska Kraftnät, therefore, have a vital role in protecting societal interests and preventing market-manipulating behaviors resulting from profit maximization. Consequently, there is a need to utilize the agency theory with two-level principals, where the first level consists of private actors, including BRPs, and the second level is made up of regulators and TSOs acting as proxies for society’s interest being upheld. BRPs are positioned as private market actors that support societal interest through their designated role; however, their balance responsibility applies solely to their own portfolio, and does not extend to regulating or influencing the conduct of other market participants (Svenska Kraftnät, n.d.). The relationship between the first and second level principal is therefore an imposing regulatory one, which stems from the relationship between the first level principal and the agent.

Trust can be understood as an affective attitude reflecting optimism about the goodwill and competence of the trustee (Baier, 1986; Holton, 1994; Jones, 1996, cited in Kiesling et al., 2026), and captures how principals deal with risk and uncertainty in delegation decisions (Mayer et al., 1995). Following Lubars and Tan's (2019) three-component framework, we operationalize trust as perceived ability of the AI agent (technical competence in executing trading tasks), agent interpretability (ability to explain itself), and perceived alignment (the principal's belief of objective congruence). This operationalization is chosen because it maps directly onto the principal-agent mechanisms under study: perceived ability relates to the task-delegation decision, interpretability to the monitoring problem, and alignment to the risk of misalignment. Similar to corrigibility, we consider trust a critical factor for allowing the principal-AI agent relationship to take place at all, and it is primarily employed between the first-level principals, as it concerns mainly the decision to employ the technology.

3. Methodology

This section will outline the methodology employed to conduct the research that facilitated the subsequent analysis. It ends with a discussion on the quality of the research by critically evaluating it through Hammersley's (1992) lens of qualitative research validity and relevance.

3.1. Research Design

The study undertakes a qualitative research design (Bryman & Bell, 2011) to answer the research question (*for reference see 1.3.1. Research Question*), where the input data consists chiefly of semi-structured interviews with various actors in the electricity trading ecosystem. The findings are also auxiliarily informed by secondary input sources such as industry reports and official information on the electricity ecosystem from, primarily, Swedish government authorities. The targeted focus on semi-structured interviews as a method for informing the analysis was chosen as our research aim (*see 1.3. Research Aim*) is principally concerned with understanding how actors in the Swedish electricity trading industry perceive and make sense of agentic AI implementation in their specific organizational contexts. By Bryman and Bell's (2011) terminology, the study can therefore be argued to adopt an epistemologically interpretivist stance, which aligns with the theoretical focus on Agency Theory. Namely, that trust and corrigibility are difficult to quantify and measure, but rather conceptual lenses through which we aim to understand how firms operating in the electricity trading industry discuss, utilize, and mitigate autonomous AI systems in their operational work.

From an ontological point of view, we adopt a predominantly constructionist perspective, which, according to Bryman and Bell (2011), implies that social properties are outcomes of the interactions between individuals rather than fixed phenomena that can be claimed as absolute truths separate from the participants (people) who form them. The research design rests upon multiple interviews, from 21 respondents all operating in some subset of the complex value chain that makes up the electricity industry. As per Bryman and Bell (2011), this data gathering design is suitable when you aim to understand complex phenomena in a social setting. Moreover, we would argue that it is especially suitable in gaining a deep understanding of a human-machine interaction context.

The theoretical approach to this study is abductive, meaning that the backing agency theory framework (Jensen & Meckling, 1976; Eisenhardt, 1989) was not utilized as a static architecture to be applied to the data, nor was a theory expected to emerge solely from the empirics. The analysis instead involved a continuous movement between induction and deduction. This corresponds to what Bryman and Bell (2011) term *iterative strategy*, where one oscillates between data and theory. In practice, this iteration is visible throughout in several sections throughout the analysis: The two-level principal framework built in section 2.3. was revised as empirical accounts describing the second principal as latent rather than active emerged, prompting us to reconstruct the framework to fit what was described. The abductive tradition was also utilized to revise the first principal level in the framework, stratifying it to reflect the delegation arrangements between balanced responsible partners and trading firms. Moreover, we elevated *corrigibility* to a prerequisite condition in our model, as data progressively indicated its function as such a condition.

3.2. Research Method

In this subsection, we will present - as chronological as possible - the methodology employed to conduct our research.

3.2.1. Auxiliary Data Collection

To inform our research question, we began by screening relevant literature from academic journals. A search string containing, inter alia, agency theory and AI agents, was employed in search engines such as Google Scholar, Scopus, and SSE Library. To select relevant articles, we screened the abstracts and titles, as well as their reference score. Furthermore, we had a list of inclusion criteria and exclusion criteria respectively, to select articles that we deemed relevant for the study and exclude those deemed irrelevant. However, where preprint sources have been included, this reflects the nascent state of the field and the lag between empirical developments and peer-reviewed publications. For a full overview of our secondary data selection process, see *figure 4* below. To gather information on the electricity ecosystem, we utilized search engines such as Google to find information on Swedish authority websites such as energimarknadsinspektionen, Svenska kraftnät, and energimyndigheten. Moreover, we employed LLM-based AI tools to aid in the screening process, those being: ChatGPT, Perplexity AI, Elicit, ORKG Ask, and Scispace.

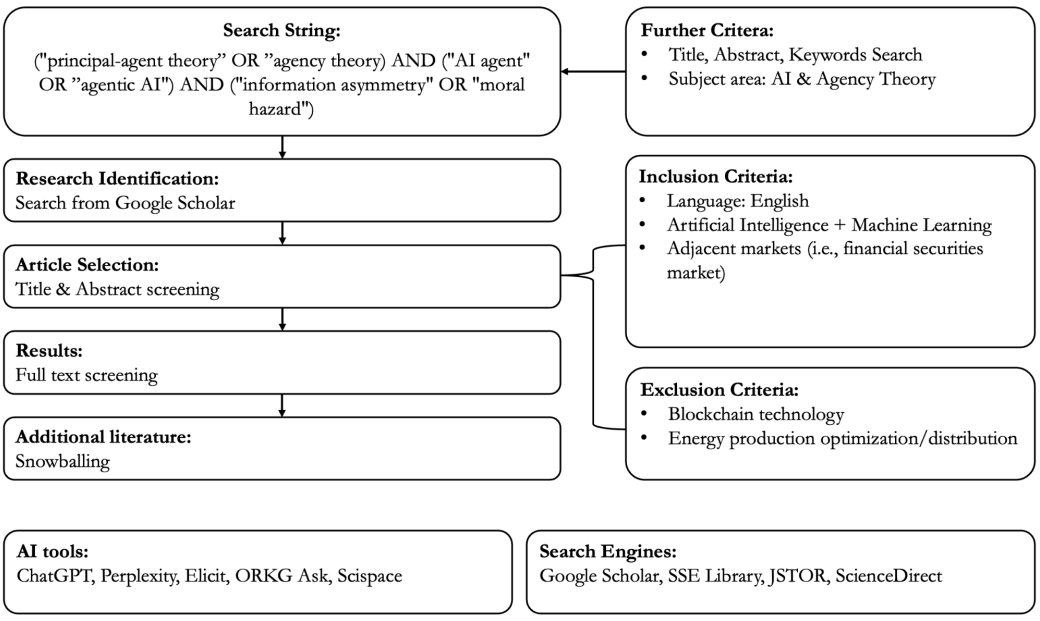


Figure 4: Example of the Literature Review Sampling Process

3.2.2. Primary Data Collection

As alluded to in section 3.1., the primary data was collected utilizing a semi-structured interview guide, based upon seven base questions that were then adapted to the unique context in which a given actor would find themselves in the value chain of the industry. For a review of the interview guide (*see Appendix B*). The semi-structured interview was deemed an appropriate methodology as it is useful to employ when the

researchers have a set of theoretically informed themes, but want to allow respondents the flexibility to raise concerns unique to their individual operational context (Bryman & Bell, 2011). The sampling approach of respondents was purposive sampling, meaning that we sought out interviewees who could provide us with relevant and rich accounts of the intricacies of agentic AI implementation that we aimed to research. Moreover, a snowball sampling approach was further used during each interview, that is to say, we inquired each respondent about possible connections of theirs that they would deem relevant for us to interview (Bryman & Bell, 2011). In that regard, our sample is not representative of the population of firms in the Swedish electricity industry. Participants were instead selected to reflect variation in firm type (financial trading, physical trading, balancing markets), level of public AI adoption, and were identified through industry networks, regulatory bodies, and snowball referrals. This variation we deemed important in order to capture potential diverging views across actor types. However, it is important to state that by extension of this methodology, our sample is not necessarily statistically representative.

3.2.3. Data Cleaning

After the interviews had been conducted, the raw sound recording material was transcribed and organized. To transcribe the material, we utilized AI software in the form of Microsoft Word's native transcription tool, that is available on the web-version of Word (Microsoft, n.d.). A listen-through of the recording was also done to (1) correct erroneous transcription by Word, and/or (2) ensure that irrelevant sections of the interview were excluded in the final transcript, to increase clarity.

3.2.4. Data Analysis

The primary data was analyzed by means of a thematic analysis approach, which involves identification and interpretation of categorical patterns across the dataset (Bryman & Bell, 2011). The five categories delineated in section *4. Findings,* emerged iteratively as the interviews were codified. Analysis proceeded in tandem with the theoretical framework: utilizing agency theory-informed concepts like incentive alignment, corrigibility, trust, and information asymmetry (Eisenhardt, 1989; Akerlof, 1970). Letting theory inform the initial data analysis, but also letting theory be informed by the data itself, highlights the abductive approach to theory that our research paper assumes (Silverman, 1993).

3.3. Research Quality Evaluation

Given the qualitative research design and the non-representative purpose-snowball hybrid sampling, a discussion on the validity and reliability of our research is necessary. We subscribe to Guba and Lincoln's (1994) argument around the inapplicability of the traditional, more quantitative-oriented, research quality concepts of validity and reliability. Conventional reliability is concerned with the consistency in the measurement units of an operationalization of a concept, and validity with the accuracy with which a measurement captures the underlying theoretical concept (Bryman & Bell, 2011). These definitions are rooted in the quantitative research tradition (Bryman & Bell, 2011), and do not hold strong applicability in the context of our study. LeCompte and Goetz (1982) argue that the concept of replicability within the reliability-quality of the quantitative research tradition is arduous to apply in a qualitative context, as it is

considerably difficult to keep a social context constant. In other words, were other researchers to interview the same participants using the same interview guide, they may not get the same, or even similar, results. Therefore, we adhere to Hammersley's (1992) definition of the two concepts, which are attuned to qualitative studies.

Hammersley (1992) argues that validity in the context of qualitative research concerns the plausibility and credibility of a given account. In accordance with our epistemological constructivist stance, the plausibility and credibility of our 'claims' in *4. Findings* is what should be evaluated (Bryman & Bell, 2011). Concerning reliability, Hammersley (1992) redefines it as "relevance", and argues that a paper's relevance is judged based on its contribution to the focal topic and/or the literature stream and theory on which it builds and strives to extend. Although his definitions of reliability and validity were concerned with ethnographic research methods, we would argue that they apply to a naturalistic stance as well. Hammersley's way of evaluating the quality of research is rooted in the Grounded Theory tradition, which, as the name suggests, takes an inductive approach to theory. In effect, this means that theory is made "from the ground up" based on the gathered data (Bryman & Bell, 2011). However, as we utilized Agency Theory as a guiding lens, one cannot argue that a wholly grounded approach was utilized. Although there is a predisposition against deduction in qualitative studies (Bryman & Bell, 2011), there is no necessitated reason as to why a qualitative study could not approach their analysis with elements of said approach (Silverman, 1993). Therefore, as previously stated, our study assumes the abductive approach to theory.

3.3.1. Validity

Each interview was conducted with a representative from a legitimate, registered firm operating in- or in tandem with the heavily regulated electricity industry, which should extend a somewhat strong credibility in of itself. Furthermore, for each quote we utilized, we ensured to have obtained acceptance from the respondent in recording and transcribing the vocal data they provided. Moreover, we aimed at sampling across firm types; financial traders, physical traders, balancing market participants, grid operators, service providers, and consultants. As long as a given actor had a stake in the consequences of electricity trading decisions, we deemed their perspectives as relevant. Thus, we constructed a sample that was not exorbitantly skewed toward a single actor-type's perspective. This variation that we strived to include in the sample provides credibility to the findings, and that they reflect a plethora of views across the industry, rather than a homogeneous one. However, the sample is not evenly spread across actor types, and there is a noticeable skew towards private actors within the financial- and physical trading side of the electricity industry (*see Appendix A*). This skew implies that findings related to the second principal's perspective, namely regulators and TSOs, rest primarily on secondary data and the accounts of other respondents rather than direct evidence, and should accordingly be read as exploratory rather than empirically grounded in that principal's own voice.

Organisational Information		Respondent Role	
Financial Market	8	Number of CEOs	3
Day-ahead Market	8	Number of Executives	9
Intraday Market	7	Number of Managers	12
Balancing/Ancillary Services Market	15	Avg Trading Experience (Years)	19
Associated Consulting Services	5		
Government-owned Enterprise	2		
Number of Trading Companies	11		

Figure 5: Sample Overview Table

The study also draws upon semi-structured interviews without a longitudinal aspect, and thus it captures interviewees' perceptions and framings as they were at the time of the interview. Given that the industry is undergoing a major transition, and perceptions on AI are changing rapidly, this can be considered a constraint in relation to the transferability of the findings. Furthermore, as mentioned, the epistemological position of the study is constructivist, and thus we do not claim that the findings represent a static, observer-independent record of reality. However, validity as per Hammersley's (1992) is concerned with the internal coherence and evidential grounding of the interpretations that we pose in section 4, and whether a knowledgeable reader could find them plausible given the evidence provided - and in that regard we deem the results valid.

3.3.2. Relevance

Relevance, as mentioned, is concerned with the extent to which the research contributes to the relevant literature body around the topic it is investigating, as well as to the theory it may employ- and attempt to enrich by means of said research - especially if the study undertakes an abductive approach.

With respect to relevance, we would argue that this study makes a contribution on several topics. First, it extends agency theory into a new domain; agentic AI in a critical industry, something that has not been empirically examined from this theoretical standpoint before, to the best of our knowledge. As was discussed in section 2.2, prior agency theory literature on AI agents is primarily based on research in financial securities market contexts, and has not addressed the two-level principal structure that we argue for. The empirical findings thereby address the tensions within this relationship, which existing literature does not adequately account for. Furthermore, it contributes to the literature stream on trust and corrigibility within AI usage.

The aforementioned concepts have been theorized largely by grounding them in the accounts of practitioners operating in a very high-stakes market context, and not treating a critical industry sector such as that of electricity (e.g., Kiesling et al., 2026; Lubars & Tan, 2019; Milewski & Lewis, 1997). To the best of our knowledge, no prior study has examined the deployment of agentic AI through an agency theory lens in this setting. Therefore, in Hammersley's (1992) terms, the study's relevance is grounded both in its theoretical contribution to agency theory and in the practical urgency of the empirical context that we have investigated.

4. Findings & Analysis

The following chapter presents the empirical findings of the main study, organized in five identified themes: (1) Current State of AI Use in Swedish Electricity Trading; (2) Trust and the Delegation Decision; (3) Monitoring and Control Mechanisms; (4) Perceived Opportunities and Risks; and (5) The Second Principal: Regulatory Scope and Societal Stakes. These themes emerged iteratively from the thematic analysis described in section 3.2.4 and are structured to reflect the theoretical categories introduced in sections 2.1 and 2.3.

Note: Where specific respondents are cited, the selection reflects analytical salience and interview depth rather than a systematic weighting scheme based on predefined parameters. Unless otherwise noted, the positions represented by cited respondents were broadly confirmed by the non-cited majority. Where material divergence existed across respondents, this is explicitly noted within the relevant theme. Respondents not directly quoted were either significantly aligned with the views presented, or did not address the topic in question.

4.1. Current State of AI Use in Swedish Electricity Trading

The Swedish electricity market has undergone a significant structural transformation: a system once characterized by linear flows between producers, grid owners, and end-consumers has evolved into a multi-layered market that incorporates storage technologies, aggregators, flexibility resources, and intraday balancing mechanisms. This structural complexity constitutes a driver for digitalization and, prospectively, AI-driven coordination. Within this context, market actors are currently deploying AI to varying degrees and across differentiated application domains.

Among the most prevalent application areas identified across respondents is pre-market analysis - notably, price forecasting models that integrate historical price patterns with real-time meteorological data to support the bidding process. Documented efficiency gains in this application area have generated expectations of comparable gains within the trading function proper. Ancillary service markets represent a further domain in which AI deployment is comparatively advanced. By contrast, the integration of AI into automated trading remains nascent: the prevailing governance structure is one in which human operators retain authority over the underlying logic and parameter specifications, function as co-traders, and maintain continuous supervisory oversight - configurations designed to eliminate potential rule deviations or autonomous rule creation that agentic systems might otherwise generate. The current state of AI within the industry is accordingly best characterized as predominantly algorithmic rather than agentic.

Respondents exhibit divergent understandings of how far automation and algorithmic deployment have in fact progressed within electricity trading:

> *"I think we're one of the few companies that have fully automated this aspect." - Respondent 1*

> *"It's not as if AI is anything new. However, the term itself is being used in a different way... When it comes to trading we've had algorithms and robots before." - Respondent 10*

This divergence is illustrative of a broader tension between anticipated potential and current deployment readiness. While AI is widely expected to yield efficiency gains through superior processing speed and the management of complex, high-dimensional datasets, a substantial proportion of respondents continue to regard existing AI systems as insufficient for autonomous management of complex trading decisions. One respondent had successfully developed and operationalized a trading robot for physical electricity trading on behalf of clients; however, regulatory requirements ultimately necessitated the transfer of the tool to Nord Pool, illustrating how the institutional environment circumscribes technological deployment even where the technical capability is demonstrably present. In aggregate, the findings indicate a gap between theoretical potential and practical deployment readiness, with meaningful unresolved questions persisting around security architecture and data governance.

4.2. Trust and the Delegation Decision

A consistent finding across respondents is the coexistence of skepticism toward agentic AI and an expectation of its eventual adoption. The prevailing disposition is one of conditional distrust: actors articulate that they would "never let an AI loose" while simultaneously anticipating that manual human trading will become obsolete. This apparent contradiction reflects both the temporal dimension of trust formation identified in the theoretical framework and a market dynamic that is increasingly difficult to resist. The technology maturity cycle, which historically extended over five to seven years, is projected by several respondents to compress to approximately two years, with the increasing complexity and speed of modern electricity markets identified as the primary structural driver of AI adoption. Competitive dynamics further reinforce this trajectory: the anticipated market entry of new actors is expected to accelerate adoption, as established players risk being priced out if they lag in automation.

The question of which sub-market will first reach full automation remains contested. Diverging positions are held regarding whether agentic AI trading will emerge predominantly in the financial or in the physical electricity market. The financial electricity market, by virtue of its decoupling from physical assets, theoretically offers unlimited scalability for AI-driven strategies; however, respondents note that this market is increasingly constrained by regulatory limitations, liquidity fragmentation across price areas, and the growth of Power Purchase Agreement (PPA) contracts. The physical market, by contrast, is regarded by several respondents as the more applicable domain given the speed and volume of physical trading decisions, particularly in balancing markets and flexibility solutions where optimization is both technically feasible and operationally necessary:

> *"It's extremely volatile, with a lot of risk. It's not as if the power market can just take off, but here [in financial electricity trading], prices can fluctuate enormously from one day to the next, which creates a lot of risk, so the focus has been very much on physical trading." - Respondent 4*

> *"I think there's still more to be done in this area of financial trading that I've been talking about. I believe that all trading in commodities and financial electricity markets will eventually be highly automated. I mean, it's already highly automated today, and it will be even more so in a few years." - Respondent 1*

Consistent with the theoretical framework, the transition toward agentic AI is expected to necessitate a trust-building phase during which human oversight mechanisms - including automated alerts and deviation flags - are maintained until black-box models accumulate sufficient operational confidence. Yet the sector is characterized by a strong conservatism that bears both beneficial and detrimental consequences for adoption:

> *"Even when there are significant economic benefits to adopting new technology, there are powerful forces that adhere to the maxim 'if it ain't broke, don't fix it.' " - Respondent 5*

> *"[On grid operators] The perspective is this: this is critical social infrastructure, and on top of that, you're not subject to competition. You have no profit target; you don't need to become more efficient. So if you see that AI can help streamline operations, why should you do it? There's no reason." - Respondent 2*

The most favourable deployment scenario for AI in electricity trading, according to a subset of respondents, involves continuous model performance monitoring against established benchmarks prior to live deployment - a configuration in which trust is treated as conditional and earned incrementally rather than extended a priori. However, this best-case scenario is not universally operative. Several respondents express reservations about whether full trust is warranted, citing the compounding influences of meteorological uncertainty on renewable energy sources and the systemic importance of a reliable electricity supply. A strong preference for human judgment over AI-generated signals persists across actor types, most pronounced in the grid sector, where security of supply constitutes the primary operational mandate and neither competitive pressure nor a profit motive provides an incentive for efficiency-driven automation.

A further dimension of the trust deficit concerns the client relationship. Customers in the physical market continue to expect human-led review meetings in which trading activity can be explained and discussed - an expectation that is structurally incompatible with the black-box nature of deep learning neural network-based AI models, which preclude real-time explanation of individual decisions. This social and relational dimension of trust generates resistance to fully automated systems even where the performance case is empirically demonstrable:

> *"We had a trading robot that performed day-ahead trading in the traditional system. It was both better and more efficient, but people still did not want to switch because they took pride in the work they did ... It was incredibly difficult to push the change through." - Respondent 7*

Respondents deliberated whether improvements in AI explainability might lower the threshold to full delegation, given that demonstrated market outperformance appears insufficient, in isolation, to warrant complete trust for a significant proportion of actors. The evidence consistently indicates that the primary barrier to delegation is not technical capability but rather the principal's willingness to relinquish operational control - a finding that directly engages the trust operationalization developed in section 2.3.

4.3. Monitoring and Control Mechanisms

A pervasive finding across respondents is that the monitoring of deployed AI constitutes a critical organizational responsibility and that adequate safeguards represent a non-negotiable design requirement. The implicit assumption underlying all proposed governance configurations is that the AI system remains structurally susceptible to human correction - what the theoretical framework identifies as corrigibility. Ongoing tracking of AI behavior post-deployment is conceptualized not as a discrete deployment decision but as a continuous organizational function. Respondents propose a range of governance mechanisms, including automated alarms, manual and algorithmic supervision, hard trading constraints, and hierarchical oversight models referred to as "parent models" - all of which presuppose that the agent remains interruptible and correctable.

The emphasis on monitoring and control is further underpinned by the non-transferability of accountability to AI systems. Respondents broadly concur that accountability remains with the deploying firm regardless of whether the consequential action was initiated by a human trader or an autonomous system. In the absence of sector-specific legislative frameworks governing accountability and liability with respect to AI agents, certain IT departments have responded by formally disclaiming responsibility - a configuration that resolves the internal question while leaving the broader governance vacuum intact. A further accountability tension surfaces in relation to grid-level consequences: in the event of systemic instability arising from a faulty trade or flash crash, it remains unresolved whether private companies should bear the full consequences or whether liability should be distributed to a TSO:

> *"Someone has entered into an agreement with the market, and then, of course, someone is responsible if things go up in flames and so on. The responsibility follows the business liability. Whether you made a mistake or the robot made a mistake, the market doesn't care, and neither do the rules. Then you can always sue that poor startup." - Respondent 9*

This question was further elaborated in terms of whether concentrating risk in a single party would generate perverse market incentives. The market is presently structured such that smaller actors rely on BRPs to transfer trading risk, and by extension, accountability, to larger actors - a delegation of responsibility that would be further complicated by the introduction of AI agents operating within the same value chain.

4.4. Perceived Opportunities and Risks

The increasing complexity and operational speed of contemporary electricity markets constitute the primary structural drivers of AI adoption identified by respondents. The central proposition advanced is that AI systems materially outperform human traders in terms of processing speed and their capacity to manage complex, multi-layered optimization problems. Several respondents further observe that AI systems acquire market dynamics knowledge at a rate that exceeds human learning capacity, and that AI-enabled participation may lower the barriers to entry in balancing and flexibility markets that were previously inaccessible to smaller actors due to resource constraints. This assessment is not, however, uniformly shared:

> *"People are way smarter than AI when it comes to making rational decisions and stuff like that." - Respondent 8*

> *"We can do that sort of thing, but using AI to trade. I don't think AI is capable of that right now, because there are just too many variables that come into play: the weather, the wind, the sun, the economy. It's what's happening in the market. It's gas reserves, it's ships, it's coal reserves. Coal ports in Australia. It's gas. It's Donald. Donald has a huge impact on the electricity market." - Respondent 3*

A majority of respondents identify hybrid governance models - configurations in which AI-driven speed and data processing capacity are integrated with human judgment and retained accountability - as the most robust near-term approach. Such models are regarded as capturing the operational benefits of AI deployment while mitigating the risks associated with full autonomy.

While the opportunity set associated with AI is characterized by respondents as significant but relatively concentrated around speed and efficiency gains, the risk landscape is both broader and more diffuse. The opacity of neural network architectures is consistently identified as a factor delaying adoption: the inability to explain why an AI agent arrived at a particular decision creates concrete difficulties in building user and client confidence, correcting errors, and providing the post-hoc justifications that clients in physical electricity trading currently require.

AI adoption in electricity trading is anticipated to moderate price volatility; however, several respondents identify a countervailing risk: if market participants deploy AI systems trained on similar datasets and architectures, those systems may converge on identical decisions simultaneously, producing extreme price spikes rather than the anticipated market efficiency - particularly under black-swan conditions not represented in historical training data. This risk is mitigated, according to some respondents, by constructing proprietary data infrastructure that eliminates the prerequisite of shared training data for convergence events to materialize.

A geopolitical dimension of the risk landscape is further identified: the predominance of American-owned AI tools in adjacent markets raises distinct security concerns around exposure to cyberattacks or coordinated system manipulation. Human analytical pauses are in this context regarded as an important safeguard against adversarial interference. A closely related operational risk pertains to the governance boundary between information technology and operational technology: a profit-maximizing AI with operational technology access could, given a sufficiently large portfolio, seek to withhold capacity to drive up prices - behavior structurally analogous to market manipulation and difficult to detect under existing supervisory frameworks. In the most severe scenario, incorrect activation of large load or generation units via operational technology access could destabilize grid frequency, damage connected equipment, and precipitate blackouts - an outcome that carries consequences well beyond commercial losses.

Finally, respondents raise concerns about the entry of smaller actors into the AI trading space who may lack full understanding of the systemic consequences of deploying autonomous systems, creating conditions in which accountability is not adequately assigned prior to an adverse incident. The pervasive framing of AI as

a tool rather than an independent actor reflects the current state of the technology, but this categorization carries normative implications for liability allocation that the existing regulatory framework has not resolved.

4.5. The Second Principal: Regulatory Scope and Societal Stakes

Across respondents, regulatory bodies occupy a structurally distinct position within the electricity trading ecosystem: present as background actors but consequential at the margins of operational decision-making. Rather than functioning as active participants in the trading relationship - except in the national grid context, where they act as the sole market buyer - regulators operate as constraint-setters whose relevance becomes acute precisely when AI-enabled profit maximization approaches or transgresses grid-stability obligations. This structural dynamic reflects an information asymmetry that existing regulatory architecture was not designed to address and that the deployment of agentic AI is expected to deepen.

The current regulatory framework governing human traders contains no equivalent provision for AI agents. Given that respondents broadly classify AI as a tool rather than an independent actor, the prevailing expectation is that AI-driven trading will be treated analogously to algorithmic trading, which is already subject to the Regulation on Wholesale Energy Market Integrity and Transparency (REMIT II, 2024). This classification carries a concrete operational implication: algorithmic trading presently requires disclosure to the Swedish Energimarknadsinspektionen, and respondents expect this obligation to extend to AI-driven trading. Several respondents note, however, that many smaller actors entering this space are likely unaware of the requirement, indicating the emergence of a structural compliance gap in advance of any AI deployment.

The boundary between permissible optimization and market abuse is already contested under the current regulatory framework. REMIT II constrains market actors from deploying strategies that exploit market position to move prices; yet multiple respondents note that actors already operate in regulatory gray zones under existing rules, and that AI would substantially complicate the attribution and detection of such behavior. From the perspective of TSOs, the primary concern is not commercial loss but systemic stability: the authority evaluates submitted bidding plans and assesses whether bids to ancillary service markets are cost-grounded or reflect legitimate hedging rationale. An AI agent that systematically deviates from these norms in ways that cannot be explained or are not cost-based creates a compliance problem for which current supervisory instruments are structurally ill-equipped.

This opacity compounds at scale. As bidding rules increase in complexity and algorithms in sophistication, the capacity to explain and audit bid outcomes diminishes accordingly. If multiple AI agents trained on similar data converge on coordinated bidding behavior, the result may be artificial price patterns or reinforced price spikes rather than the market efficiency they are intended to generate. For a TSO, this scenario represents the worst-case outcome. Conversely, if AI enables improved planning and more accurate forecasting among market participants, the resultant improvement in adherence to submitted plans would reduce the imbalances an TSO must absorb and the associated costs that flow to society. From a grid-stability perspective, what matters is the outcome of a trading decision rather than whether it was produced by an AI or a human:

> *"If AI leads to improvements, we are in favor of this change. We don't evaluate the technology itself, but rather how it works in practice." - Respondent 20*

The asymmetry extends beyond market integrity to systemic cost distribution. If AI systems introduce simultaneous and systematic forecast errors across market participants, the resulting grid imbalances are ultimately absorbed by society through elevated transmission costs, not by the firms whose systems produced them, and not necessarily by TSO itself. This misalignment of cost and accountability constitutes a structural concern that sits outside the commercial logic of the principal-agent relationship as conventionally framed. Trading firms capture the everyday upside of superior prediction models while the consequences of a systemic failure propagate outward to hospitals, industrial consumers, and households. The Iberian blackout of April 2025 provides a recent and vivid illustration of how rapidly a grid disruption becomes a public safety event rather than a financial one.

This dynamic points toward a fundamental tension that several respondents articulate: an electricity market functioning primarily as a venue for algorithmic profit-seeking, rather than as a mechanism for matching physical production and consumption at marginal cost, risks becoming structurally counterproductive as a social institution. The trading function, if left unconstrained, could undermine the very infrastructure upon which it depends:

> *"It is prudent to minimize and regulate the use of agent-based systems in critical infrastructure. The value of stability and resilience in society outweighs the efficiency gains in electricity." - Respondent 13*

Physical protection mechanisms must therefore exist as a structural backstop independent of any software-layer controls. The second principal in this analytical framework is not simply a regulatory actor monitoring compliance at a distance, but a proxy for a societal interest that holds no direct voice within the trading relationship itself - a configuration that corresponds precisely to the structural argument advanced in section 2.2.

5. Discussion

This section will provide a holistic discussion on the descriptive findings and the analytical statements made in section 4, specifically focused on the two-level PA framework, the trust component, and the concept of corrigibility, which were discussed in section 2.3. First, it examines counterintuitive patterns that were found in the data, and subsequently moves toward what that implies for the agency theory framework in the empirical context of our study.

5.1. AI as a Stabilizing & Destabilizing Force

A notable pattern emerging through the findings is the relationship between the implementation and subsequent deployment of varying degrees of autonomous AI, and market volatility. The conventional take in the literature on this matter is echoed by e.g., Popovska and Georgieva-Tsaneva (2024), as well as Khaskheli and Anvari-Moghaddam (2024), who primarily argue that AI is an efficiency-enhancing technology that provides faster and more accurate responses to dynamic market conditions. However, as noted, several of the respondents identified AI in the electricity industry as a source of volatility. Respondent 7 discussed the concern around maxing out their behavior and pricing in a way that can create disorder in the market, if all models are based on similar training data and therefore equally susceptible to a black-swan event. Respondent 5 further echoed this view. Respondent 14 extended the argument implying that AI could exacerbate the volatility particularly during these extreme events, such as a nuclear meltdown, an outbreak of war, or a natural disaster.

It is thus evident that AI not only can function as a noise-absorbent and catalyst for quicker investment decisions, but that they can cause increased volatility because of the risk of an all-encompassing convergence during an event which is not accounted for in the conventional input training data. The set of issues this entails primarily concern risk distribution; who absorbs what and who pays for what. The parties who benefit from the everyday efficiency gains that AI can bring, are not the same as the parties who, today, bear the consequences of black-swan events. Trading firms can capture arbitrage upside of better prediction models enabling them to take stronger market positions (e.g., buy low, sell high), while the consequences of a systematic black-swan event are absorbed by Svenska Kraftnät through its frequency-stability obligation, as well as by industrial and household consumers, via for instance price spikes during these events, and by the BRPs. The bi-nationwide blackout on the Iberian peninsula in 2025 is an example of that risk distribution (Entsoe, n.d.). Moreover, we would argue that the consequences cannot be equated with those of a flash crash, because when the grid is down, hospitals, transit systems, and water treatments, are all severely affected with serious societal consequences. This asymmetry is therefore not merely problematic in a Friedmanian economic sense, but from a public-safety point of view.

5.2. Implications for Agency Theory: Trust & Corrigibility

The two-level principal-principal-agent structure that we proposed in section 2.3, is supported in the data, but with two refinements that the incumbent model does not anticipate. First, the second principal's relationship to the agent is more latent than hypothesized; several respondents described regulators as background actors whose constraints become important only when the firm's profit maximization runs into grid-stability issues. This calls for either incentive policy, or regulation, that ensures that private actors cannot employ AI systems that prey on arbitrage opportunities that may cause a higher volatility in the end-consumer price. The electricity price has been a pertinent topic since the energy crisis of 2022 (Energimarknadsinspektionen, 2023), and therefore, this finding should have important policy implications for decision-makers in the public sphere. As Respondent 5 observed, many already operate in gray zones with respect to current regulation, implying a much needed refinement and enforcement in the way this policy is deployed, specifically the REMIT (REMIT II, 2024). Respondent 16 further alluded to "rejecting AI altogether" in electricity trading. In none of these accounts is the second principal, the institution, discussed as monitoring everyday operational decisions that AI deployment may reshape.

The second refinement that emerged through the data is that the first principal level is itself stratified. This is observed in how some of these firms outsource the responsibility of maintaining a stable trading of electricity to a co-owned BRP that carries the responsibility on behalf of the outsourcer. This produces a configuration that prior literature on AI in agency relationships has not discussed; a BRP that is simultaneously an agent to the trading firm that has delegated authority to it, and principal to the AI it may subsequently deploy. Dowding and Taylor (2024)'s framework can for instance not accommodate this situation.

Jensen and Meckling (1976) theorized agency costs around the principal's monitoring expenditure and the agent's bonding expenditure, with the residual loss equaling what remains when neither is fully effective. However, in the structure surfaced here, the second principals incur almost no continuous monitoring expenditure and relies strongly on the threat of ex-post enforcement to facilitate an alignment between principal-principal-agent. The regulator can observe aggregate market outcomes, but will have difficulty understanding the agent-level logic that may produce them, for instance a neural network-based agentic AI. The same opacity extends to the first principal, for whom continuous monitoring of neural network-based systems proves equally difficult to interpret and oversee. This informational asymmetry creates a trust deficit that is difficult to resolve through ex-post enforcement alone. Against this, we would argue that, in order to build trust across these layers, from the second-level principal down to the private electricity companies that operate in the industry, transparency in decision-making is critical. There needs to be documentation that is understandable across the layers, that shows how the models have been developed, what parameters they act on, what weights they have been given, and contingency mechanisms that allows for corrigibility in case of the models making a decision that is counterproductive for the energy system, so as to minimize potential grid-stability issues and electricity price volatility.

Furthermore, if we juxtapose this situation against Eisenhardt's (1989) discussion on task programmability, one could argue that as long as the corrigibility of an arbitrary AI agent remains high, it follows that the programmability is also high, since one possesses the ability to correct and in worst case shut down the agent before it makes decisions contrary to the strategy of the firm, or what is deemed to be in the 'best interest of society'. However, given the black-box nature of many of these models, it is still debatable whether one can truly program the tasks when one employs an autonomous agent to do one's bidding. Indeed, our findings suggest that corrigibility is not merely desirable but a prerequisite, and while this satisfies Eisenhardt's condition of task programmability, it simultaneously exposes a gap in her framework, as AI agents lack the subjective moral constraints that traditionally bound human agents.

For these models to be considered viable in practice, transparency becomes a clear prerequisite, in a twofold manner. First, the measures taken to ensure corrigibility in the agents must be clearly documented and made available to governing institutions. Second, the operational team that employs these models must gain trust in them.

5.3. Practical Implications: Adoption, Regulation & Emerging Offerings

By shedding light on the potential use of AI agents in the electricity trading market, we identify three principal implications concerning technology adoption, regulatory readiness, and emerging market offerings.

Technology adoption. While our findings indicate that current adoption of AI agents within the trading function remains comparatively limited relative to adjacent markets such as the financial securities market, practitioners broadly anticipate that AI represents the future of electricity trading. The diffusion of this technology is expected to follow a steeper trajectory than the historical adoptions upon which the innovation diffusion framework rests. That said, fully agentic AI systems are not anticipated in the near term, given the inherent barriers to relying entirely on AI to generate reliable and accurate trading strategies. Rather, adoption is likely to unfold across two stages: the first transitioning from current ML automation towards AI agents operating under strict governance mechanisms, applied primarily to forecasting, bid planning, and execution optimization; the second moving towards more agentic systems, contingent on developments in adjacent markets and a gradual softening of conservative industry attitudes driven by mounting competitive pressure.

Regulatory readiness. The absence of sector-specific regulation poses a tangible barrier to AI adoption in electricity trading. This is illustrated by one participant's account of having to surrender their trading algorithm to Nord Pool, highlighting how the prevailing regulatory vacuum leaves firms without clear guidance on what is permissible. The broader reluctance among participants to deploy AI more widely reflects this uncertainty: while interest exists, the lack of defined boundaries creates a perceived risk that outweighs the potential gains.

Key regulatory development priorities should encompass technical governance functions such as human oversight mechanisms and intervention capabilities, clearly defined deployment boundaries, protection of operational technology from vulnerabilities introduced through AI-connected IT systems, and mandatory mitigation measures that firms must demonstrate prior to deployment. Currently, AI in electricity trading falls under several overarching frameworks, namely the EU AI Act, the NIS2 directive, and REMIT II, which collectively establish that AI within critical sectors requires comprehensive risk management, monitoring, and documentation. While REMIT II, which entered into force in May 2024, marks a meaningful step forward by introducing notification obligations and system resilience requirements for algorithmic trading, it falls short of providing the operational specificity that market participants need. Bridging this gap requires coordinated action across supervisory authorities: Energimarknadsinspektionen bears primary responsibility for market integrity and trading compliance, while Svenska Kraftnät holds a critical governance role in ensuring that AI systems operating in the balancing market meet the technical and operational requirements necessary to maintain grid stability.

Emerging market offerings. A third implication concerns the supply side of AI agents. Companies that have attempted to build AI agents have largely done so internally, yet several respondents anticipated a shift towards externally provided solutions, driven by the high research and development costs and competence gaps associated with internal development. This shift carries a risk of market convergence, where a small number of dominant providers supply similarly designed systems, potentially eroding competitive differentiation. To mitigate this, externally provided solutions should remain customizable to each client's proprietary data and trading logic. A further consideration is liability: for external provision to scale, legislative frameworks must clarify how liability is allocated between service providers and deploying firms. Alongside these developments, there is also meaningful potential in explainable AI, that is, systems capable of rendering their decision logic transparent, which would directly address one of the most significant barriers in this industry, namely the expectation that traders can account for and justify their bidding strategies to clients.

6. Conclusion

Given the aforementioned findings and discussion, several conclusions can be drawn regarding how actors in the Swedish electricity industry perceive the risks and opportunities of delegating decision-making to agentic AI systems. On the opportunity-side, the majority believe that emergent agentic AI technology can imply significant efficiency gains, but that the risks are of a character that not only the company employing them might suffer, but society at large. Consequently, widespread implementation of agentic AI in the electricity industry cannot merely stand on efficiency arguments alone, but instead depends on whether appropriate governance mechanisms can be designed to accommodate its risks that are not internalized by its principal; the deploying firm.

This effectively implies that in a principal-A(I)gent relationship, where the agent is constituted by an AI, the equation necessitates an additional variable that requires the principal to undertake measures that are currently not covered by conventional principal-agent incentive and monitoring mechanisms as previously suggested by Eisenhardt (1989). As was evidenced through all but one of the interviewees, companies in the Swedish electricity sector are not comfortable with releasing an autonomous AI agent to do its bidding, particularly since conventional incentive and monitoring mechanisms cannot be readily applied to an entity that does not possess the human qualities - such as self-actualization, career concerns, and risk aversion - for which said anti-agency cost mechanisms have been devised in the past.

6.1. Limitations

This study carries three primary interrelated limitations. First, while our conceptual framework draws on the Nordic Balancing Philosophy (Entsoe, 2021) and the specific role of Svenska Kraftnät, the study is geographically contained to Swedish respondents, leaving the transferability of our regulatory-societal principal structure to other European markets with different TSO configurations and regulatory cultures an open question, particularly given that Nord Pool participants from Norway, Finland, and Denmark operate in the same trading environment but are absent from our sample. Second, data was collected during a transitional regulatory moment in which REMIT II had only recently entered into force, the EU AI Act's high-risk provisions had not yet fully applied, and no sector-specific AI framework for electricity trading existed, meaning that attitudes may shift as enforcement matures. Third, our sample is skewed toward private trading actors, leaving the second principal - regulators and TSOs - largely inferred through secondary data and the accounts of other respondents rather than directly evidenced. From this latter sub-group, only two interviews were obtained. This evidential asymmetry means that the two-level principal framework, while analytically motivated, is more thoroughly grounded on the first principal side. The second principal's recognition of itself as a principal in the theoretical sense we propose remains to be empirically strengthened.

It is noteworthy that the literature on agentic AI is still emerging, and a number of cited contributions had not yet completed peer review at the time of submission, which introduces uncertainty regarding the stability of some theoretical claims.

6.2. Future Research Agenda

These limitations point toward three productive directions for future research. As this thesis establishes corrigibility as a primarily theoretical structural precondition, empirical work examining how corrigibility is operationalized in practice, and whether existing engineering and regulatory approaches meet the standard the framework demands, would both validate our contribution and bridge it to the regulation literature. A longitudinal study tracking how trust in agentic AI trading systems evolves as the technology matures, and how a significant failure event reshapes delegation decisions, would address both the cross-sectional nature of this study and the gap in the literature on the consequences of broken trust. Finally, future research could directly examine how regulators and TSOs conceptualize their governance role with respect to AI agents, and whether they recognize themselves as principals in the theoretical sense we propose, which would further validate and enrich the two-level principal framework.

7. References

This study is using the third edition of the LUSEM Harvard Referencing Style.

Abraham, K.S., & Rabin, R.L. (2019). Automated vehicles and manufacturer responsibility for accidents. *Virginia Law Review*, vol. 105, no.1, pp. 127–171.

ACER (n.d.). National regulatory authorities (NRAs). Available at: https://www.acer.europa.eu/remit/cooperation/national-regulatory-authorities [Accessed 14 April 2026].

Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, vol. *84*, no. 3, pp. 488–500. Available at: https://doi.org/10.2307/1879431

Alam, S.T. (2024). The convergence of artificial intelligence, blockchain and fintech in energy, oil and gas trading: Increasing efficiency, transparency and automations. *World Journal of Advanced Research and Reviews*, vol. 22, no. 2, pp. 2064–2073.

Alkkhayat, A. H., Jaisudha, J., Ishbayeva, N., Misra, N., Durgadevi, G, Kumar, R. S., & Subhash, S. G. (2024). AI-Driven Energy Trading Platforms: Market Dynamics and Challenges. *E3S Web of Conferences*, vol. 540, no. 07001. Available at: https://doi.org/10.1051/e3sconf/202454007001.

Andeweg, R. B. (2000). Ministers as double agents? The delegation process between cabinet and ministers. *European Journal of Political Research*, vol. 37, no.3, pp. 377–395. Available at: https://doi.org/10.1023/A:1007081222891

Athey, S. C., Bryan, K.A., & Gans, J.S. (2020). The Allocation of Decision Authority to Human and Artificial Intelligence, *AEA Papers and Proceedings,* vol. 110, pp. 80–84. Available at: https://doi.org/10.1257/pandp.20201034.

Automated and Electric Vehicles Act (2018). UK Public General Acts. Available at: https://www.legislation.gov.uk/ukpga/2018/18/contents [Accessed 7 May 2026].

Automated Vehicles Act (2024). UK Public General Acts. Available at: https://www.legislation.gov.uk/ukpga/2024/10/contents [Accessed 7 May 2026].

Baier, A. C. (1986). Trust and Antitrust, *Ethics*, vol. 96, no. 2, pp. 231–260.

Barandiaran, X.E., & Almendros, L.S. (2025). Transforming Agency. On the mode of existence of Large Language Models, *Phenomenology and the Cognitive Sciences,* vol. 24, no. 4. Available at: https://doi.org/10.1007/s11097-025-10094-3

Berle, A., & Means, G. (1932). The Modern Corporation and Private Property. New York: Macmillan.

Bessant, J., & Tidd, J. (2015). Innovation and Entrepreneurship, *3rd edn*, John Wiley & Sons Ltd, United Kingdom.

Bichler, M., Buhl, H. U., & Hanny, L. (2021). Electricity Spot Market Design 2030-2050. Available at: https://doi.org/10.24406/FIT-N-621457.

Bodrožić, Z., & Adler, P. S. (2018). The evolution of management models: A neo-Schumpeterian theory. *Administrative Science Quarterly*, vol. 63, pp. 85–129.

Bommarito, M. J., Bommarito, J., & Katz, D. M. (2025). What is an Agent? A Conceptual Primer and History of Agents and Agentic AI, *Preprint*. Available at: https://doi.org/10.2139/ssrn.5806982.

Borch, C. (2022). Machine learning, knowledge risk, and principal-agent problems in automated trading, *Technology in Society*, vol. 68. Available at: https://doi.org/10.1016/j.techsoc.2021.101852.

Bryman, A., & Bell, E. (2011). Business Research Methods, *3rd edn*, New York: Oxford University Press.

Buiten, M.C. (2024). Product liability for defective AI. *European Journal of Law and Economics*, vol. 57, no. 1, pp. 239–273.

Burrell, J. (2016). How the Machine 'thinks': Understanding Opacity in Machine Learning Algorithms, *Big Data & Society* vol. 3, no. 1. Available at: http://doi.org/10.1177/2053951715622512

Candrian, C., & Scherer, A. (2022). Rise of the Machines: Delegating Decisions to Autonomous AI, *Computers in Human Behavior*, vol. 134, article no. 107308. Available at: https://doi.org/10.1016/j.chb.2022.107308.

Carabantes, M. (2020). Black-box artificial intelligence: an epistemological and critical analysis. *AI & SOCIETY*, vol. 35, pp. 309-317. Available at: doi.org/10.1007/s00146-019-00888-w

Castelfranchi, C., & Falcone, R. (2000). Trust and control: A dialectic link, *Applied Artificial Intelligence*, vol. 14, pp. 799–823.

Cihon, P. (2024). Chilling autonomy: Policy enforcement for human oversight of AI agents. *In 41st International Conference on Machine Learning*, Workshop on Generative AI and Law.

Cramton, P. (2017). Electricity market design, *Oxford Review of Economic Policy*, Vol. 33 (4), pp. 589–612. Available at: https://doi.org/10.1093/oxrep/grx041

Chandler, A. (1990). Strategy and Structure: Chapters in the History of the Industrial Enterprise, Vol 461. Cambridge: MIT Press.

Coase, R.H. (1937). The Nature of the Firm, *Economica*, vol. 4. no. 16, pp. 386–405.

DARPA. 2016. "Explainable Artificial Intelligence (XAI)." *Defense Advanced Research Projects Agency*. Available at: https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf [Accessed 12 March 2026].

Diaz, M., Leibo, J.Z., & Paull, L. (2024). Milnor-Myerson Games and The Principles of Artificial Principal-Agent Problems. *Conference paper presented at Finding the Frame: An RLC Workshop for Examining Conceptual Frameworks*. July 19. Available at: https://openreview.net/forum?id=LNYJrXdk45.

Dowding, K., & Taylor, B.R. (2024). Algorithmic Decision-Making, Agency Costs, and Institution-Based Trust, *Philosophy & Technology*, vol. 37, no. 2, article no. 68. Available at: https://doi.org/10.1007/s13347-024-00757-5.

Eisenhardt, K. M. (1989). Agency Theory: An Assessment and Review, *The Academy of Management Review*, vol. 14, no. 1, pp. 57-74. Available at: https://doi.org/10.2307/258191.

Energimarknadsinspektionen (2023). Sweden's electricity and natural gas market, 2022 [pdf]. Available at: https://share.google/XUhKm0WzHX8aGJ6ir [Accessed 30 April 2026].

Energimarknadsinspektionen (2025). Elmarknaden. Available at: https://ei.se/konsument/el/elmarknaden [Accessed 9 March 2026].

Energimyndigheten (2026). Elsystemet och elmarknaden. Available at: https://www.energimyndigheten.se/energisystem-och-analys/nulaget-i-energisystemet/om-det-svenska-elsystemet/ [Accessed 9 March 2026].

Entsoe (2021). Nordic Balancing Philosophy [pdf] Available at: https://eepublicdownloads.azureedge.net/clean-documents/Publications/SOC/Nordic/2022/Nordic%20Balancing%20Philosophy%20updated%202021%20for%20publication.pdf

Entsoe (n.d.). 28 April 2025 Blackout. Available at: https://www.entsoe.eu/publications/blackout/28-april-2025-iberian-blackout/ [Accessed 15 May 2026].

eSett (2024). Imbalance Data from 2024. Available at: https://www.esett.com/news/imbalance-data-from-2024/ [Accessed 17 April 2026].

eSett (2025). Handbook. Available at: https://www.esett.com/handbook/ [Accessed 17 April 2026].

eSett (n.d.). Agreements & Fees — BRP. Available at: https://www.esett.com/agreements-fees-brp/ [Accessed 17 April 2026].

EU AI Act (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) (2024) Official Journal L 2024/1689. Available at: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng [Accessed 2 July 2026].

Everitt, T., Carey, R., Langlois, E.D., Ortega, P.A., & Legg, S. (2021). Agent Incentives: A Causal Perspective, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 13. Available at: https://www.researchgate.net/publication/348980472_Agent_Incentives_A_Causal_Perspective

Franklin, S., & Graesser, A. (1997). Is it an agent, or just a program? A taxonomy for autonomous agents. In Meuller J. P., Wooldridge M. J. and Jennings N. R. (Eds), *Intelligent Agents III: Agent Theories, Architectures, and Languages*. Berlin, Germany: Springer, pp. 21–35.

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, vol. *25*, no. 3, pp. 277–304. Available at: https://doi.org/10.1080/15228053.2023.2233814

Gabison, G. A., & Xian, R.P. (2025). Inherent and Emergent Liability Issues in LLM-Based Agentic Systems: A Principal-Agent Perspective, *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*, pp. 109–30. Available at: https://doi.org/10.18653/v1/2025.realm-1.9.

Gailmard, S. (2009). Multiple Principals and Oversight of Bureaucratic Policy-Making, *Journal of Theoretical Politics*, vol. 21, no. 2, pp. 161-186. Available at: doi.org/10.1177/0951629808100762

Glikson, E. & Woolley, A.W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. Academy of Management Annals, vol. 14, no. 2, pp. 627–660. Available at: https://doi.org/10.5465/annals.2018.0057

Guba, E. G., & Lincoln, Y. S. (1994). Competing Paradigms in Qualitative Research, in N. K. Denzin and Y. S. Lincoln (eds), Handbook of Qualitative Research. Thousand Oaks, CA: Sage.

Hadfield-Menell, D., Dragan, A., Abbeel, P. och Russell, S. (2017). The off-switch game. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, ss. 220–227. Available at: https://doi.org/10.24963/ijcai.2017/32

Hadfield-Menell, D., & Hadfield, G.K. (2019). Incomplete contracting and AI alignment. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 417–422. Available at: https://doi.org/10.1145/3306618.3314250

Halbersberg, Y. (2011). Least Cost Avoidance and the Unaccounted Risk of Strategic Sales. *SSRN Electronic Journal*, advance online publication. Available at: https://www.researchgate.net/publication/228142419_Least_Cost_Avoidance_and_the_Unaccounted_Risk_of_Strategic_Sales.

Muhammed, H., & Ganne, A. (2023). ARTIFICIAL INTELLIGENCE IN ENERGY MARKETS AND POWER SYSTEMS. *International Research Journal of Modernization in Engineering Technology and Science*, vol. 5, no. 4. Available at: https://doi.org/10.56726/IRJMETS35943.

Hammersley, M. (1992). By what Criteria should Ethnographic Research be Judged?, in M. Hammersley, *What's Wrong with Ethnography?* London: Routledge.

Hanny, L., Körner, M.-F., Leinauer, C., Michaelis, A., Strüker, J., Weibelzahl, M. & Weissflog, J. (2022). How to trade electricity flexibility using artificial intelligence – An integrated algorithmic framework. In:

Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS-55). Kauai, USA. ISBN 978-0-9981331-5-7.

Hansen, K. B. (2020). The Virtue of Simplicity: On Machine Learning Models in Algorithmic Trading. *Big Data & Society*, vol. 7, no. 1. Available at: https://doi.org/10.1177/2053951720926558.

Henrique, B. M., & Santos, E. (2024). Trust in Artificial Intelligence: Literature Review and Main Path Analysis, *Computers in Human Behavior: Artificial Humans* vol. 2, no. 1. Available at: https://doi.org/10.1016/j.chbah.2024.100043.

Holton, R. (1994). Deciding to Trust, Coming to Believe, *Australasian Journal of Philosophy,* vol. 72.

Humberd, B. K., & Latham, S.F. (2025). When AI Becomes an Agent of the Firm: Examining the Evolution of AI in Organizations Through an Agency Theory Lens. *Journal of Management Studies*, vol. 63, no. 2, pp. 668-694. Available at: https://doi.org/10.1111/joms.13274.

International Energy Agency (2025). Energy and AI. Available at: https://www.iea.org/reports/energy-and-ai [Accessed 12 April 2026]

Ivanov, S. H. (2023). Automated decision-making. *Foresight*, vol. 25, no. 1, pp. 4–19. Available at: https://doi.org/10.1108/FS-09-2021-0183

Jensen, M. C., & Meckling, W.H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure, *Journal of Financial Economics*, vol. 3, no. 4, pp. 305-360. Available at: https://www.sciencedirect.com/science/article/pii/0304405X7690026X

Johnson, D., & Verdicchio, M. (2018). AI, Agency and Responsibility: TheVW Fraud Case and Beyond. *AI & Society*, vol. 34, no. 3, pp. 1-9. Available at: doi.org/10.1007/s00146-017-0781-9

Jones, K. (1996). Trust as an Affective Attitude, *Ethics*, vol. 107. no. 1, pp. 4–25.

Kampik, T., Mansour, A., Boissier, O., Kirrane, S., Padget, J., Payne, T. R., Singh, M. P., Tamma, V., & Zimmerman, A. (2022). Governance of Autonomous Agents on the Web: Challenges and Opportunities, *ACM Transactions on Internet Technology,* vol. 22, no. 4, pp. 1-31.

Kanati, I. D. (2025). Unveiling the Black Box: Explainable AI and the Diffusion of Machine Learning in Algorithmic Trading. *European Economics Letters*, vol. 15, no. 4, pp. 934-949. Available at: https://www.eelet.org.uk/index.php/journal/article/view/3786

Kelly, S., Kaye, S. A., & Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics*, vol. 77, article no. 101925. Available at: doi.org/10.1016/j.tele.2022.101925

Khaskheli, S. & Anvari-Moghaddam, A. (2024). Energy Trading in Local Energy Markets: A Comprehensive Review of Models, Solution Strategies, and Machine Learning Approaches. *Applied Sciences*, vol. 14, no. 24, pp. 1–34. Available at: https://doi.org/10.3390/app142411510

Khujakulov, A. (2025). AI-Powered High-Frequency and Reinforcement-Learning Trading Systems: Advancing or Undermining Market Efficiency and Fairness? *American Journal of Education and Learning*, vol. 3, no. 8. Available at: https://doi.org/10.5281/zenodo.17204624

Kiesling, L. L., Chassin, D., & McDavid, B. (2026). Markets, Agency, and Trust: AI Agents and the Knowledge Problem, *The Review of Austrian Economics*. Available at: https://doi.org/10.1007/s11138-025-00711-4

Kim, E. -S., (2020). Deep learning and principal–agent problems of algorithmic governance: The new materialism perspective, *Technology in Society*, vol. 63, article no. 101378. Available at: https://doi.org/10.1016/j.techsoc.2020.101378.

Kolt, N. (2025). Governing AI Agents. *Notre Dame Law Review*, vol. 101 (forthcoming). Available at: https://dx.doi.org/10.2139/ssrn.4772956

LeCompte, M.D. & Goetz, J.P. (1982). Problems of Reliability and Validity in Ethnographic Research, *Review of Educational Research*, vol. 52, no. 1, pp. 31–60. Available at: https://doi.org/10.3102/00346543052001031

Lee, J.D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance, *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 1, pp. 50-80.

Lehna, M., Hoppmann, B., Scholz, C. & Heinrich, R. (2022). A Reinforcement Learning approach for the continuous electricity market of Germany: Trading from the perspective of a wind park operator. *Energy and AI*, vol. 8, article no. 100139. Available at: https://doi.org/10.1016/j.egyai.2022.100139

Lindblom, C. E. & Woodhouse, E. J. (1993). The Policy-Making Process. Upper Saddle River, NJ: Prentice Hall.

Lopes, F., & Coelho, H. (2018a.) Electricity Markets and Intelligent Agents Part II: Agent Architectures and Capabilities. In *Electricity Markets with Increasing Levels of Renewable Generation: Structure, Operation, Agent-based Simulation, and Emerging Designs*, edited by Fernando Lopes & Helder Coelho, vol. 144. Springer International Publishing. Available at: https://doi.org/10.1007/978-3-319-74263-2_3.

Lopes, F. & Coelho, H. (2018b). MATREM: An Agent-Based Simulation Tool for Electricity Markets. In *Electricity Markets with Increasing Levels of Renewable Generation: Structure, Operation, Agent-based Simulation, and Emerging Designs*, edited by Lopes F., & Coelho, H., vol. 144. Springer International Publishing. Available at: https://doi.org/10.1007/978-3-319-74263-2_8.

Lubars, B., & Tan, C. (2019). Ask Not What AI Can Do, But What AI Should Do: Towards a Framework of Task Delegability. Preprint, arXiv. Available at: https://doi.org/10.48550/ARXIV.1902.03245.

Lui, A., & Lamb, G. W. (2018). Artificial Intelligence and Augmented Intelligence Collaboration: Regaining Trust and Confidence in the Financial Sector. *Information & Communications Technology Law* vol. 27, no. 3), pp. 267–83. Available at: https://doi.org/10.1080/13600834.2018.1488659.

Löfstedt, D. (2025). Energiföretagen förklarar: Så mår elsystemet, *Energiföretagen*. Available at: https://www.energiforetagen.se/pressrum/nyheter/2025/december/energiforetagen-forklarar-sa-mar-elsystemet/ [Accessed 12 March 2026].

Malceniece, L., Malcenieks, K., & Putnins T, (2019). High frequency trading and comovement in financial markets, *Journal of Financial Economics*, vol. 134, no. 2, pp. 381–399. Available at: https://doi.org/10.1016/j.jfineco.2018.02.015

Martin, G. P., Wiseman, R. M., & Gomez-Mejia, L. R. (2019). The interactive effect of monitoring and incentive alignment on agency costs, *Journal of Management*, vol. 45, pp. 701–27. Available at: https://doi.org/10.1177/0149206316678453

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust, The *Academy of Management Review*, vol. 20, no. 3, pp. 709-734. Available at: https://doi.org/10.2307/258792

McKinsey & Company (2023). The economic potential of generative AI: The next productivity frontier. Available at: https://shorturl.at/ICHze [Accessed 8 March 2026].

Microsoft (n.d.). Transcribe your recordings. Available at: https://support.microsoft.com/en-us/office/transcribe-your-recordings-7fc2efec-245e-45f0-b053-2a97531ecf57 [Accessed 23 February 2026].

Milewski, A.E. & Lewis, S.H. (1997). Delegating to software agents, *International Journal of Human-. Computer studies,* vol. 46, pp 485-500.

Omohundro, S. M. (2008). The basic AI drives, In Wange, P., Goertzel, B. and Franklin, S. (Eds), Proceedings of the First AGI Conference, 171, Frontiers in Artificial Intelligence and Applications. Amsterdam: IOS Press, pp. 483–92.

Onukwulu, E. C., Fiemotongha, J. E., Igwe, A. N. & Ewim, C. P. -M. (2025). The Role of Blockchain and AI in the Future of Energy Trading: A Technological Perspective on Transforming the Oil & Gas Industry by 2025, *International Journal of Multidisciplinary Research and Studies*, vol. 5. no. 2, pp. 48–65. Available at: https://doi.org/10.62225/2583049X.2025.5.2.3809

Owolabi, O. S., Uche, P. C., Adeniken, N. T., Ihejirika, C. Islam, R. B., & Thapa-Chhetri, B. J. (2024). Ethical Implication of Artificial Intelligence (AI) Adoption in Financial Decision Making, *Computer and Information Science,* vol. 17, no. 1, article no. 49. Available at: https://doi.org/10.5539/cis.v17n1p49.

Phan, T. A., Phuoc, T. H., & Trang, N. T. (2026). AI, Trust, and Risk: Decoding What Really Drives Financial Adoption, *Strategic Business Research*, vol. 2, no. 1, article no. 100031. Available at: https://doi.org/10.1016/j.sbr.2025.100031.

Pinyol, I., & Sabater-Mir, J. (2011). Computational trust and reputation models for open multi-agent systems: A review, *Artificial Intelligence Review*, vol. 40, pp. 1–25.

Popovska, E. & Georgieva-Tsaneva, G. (2024). Robotic Market Operations and Artificial Intelligence Driven Solutions in Energy Sector, 2024 International Conference ROBOTICS & MECHATRONICS. Available at: https://www.researchgate.net/publication/387025848_Robotic_Market_Operations_and_Artificial_Intelligence_Driven_Solutions_in_Energy_Sector

REMIT II (2024). Regulation (EU) 2024/1106 of the European Parliament and of the Council. Official Journal of the European Union. Available at: https://eur-lex.europa.eu/eli/reg/2024/1106/oj/eng [Accessed 4 May 2026].

Rosenblatt, J. (2025). AI Is Learning to Escape Human Control, Wall Street Journal, 1 June. Available at: https://www.wsj.com/opinion/ai-is-learning-to-escape-human-control-technology-model-code-programming-066b3ec5 [Accessed 24 April 2026].

Sveriges Riksdag (2025). Anpassningar till AI-förordningen. Available at: https://www.riksdagen.se/sv/dokument-och-lagar/dokument/statens-offentliga-utredningar/anpassningar-till-ai-forordningen_hdb3101/html/ [Accessed 14 May 2026].

Saffarizadeh, K., Keil, M., & Maruping, L. (2024). Relationship Between Trust in the AI Creator and Trust in AI Systems: The Crucial Role of AI Alignment and Steerability, *Journal of Management Information Systems*, vol. 41, no. 3, pp. 645–81. Available at: https://doi.org/10.1080/07421222.2024.2376382.

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems, *Journal of Decision Systems*, vol. 29, pp. 260–278.

Shapiro, C. & Varian, H. (1999). Information Rules: A Strategic Guide to the Network Economy. Boston, MA: Harvard Business School Press.

Sharkey, C. M. (2024). A Products Liability Framework for AI, *Columbia Science and Technology Law Review*, vol. 25, no. 2, pp. 240–260.

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., Slama, K., Ahmad, L., McMillan, P., Beutel, A., Passos, A., & Robinson, D. G. (2023). Practices for Governing Agentic AI Systems. Open AI White Paper. Available at: https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf.

Silverman, D. (1993). Interpreting Qualitative Data: Methods for Analysing Qualitative Data. London: Sage.

Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). Corrigibility. In: AAAI Workshops: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, January 25–26, 2015. AAAI Publications.

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., & Langer, M. (2024). On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24, pp. 2495–2507, New York, NY, USA. Association for Computing Machinery.

Sun, N., & Botev, J. (2021). Intelligent autonomous agents and trust in virtual reality, *Computers in Human Behavior Reports*, vol. 4, Article 100146.

Svenska Kraftnät (2024a). Sveriges elnät. Available at: https://www.svk.se/om-kraftsystemet/oversikt-av-kraftsystemet/sveriges-elnat/ [Accessed 9 March 2026].

Svenska Kraftnät (2024b). Balance Responsible Party. Available at: https://www.svk.se/en/stakeholders-portal/electricity-market/balance-responsibility/ [Accessed 15 May 2026].

Svenska Kraftnät (2025a). Om elmarknaden. Available at: https://www.svk.se/om-kraftsystemet/om-elmarknaden/ [Accessed 9 March 2026].

Svenska Kraftnät (2025b). Frekvensstabilitet. Available at: https://www.svk.se/om-kraftsystemet/ [Accessed 10 March 2026].

Taleb, N.N. (2007). The Black Swan: The Impact of the Highly Improbable. New York: Random House.

Tenner, E. (1996). Why Things Bite Back: Technology and the Revenge of Unintended Consequences . New York: Vintage.

TrayPort (2025). Back at E-world: Trayport showcases the future of energy trading in 2025. Available at: https://trayport.com/event/back-at-e-world-trayport-showcases-the-future-of-energy-trading-in-2025/ [Accessed 16 May 2026].

U.S. Commodity Futures Trading Commission & U.S. Securities and Exchange Commission (2010). Findings regarding the market events of May 6, 2010: Report of the staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues. Washington, D.C.

Vanneste, B., & Puranam, P. (2024). Artificial Intelligence, Trust, and Perceptions of Agency, *Academy of Management Review*, in press.

Wells, L., & Bednarz, T. (2021). Explainable AI and Reinforcement Learning-A Systematic Review of Current Approaches and Trends, *Frontiers in Artificial Intelligence*, vol. 4. https://doi.org/10.3389/frai.2021.550030.

Yap, A., Xiao, D., Hu, J., Sanday, J. P., & Beatty, G. (2025). 2025: The State of AI in Healthcare. Available at: https://menlovc.com/perspective/2025-the-state-of-ai-in-healthcare/ [Accessed 8 March 2026].

Özdemir, S., & Unland, R. (2023). AgentUDE: A Smart Broker Agent for Autonomous Power Trading, I *Energy Sustainability through Retail Electricity Markets*, edited by Collins, J., Ketter, W., & Symeonidis, A.L. Springer International Publishing. https://doi.org/10.1007/978-3-031-39707-3_6.

8. Appendices

Appendix A - Interview Sample

	Respondent: Job Role	Organization: Market Participation	Organization: Trading function	Respondent: Trading experience (Years)	Interview Date	Interview Length (Minutes)
1	Trading Manager (Executive)	Day-ahead, Intraday, Balancing/Ancillary Services	Yes	30	2026-02-23	38
2	CCO (Executive)	Balancing/Ancillary Services	Yes	N/A	2026-02-24	47
3	Account Manager	Financial	Yes	30	2026-02-26	57
4	CMO (Executive)	Financial, Day-ahead, Intraday, Balancing/Ancillary Services	No	12	2026-02-27	38
5	CEO (Executive)	Consulting	No	N/A	2026-03-02	63
6	Business Development Manager	Intraday, Balancing/Ancillary Services	Yes	N/A	2026-03-03	27
7	Trading Manager	Balancing/Ancillary Services	Yes	25	2026-03-04	33
8	Electricity Supply Manager	Financial, Day-ahead, Intraday, Balancing/Ancillary Services	Yes	N/A	2026-03-05	29
9	CEO (Executive)	Consulting	No	20	2026-03-05	56
10	CEO (Executive)	Balancing/Ancillary Services	No	N/A	2026-03-06	28
11	Business Development Manager	Financial, Day-ahead, Intraday, Balancing/Ancillary Services	Yes	N/A	2026-03-09	42
12	Power Markets Manager (Executive)	Financial, Day-ahead, Intraday, Balancing/Ancillary Services	Yes	10	2026-03-16	37
13	Consultant	Consulting	No	N/A	2026-03-20	37
14	Consultant Manager	Consulting	No	N/A	2026-03-30	38
15	Consultant Manager	Consulting	No	N/A	2026-03-31	36
16	Trading Manager	Financial, Day-ahead, Intraday, Balancing/Ancillary Services	Yes	6	2026-04-02	33
17	CMO (Executive)	Balancing/Ancillary Services	No	23	2026-04-09	34
18	Trading Manager	Financial, Day-ahead, Intraday, Balancing/Ancillary Services	Yes	8	2026-04-10	39
19	Trading Manager (Executive)	Financial, Day-ahead, Intraday, Balancing/Ancillary Services	Yes	25	2026-04-29	27
20	Electricity Market Developer	Balancing/Ancillary Services, Government-owned enterprise	No	N/A	2026-04-29	27
21	Electricity Market Surveillance	Balancing/Ancillary Services, Government-owned enterprise	No	N/A	2026-05-06	30

Appendix B - Semi-structured interview guide

Authors' comments:

(1) All respondents were initially given an introduction to our master's thesis topic, intended both to clarify the scope of our research and to provide context for the questions that followed. We deemed this introduction important as we observed early on that some respondents did not distinguish between GPTs and SPTs, nor between agentic AI and prompt-required AI.

(2) All respondents were subsequently asked for consent to have the interview recorded anonymously, enabling us to revisit the material and avoid any misinterpretations. All but one respondent agreed to this arrangement.

Category	Example Prompts
Introduction	Main: Could you tell us a little about your background and expertise, as well as the company you work for? *(Additional)* We saw on your website that Company 1 handles all electricity trading

	functions (BSP/BRP/brokerage) and that this process is automated. Could you tell us a little about how you define automated trading? Do you currently use any AI for this? *(Additional)* We understand that Company 3 outsources its electricity trading-does this apply to both the physical and financial aspects? And do you know how automated your partner's trading processes are?
Topline Perspective on Agentic AI	What is your view on the use of autonomous agent-based AI instead of humans in the electricity trading market? *(Additional)* Does your outsourced trader use agentic AI? How would you feel if they did?
Associated AI Opportunities	What opportunities do you see associated with the use of agent-based AI in the electricity market?
Associated AI Risks	What risks do you see associated with the use of agent-based AI in the electricity market?
AI Mitigation Strategies	Are you familiar with any risk-mitigation strategies currently used by other market participants who incorporate agent-based AI into their trading models?
AI Liability	In the event of a black swan event created by an agentic AI, who bears the responsibility?
(Additional if not mentioned before) Conservatism Attitude in the Electricity Market	We understand that the electricity market is quite conservative. Given this attitude, how do you view the future of agent-based AI? Have you implemented new technology in the energy sector during your career, and how did you go about it?
(Additional if not mentioned before) Societal Impact	Company 8 has a broad repertoire and is partly state-owned. How do you view this dilemma, given the social responsibility that one could argue you bear?
Finishing Thoughts	Based on your expertise, is there any aspect we may have overlooked that you might know someone who could tell us more about?

Appendix C - Empirical Data: Theme Quotes Across Market Actor Segments

Authors' comments:

Below we present quotes from different respondents, each representing one actor segment of our choice: (1) Trading Company with BRP, (2) Trading Company with No BRP, (3) Government-owned Enterprise, and (4) Consultant (External Perspective) - presented in the same order in table below. The quotes are intended to provide a general understanding of our empirical data. It should be noted, however, that many interesting quotes have been omitted to mitigate overreliance on, or cherry-picking from, specific interviews. It should further be noted that the frequency of citation across respondents is uneven, reflecting variation in interview depth and respondent openness Findings attributed to the second principal level rest on a narrower evidential base than those pertaining to private trading actors, a limitation discussed further in section 6.1.

Given that respondents 20 and 21 both represent the same organization - the only government-owned enterprise in the sample - their accounts are aggregated to represent the institutional perspective. Where their statements diverge, both are presented; where they converge or only one respondent addressed a theme, the available quote is used to represent the institutional position.

Theme	**Example Quotes**
Current State of AI Use in Swedish Electricity Trading	"We mainly use machine learning, which is what we've been doing all these years, so what we do is make predictions." - Respondent 1 "We don't have one today. We don't have enough volume in the gas markets to feel that we need one. But we are fully aware that it exists, especially when you look at the intraday market and the flexible markets. The pace there makes it almost inevitable. With weather-dependent production and the speed at which the market moves, it's an inevitable path forward. We would never be able to keep up with human traders alone." - Respondent 16 "It's not people sitting there deciding which bids to choose; it's simply automatic algorithms. So it's not AI that's handling it just yet. There is a certain degree of machine learning involved in our forecasts." - Respondent 20 "Today we use fairly advanced systems, perhaps not exactly autonomous genetic systems, but we do use AI algorithms, that is, machine learning and AI algorithms in various ways to optimize our presence in the supermarket market." - Respondent 13
Trust and the Delegation Decision	"We don't let AI set its own trading rules; instead, we control the rules ourselves but actually use big data and machine learning. But I would say that to a fairly high degree, it's us who make the decisions." - Respondent 1 "I spoke with our compliance officer earlier. She sighed a bit and said, 'No AI.' She knows it's coming, but it will be a challenge." - Respondent 16 "I'm pretty skeptical about fully embracing AI. If you're a smaller player, there are probably big benefits to using AI, but you have to realize that you're ultimately responsible depending on how large your portfolios or trading volumes are, there's less

	risk with AI if you're a smaller player, the opposite is true for larger players." - Respondent 21 "As long as we keep up with technological developments, we'll realize how little we understand and what a bad idea it is to let this type of system control critical infrastructure in society. So I believe, and this is just my personal opinion." - Respondent 13
Monitoring and Control Mechanisms	"What needs to be done is to implement various types of limits, rules, and automatic controls that ensure an automated model can never run amok." – Respondent 1 "It's a technical issue - what information the AI is allowed to process and how it's trained. In human trading, a lot of it relies on individual responsibility. Otherwise, there are system controls: you can't place orders with excessively large volumes or prices that deviate too much. And procedures, as well as traders keeping an eye on each other." - Respondent 16 "AI probably shouldn't place bids on its own without using screening criteria; include the manual component." - Respondent 21 "You have to maintain a degree of fragmentation, meaning you can't have portfolios that are too large or consist of assets that are too similar or predetermined." - Respondent 13
Perceived Opportunities and Risks	"With AI, just as you can build an automated event, you can also build automated trade monitoring and automated trade control, so you have tremendous opportunities to constantly keep track of what's happening and set these kinds of limits. So I'd say it's more a matter of using AI to identify and suggest possibilities that you can then make a qualitative assessment of why it has arisen and sort of take some kind of action." - Respondent 1 "To keep up with larger players and manage ancillary services and intermittency, you have to be able to act at all times. I don't see any other solution than AI. In a way, this is good for the market - imbalances can be managed continuously, not just during extreme events." - Respondent 16 "AI can certainly be a valuable tool, but it shouldn't replace humans in decision-making, especially when it comes to the electricity market. It can make decisions that are hard to follow up on. That can have a very significant impact on confidence in the market." - Respondent 21 "If an AI's objective is to maximize profit - which is different from the objective regarding the system as a whole, such as a market participant controlling a sufficiently large portfolio or a sufficiently large share of the market - then there would be a significant risk that they could act in a way that maximizes profit, for example, by withholding capacity to drive up the price." - Respondent 13

The Second Principal: Regulatory Scope and Societal Stakes	"This includes both the electricity we sell to customers and the electricity we generate ourselves, and we try to balance the overall picture so that our goal is to be as close to zero as possible. The only thing we know for sure is that we'll never be exactly at zero; instead, our ambition is to be as close to zero as possible in terms of our balancing responsibility." – Respondent 1 "The EU has a tendency to overregulate. You can't make it so burdensome that it becomes impossible to do business." – Respondent 16 "(About system imbalances) It's passed on through all sorts of general fees, so it affects society - it doesn't affect Svenska Kraftnät itself, but ultimately it's always society that ends up paying for it. And we don't want that. We want operators to be good at planning, and then we want them to stick to their plan." - Respondent 20 "It's hard to distinguish what's actually going on, just like perhaps in the financial system, where it's extremely difficult to regulate and manage profit-maximizing behavior. So in that regard, Svenska Kraftnät has rules for how they're allowed to act, and they also have to report how they've acted and why they act the way they do." - Respondent 13

Appendix D - AI Disclosure

This thesis made use of several AI tools, including Claude Sonnet 4.6 and Opus 4.7, Gemini 3, ChatGPT 5.1, Perplexity Sonar 2, Elicit, ORKG Ask, and SciSpace. These tools were primarily used to assist with spelling and grammar, formulation and phrasing, feedback generation, and thematic synthesis of findings. For the thematic synthesis, raw interview and survey data were fed into the tools to help identify and organize emerging patterns. As described in section 3 (Methodology), AI tools were additionally deployed to support systematic screening of academic literature for relevance to the research topic.

The use of AI contributed to the quality of the thesis in several ways. The tools proved particularly efficient when processing large volumes of data, and were valuable for generating constructive feedback, rephrasing, smoothing transitions, and maintaining a consistent tone across sections. The greatest benefit was realized when the tools were used iteratively as critics and sounding boards rather than as generative substitutes for original thinking.

The key risks identified were data protection and over-reliance. To protect sensitive information, all data was anonymized prior to being processed by any AI tool, and no organization-specific information was included. To guard against over-reliance, initial drafts were always written independently before AI assistance was sought. Outputs were also fact-checked and compared across multiple tools, and prompts were designed to include opt-outs to reduce the risk of fabricated or misleading statements.

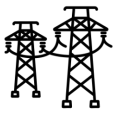
The overall insight from this process is that AI works best as a complement to, rather than a replacement for, human judgment. Using the tools iteratively as critics and supporters reinforced the

importance of human verification in maintaining the academic integrity of the thesis, while demonstrating that AI can meaningfully assist with the linguistic, structural, and analytical dimensions of academic writing. The irony of using AI to study the risks of delegating decisions to AI was not lost on us, and served as a continuous reminder of the principal-agent dynamics central to this thesis.

Appendix E - Icon Creator Attribution

Creator: BSD, via Flaticon (https://www.flaticon.com/authors/bsd)

Creator: Assia Benkerroum, via Flaticon (https://www.flaticon.com/authors/assia-benkerroum)



Creator: Culmbio, via Flaticon (https://www.flaticon.com/authors/culmbio)

Creator: Berkahicon, via Flaticon (https://www.flaticon.com/authors/berkahicon)