OmniaScience

Journal of Industrial Engineering and Management
JIEM, 2025 – 18(2): 285-304 – Online ISSN: 2013-0953 – Print ISSN: 2013-8423
https://doi.org/10.3926/jiem.8567

Human Oversight of Artificial Intelligence: An Operations Management Perspective

Jesús Salgado-Criado

Universidad Politécnica de Madrid (Spain)

jesus.salgado@upm.es

Received: November 2024
Accepted: April 2025

Abstract:

Purpose: This paper presents a theoretical framework for AI oversight and examines the key dimensions used by operational managers to define their oversight activities for AI applications.

Design/methodology/approach: The research combines a theoretical and qualitative approach. The theoretical part analyzes and proposes a framework for studying AI oversight from an operational perspective, drawing on cybernetic and control theory and recent literature on human oversight. This framework is then compared and categorized with the perceptions of managers regarding AI management and oversight.

The operational perspective views oversight not only as a safety mechanism but also as a governance mechanism that encompasses safety, ethical, and compliance requirements, as well as technical and business goals. Importantly, oversight is necessary regardless of the application's risk level.

Findings: The paper offers a more operational definition and framework for oversight, combining theoretical concepts and practical insights from industry practitioners. The theoretical framework clarifies the recursive nature of oversight within organizational control loops. The practical categorization of oversight design dimensions identifies key factors influencing the selection of resources, methods, and tools for AI applications oversight.

Research limitations/implications: The theoretical proposal is grounded in specific cybernetic and control theories, but other theoretical frameworks could be explored. The qualitative study provides a categorization of oversight dimensions, but each AI application and organization should adapt this framework to its specific needs.

Practical implications: This paper aims to assist companies in designing effective AI oversight functions that align with legal, technical, and business requirements.

Social implications: A meaningful and effective oversight of AI applications will enhance the trustworthiness of AI integration within organizations for all stakeholders, including employees, customers, investors and society at large.

Originality/value: This paper contributes to the ongoing discussion on human oversight of AI, which has been heavily focused on legal aspects since the publication of the European Union's AI Act. By adopting an operational perspective, the paper offers both conceptual and practical insights.

Keywords: human oversight, ai governance, operations management

To cite this article:

Salgado-Criado, J. (2025). Human oversight of artificial intelligence: An operations management perspective. *Journal of Industrial Engineering and Management*, 18(2), 285-304. https://doi.org/10.3926/jiem.8567

1. Introduction

Industrialization involves the mechanization of processes to enhance efficiency, reduce costs, or improve quality. In the realm of business, Artificial Intelligence (AI) is poised to industrialize cognitive processes, hitherto considered uniquely human, particularly those related to decision-making.

Decision-making permeates various corporate functions, including pricing, product development, distribution, manufacturing, research and development, innovation, finance, and human resources. Traditionally, these processes have relied on human intervention. However, AI emerges as a tool to partially or fully automate decision-making, thereby industrializing these processes.

It is important to note that the term 'Artificial Intelligence' may not fully encapsulate the transformative potential of these technologies in industrializing decision-making. This focus on AI-driven automated decision-making is often accelerated by competitive pressures, as companies seek to gain a competitive edge through increased efficiency and revenue generation.

On many occasions, the complexity of these decision support systems exceeds human cognitive capacity for understanding the underlying decision processes. However, the regulatory push for human oversight like in the European AI Act, (Regulation (EU) 2024/1689 – AI Act, 2024) appears contradictory. If AI demonstrably surpasses human capabilities, wouldn't a more rational approach involve acknowledging human limitations and delegating oversight to another digital agent, potentially one better suited for the task? Why must oversight necessarily be human-centric? In fact, several works have shown the limited efficacy of humans in the oversight role for some applications (Green, 2022).

This paper contributes in two ways: first, it introduces a conceptual operations management framework rooted in cybernetic concepts and control theory to clarify the role of oversight in organizational operations. Second, this framework is complemented with practical design criteria informed by interviews with managers overseeing AI, to establish specific requirements for oversight.

In this paper, section 2 will introduce the motivations for this study. Section 3 will introduce the theoretical framework to conceptualize human oversight in Operations Management. Section 4 will discuss the methodology for qualitative research. Section 5 will discuss the results of this research and section 6 presents conclusions and further research suggested in this area.

2. Motivation

Several factors motivate this research: the tension between legal, technical, and business requirements, the underrepresentation of operational practices in academic discourse on AI oversight, and the need for a more multidisciplinary approach to address the challenges posed by AI.

2.1. The Tension between Legal, Technical and Business Requirements

A fundamental tension exists between the legal imperative for responsibility and accountability, centered in human agency and the technological imperative for effectiveness and risk management. If AI systems surpass human capabilities, particularly in high-stakes scenarios, it raises the question of whether delegating oversight to another digital agent might be a more rational and ethical approach than delegating oversight to a human agent or group thereof. This raises the question of whether human oversight is always necessary.

What are the reasons behind this regulatory determination of including human oversight requirements on AI Systems? AI Act states in its article 14 that human oversight "shall aim to prevent or minimize the risks to health,

safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse".

One factor for this requirement might stem from the apprehension that autonomous machine operation, absent human involvement, inherently poses a greater risk, but this is not sustained by experience, as indicated by some studies that contend that assigning human supervision can lead to a false sense of security (Green & Kak, 2021).

The notion that human control is inherently less risky than algorithmic control is questionable. A human operator, influenced by factors such as unscrupulous management, poor health, or inadequate training, may exhibit less predictable behavior than a programmed oversight system with fixed parameters. In fact, a predictable system may be preferable from a risk management perspective.

Looking into the future potentially massive adoption of AI agents, capable to autonomously establish their own intermediate or final goals, the idea that a human overseer can effectively monitor and control a potentially malevolent AI system might appear unrealistic. Such a system would likely prioritize eliminating the human threat. Therefore, relying solely on human oversight as a safeguard against AI risks is regarded as insufficient.

Another factor is the requirement for human accountability in the event of adverse consequences arising from AI decisions. While Automated Decision-Making Systems (ADMS) raise questions of accountability, we are already accustomed to similar concepts in other products. For instance, if an elevator malfunctions and injures passengers, the manufacturer, maintenance company, or even the user may bear responsibility. This demonstrates that even in complex technological systems –where high-risk decisions are not made by humans– accountability can still be assigned to individuals or entities. Therefore, the need for a human to be directly responsible for AI decisions, rather than considering a more nuanced approach to artificial responsibility, is questionable. In this vein, the oversight function has been criticized as a potential scapegoat to deflect responsibility from companies.(Wagner, 2019)

The tensions described above mirror the input and output legitimacy debate in democratic theory. Mena and Palazzo (2012), Schmidt (2020) and Boedeltje and Cornips (2004), where input legitimacy contends that decision-makers should be legitimate (input legitimacy), while output legitimacy emphasizes that beneficial outcomes should be secured.

2.2. The Need to Adapt to Current Operational Practices

The oversight function, particularly in low-stakes applications, is already being implemented in organizations. The practices and insights gained from these experiences can provide valuable guidance for addressing the emerging requirements of the AI Act and industry standards. Existing research, such as that of (Schröder & Schulz, 2022), and commercial monitoring tools (Nigenda, Karnin, Zafar, Ramesha, Tan, Donini et al., 2022) offer a foundation for developing effective oversight.

A comprehensive approach to oversight will need to consider not only regulatory compliance but also the alignment of the oversight function with the specific business, operational, and technical goals of the underlying AI system.

The AI Act mandates that providers identify "appropriate human oversight measures" before market introduction or deployment. It emphasizes the need for in-built operational constraints that cannot be overridden by the system itself. In practical terms, deployers will need to parameterize control variables, such as acceptable fairness measures like demographic parity, which will be monitored by the human oversight component. Recognizing that perfect fairness is often unattainable, deployers will be responsible for defining acceptable operating ranges for these variables in their specific contexts where their applications will operate. The continued assessment of these operating ranges is critical, as these parameters may evolve over time based on experience. The AI Act does not specify precise values for these parameters, and generic standards are unlikely to provide detailed guidance. This will allow deployers considerable flexibility to operationalize the generic principle of oversight.

2.3. Multidisciplinary Approach

There is a risk, as with many AI governance issues, that the conversation surrounding human oversight in AI becomes compartmentalized within academic discourse. Fragmentation across disciplines–focusing solely on

technical aspects (human-computer interaction), legal and ethical considerations (potential harms to citizen rights), or specific areas of management (operations, data science, product marketing, finance)–could hinder progress towards a systemic view. A multidisciplinary approach is essential for a more comprehensive understanding of how to effectively govern advanced digital innovations designed to support decision-making at individual, organizational, community, and societal levels. This broader perspective can enrich academic discourse and ultimately lead to better outcomes. Today, the main perspective under which this issue is being treated is mainly legal (Enqvist, 2023) and ethical (Díaz-Rodríguez, Del Ser, Coeckelbergh, López de Prado, Herrera-Viedma & Herrera, 2023).

Sterz, Baum, Biewer, Hermanns, Lauber-Rönsberg, Meinel et al. (2024) make an important contribution in this respect, synthesizing insights from psychological, legal, philosophical, and technical domains. We argue that management science and industrial engineering research has a crucial role to play in this discussion and this is the motivation of this paper. Industrial engineering is inherently an interdisciplinary domain, directly involved in the design or improvement of a system of people, machines, information, and money to achieve some goal with efficiency, quality, and safety (Boardman & Fraser, 2020), human factors, ergonomics and safety are common concerns in industrial engineering that are also present in the analysis of an effective oversight of AI applications. Management itself is grappling with the emergence of AI technologies that have the potential to augment decision-making capabilities and, in some cases, completely automate decisions traditionally made by human managers. This inherent tension between human managers and AI systems can lead to conflicts as organizations navigate the integration of these technologies within their operations. (Leyer & Schneider, 2021).

3. Literature Review and Theoretical Framework

Many high-level legal documents, such as national and international law, UN resolutions, and those from agencies like UNESCO, often employ terms with multiple potential interpretations. This ambiguity can render these documents ineffective until subject to legal interpretation, such as in a court of law. While this practice is accepted within the legal profession and partially addressed through standardization efforts, this disconnect can hinder the effective implementation of regulation and industrial practices in alignment with regulatory frameworks designed to protect citizens. This could be the case of the term "human oversight" in the context of AI systems.

Human oversight emerges as a central theme in legal, academic, and industry discussions concerning AI governance. It is frequently cited as a critical mechanism for ensuring responsible development and deployment of AI and digital applications. The need for human oversight of AI is increasingly appearing in most international regulatory frameworks, for example, in the European Union's General Data Protection Regulation (GDPR) (European Union, 2016) in Article 22 and the proposed European AI Act (Regulation (EU) 2024/1689 – AI Act, 2024) in Article 14 or in international frame-works like the UNESCO ethical AI recommendations (UNESCO, 2022). The UNESCO document combines the direct responsibility for the decisions in case they are made by a human or the responsibility of the person who made the decision to give control to the AI (articles 35 and 36), specifically excluding the possibility to cede life and death decisions to AI systems. However, these types of political documents do not specify in detail what human oversight means, leaving the way free to the interpretations of civil organizations (Digital Future Society, 2022) industry and academia (Laux, 2023). A common concern underlying these efforts is the effective control, security, and alignment of these AI-based systems with the interests of individuals, organizations, and society. Concerns about human control and oversight are the primary drivers for safety, as illustrated by the Bletchley Declaration signed by the countries attending the AI Safety Summit in November 2023 (Various States, 2023). This emphasis is particularly warranted in the nascent stages of this technology, where our understanding of its potential and risks remains limited.

One way to address the ambiguity surrounding terms like "human oversight" in AI regulations is through standards development. The AI Act mandates the European Commission to issue standardization requests, as outlined in Article 40. Following this, on May 22, 2023, the European Commission tasked CEN and CENELEC with developing new European standards to support the AI Act by April 30, 2025. Additionally, the development of ISO/IEC AWI 42105, an international standard for human oversight of AI systems, is underway. The findings of this research may contribute to these standardization efforts.

Despite the growing importance of human oversight in AI, and in the absence of standards that clarify its definition and implementation, a clear and comprehensive understanding of human oversight remains elusive. Existing literature lacks a precise understanding of when oversight is most appropriate, the optimal methods for its implementation, and the effectiveness of different approaches in mitigating AI-related risks.

Addressing this gap requires the development of an operational definition of human oversight, a robust classification system for oversight methodologies, and a thorough evaluation of their effectiveness in mitigating potential harms of different nature. The concept of oversight, whether human or automated, is inherent to all applications due to their susceptibility to flaws and performance degradation. Such degradation can lead to various harms, ranging from financial losses to societal or individual harm caused by biased or inaccurate decision-making systems.

3.1. The Critical Function of Oversight in Organizations

Prior to examining the role of human agency in oversight, it is imperative to delineate the fundamental function of oversight within organizational processes. This sequential approach is warranted because the efficacy of oversight itself constitutes the primary analytical concern. The question of who executes this function, whether a human individual, a collective of individuals, an institutional entity, an algorithmic agent, or a hybrid combination thereof, should be considered as a subsequent, separate inquiry. At the operational or decision-making level, human oversight may be preferable if human over-sight capabilities surpass those of algorithmic oversight in all relevant dimensions (speed, reliability, cost, precision). Conversely, machine oversight at the operational level should be warranted if machines demonstrably outperform humans across all criteria. In scenarios where neither humans nor machines exhibit absolute superiority, a hybrid approach that leverages the strengths of both might be optimal. Empirical evidence demonstrates that algorithmic decision-making, while powerful, is inherently subject to limitations. Consequently, effective management is essential to mitigate potential deviations from intended outcomes. These limitations include, but are not limited to, the literal nature of algorithms, occasional challenges in interpretability, and the necessity for managers to navigate trade-offs while considering intangible or soft goals (Luca, Kleinberg & Mullainathan, 2016).

We underscore from the outset the paramount importance of oversight: all control or governance processes require direct or indirect oversight. Effective oversight mechanisms, human or otherwise, are crucial to ensure the model's proper functioning within established parameters of accuracy, safety, and–depending on the application–other performance requirements and limitations. This emphasis is particularly warranted in the nascent stages of this technology, where our understanding of its potential and risks remains limited.

3.2. AI Automated Oversight

AI agents and advanced digital applications are increasingly assuming partial or full oversight roles in various domains. Examples of AI-driven oversight include the supervision of nuclear plant operators (Ahn, Bae, Min & Lee, 2022) and the detection of judicial bias (Kleinberg, Ludwig, Mullainathan & Sunstein, 2018). Regulatory frameworks like the EU NIS2 directive (Directive (EU) 2022/2555 – NIS 2, 2022) and DORA regulation (Regulation – 2022/2554 – DORA, 2022) also encourage the use of automated tools for cybersecurity and operational resilience.

Historical nuclear accidents, such as Three Mile Island, Chernobyl, and Fukushima, highlight the limitations of human oversight in complex automated systems. These accidents were often caused by operator errors or misinterpretations of automated systems functionalities (Schmitt, 2012).

Similar challenges arise in industries like banking, where anti-money laundering teams face the burden of processing numerous alerts. Automating the handling of low-priority alerts can help prioritize critical cases and optimize resource allocation.

Regulatory frameworks, industry standards, and empirical research consistently highlight the benefits of automating risk event responses, particularly when time-critical actions are required.

3.3. Cybernetic and Control Theory Metaphors

We argue with (Wahlström & Rollenhagen, 2014) the need to use a systemic approach when analyzing systems safety and systems governance in general. Along with these authors we will use control and cybernetic metaphors to analyze this governance mechanism of human oversight of AI.

Drawing on concepts from cybernetic governance (Deutsch, 1963), viable system theory (Beer, 1995) and the work of (Schuh & Kramer, 2016) this section of the article explores the definition of the oversight function at multiple governance levels, with a specific focus on AI models. Although cybernetic governance has not been further developed, it is a de-facto framework whose concepts of information feedback-based governance is present in most management and policy domains expanding from the fields of engineering, biology and neurology where the principles of feedback and control through communication were originally observed by Norbert Wiener and Arturo Rosenblueth (Wiener, 1961).

Figure 1 illustrates the control, actuation, and supervision functions within a generic feedback loop, encompassing the system object of control and the control system (or management). For the viability of a control system and the overall system, it is assumed that a model exists that adequately represents the behavior of both the system under control and its environment (Schwaninger, 2010). In the case of an AI-based control block, this environmental model is implicitly learned during the AI model's training process.

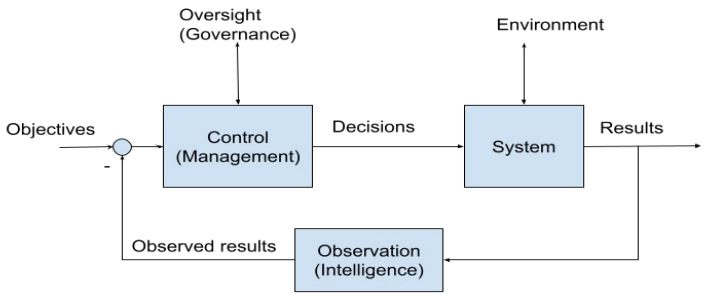

Figure 1. Control loop and oversight function.

Both the oversight function and the control or governance function itself are already carried out in many cases by individuals or human teams, but it is crucial to distinguish between control and oversight, as these terms are often used interchangeably. In this paper, control (or management) refers to the organizational components that directly influence the behavior of a system. This system may be internal or external. For instance, a factory's operations, governed by processes, personnel, and norms, constitute a control system.

Oversight, in contrast, is the mechanism by which an organization ensures the adequacy of the control system's performance. A quality assurance committee that monitors and corrects factory output is an example of an oversight function. Essentially, oversight can be considered meta-control, a control system for the control system itself. While organizations are adapting to AI-driven decision-making, the oversight function remains crucial, as noted by (Shrestha, Ben-Menahem & von Krogh, 2019), regardless of whether decisions are fully delegated to AI or made in a hybrid setting.

The various positions of humans within AI systems, often described using terms like "human in the loop," "human on the loop," and "human over the loop," (Singh & Szajnfarber, 2024) can be clarified through the lens of control loop theory. Humans "in the loop" normally refer to humans occupying positions within the control block, but humans are normally also present in the system block, or the observer block (See figure 1.) which is part of the feedback loop. Within the control block, human decision-making may be complemented by AI and other tools.

From this perspective, therefore, the Human-in-the-Loop (HITL) configuration is not a form of oversight or supervision, but rather a scenario where humans are directly involved in the control process. Effective oversight, on

the other hand, requires the ability to withdraw or modify control from the current system. In this framework, oversight always operates 'over the feedback loop', holding to this oversight function the responsibility for the control system. Arbitrarily placing a human in the loop with the sole purpose to assign to this human the complete control loop behaviors, would be a fundamental flaw in governance design, since the oversight function would be included within the control system, lacking the necessary perspective and independence. It could obscure accountability and deflect responsibility from the actual decision-makers.

In the literature, several scholars have proposed different taxonomies for the human intervention in decision-making. According to (Binns & Veale, 2021), human interventions are classified into the broad categories: Summarizing (the system consolidates human interventions/data from one or more decision-makers that leads to an automated decision), or supporting (the system provides information to the human decision-maker with the human then considering the system's "advice") or Triaging (the system automatically processes cases unless these are flagged for human review). All these cases describe in our framework a "human-in-the-loop" situation, and not a true "oversight function".

Laux (2023) distinguishes between "first-degree" and "second-degree" human oversight. First-degree oversight involves human influence over decision outcomes, often associated with HITL contexts where humans have agency, even during the design phase. In our framework, this would be considered part of the control system. Second-degree oversight, on the other hand, occurs after decisions are made and involves correction or modification. This aligns more closely with our conception of the oversight function, indicating that oversight is relevant even in partially automated decision-making scenarios.

Even though HITL configurations involve human agency and can address accountability concerns, a supervisory or oversight function remains essential. Human operators working with algorithms may be susceptible to limitations, such as fatigue or stress, which can compromise their ability to make sound decisions. Therefore, an additional layer of oversight is necessary to mitigate these risks.

In summary, in a Human-in-the-Loop (HITL) setting, where a human is involved in every decision cycle, that individual is not strictly an overseer but rather an integral part of the control system itself. In our view, an overseer is an external observer of the entire control loop, possessing the authority to intervene when the complete system –including the control block– deviates from its intended behavior.

Throughout this article, we will use the term AI "overseer" to denote the person that performs the function of oversight of a standalone AI application or oversights a team applying AI for its decision-making. It does not have to be identified with a specific job title but more with a function or activity developed perhaps by existing roles in the company. A production manager might be an overseer if she is monitoring the performance of a production facility or team that uses an AI application for its operations. A company commercial director is an AI overseer if her team is employing AI models in its pricing strategy. The board of directors act as ultimate overseer body for a company that relies on AI applications, part of its core operations.

A cybernetic perspective highlights the recursive nature of the control-oversight relationship. Oversight function operates as a meta-control mechanism, regulating the entire control loop. In turn, the oversight function and the incumbent control loop can be part of a higher-level control loop. Organizations can be conceptualized as complex networks of nested control loops interconnected by oversight mechanisms.

This multi-layered approach to oversight and governance extends beyond organizational boundaries. At the societal level, laws and regulations govern the operation of hazardous facilities and the deployment of potentially risky systems. Organizations, in turn, establish internal policies and management systems to ensure the safe design, construction, operation, and maintenance of their activities, safeguarding both the organization and its stakeholders. At the operational level, control systems, incorporating feedback loops and feedforward mechanisms, monitor outcomes and initiate corrective actions to address any identified deficiencies.

To understand how micro-level governance mechanisms within AI applications are situated within macro-level contexts, and how macro-level regulatory frameworks arise from the interactions and dynamics of lower-level elements, a multi-level perspective is essential (Klein & Kozlowski, 2012).

In contrast to these micro- and macro- level of analysis, the present research focuses on a meso-level perspective: the management of organizations operations. Within these operations, companies and their teams are actively exploring both successful and unsuccessful strategies for integrating AI applications. Operations management is also very frequently the level of governance where economic, competitive, technical, cultural and also legal, political and ethical pressures effectively converge when integrating AI applications, driving progress but also potentially generating risks. At higher levels of corporate governance, (like in the board of directors), the oversight function is normally too distant from the AI application itself. In operational teams directly using AI are usually working closely with the AI itself. In our framework, operational teams are "in-the-loop". Operations management becomes the focal point for balancing sometimes conflicting economic and strategic goals, and it bears responsibility for the entire system incorporating AI. It is critical that this company function incorporates within existing oversight functions, the ability to oversee AI decisions.

The recursive nature of the control-oversight relationship implies that, even if a specific level of oversight –e.g. operations management– is absent, a higher-level oversight mechanism –e.g. corporate governance– may inadvertently assume this role. However, such a delayed and informal oversight function may prove ineffective. In the absence of formal internal oversight within the company, external entities such as regulatory agencies or market forces can act as de facto overseers, potentially leading to organizational disruption if internal control systems are inefficient.

This is why responsible organizations prioritize the inclusion of oversight functions in all decision-making processes, human or otherwise, even when not explicitly required by regulations.

3.4. Monitoring Challenges

As it can be seen from our main cybernetic framework, both the control and the oversight functions largely depend on the ability to clearly monitor the system performance. Specifically in the oversight area, the specific definitions of system performance pose challenges. For example, legal frameworks often employ vague terms to justify the imposition of measures like human oversight. For example, AI Act says that the goal of human oversight is to prevent or minimize "risks to health, safety and fundamental rights," leaving significant room for interpretation. Translating these broad concepts into precise, formal definitions and operational limits remains a challenge.

A particularly complex area for system monitoring is the fairness of decisions involving resource allocation, such as loan approval or human resources recruitment. While extensive research has been conducted on fairness measures, a universally accepted definition remains elusive. In fact, several fairness metrics are inherently incompatible, making simultaneous optimization impossible (Verma & Rubin, 2018). Likewise, fairness metrics in Large Language Models, exemplified by potential representation discrimination or stereotyping also pose some measuring challenges (Gallegos, Rossi, Barrow, Tanjim, Kim, Dernoncourt et al., 2024).

In the more technical realm of accuracy measurement, specific indicators like accuracy, precision, and recall are commonly used. However, monitoring these metrics in real-world production environments presents significant challenges. These challenges stem from the inherent trade-offs between different performance metrics and the scarcity of post-decision data in production environments. For example, in certain contexts, accurately measuring false positives or false negatives is hindered by the difficulty in verifying the classification's correctness. For example, in human resources recruitment, incorrectly rejected candidates (false negatives) cannot be assessed, making accurate estimation of false negatives challenging. Such retrospective evaluations can be biased and unreliable.

3.5. Oversight Modes and Human Involvement

The oversight function typically operates in a supervisory role, observing the performance of the underlying control system. It responds to potential risks and opportunities that may be overlooked by the control system. The specific processes involved in oversight can vary depending on the application, but we understand the oversight function as operating in two different processing modes: monitoring mode and response mode.

In monitoring mode, oversight typically involves assessing, filtering, and alerting on cases that exceed various warning thresholds. This allows operators to prepare for potential issues when high-risk limits are breached.

Dashboards often display multiple, competing criteria for oversight, such as business objectives and safety or fairness measures. Balancing these competing priorities, overseers must navigate within the available space, often constrained by hard and soft limits. It is likely that a Pareto frontier exists for each model, representing the optimal trade-off between different objectives.

In response mode, actions may range from isolating specific cases, intervening directly (e.g. overriding AI decisions), or steering the controller towards a safe state. In extreme cases, a complete shutdown of the control system may be necessary.

A crucial question arises regarding the role of human involvement. The concept of "human oversight" could encompass the involvement of one individual or a group of individuals. The process of involvement could be defined in a more or less formal fashion. However, oversight can also be partially or completely– algorithmic, involving automated systems that monitor complex systems with rapid dynamics, such as high-frequency trading. In such cases, AI-powered tools can assist regulators in identifying potential infractions. (Finantial Conduct Authority, 2024).

Collective oversight can be observed in content moderation platforms like Wikipedia and Meta. Wikipedia employs a layered system of volunteer overseers (Wikipedia, 2022), while Meta utilizes a multi-layered approach combining AI and human reviewers (Meta, 2024). AI systems identify and flag potential violations, which are then assessed by human moderators at various levels of expertise, although this governance structure was announced to be substituted by a crowd sourced approach in January 2025. An independent oversight board (Meta Oversight Board, 2024) reviews appeals of moderation decisions. These examples illustrate the diverse nature of oversight functions, which can be layered, hybrid, and involve both individuals and groups. As will be elaborated upon below in the research discussion section, despite the theoretical possibility of fully algorithmic oversight, no AI system is reported to be oversighted solely by another digital agent. This observation reflects both the current technological state of the art, current reliability and trust in AI control systems and the prevailing organizational accountability structures. Ultimately, at the highest level of responsibility, a human individual will be accountable for the behavior of specific company functions, such as operations, marketing, or finance. And, under these conditions, proximity to the AI system enhances a human's capacity to address inquiries regarding its behavior.

4. Research Questions

In line with the preceding discussion, the following research questions emerge as relevant to the broader objective of examining the human oversight role in algorithmic decision-making: a) How do company managers overseeing data science teams, operations analysts, and human resources personnel perceive the growing adoption of Decision Support Systems (DSS) and Automated Decision-Making Systems (ADMS) in relation to the concept of human over-sight? b) Do these managers perceive human oversight as necessary? If so, under what specific conditions? And c) More generally, how do these practitioners conceptualize the role of human oversight within AI governance frameworks?

5. Methodology

This research adopted a constructivist grounded theory methodology (Charmaz, 2014), premised on the understanding that theoretical frameworks are not passively discovered but actively constructed by the researcher through interaction with participants. Guided by this approach, our investigation centered on the practice of oversight in AI applications extensively deployed across various service industries. In contrast to phenomenological qualitative research, we intentionally refrained from adopting a pre-defined theoretical foundation at the outset, aligning with the principle of 'theoretical agnosticism' (Henwood & Pidgeon, 2003). Instead, the initial stages of the study involved a process of theoretical sampling, wherein a theoretical framework was co-developed iteratively with the accumulating data and insights from interviewees.

Data collection comprised semi-structured interviews conducted with ten operational executives across the identified service industries. The participants held diverse roles, including Data Science Directors (3), Operations Directors and Managers (3), Commercial Managers (2), Research and development director (1) and a Human Resources Manager (1). These interviews, lasting approximately one hour each, were conducted online between

March and May 2023. All primary interviewees were based in Spain. The data was gathered in two sequential batches of five interviews, with initial coding procedures applied to each batch. Data saturation, indicating the point where further data collection yielded no novel codes or properties, was deemed to have been reached following the initial coding of both batches.

A key methodological decision involved the selection of AI applications for this study. Applications were drawn from the airline and hotel industries (demand forecasting and fare pricing), banking (risk management and customer churn probabilities), retail (seasonal discounts and product cataloguing), and hotel (pricing and online marketing content). From these, only banking sector respondents reported responsibility for AI Act high-risk applications. We contend that risk exists on a continuum, rendering the exclusion of low-risk applications from oversight unjustifiable. Oversight serves not only risk mitigation, including economic risks, but also opportunity capitalization. Indeed, all AI systems necessitate oversight, whether to rectify unstable states or optimize performance. This principle extends to all governance mechanisms, as the absence thereof constitutes deficient governance. Furthermore, the exclusion of low-risk applications, defined by direct harm to individuals (notwithstanding the potential for economic damage to the company from AI malfunction to indirectly impact public well-being), offers a methodological advantage. In high-risk scenarios, compliance considerations tend to dominate managerial discourse, potentially overshadowing other critical aspects of the oversight function, such as technical, operational, strategic, and commercial factors.

The analysis of the transcribed interview data was facilitated by computer-assisted qualitative data analysis software (Atlas.ti). A coding strategy informed by grounded theory principles was employed to inductively generate hypotheses and construct a coherent theoretical understanding of operational managers' perceptions regarding AI oversight. Initial coding focused on the varied processes undertaken by executives in different company functions concerning AI applications throughout their lifecycle. Notably, the early stages of analysis revealed a conceptual ambiguity surrounding the precise meaning of the concept of AI "oversight". Interviewees' interpretations ranged from active decision-making to monitoring and exception handling. To address this, a cybernetic framework was selectively sourced and adapted as a supportive conceptual tool to establish a shared vocabulary with participants regarding the concept of oversight.

It is crucial to emphasize that the cybernetic framework served as an analytical aid rather than the central research question or a pre-determined hypothesis to be tested. Instead, it provided a common language for engaging with interviewees. Throughout the research process, we maintained an open stance regarding the most appropriate theoretical lens (or lack thereof) to explain the collected data, which was subsequently coded and categorized. In addition to standard inductive methods, we also employed abductive reasoning and subjective interpretation to explore and categorize the perspectives articulated by the participants. The inherent flexibility of constructivist grounded theory proved well-suited to the exploratory nature of this inquiry and the complex social dynamics involved in how operational managers grapple with the necessity of overseeing the algorithms they utilize daily. To mitigate the inherent subjectivity of this methodology, rigorous techniques such as coding, memo writing, theoretical sampling, and the application of consensus criteria like data saturation were consistently employed.

To enhance the robustness of the findings through triangulation, the primary dataset was supplemented by two additional interviews with professionals in industries utilizing advanced digital tools but outside the specific domain of AI-driven services: a nuclear plant engineer and a cybersecurity specialist. All interviews, including these supplementary ones, were conducted in Spanish, with the subsequent coding and excerpt translation into English performed for this article.

As is inherent in qualitative research, a primary limitation of this study lies in the number of organizations represented in the interview sample. Consequently, the research was not designed to verify or falsify specific pre-existing hypotheses. Instead, following the grounded theory approach outlined, its strength lies in the generation of hypotheses and the construction of narratives directly derived from the experiences of executives actively engaged in AI oversight. The principal contribution of this research, therefore, is the development of these empirically grounded hypotheses and narratives, offering valuable insights into the perspectives of those directly responsible for overseeing AI within their operational contexts..

6. Discussion of Results

Employing a constructivist grounded theory approach, transcribed interviews were coded and conceptualized, yielding categories reflecting common managerial themes. This process culminated in the identification of five distinct dimensions of AI application oversight, forming a proposed framework for analyzing algorithmic decision-making governance: Decision significance, Perspectives of oversight, Conditions for oversight, Oversight skills and Organizational culture and accountability.

6.1. Decision Significance

The impact or consequences that decisions may have on the organization and its stakeholders is a critical dimension for the design of oversight function. This impact can be evaluated at both the individual decision and at the aggregate levels. For instance, a single decision, such as product introduction or retirement, can have significant commercial consequences. Alternatively, numerous small decisions, like product pricing, can collectively have an important impact on the organization's financial performance.

Human oversight is deemed important where decisions have asymmetrical consequences for different stakeholders. Some interviewees report that AI systems (control system in our cybernetic approach in figure 1) often prioritize a single stakeholder or criterion, typically financial performance, especially in pricing algorithms, like in maximization of revenue per available room or airplane seat. In contrast, other interviewees think the oversight function must consider a broader range of interests and criteria, like customer satisfaction (acceptability in case of excessive prices), potentially leading to trade-offs between conflicting objectives, which suggests, for these interviewees, that humans are better at this multi-criteria judgement.

The importance of decisions must be evaluated contextually, requiring human judgment that is difficult to automate. While a loan to an individual may seem significant, the well-being of a company can impact numerous families and the broader economy.

> *"During the pandemic, the bank relaxed the rules for loans. This meant we were more easygoing about risk and let businesses de-lay payments. A lot of small companies were able to survive because of this."* (D14, 9:18)

In situations where individual decisions have asymmetrical risk profiles, meaning that false positives and false negatives have different levels of impact or frequency, oversight functions often intervene, especially when the decision conflicts with the interests of another stakeholder, like customers. For example, in loan requests, false positives can lead to lost customers who may seek financing from competitors. In such cases, the oversight function, when a special case is detected, may request additional information to mitigate risks, often placing the burden of proof on the stakeholder. In the banking industry, a common practice is to request further financial information from loan applicants in cases of potential fraud.

The significance of decisions, particularly in worst-case scenarios, should inform the allocation of humans to control and oversight functions. The nature of human involvement in AI applications can vary widely, ranging from direct involvement in decision-making (control system, human-in-the-loop) to monitoring aggregate performance indicators or individual alarms (oversight function in our cybernetic framework in figure 1, or human-over-the-loop). The choice of human involvement depends on the decision's significance and frequency. If the frequency of cases where human involvement is needed is high, then humans are involved directly in the control system as part of the decision-making process, and an additional oversight function is then required.

The composition of the oversight team displays differences among the interviewees. For instance whether the oversight is individual or group-based, is contingent upon the specific task. Group oversight is often preferred for monitoring overall system behavior, while individual oversight is more suitable for scrutinizing specific decisions. This decision hinges on the nature of the impact: whether individual decisions carry significant risk, as in investment decisions (like retail outlets investment/divestment decisions), or whether the aggregate impact of multiple decisions is more critical, as in pricing decisions for the complete inventory of a retail chain.

Furthermore, a risk analysis is typically conducted in the oversight function to identify and quantify the potential consequences of incorrect decisions. For instance, a bank manager highlighted the risks associated with misclassifying a customer's expression of dissatisfaction, such as:

> *"The customer was writing to customer support "You must be proud of what you just did" with irony. The sentiment classification model classified this inter-action as "joyful", when clearly, we were losing a customer."* (D14, 14:73)

Somewhat related with decision significance, the reversibility or irreversibility of a decision can influence its perceived importance. High-significance decisions, whether business-related or ethical, become even more impactful when they are irreversible. In these cases, human oversight and accountability is more demanded by managers.

In summary, the individual or aggregate significance, or impact, of the decisions that AI systems are making or supporting, constitutes a primary rationale for the design of the oversight function and the involvement of human actors within this function.

6.2. Oversight and Control Goals

Oversight plays a crucial role in ensuring that AI systems adhere to their original goals. Effective oversight ensures that algorithms consistently meet expectations in terms of quality, accuracy, and processing speed. Oversight functions should not be limited to safety and compliance concerns but should also consider operational, technical, and business objectives. A balanced approach is necessary to address potential risks while maximizing the benefits of AI systems.

The articulation of business and strategic objectives typically occurs within a project investment concept. Subsequently, various technical goals and the anticipated goals and impacts for diverse stakeholders are delineated. Interviewees emphasized the critical importance of associating measurable criteria with all stated objectives, specifically through key performance indicators (KPIs), which are subsequently monitored within the oversight function. For example, in a hotel revenue management model, metrics such as revenue per available room (RevPAR) and occupancy rate serve as essential indicators of financial performance. Concurrently, customer satisfaction and online reputation metrics related to pricing provide insights into the operational and quality alignment between the service delivered and the pricing offered to clients.

Additionally, oversight must address potential ethical, safety, and compliance risks associated with AI applications. Risk management is a key aspect of oversight, and oversight design analysts should define acceptable operating parameters and red lines to mitigate risks.

It's important to note that oversight extends beyond safety concerns, as mandated by regulations. For instance, while the AI Act mandates human oversight to "prevent or minimize the risks to health, safety or fundamental rights," the oversight function as implemented in practice by organizations will also (and primarily in many cases of low-risk applications) aim to ascertain the fulfillment of business and technical objectives.

Even if goals and objectives have been defined and agreed beforehand during design and development phases, discrepancies can arise when the AI system is in production between the objectives of various teams and the pre-programmed goals of the AI model. In such instances, the ability to modify model priorities is a valuable tool that the overseer should possess and utilize. As one Data Science Director for a retail chain stated:

> *"The algorithm is set up to optimize margins, but there are times when store managers need to clear out excess inventory. In those cases, you need to directly change the model".* (D11, 425)

Short- and long-term objectives also might get into conflict. In these cases, an oversight function may provide priorities. A hotel marketing manager discussed with us these dilemmas between commercial tactical goals and marketing strategic goals:

> *"The algorithm [that was selecting photos for online travel agencies] was choosing too frequently young girls in the hotels swimming pools. Although conversion was good, we had to tweak it a little to align with the brand image".* (D16, 38:40)

Conflicting objectives often arise between different teams and the pre-programmed goals of AI models. These conflicts become more pronounced in high-risk applications, where they can clash with business priorities, especially when health, safety, or fundamental rights are involved.

To sum up, the goals of the oversight function, coincident with the goals of the control system, is another dimension that drives the design of the oversight function of an AI system. These goals need to be specified and agreed by affected stakeholders and teams in the company, in early phases of design. The design of the system and the oversight function should allow for flexibility. Tradeoffs between conflicting goals are clearly a feature of the oversight team and function.

6.3. System and Model Complexity

Several characteristics of the system under control and the control itself are certainly driving the methods in which the system might be oversighted, including the processes and the tools to be used. Although not all interviewees concur in the same aspects, five main characteristics of the system under control have been identified during the interviews: time dynamics, concurrency of decisions, data complexity, observability and controllability of the complete system and explainability and transparency of the control system and underlying cybernetic model, explained in section 3.3.

Time dynamics significantly influence the design of oversight procedures. In slow systems, overseers can monitor results through dashboards and manually adjust parameters if necessary. However, in fast-paced systems, automated alarm and containment systems are crucial to prevent rapid instability. These systems trigger immediate corrective actions to stabilize the system temporarily, followed by a more permanent solution through model redesign and redeployment.

Concurrency of independent decisions refers to the number of actual decisions or results the controller is making per unit of time for independent units of decision. For instance, the pricing system of a hotel chain might be making one pricing decision per day for each rate type and room type for each hotel in the chain, for each date in the future. Clearly the human oversighter cannot check each of these concurrent decisions and will only analyze a specific situation if flagged as an exceptional condition.

Data complexity significantly influences the design of the oversight function. When dealing with numerous variables and data points and complex variable interactions, automated alarm and recovery systems become essential. This is particularly relevant in multivariable optimization, where models optimize functions with many variables and constraints. In such contexts, the human-computer interface assumes paramount importance, as it serves to distill data complexity into a format comprehensible to human users.

Observability refers to the ability to completely define the internal state of the system from the data obtained from it. For Instance, a system that is not completely observable, will always have to deal with uncertainty of the measures, and be very focused on model performance drifts. This for instance is the case when you do not have complete visibility on the market drivers, like competitor reactions.

Controllability refers to the ability to completely manage the state of the system and drive it to the desired state (to have all leverages). Controllability will affect accountability at least partially, if the algorithmic decision is taken by one specific area of the company, but the desired outcome depends from uncontrollable variables of the environment or other decisions made by other areas in the company. Like when the system decides the price but not the quality or the fashion ability of a dress in a retail context.

> *"Price can be very adequate, but if the dress is horrible it will not work well"* (D11, 24:35).

In these contexts, it is regarded as crucial that the oversight function have fluid communication channels with relevant operational areas of the company that might have more exposure to external data or have more capability to both explain and give hints on how to manage specific situations.

Also affecting controllability, is the Variability of external conditions affecting the system under control. This refers to data drifts that separate the production system from training data sets or separate with respect to previous time epochs, because of seasonality, changes in customer preferences or because of data quality issues. Practitioners give

much importance to the time variability and variety of the underlying system or environment that we want to control.

6.3.1. Explainability and Transparency

A significant challenge lies in training human overseers, particularly when AI models exhibit superior performance in specific tasks. Interviewees frequently alluded to the "explainability-risk-effectiveness" dilemma: As problem complexity increases, AI models can become more effective, but explainability diminishes, potentially elevating risk of overreliance in the model. The primary objective is to achieve an optimal level of explainability such that overseers and the incumbent parts of the organization possess sufficient confidence in AI-derived suggestions, because they can understand and make sense of them.

Overseers must acknowledge limitations in both data and models, a critical aspect of responsible AI development and oversight. While explainability holds significant importance, it may not always be a prerequisite. In certain low-risk scenarios, sacrificing explainability for demonstrably superior outcomes may be deemed acceptable. The ethical implications of such a trade-off are significant.

Explainability is often considered a fundamental principle in AI. However, a deeper examination reveals an asymmetry between explainability and the nature of potential errors towards specific stakeholders. For instance, in loan approval processes, the need for explanation is often greater for negative decisions (denials) than positive decisions (approvals) from a customer perspective but maybe it may require explanation to management or other stakeholders who could be affected by false positives.

When models lack explainability, a common tactic is to avoid explicit decision communication and instead guide the process towards additional documentation requests. This approach, while potentially mitigating immediate negative impacts, can introduce opacity into the process and hinder transparency. Consequently, customers may remain unaware of the status of their requests, leading to double harm: a lack of both explainability and transparency.

In brief, both system and model complexity affect the capabilities of effective oversight. To acknowledge the complexity and dynamics of the system and the model and also the limitations of the oversight function to completely sensemaking of the situation is a paramount dimension to consider when designing an appropriate oversight of AI.

6.4. Oversight Skills and Capabilities

Before analyzing this dimension, it is important to stress that oversight function is in all situations assigned to a human (even if the control system is automated through an AI agent making all decisions), mainly because of accountability reasons (see 6.5). In this context of human oversight, practitioners interviewed show a common concern regarding the necessary capabilities for the algorithmic oversight supervisory function. Although not all interviewees coincide on a common set of skills, these capabilities encompass normally a multi-faceted range, including domain knowledge, technical expertise, sound judgment, independence, interpersonal skills (soft skills) and resilience to stress, since individuals interacting with algorithms are susceptible to emotional responses. To ensure well-founded decision-making oversight, training is crucial to equip these individuals with the ability to manage their emotions effectively. A significant challenge lies in training human overseers, particularly in situations where AI models surpass human performance in specific tasks. The design of reports presenting AI-derived suggestions should be meticulously crafted to minimize the inadvertent introduction of biases that could influence the decision-maker. Furthermore, both decision-makers at control level and oversight personnel should undergo training to adeptly detect and mitigate potential biases that may be embedded within the presentation of results or performance dashboards.

Domain knowledge refers not only about knowing the system dynamics through experience with the business function, but also the technical knowledge on the model and awareness of its limitations: data is not always complete or system is not completely observable, some processes are not entirely automated and can go wrong. As a fashion analyst puts it:

> *"There are a lot of variables you can't control. Sure, you have your own numbers, but you don't know what the other guys [competition] are doing, or what the weather (or market) will be like."* (D11, 10:51)

This dual knowledge (data science and business domain) is deemed to be the capability of the future in all positions:

> *"I think that just like Excel is completely integrated into organizations, and all teams are autonomous and do their analyses in Excel, in the future Data Science should be like Excel 2.0. We should aim for that."* (D13, 00:02.)

Sound judgment and independence, warranted by the organization, are crucial for effective oversight, as organizational pressures can bias the oversight activity. Strong interpersonal skills and interdisciplinary knowledge are also valuable for overseers, as they need to communicate and collaborate with various stakeholders (internal or external to the company) throughout the monitoring, investigation and improvement processes.

These abilities can be honed through training in simulated environments, similar to those used in industries like nuclear energy. Past event logs provide valuable real-world data for these simulations. Training is particularly important for managing emotions during stressful situations.

In Decision Support Systems (DSS) with a Human-in-the-Loop (HITL) setting, overseers and operators (decision makers in the control system) can alternate roles, promoting a deeper understanding of each other's responsibilities.

Resilience to stress, derived from the concentration of responsibility, is another essential quality for overseers and HITL operators. Here, a retail chain analyst describes her experience overseeing the pricing model:

> *"[Sales discount period] is fun but a very stressful period. Because it is a moment in which the analyst has a direct effect in the business"* (D15:12)

The oversight function interacts with a complex interplay of factors, including the decisions themselves, the input data, contextual intelligence, model understanding, and the decision-making or recommendation generation process. When planning for algorithm monitoring and results analysis, careful consideration is given to issues like data quality, uncertainty, and model explainability. Perceived reliability of the AI system is influencing over time the capability of the overseer to pay attention to the task. During the initial stages of implementation, oversight is conducted with greater scrutiny and frequency, involving detailed analysis of a larger number of cases. In this phase, *under-reliance* on the model may hinder the decision-making process.

As the model demonstrates consistent and reliable performance, oversight activities can be transitioned to a more aggregated level. At this point, the risk lies in *over-reliance*, where the analyst may fail to adequately scrutinize the model's performance. To ensure fair and unbiased supervision, overseers need to be aware of these emotional responses and take steps to manage them effectively.

As AI and advanced digital models become increasingly integrated into organizational decision-making processes, the need for model and algorithm oversight will grow across all functions. This will require a deeper understanding of AI principles and skills from all employees.

In summary, as reported by interviewees, different and multiple skills and capabilities are needed for an effective human oversight. Building these different skills and capabilities is a paramount dimension for the correct design of AI oversight.

6.5. Organizational Culture and Accountability

Organizational culture influences all aspects of AI implementation in businesses. A fundamental challenge in defining roles and responsibilities for AI applications pertains to the attribution of accountability for decisions influenced or made by these AI applications. As already mentioned, accountability attribution is the main reason why oversight function is assigned to a person in all instances consulted with practitioners. There is a need to assign responsibility to a human on the performance of an AI within a company. If it is not assigned at one specific level, a human will cover this role at a higher instance within the company organization.

Isolating the specific impact of an algorithmic decision, particularly when it co-occurs with other contributing factors, can be a complex task. For instance, a decline in demand might be attributed to external market fluctuations beyond the AI or analyst's control, but it could as well be a direct consequence of an AI-driven pricing decision.

Given this inherent complexity, a purely consequentialist approach to evaluating the oversight function should be avoided, since the environment is not completely controllable.

> *"The airlines sector has an enormous amount of variations that you cannot find in any mathematical model. Who would expect the pandemic? […] it does not have any history to make decisions. So you have to be very on top of the decisions that the system makes"* (D17:08)

Organizational processes and culture significantly influence the oversight function and should be considered in higher-level governance and oversight. A risk-averse organizational culture is likely to produce risk-averse operators and overseers. These individuals should be evaluated with this cultural context in mind. Organizations often have also varying levels of risk tolerance towards AI adoption across different departments. For example, marketing and R&D departments may be more open to experimentation compared to finance or operations. This divergence becomes apparent when speaking with different positions within the same company.

Companies already navigate various dilemmas, such as balancing tactical and strategic goals. Introducing AI adds a new layer of complexity, forcing companies to choose between prioritizing business goals or ethical considerations. This tension is present from the initial design and specification stages to ongoing operations.

Companies therefore need to recognize the inherent limitations of AI. Achieving perfect accuracy, with zero false positives or negatives, is impossible. This reality creates a conflict between the ideal of both business and risk optimization and the practical constraints of AI technology.

Incentives play a crucial role in shaping behavior. Currently, analysts are often evaluated based on business performance, even though a complete evaluation is often challenging. Comparing overseers is only feasible when they work in identical positions and contexts.

The introduction of new overseers can pose risks, particularly if there is high turnover. To mitigate these risks, companies must ensure that new overseers are adequately prepared. Additionally, operational pressures may sometimes conflict with safety considerations.

Attributing to a specific person or team the performance of a system developed and influenced by multiple teams and variables (sometimes external to the company) can be complex, especially in the context of high team turnover and organizational changes. One effective strategy is to involve analysts and line managers in the maintenance of the AI system, including the design of user interfaces, procedures, alerts, manuals, and other documentation that impacts the interaction between overseers and the AI application or oversight tools.

Beyond monitoring and reacting to events, overseers play a crucial role in proactively suggesting improvements to design teams, proposing changes to models, to control structures and automated procedures, and defining strategies for oversight and overall performance criteria. These actions contribute to a positive and productive work environment.

A strong data governance framework, ensuring data quality, availability, and traceability, is essential for effective oversight. This enables overseers to make informed decisions and take appropriate actions.

In summary, the prevailing organizational culture significantly shapes the evaluation of the oversight function and the assignment of accountability. Furthermore, as overseers are integral members of the organization, their performance will inherently reflect the established cultural norms and values.

7. Conclusions and Further Research

The preceding analysis indicates that a systems perspective grounded in cybernetic and control theory principles, widely applied within industrial engineering and operations management, elucidates the role of the oversight function, particularly in relation to the control or decision-making function. This framework is anticipated to facilitate future interdisciplinary discourse across fields such as law, business, engineering, governance, and AI ethics.

We have identified five key dimensions that significantly influence the design and performance of an AI oversight function:

- **Decision significance** of the AI system for the organization and its stakeholders.
- **Oversight and control goals**, often spanning in various dimensions (economic, operational, strategic, legal, ethical)
- **System and model complexity**, affected by time dynamics, concurrency of decisions, data complexity, observability and controllability issues of the complete system and explainability and transparency of the AI model.
- **Oversight skills and capabilities** of overseers like domain knowledge, technical expertise, sound judgment, independence, interpersonal skills and resilience to stress.
- **Organizational culture** that will influence the performance of oversight **and accountability**, which is the main justification for assigning humans to oversight function.

These dimensions impact the design and implementation of oversight in various areas, including the selection of methods, metrics, and tools, as well as the organizational aspects of the oversight function. Ultimately, these dimensions will determine the most suitable resources, human or otherwise, to carry out the oversight tasks.

The development of a robust oversight function for AI systems is already a practical reality within the field of operations management. It is widely used to ensure that automated decision-making systems function as intended. Companies that have integrated AI models into their decision-making processes are currently implementing oversight practices in their daily operations.

Most AI applications and advanced digital technologies are considered low-risk from a regulatory perspective. These routine operational scenarios offer valuable insights for policymakers and standard-setting bodies in their efforts to clearly define the essential role of oversight.

Given the global and systemic importance of oversight and its associated procedures, a standardized approach to its design is recommended. Oversight, whether human, automated or hybrid, is crucial not only for protecting health, safety, and fundamental rights, as required by regulations, but also for ensuring the timely response to performance degradation within an organization, which in turn affects the wellbeing of its stakeholders.

Performance encompasses compliance with operational, business, legal, and ethical requirements and boundaries. It is a multidimensional concept, with each dimension having specific performance indicators and limits. Defining these dimensions for each specific application will require collaboration between industry stakeholders, including both providers and deployers as defined by the AI Act.

Our proposed framework offers a potential foundation for establishing organizational structures that facilitate the effective implementation of a responsible human oversight function, ensuring it fulfils its quality control objectives. Ideally, such specifications should be universally applicable, regardless of whether the oversight agent is human or digital. While each dimension of the proposed framework undoubtedly merits further research, due to the limitations of this exploratory research into low-stakes applications, we posit that the development of a practical guide specifically tailored to AI and data science teams on designing this oversight function would be a highly valuable resource for practitioners across all industries. This would extend beyond industries mandated to adopt such oversight functions due to high-risk applications, offering a valuable tool for a broader range of practitioners.

There is a risk, as with many complex problems, that the conversation about human supervision in AI becomes compartmentalized within academic discourse and practice. Fragmentation between disciplines, focusing solely on technical aspects (human-computer interaction), legal and ethical considerations (potential damage to citizens' rights) or specific management areas (operations, data science, product marketing, finance), could hinder progress in this regard.

A systemic and multidisciplinary approach, such as the cybernetic information governance framework adopted in this article, is essential for a more complete understanding of how to effectively govern advanced digital innovations designed to support decision-making at multiple levels, from individual, through organizational, community and societal levels. This broader perspective can enrich academic discourse and ultimately lead to better outcomes for our society.

Regulatory and standardization efforts related to AI governance, particularly human oversight, should consider integrating regulatory requirements into current company practices, even for those not operating in high-risk applications. This proactive approach can streamline future compliance efforts and reduce costs by minimizing the need for significant changes to existing processes. It involves adding parameters to the oversight function to address legal and ethical implications, and avoids the need for a separate safety oversight function that could potentially conflict with existing business oversight functions.

7.1. Further Research and Development

To further develop this framework, real-world case studies would be valuable. Additionally, to train qualified overseers, the design and development of simulation tools could be beneficial. These simulators could be used to test different oversight strategies in challenging scenarios, highlighting potential limitations of overseers and operators. Particular attention should be paid to situations where intuition might lead to incorrect conclusions. Moreover, improving the user experience (UX/UI) of human oversight interfaces is crucial.

Regulatory sandboxes offer an excellent opportunity to test human oversight concepts for specific applications. Both organizations deploying AI and regulatory bodies can learn from these experiments to better understand the role, capabilities, and limitations of human overseers. By collaborating within regulatory sandboxes, valuable insights can be gained.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

References

Ahn, J., Bae, J., Min, B.J., & Lee, S.J. (2022). Operation validation system to prevent human errors in nuclear power plants. *Nuclear Engineering and Design,* 397, 111949. https://doi.org/10.1016/j.nucengdes.2022.111949

Beer, S. (1995). *Brain of the Firm.* John Wiley & Sons.

Binns, R., & Veale, M. (2021). Is that your final decision? Multi-stage profiling, selective effects, and Article 22 of the GDPR. *International Data Privacy Law,* 11(4), 319-332. https://doi.org/10.1093/idpl/ipab020

Boardman, B., & Fraser, J. (2020). *Introduction to Industrial Engineering.* Mavs Open Press Open Educational Resources. Available at: https://mavmatrix.uta.edu/oer_mavsopenpress/20

Boedeltje, M., & Cornips, J. (2004). Input and output legitimacy in interactive governance. *NIG Annual Work Conference*. Article NIG2-01. Available at: https://repub.eur.nl/pub/1750/

Charmaz, K. (2014). *Constructing Grounded Theory.* SAGE Publications Ltd. Available at: https://uk.sagepub.com/en-gb/eur/constructing-grounded-theory/book255601

Deutsch, K.W. (1963). *The Nerves of Government: Models of Political Communication and Control.* Free Press of Glencoe.

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion,* 99, 101896. https://doi.org/10.1016/j.inffus.2023.101896

Digital Future Society (2022). *Towards a meaningful human oversight of automated decision-making systems.* Digital Future Society. Available at: https://digitalfuturesociety.com/report/towards-a-meaningful-human-oversight-of-automated-decision-making-systems/

Directive (EU) 2022/2555 – NIS 2 (2022). Available at: https://eur-lex.europa.eu/eli/dir/2022/2555

Enqvist, L. (2023). 'Human oversight' in the EU artificial intelligence act: What, when and by whom? *Law, Innovation and Technology,* 15(2), 508-535. https://doi.org/10.1080/17579961.2023.2245683

European Union (2016). *Regulation – 2016/679 – EN – gdpr – EUR-Lex*. Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj

Finantial Conduct Authority (2024). *Market Abuse Surveillance TechSprint*. FCA. Available at: https://www.fca.org.uk/firms/techsprints/market-abuse-surveillance-techsprint

Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F. et al. (2024). Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3), 1097-1179. https://doi.org/10.1162/coli_a_00524

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review,* 45, 105681. https://doi.org/10.1016/j.clsr.2022.105681

Green, B., & Kak, A. (2021). *The false comfort of human oversight as an antidote to A.I. harm*. Available at: https://slate.com/technology/2021/06/human-oversight-artificial-intelligence-laws.html

Henwood, K., & Pidgeon, N. (2003). Grounded theory in psychological research. In *Qualitative research in psychology: Expanding perspectives in methodology and design* (131-155). American Psychological Association. https://doi.org/10.1037/10595-008

Klein, K., & Kozlowski, S. (2012). *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions*. Jossey-Bass.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C.R. (2018). Discrimination in the Age of Algorithms. *Journal of Legal Analysis,* 10, 113-174. https://doi.org/10.1093/jla/laz001

Laux, J. (2023). Institutionalised distrust and human oversight of artificial intelligence: Towards a democratic design of AI governance under the European Union AI Act. *AI & Society,* 39, 2853-2866. https://doi.org/10.1007/s00146-023-01777-z

Leyer, M., & Schneider, S. (2021). Decision augmentation and automation with artificial intelligence: Threat or opportunity for managers? *Business Horizons,* 64(5), 711-724. https://doi.org/10.1016/j.bushor.2021.02.026

Luca, M., Kleinberg, J., & Mullainathan, S. (2016). *Algorithms Need Managers, Too*. Available at: https://hbr.org/2016/01/algorithms-need-managers-too

Mena, S., & Palazzo, G. (2012). Input and Output Legitimacy of Multi-Stakeholder Initiatives. *Business Ethics Quarterly,* 22(3), 527-556. https://doi.org/10.5840/beq201222333

Meta (2024). *Transparency Center*. Available at: https://transparency.meta.com/en-us/

Meta Oversight Board (2024). *Oversight Board | Improving how Meta treats people and communities around the world*. Available at: https://www.oversightboard.com/

Nigenda, D., Karnin, Z., Zafar, M.B., Ramesha, R., Tan, A., Donini, M. et al. (2022). Amazon SageMaker Model Monitor: A System for Real-Time Insights into Deployed Machine Learning Models. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (3671-3681). https://doi.org/10.1145/3534678.3539145

Regulation – 2022/2554 – DORA (2022). Available at: https://eur-lex.europa.eu/eli/reg/2022/2554/oj

Regulation (EU) 2024/1689 – AI Act (2024). Available at: http://data.europa.eu/eli/reg/2024/1689/oj/eng

Schmidt, V.A. (2020). Conceptualizing Legitimacy: Input, Output, and Throughput. In Schmidt, V.A. (Ed.), *Europe's Crisis of Legitimacy: Governing by Rules and Ruling by Numbers in the Eurozone*. Oxford University Press. https://doi.org/10.1093/oso/9780198797050.003.0002

Schmitt, K. (2012). Automations influence on nuclear power plants: A look at three accidents and how automation played a role. *Work,* 41(Supplement 1), 4545-4551. https://doi.org/10.3233/WOR-2012-0035-4545

Schröder, T., & Schulz, M. (2022). Monitoring machine learning models: A categorization of challenges and methods. *Data Science and Management,* 5(3), 105-116. https://doi.org/10.1016/j.dsm.2022.07.004

Schuh, G., & Kramer, L. (2016). Cybernetic Approach for Controlling Technology Management Activities. *Procedia CIRP,* 41, 437-442. https://doi.org/10.1016/j.procir.2015.12.102

Schwaninger, M. (2010). Model based management (MBM): A vital prerequisite for organizational viability. *Kybernetes,* 39(9/10), 1419-1428. https://doi.org/10.1108/03684921011081105

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

Singh, A., & Szajnfarber, Z. (2024). *Preposition Salad: Making Sense of Human-in/on/over-the-Loop Control for AI Systems*. Social Science Research Network. https://doi.org/10.2139/ssrn.4921359

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P. et al. (2024). On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (2495-2507). https://doi.org/10.1145/3630106.3659051

UNESCO (2022). *Recommendation on the Ethics of Artificial Intelligence.* UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137?posInSet=7&queryId=005b467f-7d34-4a7b-be64-515583118655

Various States (2023). *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023*. GOV.UK. Available at: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness* (1-7). https://doi.org/10.1145/3194770.3194776

Wagner, B. (2019). Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems. *Policy & Internet,* 11(1), 104-122. https://doi.org/10.1002/poi3.198

Wahlström, B., & Rollenhagen, C. (2014). Safety management – A multi-level control problem. *Safety Science,* 69, 3-17. https://doi.org/10.1016/j.ssci.2013.06.002

Wiener, N. (1961). *Cybernetics Or Control and Communication in the Animal and the Machine.* MIT Press.

Wikipedia (2022). *Wikipedia: Editorial oversight and control.* Available at: https://en.wikipedia.org/w/index.php?title=Wikipedia:Editorial_oversight_and_control&oldid=1104039418