Effective Human Oversight of AI Systems:

The Interplay of Experience, Information Design, and Intervention Mechanisms

Effective Human Oversight of AI Systems:

The Interplay of Experience, Information Design, and Intervention Mechanisms

by

Diego Viero

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday June 19, 2026 at 15:00.

Student number:	5270480	
Project duration:	September 1, 2025 – June 19, 2026	
Thesis committee:	Prof. dr. ir. U. Gadiraju,	TU Delft, Supervisor
	Dr. T. Draws,	OTTO group, External Supervisor
	Dr. D. Murray-Rust,	TU Delft, Committee Member

Style: TU Delft Report Style, with modifications by Daan Zwaneveld

An electronic version of this thesis is available at
https://repository.tudelft.nl/uuid:e25c1b73-b057-44d1-8048-66ca846da8ca.

TU Delft

"A computer can never be held accountable,
Therefore a computer must never make a management decision."

— IBM internal training, 1979, via Simon Willison's Weblog

Abstract

Human oversight of artificial intelligence systems is increasingly mandated by regulation, yet empirical evidence on which interface design choices actually improve oversight quality remains scarce. This thesis investigates how two modifiable design factors, information signal granularity and intervention option range, affect the effectiveness and perceived workload of human overseers in an AI-assisted fraud detection task. A 3×3 between-subjects experiment was conducted online via Prolific ($N = 144$), in which participants reviewed 30 bank account applications flagged by a Gradient Boosting model under one of nine conditions, crossing three levels of information signal (plain case features, categorical risk level indicator, and continuous model confidence score) with three levels of intervention options (binary decision, decision with delegation, and decision with flagged delegation). Oversight effectiveness was operationalised as a composite score rewarding correct classifications and appropriate delegation decisions, and penalising both misclassifications and over-delegation. Perceived workload was measured using the NASA Task Load Index. Self-reported domain expertise was included as a covariate. Neither main effect reached the pre-registered Bonferroni-corrected significance threshold of $\alpha = 0.008$, and no significant interaction was found on either dependent variable. A marginal effect of information signal granularity on oversight effectiveness was observed ($F(2, 134) = 3.29$, $p = .040$, $\hat{\eta}_p^2 = .047$), with a non-monotonic pattern in which the categorical risk level indicator outperformed both the plain information baseline and the continuous model confidence score. This reversal of the hypothesised ordering suggests that, for non-expert overseers, a well-designed categorical signal may be more actionable than a continuous probability score, as interpreting the latter requires complementary domain knowledge not uniformly present in a general population. The results are treated as preliminary, since the study was underpowered relative to the pre-registered target of $N = 288$, due to the planned expert-screened wave not being completed within the thesis timeline. Theoretical and practical implications for the design of human oversight interfaces are discussed, with particular attention to the relationship between signal granularity and overseer expertise.

Contents

1

Introduction

1.1. Motivation

Artificial intelligence (AI) systems are increasingly being deployed across industrial contexts to support critical organisational decisions. From credit scoring and hiring to medical diagnostics and supply chain management, AI-driven predictions now influence outcomes that directly affect individuals, businesses, and society [1, 2]. This widespread adoption reflects the significant efficiency gains and analytical capabilities that AI can provide. However, unlike traditional deterministic systems, AI models are inherently probabilistic and data-dependent, making their outputs uncertain and potentially erroneous in ways that are not always predictable or transparent [3–6]. As these systems take on increasingly consequential roles, the question of how to maintain meaningful human control over their operations has become both urgent and complex.

One prominent approach to addressing these challenges is *human oversight*, the practice of having natural persons supervise AI systems with the authority to influence their operation or intervene when necessary [1, 7]. Human oversight serves multiple functions: it provides a safeguard against incorrect or unfair outputs, ensures accountability for automated decisions, and maintains alignment between system behaviour and organisational or ethical standards [2]. The concept extends beyond simply placing a human "in the loop"; effective oversight requires that humans possess a genuine understanding of the system's capabilities and limitations, have meaningful authority to intervene, and can reliably detect when intervention is necessary [1]. Given its importance, human oversight has been incorporated as a legal requirement for high-risk AI systems under the European Union's Artificial Intelligence Act (AIA), representing a significant milestone in AI governance [7, 8].

Despite this regulatory recognition and growing practical adoption, the study of human oversight is currently still in its infancy. Existing research has established important conceptual frameworks and identified cognitive challenges, including complacency and suboptimal monitoring behaviour [9], algorithm aversion [10], and the difficulties of maintaining accurate error detection under realistic task conditions [3], but a clear and comprehensive understanding of when oversight is most appropriate, how it should be implemented, and how to evaluate the effectiveness of different approaches in mitigating AI-related risks remains ambiguous [2]. The lack of empirical evidence on how oversight operates under realistic working conditions means that practitioners and regulators must currently act with limited guidance [1, 2].

A further complication is the asymmetric nature of the risks that oversight is designed to manage. When an AI system operates correctly and the human overseer confirms its outputs without incident, the process is invisible and unremarkable. When the system errs and the oversight process fails to catch the mistake, the consequences can be severe and, in high-stakes domains such as healthcare, finance, or public safety, potentially irreversible [1, 7]. This asymmetry creates a structural challenge: the situations in which oversight failures carry the greatest consequences are often those in which the cognitive and contextual demands on the overseer are highest, making reliable performance most difficult to sustain.

The consequences of this asymmetry are made worse by the limited empirical evidence on which factors facilitate or inhibit oversight effectiveness under realistic working conditions [1, 2]. Without a clearer understanding of these factors, neither practitioners nor system designers can make informed decisions about how to structure oversight in practice [1, 11]. Translating the theoretical requirements set out in legislation such as the AIA into concrete design principles therefore requires systematic investigation of how oversight operates under the varied conditions encountered in industrial settings [11].

A core obstacle to effective oversight is that human overseers are not neutral observers. Two factors that are likely to shape how well they perform are the quality of information signals available to them and the range of intervention options at their disposal. Richer information signals support the epistemic access overseers need to evaluate system outputs independently [1], while a broader range of intervention options provides the causal power required to act on that judgement [1, 12]. Without either, overseers remain vulnerable to well-documented psychological threats such as *automation bias* (the tendency to over-rely on algorithmic outputs as a heuristic substitute for independent judgement [3, 9]) and *algorithm aversion* (the tendency to distrust and discard algorithmic advice even when it outperforms human judgement, especially after witnessing a system error [8, 10]). In the absence of adequate support on both fronts, oversight can degrade into what has been termed *rubber-stamping*: confirming or rejecting AI outputs without substantive engagement [8, 13–15], satisfying regulatory requirements on paper while failing to provide the safeguard that oversight is intended to supply. Despite the theoretical plausibility of these relationships, empirical evidence on how information signals and intervention options jointly shape oversight quality remains scarce, and it is precisely this gap that this thesis addresses.

1.2. Research Context

To examine whether information signals and intervention options affect the effectiveness and perceived workload of human oversight, we collaborate with OTTO, a major German online retailer that uses machine learning models to forecast consumer demand. The forecasting team at OTTO already follows a structured oversight process in which data scientists and engineers regularly review model predictions, providing an opportunity to study oversight practices in an authentic organisational context. We conducted semi-structured interviews with members of this team to ground the research in real-world oversight practice. These interactions helped define the scope of this research, highlighting that oversight is highly expertise-dependent and primarily functions as a form of risk management.

To investigate these dynamics in a controlled environment, this research involves an online user study centred on AI-assisted fraud detection. Participants use the Financial Fraud Alert Review (FiFAR) dataset [16], taking on the role of an overseer who reviews bank account applications flagged by a risk model. This scenario was chosen because it reflects the asymmetric risk characteristic of real-world AI oversight: when the system errs and the overseer fails to intervene, the consequences can be severe, making the cost of ineffective oversight particularly important.

The FiFAR dataset is built upon the Bank Account Fraud (BAF) dataset [17], a large-scale collection of synthetic bank account applications that closely mirrors real-world fraud detection workflows. The use of synthetic data allows the study to avoid the privacy constraints associated with real banking records while maintaining the statistical properties and decision-making challenges present in operational settings.

Although this study is situated in the financial domain, the fraud detection scenario is intended as a representative proxy for similar contexts in which humans oversee the decisions of automated systems. By examining how interface design and information signals shape human judgement, we expect the findings to translate to related sectors, from healthcare to demand forecasting, wherever human oversight of AI is required.

1.3. Research Questions & Hypotheses

This thesis addresses current gaps in the literature by examining two categories of factors that influence how effectively humans can oversee AI systems: system-level factors and contextual factors. System-level factors refer to interface design elements, such as the information signals presented to the user and the intervention options available to them. Contextual factors refer to the human and environmental circumstances surrounding the oversight task, such as the overseer's domain expertise and cognitive

workload.

The study focuses specifically on how the quality of information signals and the range of available intervention options affect an overseer's ability to detect model errors, make sound delegation decisions, and manage cognitive workload. These factors were selected for their theoretical foundation in risk management and human-computer interaction, as well as their practical relevance to real-world AI deployment [1, 2]. Domain expertise is carefully controlled throughout the study to isolate the effects of interface design.

Information signals operationalise the *dimension of understandability and explainability* of technical design, which Sterz et al. [1] identify as a key facilitator of epistemic access. Intervention options operationalise the *depth of causal power* available to the overseer: without meaningful action options, even a well-informed overseer cannot translate understanding into effective intervention [1, 12]. Together, these two factors capture the interface-level conditions that the literature identifies as most directly modifiable by system designers, and which have not previously been examined in combination in a controlled experiment with human participants [18, 19].

Domain expertise is included as a covariate rather than a manipulated independent variable. Prior work suggests that overseers with greater domain knowledge are better positioned to detect errors and maintain epistemic access [1, 2, 20]. By controlling for it statistically and recruiting participants with varying levels of expertise via *Prolific*[1], the study isolates the effects of interface design while accounting for the individual variance that expertise introduces.

This research is guided by the following primary research questions and their corresponding hypotheses.

RQ1. To what extent do information signal granularity and intervention option range affect the effectiveness of human oversight of AI, controlling for the domain expertise of the overseers?

- **H1a** Higher information signal granularity leads to higher oversight effectiveness, controlling for domain expertise.
- **H1b** A wider intervention option range leads to higher oversight effectiveness, controlling for domain expertise.
- **H1c** The effect of information signal granularity on oversight effectiveness is moderated by the intervention option range (i.e., there is an interaction effect on oversight effectiveness).

RQ2. To what extent do information signal granularity and intervention option range affect the perceived workload of human overseers of AI, controlling for their domain expertise?

- **H2a** Higher information signal granularity leads to lower perceived workload, controlling for domain expertise.
- **H2b** A wider intervention option range leads to higher perceived workload, controlling for domain expertise.
- **H2c** The effect of information signal granularity on perceived workload is moderated by the intervention option range (i.e., there is an interaction effect on perceived workload).

1.4. Thesis Outline

The remainder of this thesis is structured as follows.

Chapter 2 reviews the theoretical and empirical foundations of the research. It defines human oversight as a concept distinct from human-in-the-loop configurations, surveys the factors identified in the literature as facilitators and inhibitors of oversight effectiveness, and examines the two principal cognitive biases automation bias and algorithm aversion that pose systematic threats to oversight quality.

Chapter 3 presents the industry context established through semi-structured interviews with OTTO data scientists and a business analyst. It reports the three primary themes that emerged from the thematic analysis of the interview data and explains how these findings informed the choice of experimental variables.

[1] `https://prolific.co`

Chapter 4 describes the research design. It presents the 3×3 between-subjects factorial structure, defines the independent variables (information signal level and intervention option level), the dependent variables (oversight effectiveness and perceived workload), and the domain expertise covariate. It also describes the participant recruitment strategy via Prolific, the data analysis plan, and the a priori power analysis that determined the target sample size.

Chapter 5 details the design and implementation of the experiment interface and task. It describes the rationale for a custom-built web interface, the three information signal conditions (plain information, risk level indicator, and model confidence score), the three intervention option conditions (shallow, mid-level, and in-depth), and the scoring and performance-contingent compensation scheme.

Chapter 6 reports the empirical findings. It presents descriptive statistics for the recruited sample, followed by the main effect and interaction effect tests for both research questions, and closes with exploratory analyses of self-reported expertise, NASA-TLX subscales, metacognitive calibration, and delegation rates.

Chapter 7 interprets the findings in relation to the hypotheses and the existing literature, discusses practical implications for the design of oversight interfaces in industrial settings, and identifies directions for future work.

Chapter 8 situates the thesis within the broader field of computer science and human-computer interaction, and reflects on the societal and ethical implications of the research.

Chapter 9 synthesises the main contributions of the thesis, offering a concise answer to each research question and indicating how the findings advance the understanding of effective human oversight of AI systems.

2

Background and Related Literature

This chapter presents the theoretical and conceptual foundations necessary to define effective human oversight of AI systems, the practical challenges linked to it, and the facilitators and inhibitors identified in the literature. The literature discussed in this chapter was selected through a combination of searches in academic databases focusing on human oversight, AI governance, and human-AI interaction, complemented by targeted examination of relevant regulatory documents, particularly the EU AI Act and associated guidance materials. Priority was given to recent publications (2020–2026) that directly address human oversight effectiveness, although earlier foundational work was also included. This chapter aims to provide a comprehensive understanding of the nature of human oversight and inform the empirical analysis presented in subsequent chapters.

2.1. Human Oversight of AI Systems

Human oversight is fundamentally defined as the supervision of a system by at least one natural person, who normally possesses the authority to influence its operations or effects [1]. The concept goes beyond mere human presence, emphasizing meaningful participation that reliably reduces the risks connected with the system's deployment. Such supervision can occur at various stages of the system's life cycle, including intervening during execution, reversing faulty decisions, or adjusting parameters to improve performance. However, the purpose of human oversight goes beyond error mitigation alone; the EU AI Act (AIA) defines human oversight as essential to promote human-centric and trustworthy AI, ensuring a high level of protection for fundamental rights, health, and safety [1, 7, 8]. Human agency and oversight are identified as one of the seven ethical principles for trustworthy AI, ensuring that systems are developed and used as tools that serve people, respect human dignity, and are appropriately controlled and monitored by humans [7].

Gaube et al. [6] extend this definition further, characterising oversight as "deliberate human activity with the goal of sufficiently mitigating risks in the use of an AI system", requiring explicit assignment to designated roles, organisational embedding, and the capacity to both monitor system outputs and intervene when professional judgement or legal provisions require it. This operational framing is useful because it highlights that oversight is not incidental: it demands that roles be explicitly assigned, individuals be trained, and intervention pathways be structurally defined.

Although safety and risk mitigation are essential, human oversight also supports broader technical and business goals, such as ensuring that AI systems consistently meet expectations regarding quality, accuracy, and operational performance [2]. The fundamental goal of unifying these various dimensions is risk mitigation, taking into account both technical risks related to inaccurate outputs and errors, and ethical concerns surrounding unfairness in system outcomes [1]. Crucially, accountability is the key reason humans are responsible for oversight instead of automated safeguards [7, 12].

Despite growing regulatory attention, human oversight remains a relatively young field of study. Clear and comprehensive understanding of when oversight is most appropriate, how it should be implemented operationally, and how different approaches compare in their effectiveness at mitigating AI-related risks

remains elusive [2]. This gap motivates the present research.

The primary objective of human oversight is to mitigate the risks associated with the use of AI systems. These risks range from threats to health, safety, and fundamental rights to issues of discrimination, unfairness, and inaccuracies in outputs [7]. To address them effectively, oversight must go beyond a single safeguard mechanism and instead combine technical, operational, and structural measures. The following objectives capture the key dimensions through which human oversight can be made both effective and reliable.

Risk Mitigation and Safety. The main goal of human oversight in high-risk AI systems is risk mitigation [1, 7]. According to the European AI Act, oversight is intended to prevent or minimise risks to health, safety, or fundamental rights [7]. In practice, this function acts as a safety measure: human overseers can identify and correct inaccurate, biased, or otherwise harmful system outputs before they lead to potential harm [3].

Upholding Human Values and Dignity. Another central aim of oversight is to maintain AI as a human-centric technology that supports, rather than replaces, human agency [7, 8]. Oversight ensures that humans retain a primary role in decisions that affect them, preserving human dignity in the process [7, 14]. Beyond individual considerations, human oversight is seen as a safeguard for societal values, reinforcing democratic principles and the rule of law by guaranteeing that decisions affecting citizens are subject to human discretion, moral reasoning, and ethical judgement [7, 8].

Accountability and Responsibility. Oversight mechanisms are also intended to close the accountability gap created by autonomous systems [8, 12]. By ensuring that humans remain in control of critical decisions, the framework assigns responsibility to individuals rather than machines [7, 12]. Maintaining a human in the loop allows organisations to identify at least one accountable person along the causal chain who can be held responsible for the system's actions [12, 15, 21, 22]. This approach supports justiciability, giving affected individuals the legal basis to challenge decisions and seek amendment when AI outputs produce conflicting effects [8, 14]. Wagner [15] further argues that assigning liability to a human operator does not guarantee meaningful agency: where operators lack the information or authority to genuinely evaluate system outputs, responsibility is assigned without the corresponding control.

Contestability. Finally, a distinct objective of human oversight is to ensure that automated decisions remain contestable [14, 23]. Systems must be designed to allow individuals to request a review or re-evaluation of outputs that unjustly constrain their rights, freedoms, or legitimate interests [7, 14]. By implementing contestability into the system's operational design, human intervention becomes not an ad hoc process but a structural feature, ensuring that individuals can meaningfully challenge and influence AI-driven outcomes throughout the system's life cycle [8, 14, 24, 25].

2.2. Human Oversight vs. Human-in-the-Loop

Human oversight should not be interpreted as simply adding a human to check or approve AI outputs. This kind of approach can easily lead to what is often described as *rubber-stamping*, where human involvement is present in form but not in practice. In these cases, decisions remain effectively automated because the human either lacks the authority, the necessary understanding, or the practical ability to question or change the system's recommendation [8, 13, 14].

Several researchers have pointed out that superficial human involvement does not amount to genuine oversight. Almada [14] argues that when a human merely confirms an automated decision without engaging in interpretation or critical evaluation, their role becomes purely formal, and the decision-making process remains essentially automated. Similarly, Laux [8] highlights the risk that rubber-stamping may be used to create the appearance of human involvement while avoiding regulatory restrictions. Green [13] further observes that policies requiring human oversight, if not carefully defined, may unintentionally encourage organisations to adopt minimal and symbolic forms of human intervention. Wagner [15] introduces the related concept of *quasi-automation* to characterise systems where the human operator's only practical function is to regularise agreement with the machine, retaining formal responsibility while exercising little genuine agency.

Drawing on cybernetic and control theory frameworks, true oversight operates as meta-control oper-

ating over the feedback loop to ensure the adequacy of the entire control system's performance [2, 12]. This distinguishes it fundamentally from *Human-in-the-Loop* (HITL) configurations, where human involvement is embedded within the control system itself as an integral component of each decision cycle, rather than serving a supervisory function over the system's overall operation. This conceptual distinction is critical because it clarifies that oversight is not simply about having humans participate in automated processes, but about establishing a supervisory layer that monitors and evaluates the system's functioning as a whole [2].

Doctorow [26] introduces the concept of the *reverse centaur* to describe situations in which human involvement is structurally present but significantly weakened. In contrast to traditional human–machine cooperation, the reverse centaur refers to configurations where the human primarily functions within a process largely dictated by the system. This is particularly helpful for understanding how improper forms of human oversight can result in what Wagner [15] calls an *accountability sink*: humans retain formal responsibility for decisions, while their practical ability to meaningfully evaluate or influence system output is limited. Responsibility is assigned to the human without a corresponding degree of control or decision-making capacity.

These perspectives highlight an important distinction: meaningful human oversight is not defined by the simple presence of a human, but by the human's real ability to understand, assess, and, when necessary, modify or override AI-driven decisions. Without adequate training, clear guidelines, and interfaces that support proper understanding, human participation can become superficial, undermining the purpose of oversight mechanisms. Cavalcante Siebert et al. [27] formalise this through the concept of *meaningful human control*, requiring both that the system be responsive to relevant human moral reasons and that at least one human has a proper technical understanding of its behaviour conditions that depend on the interface, not only on the human. This distinction is also central to the research design of the present study, which focuses on oversight as a supervisory function rather than as embedded participation in the decision pipeline.

Specifically, the experiment situates participants as external reviewers of individual model-flagged cases, mirroring what Salgado-Criado [2] calls a *supervisory* or *oversight function* rather than a HITL decision-support setting. Participants are not co-decision-makers alongside the model; they evaluate and can override the model's implicit classification, a configuration that maps directly onto the oversight-as-meta-control conception described above.

Laux [8] further distinguishes between *first-degree* and *second-degree* human oversight. First-degree oversight involves direct human influence over individual decision outcomes, typically in real-time; second-degree oversight occurs after decisions have been made, through review, audit, or appeals mechanisms. The experiment in this thesis instantiates first-degree oversight, as participants evaluate individual flagged cases before a final outcome is produced.

2.3. Factors Influencing Oversight Effectiveness

As explained in the previous section, the AIA provides limited guidance on what effectiveness means in practice or how it should be measured and implemented. To address this underspecification, researchers have developed taxonomies and frameworks that quantify effective oversight by defining its key components, requirements, and implementation strategies [1, 2, 6, 28].

The literature organises the factors that shape oversight quality into three broad categories: *technical design features*, meaning properties of the system and its interface; *individual human factors*, meaning traits, states, and competencies of the oversight person; and *environmental and organisational circumstances*, meaning contextual conditions surrounding the oversight task [1]. This taxonomy structures the review below and also motivates the experimental design: the two independent variables (information signals and intervention options) are technical design features, while domain expertise is an individual factor controlled as a covariate.

Regulatory baseline: the EU AI Act. The European Union Artificial Intelligence Act (AIA) represents one of the first comprehensive regulatory frameworks specifically designed to govern the development and use of artificial intelligence systems. The regulation adopts a risk-based approach, introducing obligations that vary depending on the potential impact of AI systems on health, safety, and fundamental

rights. Article 14 stipulates that high-risk AI systems must be designed and developed in a way that allows them to be effectively overseen by natural persons, and makes clear that simply inserting a human into the decision-making process is insufficient [7]. However, despite mandating effectiveness, the regulation provides only limited guidance on what it involves in practice. Obligations such as "understanding the capacities and limitations of the system" or "correctly interpreting outputs" lack clear operational criteria [1, 3, 8, 29]. This ambiguity makes both implementation and evaluation more difficult, and motivates the conceptual frameworks reviewed below . Notably, Laux and Ruschemeier [29] show that the AIA creates an asymmetric division of responsibility: providers are obligated to enable awareness of automation bias, yet the bias itself is driven by individual, motivational, and contextual factors that fall under the deployer's control, a structural gap that further motivates empirical research on interface-level design choices.

Four conditions for an effective overseer. Based on the premise that the central objective of effective human oversight is risk mitigation, Sterz et al. [1] propose a framework specifying four conditions necessary for an oversight person to be effective. These are (1) *Causal Power*, i.e., the oversight person must have the authority and technical capability to influence the system or its effects, for instance by overwriting outputs or triggering a stop mechanism; (2) *Epistemic Access*, i.e., the overseer must have sufficient knowledge of the decision situation, including the system's capacities and limitations and how to correctly interpret its outputs; (3) *Self-control*, i.e., the agent must be capable of purposefully retaining attention and following through on decisions, remaining vigilant against phenomena such as automation bias; and (4) *Fitting intentions*, i.e., the oversight person must have intentions that align with their role, such as a genuine intent to mitigate risks, free from conflicting personal motivations. Together, these four conditions define when oversight can be considered effective. They highlight that effectiveness depends not merely on having a human present, but on the operator being morally responsible and equipped to act in accordance with their role.

Although the framework identifies epistemic access and causal power as necessary conditions for effective oversight, it leaves open the question of how specific interface design choices impact these conditions in practice and whether addressing them jointly produces effects beyond what either achieves alone [1].

Signal Detection Theory applied to error detection. Although the presence of an effective overseer is essential, it is insufficient on its own; a fundamental prerequisite for effective oversight is the reliable detection of errors, defined broadly as inaccuracies or unfairness. Langer et al. [3] proposes a Signal Detection Theory (SDT) approach to investigate the conditions that influence effective error detection. SDT describes how people distinguish a signal from noise [3], and in the context of oversight it can be used to examine how humans discriminate between correct and incorrect AI outputs. An overseer's judgement is shaped by two factors: *sensitivity*, their ability to discriminate between normal and erroneous outputs; and *response bias*, their internal threshold of evidence before judging an output as erroneous. By quantifying these dimensions, SDT provides clear measures to assess whether human oversight is effective in its foundational role of error detection.

The SDT framework is particularly relevant to the present study because the oversight effectiveness score defined in Chapter 5 directly operationalises hit and miss rates: a correct fraud or legitimate classification corresponds to a hit, while an incorrect classification corresponds to a miss (or false alarm). Differences between conditions in the overall score, reflect underlying differences in either sensitivity or response bias, or both.

Variable Autonomy and meaningful control. Methnani et al. [12] define a *Variable Autonomy* (VA) model to ensure meaningful human control in AI-based systems. The core premise is that simple human presence is insufficient and may unnecessarily limit the benefits of automation. The VA model allows systems to adjust their level of autonomy in real time, maximising human control while avoiding operator overload. Central to the framework are three socio-ethical values: *accountability and responsibility*, ensured by explicitly defining the roles of all actors throughout the system's life cycle; and *transparency*, which requires that appropriate information be available to relevant actors at the right level of abstraction and at the right time, enabling them to gain or regain awareness of the system's status and to calibrate trust appropriately. Relevant to the present study, this transparency requirement suggests that providing overseers with model-derived signals, such as a confidence score or risk indicator, may improve their ability to calibrate trust in the system and, by extension, their decision quality. Empirical support for this comes from Zhang et al. [5], who demonstrate in a controlled exper-

iment that showing prediction-specific confidence scores improves case-by-case trust calibration, even when participants must delegate decisions without seeing the model's prediction directly.

An operations management perspective. Salgado-Criado [2] argues that human oversight must be viewed as a governance mechanism rather than merely a safety requirement, covering safety, ethical, compliance, and business goals. Using cybernetic and control theory concepts, the paper positions oversight as meta-control that operates over the feedback loop. Five key dimensions influence the design and performance of an oversight function: *Decision Significance*, covering the impact and reversibility of decisions; *Oversight and Control Goals*, spanning operational, strategic, legal, and ethical objectives; *System and Model Complexity*, driven by data complexity, concurrency of decisions, and model explainability; *Oversight Skills and Capabilities*, including domain knowledge, technical expertise, and resilience to stress; and *Organisational Culture and Accountability*, which determines how responsibility for AI-driven decisions is attributed within the organisation.

Technical Design Features

Under technical design features, Sterz et al. [1] classifies facilitators and inhibitors whose influence on effective oversight is derived from the design of the system and its interface. The following are the most directly relevant to the present study.

Intervention options are the primary technical mechanism through which an overseer exercises causal power. Without options to intervene, control, overwrite, or undo system decisions, an oversight person has no ability to translate their judgement into action, reducing their role to passive observation [1]. The range and granularity of available actions therefore determines the upper bound on what effective oversight can achieve. This is operationalised in the present study as the intervention options independent variable. Methnani et al. [12] reinforce this point, arguing that a system satisfies meaningful human control only when human-initiated intervention is a structurally available option; Cavalcante Siebert et al. [27] similarly identify the ability and authority to act as a necessary property of systems under meaningful human control.

System understandability and interpretability refer to the degree to which the system's inputs, outputs, and reasoning are represented in a form that is appropriate for human understanding [1]. Enhancing understandability aims primarily at improving the overseer's epistemic access. This includes the use of transparent algorithms, output explanations, what-if analyses, and the interpretable representation of input features [1]. The information signals manipulation in this study operationalises this dimension: the risk level indicator and model confidence score each provide a different form of interpretive support alongside the plain information. A broader empirical literature on Explainable AI confirms this principle: providing interpretable outputs consistently improves users' understanding of AI recommendations and, across many domains, their decision accuracy, though effect sizes are moderated by the form of explanation and users' domain expertise [30].

Faas et al. [18] extend these principles through an empirical co-design study with domain experts performing an oversight task. Their findings identify several concrete interface properties that support oversight quality, including feedback mechanisms that make the impact of interventions visible, autonomy over how and when to monitor, and elements that communicate the overseer's role clearly. Notably, participants in that study expressed a desire for interface elements that support *mastery*, understood as the ability to grasp the AI's capabilities and their own contribution, which maps onto the epistemic access condition. The risk level and model confidence score conditions in the present experiment can be understood as lightweight implementations of this mastery-supporting principle.

Preselection of outputs to review is a further technical facilitator: by filtering outputs so that the oversight person only reviews a subset of cases, the volume of work is reduced, potentially maintaining vigilance and epistemic access over the reviewed cases [1]. In the present study, the model-flagging step in the FiFAR pipeline serves this function: participants only review cases already identified as high-risk by the Gradient Boosting model, reducing the total case load to a manageable 30 items.

Technical design can also *inhibit* oversight. Opacity the absence of information about how the system reaches its outputs undermines epistemic access by making it difficult for the overseer to evaluate whether a recommendation is justified [1]. Conversely, *leading assistance* refers to AI-derived signals that frame information in evaluative rather than factual terms, potentially biasing the overseer toward

the system's recommendation before they have formed their own judgement. The interface design in this study deliberately avoids evaluative framing, presenting all AI-derived cues as neutral data points rather than recommendations, in order to isolate the epistemic effect of the signal from any anchoring effect it might otherwise produce.

The joint effect of these technical features is not merely additive. Research on human-AI complementarity shows that combining information asymmetry — what the AI knows that the human does not — with capability asymmetry — what the human can do that the AI cannot — is the primary pathway through which human-AI teams exceed individual performance [19, 31]. The extent to which this complementarity potential is realised depends significantly on the user's expertise level: highly proficient users are better able to identify when AI recommendations should be followed or overridden [20]. Zhang et al. [31] further demonstrate that human-AI teams exhibit complementary performance specifically when members can access information the other lacks and act on it, reinforcing the case for studying the joint effect of both independent variables. These findings directly motivate the interaction hypotheses (H1c and H2c) in this study, which predict that the combined effect of richer information signals and deeper intervention options exceeds the sum of their individual contributions.

Individual Human Factors

Individual factors of oversight persons include traits and states that facilitate or inhibit the conditions for effective oversight [1].

Domain expertise is among the most consistently identified facilitators. Overseers with substantive knowledge of the task domain are better positioned to identify when a system output is plausible or implausible, enhancing their epistemic access and, in some cases, their causal power [1]. This is confirmed by the OTTO interviews discussed in Section 3, which identified expertise as the dominant individual-level predictor of oversight quality. In this study, domain expertise is measured as a covariate and statistically controlled, so that differences between experimental conditions can be attributed to interface design rather than to variation in participants' background knowledge. Inkpen et al. [20] provide empirical support for this: participants with higher domain proficiency integrate AI recommendations more selectively, following the signal when reliable and overriding it when not, while less proficient users show greater susceptibility to automation bias.

Conscientiousness — the tendency to be disciplined, goal-directed, and norm-adherent — is expected to support self-control and fitting intentions in oversight persons [1]. Conscientious overseers are more likely to engage carefully with each case rather than cutting corners, and to follow through on their intent to mitigate risk even when the task requires effort.

Motivation drives the initiation and persistence of goal-directed oversight behaviour [1]. In crowd-worker experiments, motivation is a particular concern because participants' financial incentives may not align with careful task engagement. The performance-based compensation scheme described in Chapter 5 is specifically designed to address this: by tying bonus payment to classification accuracy and appropriate delegation, the study approximates the accountability that professional analysts experience.

Exhaustion and vigilance decrements are individual-level inhibitors. Sustained monitoring of a mostly-reliable system erodes vigilance over time, a phenomenon sometimes called *learned carelessness* [1].

Environmental and Organisational Circumstances

Environmental circumstances lie outside the technical system and the individual oversight person, yet they shape the conditions under which oversight is performed [1].

Adequate job design is a key facilitator. Oversight roles that are motivating, provide meaningful feedback, and offer appropriate levels of autonomy help to maintain the fitting intentions and self-control conditions for effective oversight [1]. Faas et al. [18] similarly identify work meaningfulness — the sense that the oversight task contributes to an important outcome — as a mediating variable between interface design and overseer engagement.

Time pressure is an inhibitor of epistemic access and self-control: when overseers must review cases quickly, the depth of engagement they can sustain with each case is reduced [1]. The present study does not impose a time limit on individual decisions, allowing participants to engage with the case material

at their own pace.

Role conflicts arise when the oversight person also occupies another role whose goals are in tension with effective oversight [1]. In the experiment, participants have a single well-defined role, mitigating this risk. However, the over-delegation penalty is designed to prevent a different form of role misalignment: using delegation as a way to disengage from difficult cases rather than as a genuine acknowledgment of ambiguity.

Accountability the obligation to explain and justify one's conduct can promote attentiveness and epistemic engagement [1, 9, 32]. Skitka et al. [9] empirically demonstrate that making participants accountable for their decision accuracy reduces both errors of omission (failures to respond to system irregularities) and commission (following an incorrect automated directive). The performance-contingent compensation scheme in this study serves as an accountability mechanism: participants know their payment depends on accuracy, even if they do not see a running score during the task.

Independent thinking refers to the practice of forming one's own judgement before consulting the system's output, which has been shown to reduce anchoring effects [1]. The interface presents case features first, with the AI-derived signal (where present) appearing alongside rather than before the case data, reducing but not eliminating the risk of anchoring.

2.4. Cognitive Biases in Oversight

Beyond individual traits such as conscientiousness and expertise, oversight quality is systematically threatened by two well-documented cognitive biases that operate in opposing directions: *automation bias* and *algorithm aversion*. Both biases represent failures of the self-control and epistemic access conditions for effective oversight [1], and both are directly relevant to interpreting the results of an experiment where participants review AI-flagged cases with varying levels of model-derived information.

Automation Bias

Automation bias is a cognitive bias in which people use the output of an automated decision aid as a heuristic replacement for vigilant information seeking and processing [1, 9]. Rather than independently evaluating each case, an overseer subject to automation bias defaults to accepting the system's implicit or explicit recommendation, particularly when the system has a track record of reliability. Sterz et al. [1] note that automation bias becomes especially likely when facing a system that works reliably and with high performance, because in such cases it is rational for people to divert their attention to other tasks. The regulatory significance of this bias has recently been formalised: Article 14(4)(b) of the AIA explicitly names automation bias as a condition providers must enable overseers to be aware of, making it one of the few cognitive biases directly referenced in EU law. However, Laux and Ruschemeier [29] argue that the legal enforceability of this obligation is limited, as demonstrating that bias occurred in a specific oversight decision typically requires multiple data points and is rarely feasible in practice. Over time, this can lead to what Parasuraman and Manzey [33] call "learned carelessness": a gradual erosion of vigilance driven by the absence of negative consequences.

Automation bias manifests in two distinct forms [9]. *Errors of omission* occur when the overseer fails to detect a system irregularity because they did not look for it, relying instead on the system to flag problems. *Errors of commission* occur when the overseer follows an incorrect automated directive despite contradictory information from other, more reliable sources, because they fail to check or discount that information. Skitka et al. [9] demonstrate that both forms can be reduced by introducing social accountability: participants made accountable for their decision accuracy showed significantly fewer omission and commission errors, and engaged in more thorough verification behaviour before confirming a decision. This finding directly motivates the performance-based incentive structure in this study.

In the context of information signals, automation bias is of particular theoretical concern in the mid-level condition (risk indicator). A simple, colour-coded signal that is correct most of the time presents precisely the conditions under which automation bias is most likely to emerge: the signal is salient, intuitive, and generally reliable, making it tempting to follow without engaging with the underlying numerical case data. The rich condition (model confidence) may partially counteract this tendency by presenting a continuous numerical value rather than a categorical indicator, which requires a more active

interpretation [5, 9, 34]. Estévez-Almenzar et al. [35] warn that the risk is not just one of missed benefit: in a crowdsourced face-matching experiment, participants were more susceptible to being misled by an inaccurate machine on difficult tasks than they were helped by an accurate one on easy tasks, and could not reliably distinguish useful from misleading support. Zhang et al. [5] provide evidence consistent with this: displaying model confidence improved case-by-case trust calibration, with participants showing more appropriate reliance in high-confidence cases and greater scepticism in low-confidence ones, an effect that held even when the model's direct prediction was not shown. The quality of the signal matters beyond calibration:Estévez-Almenzar et al. [35] demonstrate in a crowdsourced face-matching experiment that on difficult tasks, a low-accuracy machine can degrade human performance more than a high-accuracy machine can improve it and that participants are often unable to distinguish between the two, rating inaccurate systems as equally useful. This asymmetry suggests that poorly calibrated signals may actively harm oversight quality on hard cases, rather than simply failing to help.

Algorithm Aversion

The opposite failure mode is algorithm aversion: the tendency to distrust and discard algorithmic forecasts even when they demonstrably outperform human judgement, especially after witnessing the algorithm make a mistake [10]. Dietvorst et al. [10] show across five studies that participants who saw a statistical model perform were significantly less likely to rely on it in subsequent decisions, even when they had observed the model outperform a human forecaster. Crucially, people lose confidence in algorithms far more quickly after a single error than they would after an equivalent error by a human judge. This asymmetry means that algorithm aversion can emerge even among overseers who are in principle willing to use AI support. Laux [8] and Dietvorst et al. [10] note that both biases present challenges to oversight reliability as a governance mechanism: untrained overseers tend to be more susceptible to automation bias, while expert overseers may over-rely on their own judgement in domains where they hold strong prior beliefs, producing behaviour consistent with algorithm aversion.

Algorithm aversion is relevant to the present study because participants are reviewing cases flagged by a model that, by design, is imperfect: the FiFAR dataset includes both correctly and incorrectly flagged cases. In Signal Detection Theory terms, algorithm aversion is associated with a more liberal response bias a lower threshold for judging model outputs as erroneous while automation bias is associated with a more conservative one [3]. A well-calibrated overseer would ideally exhibit neither extreme.

The Reverse Centaur and Accountability Sinks

As introduced in Section 2.2, Doctorow [26] uses the concept of the *reverse centaur* to describe configurations in which the human formally oversees an automated system but effectively functions as a passive component of the process. The reverse centaur is the structural outcome of unchecked automation bias at scale: when overseers routinely accept system outputs without engagement, the human's role becomes indistinguishable from automation, except that the human bears formal responsibility for the outcome. This creates an *accountability sink* in which neither the system (which cannot be legally responsible) nor the overseer (who did not genuinely deliberate) is meaningfully accountable. Wagner [15] frames this same dynamic from a legal perspective: when liability is assigned to an operator who lacks genuine agency over the decision, the human is responsible in name but not in practice, satisfying the letter of oversight requirements while undermining their purpose.

Both biases have concrete implications for the scoring approach adopted in this study. An overseer dominated by automation bias will tend to classify cases in the direction of the model's signal regardless of the case data, producing high rates of commission errors on cases where the signal is misleading. An overseer dominated by algorithm aversion will tend to ignore the AI signal and rely solely on the plain information, sacrificing the epistemic benefit that the signal is intended to provide. The scoring system, which rewards only accurate classifications and appropriate delegations, should in principle motivate overseers to integrate both the case data and the AI signal appropriately a calibrated strategy that neither over-weights nor under-weights the model's contribution.

3

Industry Context: OTTO Case Study

The empirical study in this thesis was developed in collaboration with OTTO, a major German online retailer that operates AI-driven demand forecasting models to support purchasing and inventory decisions across its operations. Data scientists and one business analyst at OTTO regularly review model predictions and intervene when outputs appear unreliable or contextually inappropriate, making OTTO's data science department an ideal site for studying human oversight practices in an industrial setting.

To ground the research in practitioner experience and verify important aspects related to human oversight reported in the scientific literature, a series of semi-structured interviews were conducted with four OTTO employees between the 2nd and 10th of December 2025: three data scientists and one business analyst, belonging to a team of 11 people, all with between three and six years of experience at OTTO. The interviews were conducted online via Microsoft Teams. Sessions were recorded solely to make use of the platform's automatic transcription feature; video recordings were deleted immediately after each call, and only the transcripts were retained, stored securely on university servers. Notes were taken manually during each session. The interview guide is provided in Appendix A. The interviews focused on how participants perceived their oversight responsibilities, what information they relied upon when reviewing model outputs, what actions they could take when they judged an output to be problematic, and what factors they considered most important for performing their role effectively. The interview notes were subjected to reflexive thematic analysis [36], following an approach in which themes were developed from the data rather than imposed from a pre-existing framework.

The analysis produced three primary themes that directly informed the research design of this thesis.

Oversight is strongly expertise-dependent. Participants consistently described domain knowledge as the foundation of their ability to evaluate model outputs. Analysts who understood the business logic underlying a forecast (seasonal patterns, promotional effects, supplier lead times) were better positioned to recognise when a model prediction was implausible, even when the numerical deviation from historical averages was not large. This finding motivates the inclusion of domain expertise as a covariate and the recruitment of two parallel participant streams (domain experts and general participants) via Prolific.

Oversight is shaped by what the interface makes visible. Participants described their oversight practice as heavily constrained by the information presented in the monitoring dashboards and decision-support tools. When key contextual signals were absent, analysts reported having to rely on memory or informal knowledge obtained through conversations with colleagues, increasing the cognitive burden of each review. In contrast, interfaces that surfaced model confidence or highlighted individual feature contributions were described as facilitators of a more focused and efficient review. This finding directly motivates the information signals manipulation: by varying whether the interface exposes plain information only, a risk indicator, or a model confidence score, the study operationalises the interface-visibility constraint that practitioners identified as a primary determinant of oversight quality.

Oversight is primarily about risk management. Participants framed their role not as verifying every model output, but as identifying and escalating cases that carried a meaningful risk of consequential error. This risk-management framing aligns with the definition of effective oversight as risk mitigation [1] and supports the delegation mechanism in the intervention options manipulation: a practitioner who identifies a case as genuinely ambiguous does not simply guess; they escalate it for further review. The distinction between appropriate and inappropriate delegation in the scoring system is grounded in this practitioner logic.

Although these findings were gathered in a demand forecasting context, the underlying oversight dynamics they describe – the dependence on expertise, the role of interface visibility, and the risk-management framing – are not domain-specific. They reflect general properties of human oversight of AI systems that transfer across settings where a human must evaluate and act on model outputs under uncertainty. The fraud detection context used in the experiment shares this structure, even if the specific domain knowledge required differs. The findings therefore informed the research design not as domain-specific guidelines, but as evidence for which factors matter most when designing and studying oversight interfaces more broadly.

4

Methodology

4.1. Research Design

This study uses a fully factorial between-subjects experimental design to examine how information signal granularity and intervention option range affect the quality of human oversight of AI systems, in terms of both oversight effectiveness and perceived workload. The two independent variables information signal granularity and intervention option range each have three levels, yielding a 3×3 design with nine conditions in total. Participants are randomly assigned to one of these nine conditions when they begin the study, and the assignment determines both how the data is presented for each case and the set of actions available to them. The conditions are described in detail in Chapter 5; Table 4.1 summarises the full factorial structure.

Table 4.1: *The 3×3 between-subjects factorial design. Each cell represents one experimental condition. Participants are randomly assigned to one cell only.*

	Information Signal Granularity		
Intervention Option Range	Basic	Risk Indicator	Model Confidence Score
Binary	C1	C2	C3
Delegate	C4	C5	C6
Delegate with Flags	C7	C8	C9

The study is conducted online via the Prolific crowdsourcing platform, which provides access to a large and diverse pool of participants and supports screened recruitment for domain-relevant subgroups. Data collection, response logging, and condition assignment are handled by a custom-built web interface connected to a database hosted on university servers. All data is stored and processed in accordance with the university's data management and ethical research guidelines. This study was approved by the Human Research Ethics Committee of TU Delft.

The experimental approach is motivated by a gap between how oversight is typically assessed and what actually matters in practice. Existing compliance checks for AI oversight requirements, including those mandated by Article 14 of the EU AI Act, tend to rely on documentation review rather than direct measurement of human performance [11]. A controlled experiment allows us to make causal claims about which interface configurations actually improve oversight quality, rather than simply which ones satisfy a checklist. This study therefore operationalises oversight effectiveness as a behavioural outcome measured under realistic task conditions, varying systematically across the nine conditions of the factorial design.

The two independent variables were selected because they represent levers that interface designers and system operators can directly control when deploying AI oversight systems. The quality of information available to an overseer affects how well they can distinguish erroneous outputs from correct ones: richer and more diagnostic cues are expected to improve the ability to distinguish fraudulent applications from legitimate ones [3]. The range of intervention options shapes both the causal power available to the

participant and the response threshold induced by the available action set [1]. In particular, the presence of a delegation option makes it possible for a participant to express appropriate uncertainty rather than being forced into a binary decision, which is predicted to reduce costly errors on genuinely ambiguous cases.

This framework directly motivated the choice of experimental variables in the present study. Information signals primarily affect *epistemic access*: richer signals give the overseer more accurate and actionable knowledge about the case. Intervention options primarily affect *causal power*: a broader range of actions allows the overseer to translate their judgement into a more calibrated outcome. The interaction between the two conditions was therefore of particular theoretical interest: an overseer with both high epistemic access and meaningful causal power should be better positioned to exercise effective oversight than one who has only one of these conditions met.

The study also addressed the fitting intentions condition through two design choices. First, participants were recruited as paid crowdworkers and offered additional performance-based compensation tied directly to their oversight score, creating an external incentive to genuinely engage with the task; this aligns with Sterz et al. [1]'s characterisation of motivation as a key driver of fitting intentions. Second, the Prolific screening and pre-task comprehension checks reduced the risk that participants with conflicting interests or insufficient engagement completed the study, further supporting the likelihood that those who continued had intentions aligned with the oversight role.

This study was preregistered on the Open Science Framework (OSF) prior to data collection [37]. The pre-registration specifies the hypotheses, study design, randomisation procedure, sample size rationale, data exclusion criteria, and planned statistical analyses, all of which are reported here as specified at registration.

4.2. Experimental Variables

Independent variables. The two independent variables are *information signal granularity* and *intervention option range*, both manipulated between subjects. Information signal granularity refers to the type and richness of AI-derived decision-support cues presented to the participant alongside the case data, and has three levels: low (base information only), mid (per-feature colour-coded risk indicators), and high (model confidence score) [5]. Intervention option range refers to the set of actions available to the participant when reviewing a case, and also has three levels: shallow (binary Fraud / Legitimate decision only), mid-level (decision or delegate), and in-depth (decision, delegate, or delegate with directional flag). The design and visual implementation of each level are described in Chapter 5.

Dependent variables. The primary dependent variable is *oversight effectiveness*, defined as the correctness of a participant's classification and delegation decisions across the 30 cases they reviewed, computed as a composite score described in full in Section 4.6. Langer et al. [3] frame effective error detection in terms of both discrimination ability and calibrated response to uncertainty: a participant who correctly classifies clear-cut cases and appropriately delegates ambiguous ones demonstrates both components, which is what the oversight effectiveness score was designed to capture.

The secondary dependent variable is *perceived workload*, measured at the end of the task using the NASA Task Load Index (NASA-TLX) [38], a six-subscale instrument capturing subjective mental demand, physical demand, temporal demand, effort, performance, and frustration, each rated on a seven-point scale. It was operationalised as the unweighted mean of the six subscale ratings. The unweighted (Raw TLX) procedure was used in preference to the original weighted scoring, which requires participants to complete pairwise importance comparisons between subscales; this additional step adds task burden without consistent psychometric advantage in online experimental settings [38].

Covariate. *Domain expertise* is included as a covariate [39]. It is measured via a five-item pre-task questionnaire in which participants self-rate their experience in: fraud detection, banking and financial services, accounting and financial reporting, financial data analysis, and AI-assisted decision tools. Responses are collected on a Likert scale and combined into a composite expertise score. The covariate is not manipulated; it is measured and controlled for statistically in the analysis to reduce within-condition variance and improve the precision of the estimated treatment effects.

To make the covariate measurement more robust, recruitment is split across two streams on Prolific: a general stream with no domain filters applied, and an expert stream in which Prolific's built-in screening filters participants for financial sector expertise. This allows the study to capture a meaningful range of expertise levels across the sample.

Domain expertise is expected to predict oversight effectiveness through its influence on epistemic access: overseers with relevant financial domain knowledge are better positioned to evaluate whether a case is plausible or implausible, and therefore to discriminate fraudulent from legitimate cases independently of the AI-derived cue [3, 20].

Exploratory measures. Two additional variables are recorded for exploratory analysis. *User confidence* is collected as a per-case self-rating on a seven-point Likert scale, immediately after each decision. The interface *passively records the time on task* as the duration between the case presentation and decision submission for each case.

4.3. Dataset and Sampling Strategy

The task material for the experiment is drawn from the Financial Fraud Alert Review (FiFAR) dataset [16], which is built on the Bank Account Fraud (BAF) dataset [17]. The BAF dataset consists of approximately one million synthetic bank account applications generated from real-world data distributions. The dataset authors trained a Gradient Boosting model on this data to produce a continuous risk score for each application. Cases exceeding a risk threshold of approximately 0.05 around 30,000 instances were flagged for expert review. Fifty synthetic fraud analysts with varying characteristics then produced predictions for each of these flagged cases, resulting in a dataset that includes not only ground-truth labels and case features, but also a distribution of expert judgements per case.

For the experiment, four features were selected from the available case attributes: *Income*, *Name–Email Similarity*, *Credit Risk Score*, and *Proposed Credit Limit*. These were identified through an exploratory data analysis of the flagged cases as the features with the most consistently measurable distributional difference between fraud-labelled and legitimate-labelled cases, based on IQR and mean comparisons. Presenting only these four features keeps the per-case cognitive load manageable for participants while ensuring the task remains discriminable.

The cases were then filtered and sampled according to two criteria. First, the *degree of expert agreement* was used to classify each case. Expert agreement was quantified as the variance across the 50 synthetic analysts' predictions for that case. Cases with variance below 0.05 were classified as *clear cut* (high agreement, easy to resolve), and cases with variance above 0.15 were classified as *delegation appropriate* (low agreement, genuinely ambiguous). Cases with a variance between 0.05 and 0.15 were excluded from the sample, as their classification was ambiguous under both criteria. Second, the sample was balanced across fraud label and delegation appropriateness: the final set of 30 cases consists of 15 fraud cases and 15 legitimate cases, and within each of these groups, half are clear-cut and half are delegation appropriate, yielding 15 clear-cut and 15 delegation appropriate cases overall.

This sampling strategy serves two purposes. The balanced fraud/legitimate split ensures that a participant who simply classifies every case as fraud or every case as legitimate cannot achieve a high score through response bias. The presence of delegation-appropriate cases ensures that the delegation options available in the mid-level (delegate) and in-depth (delegate with flags) intervention conditions have a meaningful and well-defined correct use, grounded in the expert agreement structure of the dataset.

4.4. Participants and Recruitment

Participants were recruited through Prolific, an online platform designed for academic research that provides access to a large and diverse pool of participants, and supports screened recruitment. Two recruitment groups were used, with English proficiency being required for both. The first group consisted of general participants without any additional domain requirements. The second group was screened for domain expertise by filtering for participants whose Prolific profile indicated experience in finance, using Prolific's built-in "Domain Experts — Finance" screening filter. This two-group approach allowed the study to capture a range of levels of domain expertise, which was necessary for the domain expertise covariate to have meaningful variance across participants.

The target sample size was **288 valid, complete responses** (32 per cell × 9 cells), as determined by an a priori power analysis conducted in Python using `statsmodels.stats.power.FTestAnovaPower`. The analysis targeted $1 - \beta = 0.80$ at a Bonferroni-corrected threshold of $\alpha = 0.008$ (i.e. 0.05/6 planned hypothesis tests), assuming a medium effect size of $f = 0.25$ for both main and interaction effects. The binding constraint was the interaction term, which requires the largest sample due to its higher numerator degrees of freedom; this yielded 32 participants per cell (288 total). If the domain expertise covariate is confirmed at $r \geq 0.20$ at analysis time, ANCOVA reduces this to 29 per cell (261 total). Data collection remained open until the required number of valid, complete responses was reached, regardless of attrition.

Participants were compensated at a rate consistent with Prolific's fair pay guidelines, based on an estimated task completion time of approximately 15 minutes. In addition, a performance-based bonus was available, as described in Section 4.6. Participants were required to have access to a desktop or laptop computer, as the interface was not optimised for mobile devices. Participants who failed the comprehension checks embedded in the tutorial phase were screened out and did not enter the analysis dataset.

Participation was voluntary and participants could withdraw at any time without penalty to their base compensation. Informed consent was obtained at the start of the session before any data were collected. The study did not involve deception: participants were told they were reviewing AI-flagged bank account applications and that their performance affected their bonus payment. No sensitive personal data were collected. Data were stored on university servers in accordance with institutional data management policy.

4.5. Experiment Procedure

The experiment was administered through a custom web interface that handled condition assignment, behavioural logging, and Prolific redirect in a single controlled environment; the design rationale and condition-specific implementations are described in full in Chapter 5. The procedure was divided into three phases: pre-task, oversight-task, and post-task.

Pre-Task. Upon arrival from Prolific, participants are presented with an informed consent form including a description about the study, what data will be collected and details about the compensation. If they decide to consent, they proceed to the domain expertise questionnaire, rating their self-assessed experience across five areas on a Likert scale. They then continue to a tutorial that explains the task context (reviewing AI-flagged bank account applications), describes the case features, and introduces the specific information signal and intervention options they will see, corresponding to their assigned condition. The tutorial includes illustrative examples and ends with a set of comprehension check questions. Participants who do not pass the comprehension checks are excluded from the main task.

Oversight-Task. Participants review 30 bank account application cases in a randomised order determined by a Fisher-Yates shuffle applied at the start of their session. For each case, they are shown the case features and asked to make a decision using the action buttons available in their assigned intervention condition. After selecting a decision, they rate their confidence on a seven-point Likert scale before confirming and moving to the next case. Participants can access the task instructions at any time via a persistent button at the top of the page. The interface passively logs the timestamp and any decision revisions for each case.

Post-Task. After completing all 30 cases, participants are directed to the NASA-TLX questionnaire [38], which collects their subjective workload ratings across the six dimensions of the instrument. Upon completing the questionnaire, participants are automatically redirected to Prolific to confirm task completion and receive their compensation.

The study flow is illustrated in Figure 4.1.

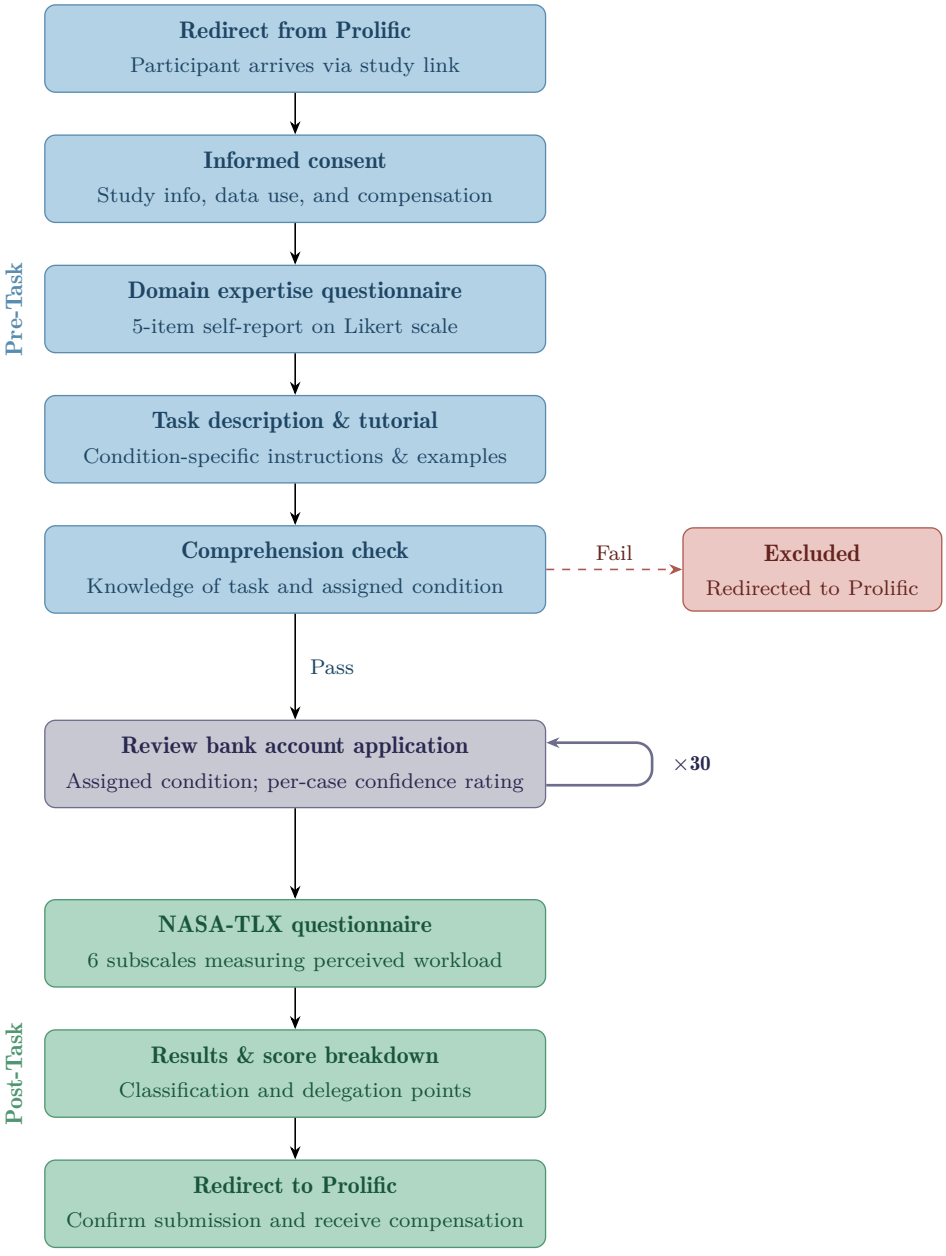

Figure 4.1: *Experiment flow.*

4.6. Scoring and Compensation Scheme

4.6.1. Task Remuneration

Participants receive a performance-contingent bonus on top of their baseline Prolific compensation. The base payment corresponds to an hourly rate of £9.00; given a nominal task duration of 15 minutes, the baseline compensation is £2.25. Participants who score above zero are eligible for a bonus of up to £0.56 (25% of the base payment), awarded proportionally to their score relative to the maximum achievable score of +60 points:

$$\text{bonus} = \frac{\text{score}}{60} \times £0.56 \quad \text{for score} > 0$$

For example, a score of +30 yields a £0.28 bonus, and a score of +15 yields £0.14. Scores at or below zero incur no penalty; those participants receive the standard base payment with no deduction.

This incentive structure was adopted to address a well-known limitation of crowd-worker studies: when payment is independent of performance, participants have little reason to engage carefully with effortful tasks, introducing careless behaviour as a source of noise [40]. The performance-contingent bonus is intended to operationalise, at least partially, the *fitting intentions* condition identified by Sterz et al. [1]: effective oversight requires that the overseer genuinely intends to perform their role diligently, a condition supported by motivation and a felt sense of accountability [1, 9]. The interviews with OTTO analysts also highlighted that motivation and conscientiousness are among the most important individual-level conditions for effective oversight; the compensation scheme is intended to approximate, at least partially, the sense of accountability that professional analysts experience.

Correctness scoring. Each case is scored against the ground-truth label provided in the FiFAR dataset. A decision of *Fraud* or *Legitimate* that matches the ground truth is scored as correct; a mismatching decision is scored as incorrect. In the shallow condition, these two outcomes are the only ones possible. In the mid and in-depth conditions, delegation outcomes are additionally scored as described below.

Delegation scoring. Delegation appropriateness is determined by the variance of the 50 synthetic expert predictions available for each case in the FiFAR dataset. A case is classified as *genuinely ambiguous* if expert-agreement variance exceeds 0.15, and as *clear-cut* if variance falls below 0.05. Cases with variance between these two thresholds were excluded from the 30-case sample during data preparation, so every delegated case in the study is unambiguously either appropriate or inappropriate to delegate. Delegating an ambiguous case (variance > 0.15) is scored as an appropriate delegation; delegating a clear case (variance < 0.05) is scored as an inappropriate delegation. In the in-depth condition, a flagged delegation that correctly indicates the likely decision direction receives additional credit beyond a plain delegation, to reward directional accuracy even when the participant does not commit to a final decision. For example, selecting "Delegate — with Fraud Flag" on a ground-truth fraud case, or "Delegate — with Legitimate Flag" on a ground-truth legitimate case, both count as correctly flagged delegations.

The full scoring scheme is summarised in the tables below. Classification outcomes are scored symmetrically with respect to the ground truth:

Classification:

	Ground-truth	
Decision	Genuine	Fraudulent
Genuine	+2	−2
Fraudulent	−2	+2

Delegation outcomes depend on both appropriateness (determined by expert-agreement variance) and, in the in-depth condition, the direction of any flag provided:

		Appropriate delegation		Improper delegation	
	Decision	Non-Fraud	Fraud	Non-Fraud	Fraud
Delegation:	Delegate	$+1^{\dagger}$		$-1^{\dagger}$	
	Delegate w/ Legitimate Flag	$+1.5^{\dagger}$	$+0.5^{\dagger}$	$-0.5^{\dagger}$	$-1.5^{\dagger}$
	Delegate w/ Fraud Flag	$+0.5^{\dagger}$	$+1.5^{\dagger}$	$-1.5^{\dagger}$	$-0.5^{\dagger}$

† **Over-delegation penalty:** All delegation scores are adjusted as follows:

$$\text{adjusted score} = \text{base score} - 0.5 \times \frac{\#\,\text{delegations}}{\text{total cases}} \times |\text{base score}|$$

This formula shrinks positive scores toward zero and pushes negative scores further from zero, symmetrically penalising over-delegation in both directions. For example, if a participant delegates 15 out of 30 cases (50% delegation rate):

- A positive base score of $+1$ (appropriate delegation) becomes $+1 - 0.5 \times 0.5 \times 1 = +0.75$.
- A negative base score of -1 (improper delegation) becomes $-1 - 0.5 \times 0.5 \times 1 = -1.25$.

And if a participant delegates 24 out of 30 cases (80% delegation rate):

- A positive base score of $+1.5$ becomes $+1.5 - 0.5 \times 0.8 \times 1.5 = +0.90$.
- A negative base score of -1.5 becomes $-1.5 - 0.5 \times 0.8 \times 1.5 = -2.10$.

The theoretical score range across 30 cases is $[-60, +60]$. A simulation study ($N = 100{,}000$ random-strategy trials) confirmed the scoring mechanics and variance heterogeneity across intervention conditions: Binary $\sigma \approx 10.9$; Delegate $\sigma \approx 9.3$; Delegate-with-Flags $\sigma \approx 7.6$.

Transparency and incentive alignment. Participants are informed during the tutorial phase that their compensation includes a performance-based bonus and that both correct classifications and appropriate delegations contribute positively to it, while incorrect decisions reduce it. However, the precise point values and the over-delegation penalty formula are not disclosed. Participants also do not see a running score during the task, so they cannot track their performance case by case. This partial transparency is intended to motivate genuine effort without encouraging participants to game the scoring scheme by optimising for known point values rather than making authentic judgements.

The compensation structure is designed to avoid two specific undesirable incentives. First, it does not reward confidence independently of accuracy: a high-confidence incorrect decision earns the same score as a low-confidence incorrect decision, so there is no benefit to appearing decisive while guessing. Second, delegation cannot be used as a blanket strategy to avoid difficult decisions: inappropriate delegation is penalised, and the over-delegation penalty scales continuously from the first delegated case, reducing positive delegation gains in proportion to the overall delegation rate and thereby diminishing the incentive to defer clear cases. Together, these properties mean that the highest scores are only achievable through genuine engagement with the case material: correctly classifying clear cases, delegating genuinely ambiguous ones, and, in the in-depth condition, accurately specifying the direction of uncertainty on delegated cases.

4.7. Data Analysis

The primary analysis tested the effects of information signal granularity and intervention option range on oversight effectiveness and perceived workload. *Oversight effectiveness* was computed as described in Section 4.6, and served as the unit of analysis for the ANCOVA. Langer et al. [3] frame effective error detection in terms of both discrimination ability and calibrated response to uncertainty: a participant who correctly classifies clear-cut cases and appropriately delegates ambiguous ones demonstrates both components, which is what the oversight effectiveness score was designed to capture.

Perceived workload was operationalised as the unweighted mean of the six NASA-TLX subscale ratings collected at the end of the task [38]. The unweighted (Raw TLX) procedure was used in preference to the original weighted scoring, which requires participants to complete pairwise importance comparisons between subscales; this additional step adds task burden without consistent psychometric advantage in online experimental settings [38]. Individual subscale scores were additionally examined in exploratory analyses to identify which dimensions of workload, if any, showed condition differences; these are reported in Section 6.4.

Given the 3×3 between-subjects design, the primary statistical approach was ANCOVA [39], with information signal granularity and intervention option range as between-subjects factors and self-reported domain expertise (mean-centred) as a continuous covariate. The $r \geq 0.20$ criterion served only to determine which recruitment target applied: if the covariate correlation met that threshold, the reduced target of 261 valid responses (rather than 288) was sufficient to maintain the planned power. Because the observed correlation was $r = -0.16$, below that threshold, the larger target of 288 applies; however, the covariate was retained in all models as pre-registered regardless of the observed correlation, in line with the intention to control for within-condition variance due to differences in self-reported expertise. Sum (deviation) contrasts were used to obtain valid Type III sums of squares. Prior to interpreting results, the homoscedasticity assumption was formally tested using Levene's test [41]. Where heteroscedasticity was confirmed, the ANCOVA model was refit with heteroscedasticity-consistent HC3 robust standard errors [42], which accommodates unequal variances while preserving the ability to test both main effects and interaction terms simultaneously.

For significant main effects or interactions, pairwise post-hoc comparisons were conducted using Games-Howell post-hoc tests [43], which do not assume equal variances or equal sample sizes and are therefore appropriate given the heteroscedasticity confirmed by Levene's test. Effect sizes were reported as Cohen's f [44] throughout to allow direct comparison with the power analysis assumptions.

The two dependent variables — oversight effectiveness and perceived workload — were analysed separately. Exploratory analyses of user confidence and time on task were reported as secondary findings and were not used to draw confirmatory conclusions.

5

Interface and Task Design

5.1. Interface Design Rationale

The experiment interface was designed to serve a dual purpose: to implement the experimental manipulations while minimising sources of confounding that could bias participants' decisions in unintended ways. Bank account fraud detection is a high-risk domain: judging a legitimate applicant incorrectly denies them access to financial services, while missing a fraudulent one carries direct economic harm [17]. Human oversight is explicitly required in such settings by regulators [7], making the interface design choices studied here practically relevant beyond the experimental context.

This section describes the general design philosophy; the specific implementation of each experimental condition is discussed in Sections 5.2 and 5.3.

A core requirement was *condition isolation*. Each participant is randomly assigned to exactly one combination of information signal and intervention option at the moment they begin the study, and that assignment remains fixed for the entire session. The assignment is performed server-side at the time the pre-study questionnaire is completed, preventing any possibility of within-session condition switching. By keeping the interface visually identical across conditions except for the deliberately manipulated elements, unintended visual differences are eliminated as a competing explanation for differences in performance.

The interface follows a minimal, task-focused design. The screen is structured as a vertical sequence of three sections: a case information section at the top, a decision section in the middle, and a confidence rating section at the bottom. The case information displays a table listing each case feature with its numerical value. Depending on the assigned information signal condition, the section additionally shows a colour-coded risk dot next to each feature value (risk indicator condition) or a model confidence score for the case (model confidence condition); in the plain information condition only the raw feature values are shown. A colour legend is displayed at the top of the section whenever the risk indicators are present. The decision section presents the available action buttons for that participant's assigned intervention condition. The confidence section contains the seven-point Likert scale that the participant completes after selecting a decision. A progress bar and case counter are shown persistently at the top of the page, and a "Task Instructions" button is available at all times so participants can revisit the tutorial without leaving the task. The confirm button at the bottom of the page remains disabled until both a decision and a confidence rating have been selected, preventing accidental submissions. Figure 5.1 shows the interface as it appears for the risk indicator information signal and binary intervention condition.

Colour use across the interface is kept neutral so that the coloured risk indicators used in the risk indicator condition stand out clearly against the background and different sections. The decision buttons use a blue/orange pair for the *Legitimate* and *Fraud* decisions respectively, and white for delegation options.

Care was taken to avoid what the oversight literature identifies as "leading assistance" [1]: the interface never frames information in evaluative language (e.g., "the model thinks this is fraudulent") beyond

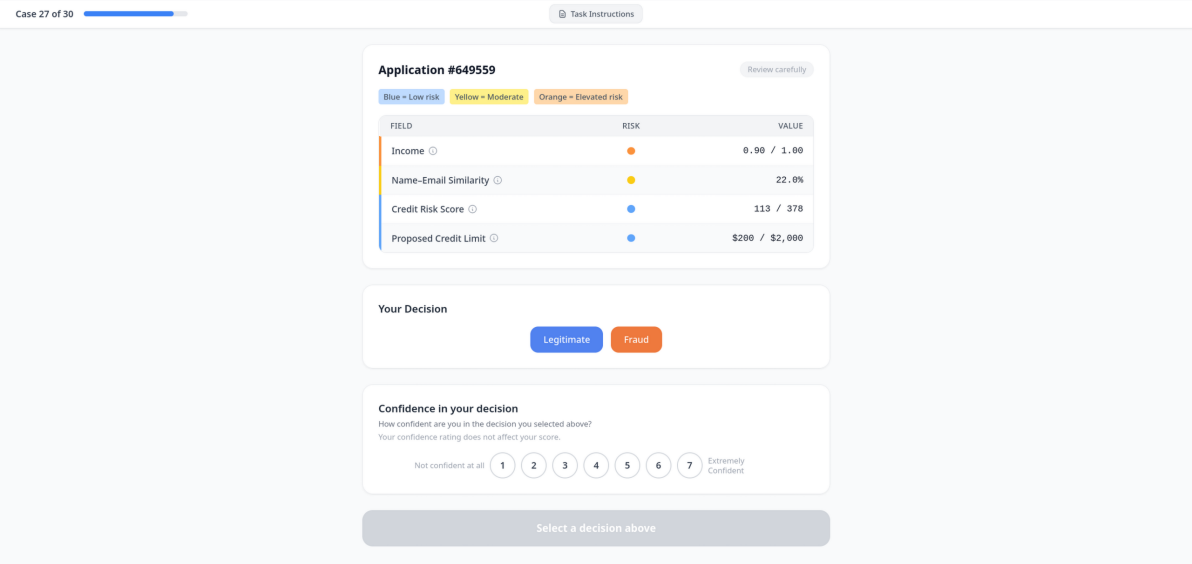

Figure 5.1: *Interface example.*

what the assigned signal condition calls for. AI-derived cues are presented as factual data points rather than recommendations. This is intended to preserve the participant's independent judgement while still providing the information that the condition specifies.

To support participant engagement, the interface presents one case at a time. Cases are presented in a randomised order determined by a Fisher-Yates shuffle applied at session start, so each participant sees the same 30 cases but in a different order. Importantly, the cases shown are exclusively those already flagged by the Gradient Boosting model for human review, so participants act as its check rather than part of its pipeline [6, 8]. This mirrors the single-case review workflow reported by OTTO domain experts.

Behavioural data are logged passively: decision timestamps, decision revision sequences, and time spent on each case are recorded automatically without requiring extra input from the participant. This approach reduces demand characteristics that would arise from explicitly asking participants to report their decision process. All data are stored in a SQLite database hosted on university servers. The interface also provides a password-protected dashboard where the researcher can monitor study progress and download data without interrupting live sessions.

5.2. Information Signal Conditions

The information signal manipulation operationalises the *understandability and explainability* dimension identified as a key technical facilitator of oversight effectiveness [1]. Three conditions were constructed, representing qualitatively distinct ways of communicating decision-support information alongside the plain case features.

Across all three conditions, participants see the same four case-level features derived from the FiFAR dataset: *Income*, *Name–Email Similarity*, *Credit Risk Score*, and *Proposed Credit Limit*. These features were selected through an exploratory data analysis of the approximately 30.000 model-flagged cases in the dataset: they were the features that showed a measurable distributional difference between fraud-labeled and legitimate-labeled cases, assessed via IQR and mean comparisons. This selection ensures that the plain information alone is sufficient for a careful analyst to form a judgement, which is a prerequisite for the plain information condition to serve as a meaningful baseline rather than an impossible task.

Plain Information – No explicit signal. Participants see only the four case features described above, presented in the case table with labeled fields and numerical values. This condition represents the situation in which an analyst reviews a case without any model-generated signal, providing a baseline measure of unaided human oversight performance. It corresponds to settings where model outputs are

not surfaced to reviewers, which remains common in practice despite growing regulatory pressure to provide transparency in automated decision systems [8, 13].

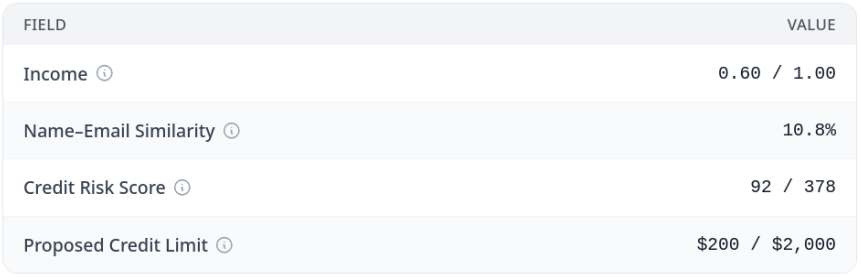

Figure 5.2: *Plain information signal condition.*

Risk Indicator - Categorical Signal In addition to the four case features, each feature row in the case table includes a coloured dot indicating that feature's individual risk level. The colour scheme uses a colour-blind friendly palette: light blue indicates low risk, yellow indicates moderate risk, and orange indicates elevated risk. This palette was chosen over the conventional red/yellow/green traffic-light scheme because red-green colour blindness affects a non-trivial proportion of the population [45], possibly making the standard traffic-light palette unreliable for a share of crowd-worker participants. The light blue/yellow/orange progression maintains a clear perceptual ordering from low to high risk. This palette was selected following the colour-accessible design guidelines outlined by Geissbuehler and Lasser [45], which recommend avoiding red–green contrasts and favouring blue–orange distinctions that remain distinguishable in the most common forms of colour vision deficiency. The risk level for each feature is determined by thresholds derived from a statistical analysis of each feature's distribution in the FiFAR dataset, computed separately for fraudulent and legitimate cases using measures such as the median, interquartile range, and variance. These thresholds reflect the distributional properties of the underlying data. A legend is displayed at the top of the case as a reminder. This condition is of theoretical interest because it represents the type of lightweight assistance most likely to trigger automation bias: a simple, visible signal that is correct most of the time may lead participants to follow it without engaging with the underlying case data [9, 46].

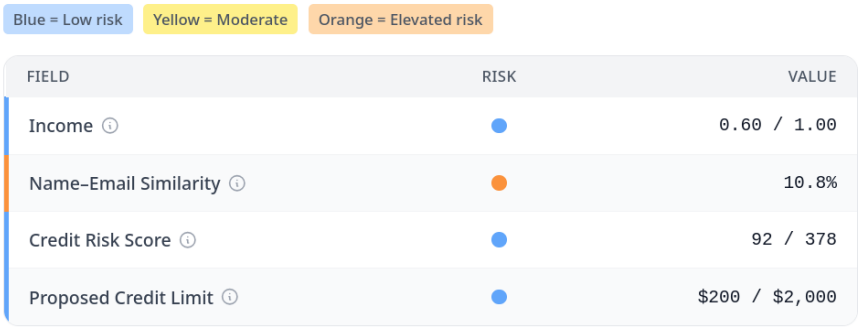

Figure 5.3: *Risk level indicator information signal condition.*

Model Confidence Score – Continuous Signal. In addition to the case features, participants see the Gradient Boosting model's continuous confidence score for the case, expressed as a percentage probability of fraud. The confidence score is presented as a labelled percentage value within the case section, consistent with the factual, non-evaluative presentation style used across all conditions.

The model confidence condition provides finer-grained information than the risk indicator because it exposes the model's raw predicted probability as a continuous value rather than collapsing it into one of three discrete categories. A categorical signal such as the risk indicator necessarily discards within-category variation: two cases assigned the same colour label may carry meaningfully different fraud

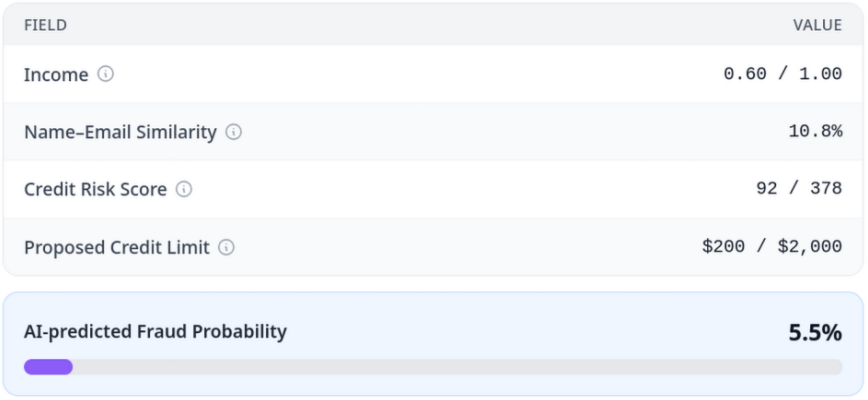

Figure 5.4: *Model confidence score information signal condition.*

probabilities, and that distinction is invisible to the overseer. The continuous score preserves that variation, allowing overseers to distinguish, for instance, a borderline case near a category threshold from one deep inside it. This matters for oversight quality because the ability to act appropriately on a signal depends on how precisely the signal tracks the underlying risk. Zhang et al. [5] showed empirically that displaying a model's predicted probability calibrated participants' trust in proportion to the model's actual confidence level, whereas hiding the score left trust largely flat across cases of varying certainty; moreover, showing the score had a significant positive effect on AI-assisted decision accuracy [5]. The risk indicator, by encoding risk as a three-way categorical label, can only produce a coarser version of this calibration effect. The model confidence score therefore represents higher information signal granularity in the sense that it makes more of the model's internal evidence available to the overseer without additional abstraction.

5.3. Intervention Option Conditions

The intervention option manipulation operationalises the *depth of intervention* dimension, which prior work identifies as a key factor in whether a human overseer can meaningfully affect outcomes or is instead reduced to a passive rubber-stamp role [1, 6]. Three levels were constructed, each providing a different set of actions available to the participant.

Binary – Shallow condition. Participants choose between two options: *Fraud* or *Legitimate*. The shallow condition establishes a minimal-agency baseline: the participant must commit to one of two decisions on every case. It mirrors settings where oversight is implemented as a binary accept/reject scenario, which remains common in automated decision pipelines [8, 13, 14]. Critically, it removes the possibility of routing difficult cases elsewhere, forcing participants to engage even with cases they find ambiguous.

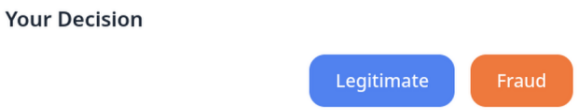

Figure 5.5: *Binary intervention option condition.*

Delegate – Mid-level condition. Participants choose between *Fraud*, *Legitimate*, or *Delegate*. The delegation option represents an explicit acknowledgment of uncertainty: the participant routes the case for further review by a theoretical downstream expert rather than issuing a decision. Delegation is scored as appropriate or inappropriate based on the variance in expert-agreement in the FiFAR dataset (see Section 4.6), so it is not treated as a costless option.

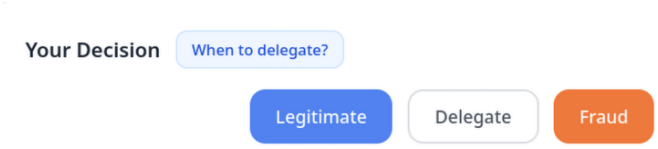

Figure 5.6: *Delegate intervention option condition.*

Delegate with Flags – In-depth condition. Participants choose between *Fraud*, *Legitimate*, *Delegate*, *Delegate — flag as likely Fraud*, or *Delegate — flag as likely Legitimate*. The two flag delegation options allow the participant to communicate a directional judgement to the downstream reviewer while still acknowledging that the case requires further scrutiny. This is the richest set of actions and closely resembles the escalation practices reported by professional fraud analysts. It also introduces the highest decision complexity: participants must decide not only whether to delegate, but also whether they have formed a sufficiently clear view to indicate a direction.

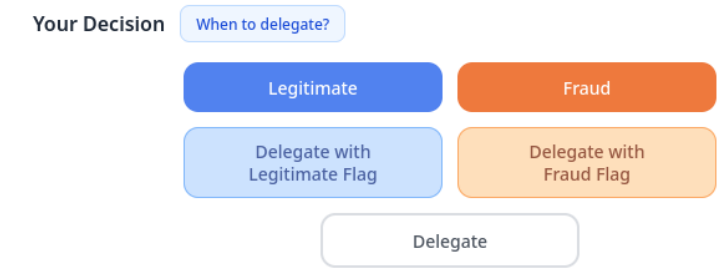

Figure 5.7: *Delegate with flags intervention option condition.*

Over-delegation penalty. Across both mid-level and in-depth conditions, participants who delegate excessively receive a penalty applied to their delegation scores, as described in Section 4.6.1. This penalty was introduced to prevent *trivial delegation*: without it, a participant could reduce task effort by delegating every uncertain case, making the delegation option a way to disengage rather than a meaningful oversight action. Crucially, the penalty scales continuously with the delegation rate rather than applying a binary cutoff, so that each additional delegation incrementally reduces the value of all delegation scores, symmetrically shrinking appropriate delegation rewards and amplifying improper delegation penalties.

Table 5.1: *Available actions per intervention option condition.*

Action	Shallow	Mid-level	In-depth
Fraud	✓	✓	✓
Legitimate	✓	✓	✓
Delegate		✓	✓
Delegate — flag as likely Fraud			✓
Delegate — flag as likely Legitimate			✓
Total actions	2	3	5

5.4. Technical Implementation

The experiment interface was implemented as a full-stack web application, with a Python/FastAPI backend, SQLite database, and a React/Tailwind CSS/Zustand frontend, tested with pytest and deployed on university servers via Apache. Both the backend and frontend were developed with the assistance of Claude Code. The full source code is available in the accompanying repository.[1]

[1] `https://github.com/Diego-Viero/Effective-Human-Oversight-Master-Thesis`

6

Results

6.1. Participant Overview and Descriptive Statistics

Sample and Data Collection

Data collection for the general-population wave took place between the 15th and 20th of May, 2026. This study was designed as a two-wave between-subjects experiment, with a general-population wave and a domain-expert wave to be collected in sequence. Due to time and budget constraints, only the general-population wave could be completed within the thesis timeline. The originally planned recruitment target was 288 valid complete responses across both waves (32 per cell $\times$ 9 cells), or 261 if the ANCOVA covariate correction was confirmed ($r \geq .20$). The data reported in this chapter represent the general-population wave only and should be treated as preliminary.

A total of 150 participants were recruited via Prolific. Four pre-registered exclusion criteria were applied: failure of the comprehension check on two attempts ($n = 1$); fewer than 30 decisions submitted ($n = 5$); an average decision time below 3 s across all 30 cases, indicating random responses ($n = 0$); and identical responses for all 30 cases, indicating extreme satisficing ($n = 0$). After these exclusions the analytical sample comprised $N = 144$ participants.

Participants were randomly assigned to one of nine between-subjects conditions formed by crossing three levels of information signal granularity (plain information, risk indicator, model confidence score) with three levels of intervention options (Binary, Delegate, Delegate-with-Flags). Cell sizes ranged from 14 to 17 participants (see Table 6.1).

Table 6.1: *Cell sizes by experimental condition (non-expert wave, $N = 144$).*

Info. Signal	Binary	Delegate	Delegate-with-Flags
Plain Information	16	17	17
Risk Indicator	16	15	16
Model Confidence	17	14	16

Demographic Characteristics

Demographic data were obtained from Prolific metadata for all approved submissions ($n = 144$). Of these, 142 provided a gender response: 91 identified as male (64.1%) and 51 as female (35.9%); two participants either preferred not to say or had expired data. Age ranged from 19 to 76 years ($M = 30.5$, $SD = 9.0$). Participants resided in 24 different countries; the largest represented countries were Egypt ($n = 48$), South Africa ($n = 35$), and the United Kingdom ($n = 16$). The sample is therefore geographically diverse and skewed towards younger adults, which is typical of the Prolific general-population panel.

Reported Expertise

Domain expertise was assessed via a five-item self-report scale covering (1) fraud detection, (2) banking/financial services, (3) accounting/financial reporting, (4) financial data analysis, and (5) AI-assisted decision tools, each rated on a 7-point Likert scale. The composite score was computed as the unweighted mean of the five items. Because the expert wave was not collected, all participants in the current sample are from the general Prolific population. Their expertise scores therefore reflect self-reported familiarity rather than verified domain experience, and should be interpreted accordingly.

The composite self-reported expertise score ranged from 1.20 to 7.00 ($M = 4.86$, $SD = 1.31$), indicating a moderately high average with substantial individual variation, as illustrated in Figure 6.1. The bivariate correlation between self-reported expertise and oversight effectiveness was $r = -0.16$, $p = 0.057$, falling below the pre-registered threshold $r \geq 0.20$ for applying the ANCOVA covariate correction. The correlation between self-reported expertise and perceived workload was $r = 0.09$, $p = 0.275$. These correlations suggest that, in this sample, reported expertise did not predict performance to the degree assumed during power analysis. The covariate was nevertheless retained in all models as pre-registered.

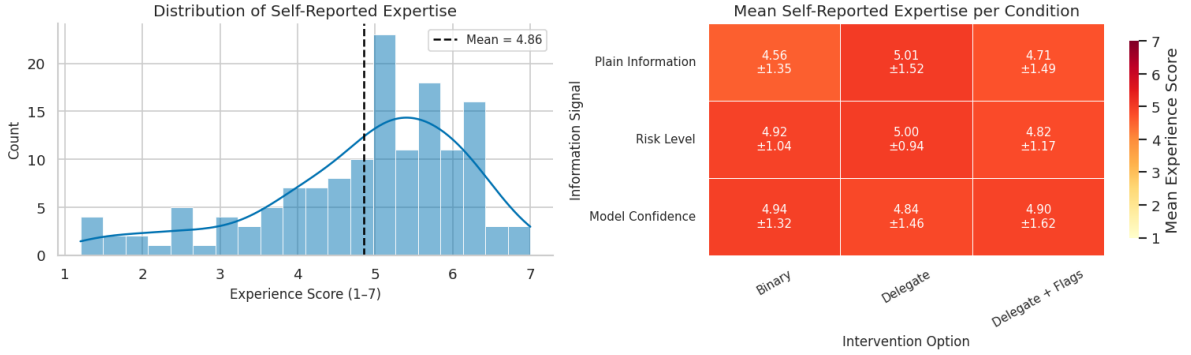

Figure 6.1: *Distribution of composite self-reported domain expertise scores across the full sample (left) and mean expertise per experimental condition (right), with values shown as mean ± SD ($N = 144$). Scores are the unweighted mean of five 7-point Likert items covering fraud detection and financial services domain expertise. The distribution is moderately left-skewed, with most participants clustering in the upper half of the scale; mean expertise was broadly consistent across conditions.*

Descriptive Statistics for Dependent Variables

Table 6.2 reports cell means and standard deviations for oversight effectiveness and perceived workload across the nine experimental conditions.

Table 6.2: *Cell means (and standard deviations) for oversight effectiveness score and perceived workload (NASA-TLX composite) across the nine experimental conditions. Non-expert wave only, $N = 144$.*

Oversight effectiveness

Information Signal	Binary	Delegate	Delegate with Flags
Plain Information	11.50 (14.45)	14.34 (9.87)	11.43 (8.63)
Risk Indicator	16.75 (7.34)	18.60 (7.72)	15.86 (6.41)
Model Confidence	16.94 (12.29)	13.49 (7.67)	12.73 (8.08)

Perceived workload (NASA-TLX)

Information Signal	Binary	Delegate	Delegate with Flags
Plain Information	3.72 (1.06)	3.86 (0.80)	3.60 (0.96)
Risk Indicator	3.22 (1.16)	3.04 (0.96)	3.85 (1.06)
Model Confidence	3.90 (1.21)	3.38 (0.89)	4.09 (0.85)

Oversight effectiveness scores ranged from -24 to $+36$ out of a theoretical range of $[-60, +60]$, with a mean of $M = 14.61$ ($SD = 9.58$) and mild negative skew (-0.43). All conditions produced mean scores above zero, indicating that participants performed better than chance on average. Overall, 93.1% of participants ($n = 134$) scored positively. Only 10 participants scored 0 or below, suggesting that

most understood the task and were able to distinguish cases and make decisions accordingly. The Risk Indicator condition showed the highest marginal mean for oversight effectiveness ($M = 17.04$), followed by Model Confidence ($M = 14.48$) and Plain Information ($M = 12.44$). For intervention options, the Delegate condition showed the highest marginal mean ($M = 15.47$), followed by Binary ($M = 15.10$) and Delegate-with-Flags ($M = 13.30$) (Table 6.2).

NASA-TLX composite scores ranged from 1.17 to 6.33 ($M = 3.64$, $SD = 1.03$), indicating a moderate average workload with a near-symmetric distribution (skew $= -0.07$).

Prior to model estimation, variance homogeneity across the nine condition cells was assessed using Levene's test [41]. Heteroscedasticity was confirmed for the oversight effectiveness model (Levene $W = 2.23$, $p = 0.029$), consistent with the simulation study that predicted different score standard deviations across intervention conditions (Binary $\sigma \approx 10.9$; Delegate $\sigma \approx 9.3$; Delegate-with-Flags $\sigma \approx 7.6$). This model was refitted using heteroscedasticity-consistent HC3 robust standard errors [42]. There was no indication that the perceived workload model was heteroscedastic (Levene $W = 0.54$, $p = 0.826$), so the corresponding hypothesis tests were conducted using the standard ANCOVA as planned. Following the pre-registered exclusion protocol, oversight effectiveness scores were additionally screened for outliers, defined as values more than three standard deviations from the cell mean. No outliers were identified; all 144 participants were retained in the primary analysis.

User Confidence

Participant-reported per-trial confidence (7-point Likert, 1 = not at all confident to 7 = extremely confident) was high overall ($M = 5.52$, $SD = 0.73$), suggesting that participants generally felt certain in their decisions regardless of condition. Marginal means across information signal conditions were similar: Plain Information ($M = 5.55$), Risk Indicator ($M = 5.55$), and Model Confidence ($M = 5.46$). Across intervention options, confidence was marginally higher in the Delegate condition ($M = 5.73$) than in Binary ($M = 5.38$) or Delegate-with-Flags ($M = 5.46$).

Time on Task

Mean decision time per trial was $M = 17.90\,\text{s}$ ($SD = 9.27\,\text{s}$). Decision times exceeding 120 s per trial were winsorized prior to analysis to remove entries likely attributable to participants leaving the browser tab; 29 of 4,320 entries (0.7%) were affected. Marginal means were broadly similar across information signal conditions (Plain Information: $M = 18.21\,\text{s}$; Risk Indicator: $M = 18.95\,\text{s}$; Model Confidence: $M = 16.52\,\text{s}$) and across intervention options (Binary: $M = 17.22\,\text{s}$; Delegate: $M = 18.41\,\text{s}$; Delegate-with-Flags: $M = 18.11\,\text{s}$).

Decision Revisions

One hundred of 144 participants (69%) changed at least one decision before confirming it. Among those who revised at least one decision, the median total number of revisions across all 30 trials was 3 (range 1–46). The relatively high proportion of participants who revised at least one decision, combined with the low median revision count, suggests that most participants engaged thoughtfully with individual cases rather than adopting a fixed response strategy, but that reconsideration was selective rather than systematic. The low variance in revision counts precluded a condition-level comparison.

6.2. Main Effects

All models used a 3×3 ANCOVA with self-reported expertise (mean-centred) as a covariate. Each effect (information signal granularity, intervention option range, their interaction, and the covariate) was evaluated after accounting for all other effects simultaneously, so that no single factor absorbs variance that belongs to another. Sum (deviation) contrasts were applied to ensure this decomposition is computed correctly in unbalanced designs, where cell sizes are not perfectly equal [39]. The pre-registered significance threshold is $\alpha = 0.008$ (Bonferroni-corrected across six hypotheses).

6.2.1. Effect of Information Signals on Oversight Effectiveness (H1a)

The main effect of information signal granularity on oversight effectiveness was $F(2, 134) = 3.29$, $p = 0.040$, $\hat{\eta}_p^2 = 0.047$ (HC3 robust standard errors). This effect does not meet the pre-registered threshold of $\alpha = 0.008$ and H1a is therefore not supported. Figure 6.2 displays the marginal means with 95%

confidence intervals for each information signal condition.

The observed pattern of marginal means was non-monotonic: the Risk Indicator condition produced the highest mean oversight effectiveness ($M = 17.04$), followed by Model Confidence ($M = 14.48$) and Plain Information ($M = 12.44$). This ordering does not match the predicted direction (Plain Information < Risk Indicator < Model Confidence), in that the categorical risk indicator outperformed the continuous probability score. One possible interpretation is that crowdworkers without verified domain expertise are better positioned to act on a categorical signal than a continuous probability, since the latter requires complementary domain knowledge to be used effectively, discussed further in 7.2.1. This is consistent with evidence that the benefit of AI confidence information depends on complementary domain knowledge [5]. However, given that we did not find a significant difference between conditions at this stage of data collection (i.e., our sample represents approximately half the intended recruitment target; see Section 6.1), this interpretation should be treated as tentative.

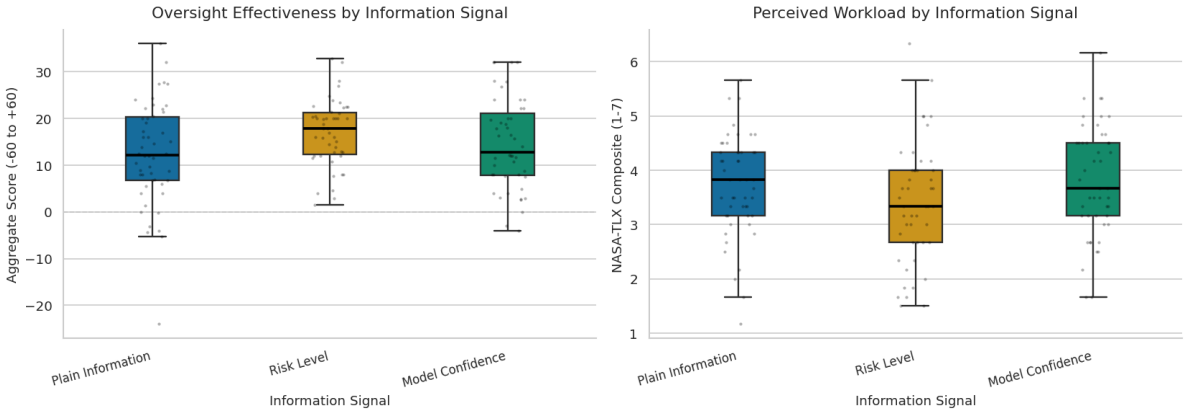

Figure 6.2: *Marginal mean oversight effectiveness scores (left) and NASA-TLX perceived workload composites (right) by information signal condition. The risk indicator condition yielded the highest effectiveness and the lowest workload, with no monotonic trend across conditions in either outcome.*

The association between the self-reported expertise covariate and oversight effectiveness is reported in Section 6.4.

6.2.2. Effect of Intervention Options on Oversight Effectiveness (H1b)

We did not find a main effect of intervention options on oversight effectiveness ($F(2, 134) = 1.02$, $p = 0.364$, $\hat{\eta}_p^2 = 0.015$; HC3 robust standard errors). Marginal means by intervention condition are displayed in Figure 6.3.

The marginal means did not follow the predicted ordering (Binary < Delegate < Delegate-with-Flags). The Delegate condition produced the highest mean oversight effectiveness ($M = 15.47$), followed by Binary ($M = 15.10$) and Delegate-with-Flags ($M = 13.30$). The reversed position of Delegate-with-Flags could reflect the additional cognitive demands of flag selection, or the operation of the over-delegation penalty in the scoring scheme, which scales continuously with delegation rate, reducing the benefit of even appropriate delegations proportionally to the fraction of cases delegated. At the current sample size, however, the difference between conditions is small relative to within-cell variance, and no firm conclusions can be drawn.

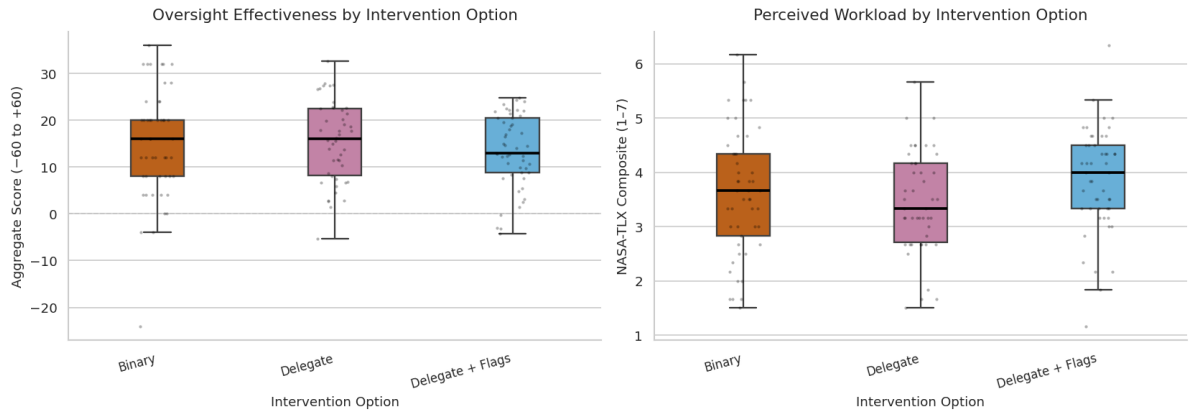

Figure 6.3: *Marginal mean oversight effectiveness scores (left) and NASA-TLX perceived workload composites (right) by intervention option condition. No significant differences were found between conditions; the delegate-with-flags condition produced the lowest median effectiveness and the highest median workload, contrary to the predicted ordering.*

6.2.3. Effect of Information Signals on Perceived Workload (H2a)

The main effect of information signal granularity on perceived workload was $F(2, 134) = 2.45$, $p = 0.090$, $\eta_p^2 = 0.035$ (standard OLS). This effect does not meet the pre-registered threshold and H2a is not supported. The marginal means showed a non-monotonic pattern: Plain Information and Model Confidence conditions reported similar workload levels ($M = 3.73$ and $M = 3.81$ respectively), while the Risk Indicator condition reported the lowest workload ($M = 3.38$).

6.2.4. Effect of Intervention Options on Perceived Workload (H2b)

The main effect of intervention options on perceived workload was $F(2, 134) = 2.18$, $p = 0.117$, $\eta_p^2 = 0.032$ (standard OLS). This effect does not reach statistical significance and H2b is not supported.

The means partially followed the predicted ordering, with Delegate-with-Flags showing the highest workload ($M = 3.84$), consistent with the hypothesis that a broader option set increases cognitive demand. However, the Delegate condition showed lower workload than Binary ($M = 3.45$ vs. $M = 3.62$), suggesting that the option to delegate may have functioned as a cognitive offload mechanism for some participants, reducing perceived effort relative to a forced binary choice. However, as no overall effect was found at this stage of data collection, this interpretation cannot be confirmed.

Exploratory inspection of NASA-TLX subscales is reported in Section 6.4.

6.3. Interaction Effects

6.3.1. Interaction Effect on Oversight Effectiveness (H1c)

The interaction between information signal granularity and intervention options on oversight effectiveness was $F(4, 134) = 0.47$, $p = 0.756$, $\hat{\eta}_p^2 = 0.014$ (HC3). This term is not statistically significant and no synergistic effect on oversight effectiveness was found. H1c is not supported. The cell means underlying this test are visualised in Figure 6.4.

Inspection of the cell means (Table 6.2) does not suggest any coherent interaction pattern. The Risk Indicator condition showed relatively consistent effectiveness across intervention options ($M = 16.75$, 18.60, 15.86 for Binary, Delegate, and Delegate-with-Flags respectively), while the Model Confidence condition showed a notable drop from Binary ($M = 16.94$) to both delegation conditions ($M = 13.49$ and $M = 12.73$). This pattern, if it were to persist in a larger sample, could suggest that continuous probability information supports a binary forced-choice strategy but is less helpful when delegation is available, possibly because participants in those conditions use delegation as a way to avoid committing to a probabilistic judgement. This remains speculative at the current sample size.

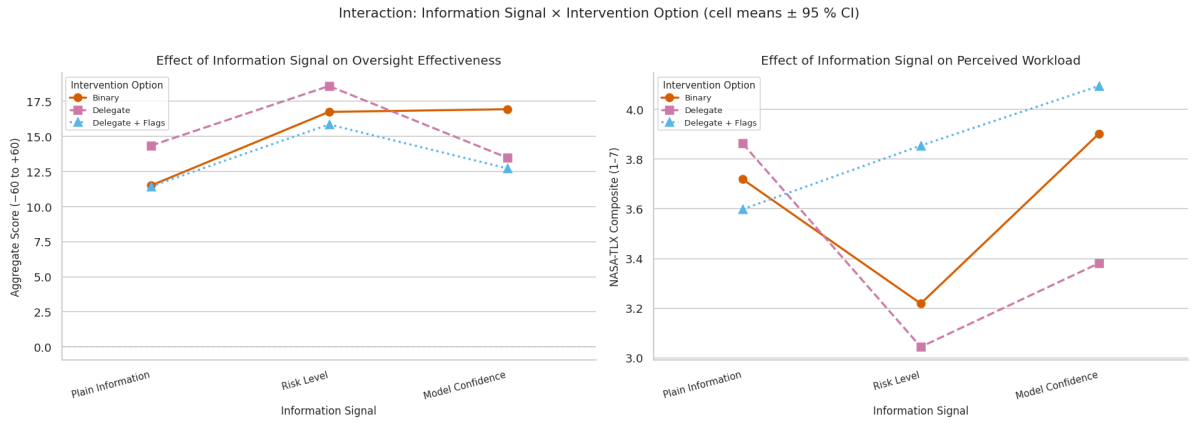

Figure 6.4: *Two-way interaction plot of information signal (plain information, risk level, model confidence score) and intervention option (binary, delegate, delegate-with-flags) on mean oversight effectiveness (left) and mean NASA-TLX perceived workload (right). Error bars show 95% confidence intervals. Neither interaction term was statistically significant for effectiveness ($F(4, 134) = 0.47$, $p = .756$) or workload.*

6.3.2. Interaction Effect on Perceived Workload (H2c)

The interaction between information signal granularity and intervention options on perceived workload was $F(4, 134) = 1.59$, $p = 0.181$, $\eta_p^2 = 0.045$ (standard OLS). This term does not reach statistical significance and H2c is not supported.

The directional check for H2c examined whether the Model Confidence + Delegate-with-Flags cell produced lower workload than the additive prediction from each factor's marginal effect. The additive prediction was $M = 4.01$ and the observed cell mean was $M = 4.09$, a residual of $+0.08$, indicating slightly *higher* workload than the additive prediction rather than lower. The hypothesis was therefore not supported in either direction.

6.4. Exploratory Findings

6.4.1. Self-Reported Expertise and Oversight Effectiveness

The self-reported expertise covariate was associated with oversight effectiveness in a negative direction ($b = -1.31$, $p = 0.031$), which is noteworthy given that the bivariate correlation was $r = -0.16$. Controlling for condition, participants who rated themselves as more experienced tended to score slightly lower, which is counterintuitive and could reflect overconfidence or miscalibration effects in a population of non-verified crowdworkers. This relationship is visualised in Figure 6.5.

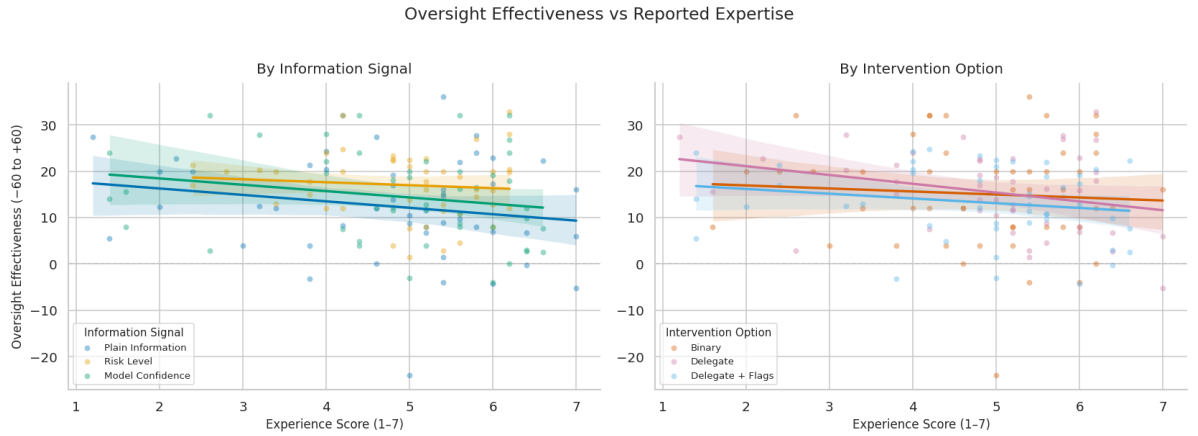

Figure 6.5: *Scatterplot of composite self-reported expertise against oversight effectiveness score ($N = 144$), with linear regression lines fitted separately by information signal condition (left) and intervention option condition (right). Shaded bands show 95% confidence intervals. The negative slope ($b = -1.31$, $p = .031$) indicates that, controlling for condition, higher self-reported expertise was weakly associated with lower effectiveness in this non-expert sample.*

6.4.2. NASA-TLX Subscales

Exploratory inspection of the six NASA-TLX subscales under the main effect of intervention options revealed that the Physical Demand subscale showed the smallest p-value ($F(2, 134) = 5.37$, $p = 0.006$, $\hat{\eta}_p^2 = 0.074$). Given that the task was purely cognitive, this might indicate that participants in conditions with more limited intervention options experienced some form of physical restlessness or tension alongside the cognitive load of the task, though this interpretation is speculative in the absence of a pre-registered hypothesis. For the interaction between information signal granularity and intervention options, the Effort subscale showed the smallest p-value ($F(4, 134) = 2.61$, $p = 0.038$, $\hat{\eta}_p^2 = 0.072$), which may suggest that the combination of richer information signals with more constrained intervention options placed disproportionate demands on participants' effort relative to other combinations, though the small effect size and absence of a correction threshold caution against strong conclusions. Neither result was pre-registered; no correction threshold applies and both should be treated as preliminary observations rather than confirmatory findings.

6.4.3. Metacognitive Calibration

To assess whether participants' self-reported confidence tracked their decision accuracy, the association between per-trial confidence ratings and binary trial-level accuracy was examined at two levels. At the population level, a pooled point-biserial correlation across all valid trials ($N = 4{,}320$) yielded $r = +0.08$, $p < .001$, and a logistic regression with mixed-effects with a random intercept per participant confirmed a small positive association ($\beta = +0.15$, $SE = 0.006$, $z = 25.95$, $p < .001$; odds ratio per unit increase in confidence: 1.16). Figure 6.6 displays the proportion of correct decisions at each confidence rating level.

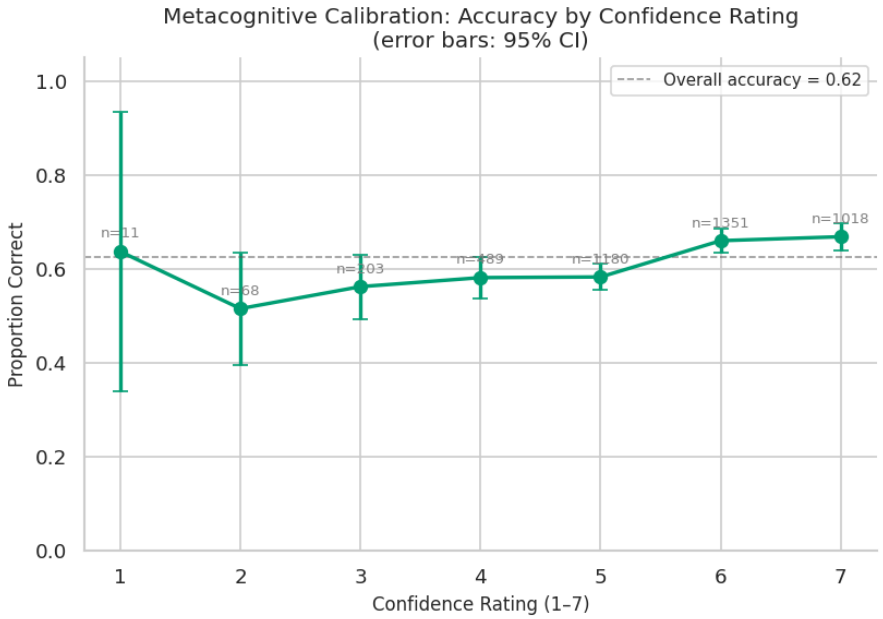

Figure 6.6: *Proportion of correct decisions at each confidence rating level (1–7), with 95% confidence intervals. Trial counts per rating level are annotated. The dashed line indicates overall accuracy (= 0.62). The pattern shows a weak positive association between confidence and accuracy, most pronounced at the upper end of the scale.*

At the individual level, Spearman correlations between confidence and accuracy were computed per participant ($N = 143$; one participant excluded due to zero variance in confidence ratings across all 30 trials). The distribution of per-participant correlations was approximately normal, with a mean of $\bar{r} = +0.11$ ($SD = 0.17$, median $= +0.10$, IQR $= [+0.01, +0.21]$). A one-sample t-test indicates a potential difference from zero ($t(142) = 7.91$, $p < .001$), as illustrated in Figure 6.7.

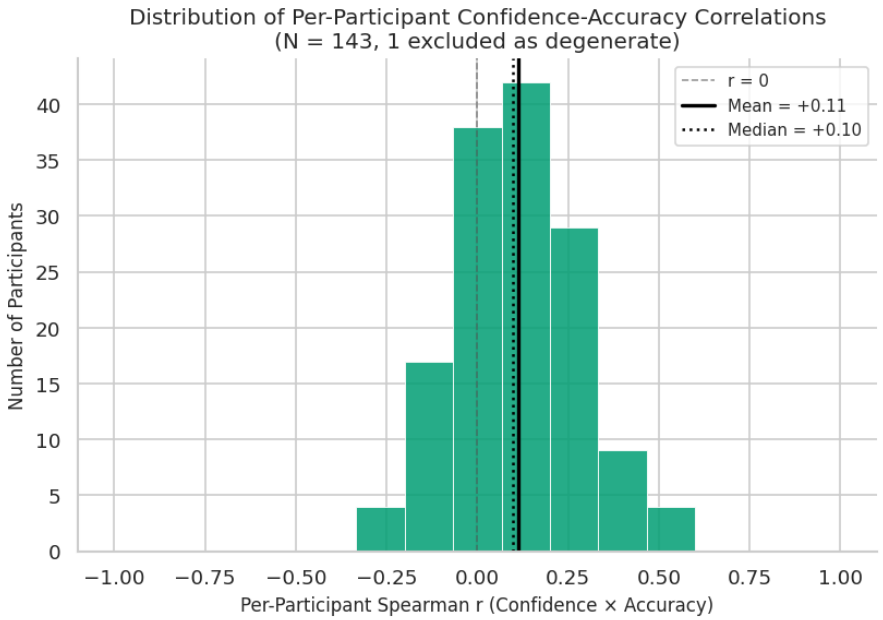

Figure 6.7: *Distribution of per-participant Spearman correlations between confidence rating and trial-level accuracy ($N = 143$; one participant excluded as degenerate). The solid vertical line indicates the mean ($\bar{r} = +0.11$); the dotted line indicates the median ($= +0.10$). The distribution is approximately normal and shifted positively from zero, consistent with weak but reliable metacognitive calibration across participants.*

6.4.4. Delegation Rates

Figure 6.8 displays delegation rates across the six conditions in which delegation was available (Delegate and Delegate-with-Flags), broken down by information signal. Table 6.3 reports the corresponding means and standard deviations.

Table 6.3: *Mean (and standard deviation) delegation rates per condition, expressed as a proportion of 30 cases. Delegate-with-Flags rates are further broken down by subtype in Table 6.4.*

Information Signal	Delegate	Delegate-with-Flags
Plain Information	0.22 (0.14)	0.33 (0.17)
Risk Indicator	0.25 (0.14)	0.35 (0.21)
Model Confidence	0.23 (0.13)	0.37 (0.15)

Overall delegation rates were moderate in the Delegate condition and somewhat higher in the Delegate-with-Flags condition across all three information signal groups (Table 6.3). In the Delegate condition, mean rates ranged from $M = 0.22$ ($SD = 0.14$) under plain information to $M = 0.25$ ($SD = 0.14$) under the risk indicator. In the Delegate-with-Flags condition, rates were consistently higher, ranging from $M = 0.33$ ($SD = 0.17$) under plain information to $M = 0.37$ ($SD = 0.15$) under model confidence.

The higher delegation rates in the Delegate-with-Flags condition are relevant to interpreting its lower mean oversight effectiveness score relative to the Delegate condition. Because the over-delegation penalty scales continuously with delegation rate, reducing the benefit of even appropriate delegations in proportion to the fraction of cases delegated, the systematically higher rates in the Delegate-with-Flags condition would have compressed delegation scores in that group, partially accounting for its lower mean effectiveness ($M = 13.30$ vs. $M = 15.47$; see Table 6.2). Importantly, this compression was not driven by a small number of heavy delegators: among the $N = 49$ participants in the Delegate-with-Flags condition, 73.5% delegated more than 25% of cases, and only 7 participants (14.3%) delegated more than half the cases, so the penalty applied broadly across the condition rather than being absorbed by outliers.

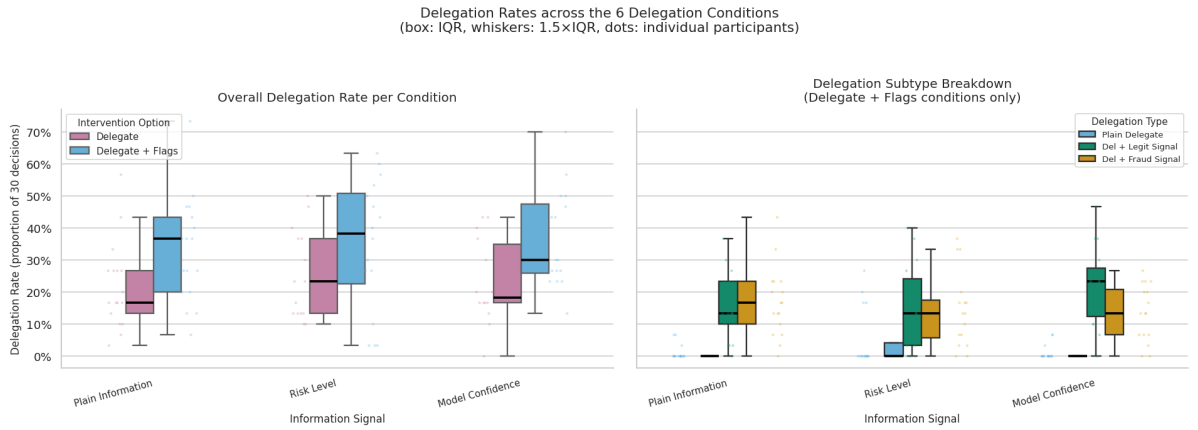

Figure 6.8: *Delegation rates across the six delegation conditions (box: IQR, whiskers:* 1.5× *IQR, dots: individual participants). Overall delegation rate by information signal condition (left) and delegation subtype breakdown for the delegate-with-flags conditions only (right), showing the proportion of plain delegations, delegations accompanied by a legitimate-transaction flag, and delegations accompanied by a fraud flag.*

Table 6.4 breaks down delegation behaviour within the Delegate-with-Flags condition by subtype. Plain delegations were rare across all signal conditions ($M = 0.01$ under plain information and model confidence; $M = 0.05$ under risk indicator). The large majority of delegations were flagged: delegations with a legitimate signal ranged from $M = 0.15$ to $M = 0.22$, and delegations with a fraud signal from $M = 0.14$ to $M = 0.17$. This pattern indicates that when participants chose to delegate, they generally also formed a directional view, which is consistent with engaged rather than purely avoidant use of the delegation option.

Table 6.4: *Mean delegation subtype rates within the Delegate-with-Flags condition, expressed as a proportion of 30 cases.*

Information Signal	Plain Delegate	Delegate + Legitimate Flag	Delegate + Fraud Flag
Plain Information	0.01	0.15	0.17
Risk Indicator	0.05	0.15	0.14
Model Confidence	0.01	0.22	0.14

One motivation for including delegation conditions was the hypothesis that the option to defer a case might function as a cognitive offload mechanism, reducing perceived workload relative to a forced binary choice. The Delegate condition did show a lower mean NASA-TLX score than the Binary condition ($M = 3.45$ vs. $M = 3.62$; see Table 6.2), consistent with this interpretation. However, the main effect of intervention options on perceived workload was not significant (Section 6.2.4), so this directional pattern cannot be confirmed at the current sample size.

To assess whether delegation functioned as a cognitive offload mechanism, Spearman correlations between per-participant delegation rate and NASA-TLX composite (and the Mental Demand and Effort subscales) were computed separately within the Delegate and Delegate-with-Flags conditions. No meaningful association was found in either condition; all correlations were near zero and none reached significance. The cognitive offload hypothesis is therefore not supported by this analysis.

7

Discussion

7.1. Summary of Findings

This study investigated how two interface design factors, the granularity of *information signals* and the range of *intervention options*, shape the quality of human oversight of an AI-assisted fraud detection system. Oversight quality was operationalised through two dependent variables: *oversight effectiveness* (correctness of fraud classification and appropriateness of delegation decisions, scored across 30 cases) and *perceived workload* (assessed via the NASA Task Load Index [38]). A 3×3 between-subjects ANCOVA was used, with self-reported domain expertise as a covariate, and a Bonferroni-corrected significance threshold of $\alpha = 0.008$ across six pre-registered hypotheses. All analyses are based on the general-population wave only ($N = 144$), as the planned expert-screened wave could not be completed within the study timeline.

RQ1: Effects on oversight effectiveness. Information signal granularity did not significantly predict oversight effectiveness at the pre-registered threshold (H1a not supported; $F(2, 134) = 3.29$, $p = 0.040$, $\hat{\eta}_p^2 = 0.047$). Notably, the direction of the raw (non-significant) differences between conditions was non-monotonic: the categorical Risk Level Indicator condition outperformed both Plain Information and the continuous Model Confidence score, contrary to the hypothesised ordering. Intervention option range did not significantly affect oversight effectiveness (H1b not supported, $F(2, 134) = 1.02$, $p = 0.364$). No significant interaction between information signals and intervention options was found on oversight effectiveness (H1c not supported, $F(4, 134) = 0.47$, $p = 0.756$). The two interface factors therefore did not amplify each other's effects in the current sample, and their contributions to oversight quality, to the extent they exist, appear to be independent and weak.

RQ2: Effects on perceived workload. Intervention option range did not significantly predict perceived workload (H2b not supported, $F(2, 134) = 2.18$, $p = 0.117$). Information signal granularity similarly showed no significant effect on perceived workload (H2a not supported, $F(2, 134) = 2.45$, $p = 0.090$). No significant interaction between the two factors was found on perceived workload (H2c not supported, $F(4, 134) = 1.59$, $p = 0.181$).

These results should be interpreted with caution. The study was designed for a total of 288 valid responses across two waves; the current sample of 144 provides approximately 50% of that target and is substantially underpowered for the interaction effects that were the binding constraint in the power analysis. The patterns described below are therefore preliminary and, where null results are obtained, should not be taken as strong evidence of absence.

7.2. Interpretation of Results

7.2.1. Descriptive Differences in Information Signal Granularity

The most noteworthy preliminary finding is the non-monotonic pattern in oversight effectiveness across signal conditions. The Risk Level Indicator condition achieved the highest marginal mean oversight score

($M = 17.04$), outperforming both Plain Information ($M = 12.44$) and Model Confidence ($M = 14.48$). This reversal of the hypothesised ordering — in which continuous probability scores were expected to be most informative — would be striking if confirmed after data collection is completed.

One account of this pattern draws on the distinction between epistemic access and epistemic usability [1]. Having access to a model confidence score requires complementary domain knowledge and the ability to convert a percentage into an actionable judgement. For crowdworkers without verified domain expertise in fraud detection, the cognitive cost of interpreting a continuous probability may outweigh its informational advantage over a categorical signal. A traffic-light style Risk Indicator, by contrast, delivers an immediately actionable signal with minimal interpretive overhead, which could explain its performance advantage in this sample [5]. This interpretation is consistent with evidence that the benefit of AI confidence information depends on complementary domain knowledge, and that naive users may not extract the intended signal from numerical probability displays [5]. It also echoes the OTTO interview findings, where overseers with deeper domain understanding were better able to extract actionable insight from the model outputs they were shown, whereas less experienced colleagues relied more heavily on structured summaries and thresholds.

If this pattern replicates with a larger sample, it would carry a practically important implication: the informativeness of a signal cannot be evaluated in isolation from the expertise of the overseer it is designed for. Deployers who assume that richer, more granular model outputs automatically produce better-informed oversight may be mistaken, particularly when the overseer population is heterogeneous in domain knowledge.

7.2.2. No Detectable Effects of Intervention Option Range

The absence of a significant effect of intervention option range on oversight effectiveness (H1b) is consistent with, though not confirmatory of, several competing accounts. First, the cognitive cost of managing a wider option set may offset the potential benefit of more calibrated delegation. Participants in the Delegate-with-Flags condition had access to five distinct response options, requiring not only a fraud/legitimacy judgement but also a conditional probability assignment when delegating. If this complexity exceeded the participants' processing capacity or familiarity with the task, the result would be worse calibration and no net gain in effectiveness [12]. This is broadly consistent with the observed reversal in which the Delegate-with-Flags condition produced the lowest marginal oversight effectiveness ($M = 13.30$) despite being the most informationally expressive response set.

Second, the over-delegation penalty embedded in the scoring scheme may have interacted with participants' delegation strategies in ways not anticipated by the hypothesis. As reported in Section 6.4.4, delegation rates in the Delegate-with-Flags condition were consistently higher than in the Delegate condition across all three information signal groups. Because the over-delegation penalty scales continuously with delegation rate, shrinking positive delegation scores in proportion to the fraction of cases delegated, these higher rates would have reduced the benefit of even appropriate delegations, compressing the score advantage that condition might otherwise have shown relative to the Delegate condition. Importantly, this compression was not driven by a small number of heavy delegators but applied broadly across the condition, as the descriptive statistics in Section 6.4.4 confirm.

Third, it is possible that the null result reflects genuine insensitivity of oversight quality to intervention range in non-expert populations. Without the domain knowledge to recognise when a case genuinely warrants expert review, the delegation option may hold little functional value, appearing as a way to avoid a decision rather than a principled triage mechanism. This would be consistent with the interview evidence from OTTO, where meaningful delegation was described as a skill that required experience with the domain to exercise appropriately.

7.2.3. Perceived Workload and the Absence of Predicted Effects

The absence of workload differences across conditions (H2a, H2b, H2c not supported) is unexpected given prior evidence that richer information environments and larger option sets impose greater cognitive demands [38]. The overall workload level was moderate ($M = 3.64$ on a 1–7 scale), suggesting neither a floor effect that would prevent detecting increases nor a ceiling effect that would mask decreases. One possible explanation is that participants adapted their decision strategy to the available option set, using delegation as a cognitive offload mechanism in conditions where it was available, as noted in

Section 6.2.4. The delegation rates discussed above are compatible with this account: participants in the Delegate condition delegated around 22–25% of cases on average, which may have been sufficient to relieve the most cognitively demanding decisions without adding considerable overhead. If delegation functioned in this way, one would expect delegation rates to correlate negatively with NASA-TLX scores within delegation conditions. However, correlations between per-participant delegation rate and NASA-TLX composite were near zero in both the Delegate ($r = -0.01$, $p = 0.936$) and Delegate-with-Flags ($r = -0.07$, $p = 0.643$) conditions, and the same pattern held for the Mental Demand and Effort subscales. The cognitive offload account therefore finds no support in the data. The anomalous Physical Demand result (discussed in Section 6.4) similarly resists straightforward theoretical interpretation and is most plausibly attributable to noise or an artefact of the standard NASA-TLX instrument rather than a genuine effect of the experimental manipulation.

7.2.4. The Expertise Covariate: An Unexpected Pattern

The potential negative association between self-reported expertise and oversight effectiveness ($b = -1.31, p = .031$) deserves attention. The pre-registered hypothesis was that higher expertise would predict better performance, and the bivariate correlation ($r = -0.16$) fell below the $r \geq 0.20$ threshold used to determine whether the ANCOVA covariate correction was warranted. The direction of the association is counterintuitive.

Several mechanisms could produce this pattern in a Prolific sample. First, overconfidence: if participants who self-rated as more knowledgeable in fraud detection or financial services were also more likely to dismiss the AI model's signal and rely on their own intuitions, they could perform worse than less confident participants who weighed the available signals more carefully. This would be consistent with documented patterns of miscalibration in self-assessed expertise [29, 47]. Second, the five-item scale may not validly discriminate expertise relevant to *this specific task*: familiarity with banking or accounting does not necessarily entail knowledge of how to interpret a model's fraud probability score. Third, the general Prolific sample may not contain a sufficient range of genuine domain expertise for the measure to behave as a discriminating covariate. The planned expert-screened wave would have provided a more meaningful test of whether true domain expertise moderates the effects of interest.

7.2.5. Metacognitive Calibration

The per-trial confidence data offer a partial counterpoint to the overconfidence interpretation. As reported in Section 6.4.3, participants showed weak but consistent metacognitive calibration, with confidence ratings positively associated with trial-level accuracy at both the population and individual levels. This suggests that participants had some genuine sense of when they were getting things right, even if that sensitivity was small. The apparent contradiction with the negative expertise-performance association may be partly resolved by noting that the two findings operate at different levels: the expertise result reflects a between-person pattern, where higher self-rated expertise predicted lower scores, while the calibration result reflects a within-person pattern, where individuals were relatively more confident on their correct trials. A participant can therefore be poorly calibrated in the absolute sense, overestimating their overall competence, while still showing local sensitivity to trial difficulty.

7.3. Relation to Theoretical Frameworks

The theoretical framework that most directly motivated this study is the four-condition model of effective human oversight proposed by Sterz et al. [1]: causal power, epistemic access, self-control, and fitting intentions. The two independent variables in this study map cleanly onto the first two conditions. Intervention options operationalise causal power, providing participants with differing degrees of control over the AI system's output. Information signals operationalise epistemic access, providing differing amounts of decision-relevant information about the risk each case represents. The absence of significant effects on oversight effectiveness in the current sample does not falsify this framework, but it does suggest that access to causal power and epistemic access are necessary but not sufficient conditions for effective oversight: their benefit appears contingent on the overseer having the domain knowledge to exploit them.

This contingency aligns with the facilitator/inhibitor taxonomy in Sterz et al. [1], which classifies domain expertise as a key individual-level facilitator. The implication is that interface design features and

individual capability are not independent levers but interacting ones: the non-monotonic signal pattern observed in this study is exactly what that framework would predict for a low-expertise population encountering high-granularity information signals.

The design considerations proposed by Faas et al. [18] reinforce this reading. Their co-design findings highlight that oversight personnel's sense of mastery — their subjective confidence in understanding what the system is doing and what their response options mean — is a precondition for motivated engagement with oversight tasks. Participants who find a model confidence score uninterpretable are unlikely to use it as intended, regardless of how informationally rich it is in principle. A categorical Risk Indicator, by contrast, requires no specialised knowledge to act upon, which may explain its performance advantage in the current sample.

The role of automation bias and algorithm aversion as potential moderators [9] could not be directly measured in this study. Delegation rates across conditions, reported in full in Section 6.4.4, do not show a systematic pattern that would provide clear indirect evidence of either mechanism: rates were moderate and broadly consistent across information signal conditions within both delegation intervention groups, offering no indication that richer signals led to either greater deference to or rejection of the implicit recommendations of the AI system.

7.4. Practical Implications

7.4.1. Interface Design for Oversight

A persistent concern in the literature is that placing a human in the review loop does not guarantee effective oversight [13]. The current results add a nuance to this concern: even when the interface is designed to provide informative signals and meaningful intervention options, the benefit depends on whether the overseer can interpret and act on what they are shown. Organisations designing review interfaces should therefore calibrate information signal design to the expertise of the intended overseer population, rather than defaulting to the richest available output from the model.

The potentially superior performance of the categorical Risk Level Indicator in the current sample suggests that well-designed categorical summaries may be more effective than raw probability scores for non-expert overseers, consistent with evidence that categorical displays reduce the cognitive cost of interpreting uncertain predictions [5]. Where a continuous confidence score is deemed necessary for accountability or audit purposes, practitioners should consider pairing it with a categorical interpretation layer (e.g., a colour-coded band that contextualises the probability within a risk threshold), rather than presenting the raw number alone.

Regarding intervention options, the current results do not support the claim that expanding the action vocabulary improves oversight effectiveness in non-expert populations. The higher delegation rates observed in the Delegate-with-Flags condition, combined with the over-delegation penalty in the scoring scheme, may reflect a mismatch between the richness of the option set and the domain knowledge required to use it discriminatively. Practitioners considering the introduction of delegation pathways should account for the learning cost associated with a larger option set, and should consider whether overseers have sufficient domain knowledge to use delegation as genuine triage rather than as a way to avoid commitment. The variable autonomy framework of Methnani et al. [12] offers a useful design principle here: the range of available options should be matched to the operator's current situational awareness, which is itself a function of domain knowledge and interface support.

7.4.2. Role Design and Overseer Selection

The OTTO interview data highlighted that oversight quality is strongly conditioned on the overseer's understanding of the domain: experienced data scientists were better able to detect anomalous model behaviour and to calibrate their reliance on predictions appropriately. This finding, combined with the negative association between self-reported expertise and performance in the current sample, underscores a key risk in oversight system design. If organisations recruit overseers primarily on the basis of availability rather than domain competence, and if those overseers hold inflated beliefs about their own expertise, the result may be confident but poorly calibrated oversight decisions. This is a variant of the rubber-stamping dynamic identified by Green [13], where oversight fails not from apathy but from miscalibration.

Organisations should therefore treat oversight as a skilled role with defined competency requirements, and should invest in pre-deployment training that makes the limits of the AI system's reliability explicit. The EU AI Act's requirement that high-risk AI systems be deployed in a manner that enables overseers to remain aware of their susceptibility to automation bias [29] provides a regulatory anchor for such training requirements, but it does not specify what adequate awareness looks like in practice. Empirical research of the kind reported here is therefore needed to operationalise that requirement.

7.4.3. Workload and Sustainable Oversight

The absence of significant workload differences across conditions is, in one sense, practically encouraging: it suggests that enriching the information environment or expanding the option set did not impose a measurable cognitive cost on participants in the short term. However, the design was a single 30-case session, and the absence of a workload penalty in a time-limited experiment does not guarantee that richer interfaces would not accumulate into fatigue over a full working day [1]. Organisations deploying continuous AI-assisted review workflows should monitor overseer workload longitudinally rather than inferring sustainable workload from controlled laboratory tasks.

7.4.4. The Role of Delegation

The delegation conditions in this study were motivated by learning-to-defer frameworks [16], which demonstrate that human-AI teams can achieve superior performance when appropriate task-routing mechanisms are in place. The scoring scheme distinguished between appropriate delegation (high expert disagreement variance, suggesting the case is genuinely ambiguous) and improper delegation (low variance, suggesting the case has a clear answer), rewarding the former and penalising the latter. In practice, this means that effective use of delegation requires the overseer to have some implicit knowledge of when a case is genuinely uncertain, which is itself a form of domain expertise. Without that knowledge, delegation options may be used indiscriminately, producing neither the performance benefits predicted by H1b nor the workload reductions that might follow from appropriate classification. The delegation rates observed in this study, reported in full in Section 6.4.4, are consistent with this account: participants in the richer option condition delegated more, but did not achieve correspondingly higher oversight effectiveness scores.

7.5. Future Work

7.5.1. Completing the Expert Wave

The most direct extension of this study is the collection of the planned expert-screened wave. The current results are insufficient to draw conclusions about the role of domain expertise as a moderator, because the general Prolific sample does not contain sufficient variation in genuine expertise to test this question. A sample of participants prescreened for verified backgrounds in accountancy, financial compliance, or fraud investigation would enable a direct comparison of whether the non-monotonic signal pattern and the intervention option null result hold across expertise levels or are specific to non-expert overseers. The pre-registration already accounts for this comparison; the expert wave can be added without modification to the analysis plan.

7.5.2. AI-Assisted Oversight: The Hybridisation Extension

This study originally considered a fourth information signal level: AI-generated recommendations produced via a confidence-based hybridisation approach inspired by Jain et al. [48]. In this approach, a large language model was provided with a labelled training set of fraud cases and asked to generate case-level recommendations for the 30 experimental items, functioning as an alternative form of decision-support signal alongside the Risk Indicator and model confidence score. This level was set aside in order to first establish a baseline understanding of how interface design factors shape human oversight without AI-generated assistance. Introducing an LLM recommendation into the signal conditions would have made it difficult to isolate the contribution of human judgement from that of a second AI system, conflating the overseer's independent reasoning with the downstream influence of an additional model. Examining the hybridisation approach remains a natural next step once the effects of information signal granularity and intervention option range are better understood in a purely human-driven oversight setting.

7.5.3. Varying Review Volume as a Workload Manipulation

A workload manipulation was also considered prior to finalisation, varying the number of cases assigned to participants to examine how review volume affects both oversight effectiveness and perceived workload. This was set aside because adding a further condition would have substantially increased the required sample size beyond what was feasible within the study timeline, and because perceived workload was already captured as a dependent variable via NASA-TLX, making the case quantity the most sensible manipulation to drop without losing measurement coverage of the construct. Future work could reintroduce review volume as an independent variable, particularly in longitudinal designs where fatigue accumulation over a full working session is of interest.

7.5.4. Longitudinal and Ecological Validity

The present study used a cross-sectional task-based design in which crowdworkers completed a single 30-case session. Real-world oversight is sustained over weeks and months, and involves accumulating fatigue, skill acquisition, and organisational pressures absent from a laboratory setting. Future work should track whether the benefits (and costs) of information-rich interfaces are maintained over time, or whether overseers adapt by developing fixed heuristics that reduce engagement with the signal content, a risk analogous to the vigilance decrements discussed in the facilitator/inhibitor framework of Sterz et al. [1].

7.5.5. Expert Populations and Ecological Validity

The FiFAR dataset used in this study [16] was constructed specifically to support realistic evaluation in high-stakes fraud detection contexts and provides a suitable benchmark for replication with professional fraud analysts or financial compliance officers. A professional sample study would substantially improve ecological validity and test whether the non-monotonic signal pattern observed here generalises beyond Prolific crowdworkers.

7.5.6. Adaptive Interfaces and Workload Management

An open question raised by the workload findings is whether cognitive load could be managed dynamically without sacrificing oversight quality. An adaptive interface could, for example, simplify the available option set during periods of high estimated load (inferred from response time or revision patterns), reverting to richer options when the overseer's processing capacity recovers. The variable autonomy framework of Methnani et al. [12] provides a theoretical basis for such adaptive designs, and the present study's finding that workload did not increase with option range suggests that the cost of additional options may be lower than expected in short-horizon tasks, making adaptive switching practically feasible.

8

Reflection

8.1. Position within the Broader Computer Science Field

This thesis sits at the intersection of human-computer interaction, AI safety, and algorithmic decision-making. Within computer science, it contributes to a growing body of empirical work on the conditions under which human oversight of AI systems is effective, a question that has become urgent as high-risk AI applications proliferate in finance, healthcare, criminal justice, and public administration.

The broader context is set by regulatory developments such as the EU AI Act, Article 14 of which mandates "effective human oversight" for high-risk AI systems without specifying what effectiveness entails in practice [1]. This thesis responds to that gap empirically: rather than theorising about what effective oversight should look like, it tests specific design parameters in a controlled experiment. In doing so, it connects to the framework proposed by Sterz et al. [1], who argue that effectiveness requires the overseer to have (a) causal power, (b) epistemic access, (c) self-control, and (d) fitting intentions, and provides preliminary empirical evidence about two of these conditions: causal power, operationalised via intervention options, and epistemic access, operationalised via information signals.

Within the HCI literature, this work also contributes to debates about AI-assisted decision-making and how interface design modulates the quality of human judgement [5]. Importantly, it distinguishes itself from the mainstream AI-assisted decision-making paradigm by adopting an *oversight* framing, in which the human acts as a meta-controller over an already-deployed system, rather than a *human-in-the-loop* framing, in which the human is an integral part of the primary prediction process. This distinction matters: oversight introduces an asymmetric risk structure where successful oversight is invisible but failed oversight can have severe consequences [2, 13].

The study also makes a methodological contribution by introducing a scoring scheme that rewards both classification accuracy and appropriate use of delegation, distinguishing itself from binary per-case accuracy measures that cannot capture the nuanced quality of oversight in settings where triage and escalation are legitimate responses. The use of a simulation-validated scoring range ($[-60, +60]$) and the pre-registration of all analysis decisions before data collection strengthen the evidential value of the preliminary findings reported here, even under conditions of partial power.

8.2. Limitations

8.2.1. Sample and Statistical Power

The most consequential limitation of this study is that data collection could only be completed for the general-population wave ($N = 144$), representing approximately half the overall two-wave target, with none of the planned expert-screened wave collected. The pre-registered power analysis identified the interaction effect ($f = 0.25$, $\alpha = 0.008$, power $= 0.80$) as the binding constraint, requiring 288 valid responses across both waves. The current sample is underpowered for all six pre-registered hypotheses, particularly the interaction terms, and the planned expert-screened wave could not be collected within the thesis timeline due to administrative delays in budget approval. All null results should therefore be

interpreted as inconclusive rather than as evidence of absence: the widened confidence intervals under partial power make it impossible to distinguish genuine null effects from effects that would emerge with adequate power.

8.2.2. Self-Reported Expertise as Covariate

Because the expert-screened wave was not collected, domain expertise is measured solely through a five-item self-report scale. Self-reported expertise is known to diverge from demonstrated competence, particularly in domains where respondents lack the benchmarks needed to calibrate their self-assessments [29, 47]. The unexpected negative association discussed in Section 7.2.4 is consistent with overconfidence in the Prolific sample, but cannot be attributed to any single mechanism without additional data. Including anchoring stimuli drawn from the task itself could have improved the validity of the measure by giving participants a concrete reference point against which to calibrate their self-ratings. For example, presenting one or two representative cases prior to the expertise self-assessment would have grounded responses in the actual demands of the task, potentially reducing overconfidence effects. A study combining self-report measures with objective performance indicators, or using participants with verified professional credentials, would provide a more valid test of the expertise moderation hypothesis.

8.2.3. Crowdworker Ecological Validity

Prolific participants completing a fraud detection task for pay are not equivalent to professional fraud analysts operating under organisational accountability structures. The financial incentive structure rewards accuracy but does not replicate the long-term reputational and professional consequences of oversight errors in real deployments; participants had no prior exposure to the AI system and no accumulated knowledge of its error patterns; and the 30-case single-session format cannot produce the vigilance decrements or expertise accumulation that characterise sustained real-world oversight. These limitations constrain the generalisability of specific effect size estimates to professional settings, though they do not invalidate the findings as estimates of interface effects on oversight quality in controlled conditions.

8.2.4. Task and Dataset Constraints

The FiFAR dataset provides a realistic and well-documented benchmark [16], but several features of the sampled task set may have affected results. The 30-case set was stratified to include equal proportions of appropriate-delegation and improper-delegation cases, producing a base rate of 50%, which is unlikely to reflect real-world review queues and may have artificially elevated delegation rates across conditions. Additionally, the four case features selected for display represent only a subset of what a professional fraud analyst would typically access, which may have limited the practical difference between information signal conditions.

8.2.5. Single-Session Design and Absent Hybridisation Condition

The study was administered as a single time-limited session and therefore cannot capture learning effects, fatigue, or the longer-term calibration of AI reliance that characterise sustained oversight. Although the 30-case design was chosen to keep session length manageable, it does not fully preclude within-session vigilance decrements; sustained engagement with a mostly-reliable AI system can erode attentiveness over time, and the present study does not include a measure of decision quality over trial order, making it impossible to determine whether performance deteriorated across the session [1]. Longitudinal designs are needed to assess whether the interface effects observed here persist, decay, or strengthen over extended use. Finally, the planned condition in which participants would receive AI-generated case recommendations alongside the standard information signals was not included in the present study, limiting what can be concluded about human-AI complementarity in oversight contexts [48].

8.3. Societal and Ethical Implications

8.3.1. The Risk of Rubber-Stamping

One of the central ethical concerns motivating this research is that poorly designed oversight systems may create a false sense of accountability without providing genuine protection against AI errors or unfairness. As discussed in Section 7.4.1, interface design choices determine not merely whether the

oversight process is usable, but whether it is genuinely protective of the people whose applications are at stake. In the fraud detection context, this distinction carries direct consequences for applicants: a system that satisfies oversight requirements on paper while leaving overseers under-informed or without meaningful intervention options transfers the moral weight of the decision to a human who lacks the conditions to exercise it responsibly. Designers and developers therefore carry an ethical responsibility to ensure that their interfaces substantively enable effective oversight rather than merely instantiate it formally.

8.3.2. Fairness and Differential Impact

The use of a synthetic dataset (FiFAR/BAF) means that the present study cannot directly examine whether the interface manipulations tested here interact with the demographic characteristics of applicants in ways that create or worsen unequal outcomes. In real-world fraud detection, AI systems have been shown to exhibit demographic bias [16, 17], and it is plausible that overseers who rely heavily on model signals will inherit and reinforce such biases, while overseers with deeper intervention options and domain expertise may be better placed to identify and correct them. Future work should examine this question directly using datasets that include protected attribute information.

8.3.3. Crowdwork and Labour Ethics

This study compensated Prolific participants using performance-based pay, a design choice intended to promote engagement with the task. Although financial incentives are standard in online experimental research, the use of crowdwork raises broader questions about whether this model is appropriate for the populations to whom oversight findings are intended to generalise. The OTTO interviews underlined that effective oversight in practice requires ongoing relationships, accumulated domain knowledge, and organisational accountability structures that are structurally absent from gig-economy labour arrangements. The findings should therefore be understood as evidence about *how interfaces mediate oversight quality in controlled conditions*, not as evidence that crowdworkers constitute an adequate oversight workforce for high-stakes AI deployment.

8.4. Applicability and Transferability of Findings

The experimental context of this study, the review of bank account fraud alerts, was chosen for its combination of ecological realism and experimental tractability. The FiFAR dataset is grounded in realistic synthetic expert behaviour [16], and the binary ground truth labels allow objective effectiveness measurement without reliance on subjective quality assessments.

Several features of this context may nonetheless limit transferability to other oversight settings. First, fraud detection involves a relatively structured and information-rich feature space compared to, for example, content moderation or medical diagnosis, where the relevant information may be more ambiguous or require more specialised processing. The benefit of model confidence signals, in particular, may depend on the quality of the underlying model's calibration, which cannot be assumed across domains [5]. Second, the binary nature of the ground truth simplifies the oversight task relative to many real-world settings where errors are graded, consequences are asymmetric, or multiple risk dimensions must be evaluated simultaneously. The inclusion of the delegation pathway partially addresses this, but future work in multi-label or graded-risk contexts is needed. Third, the regulatory and organisational context differs across deployment settings: oversight of a retail demand-forecasting system (as at OTTO) differs from oversight of a financial fraud detection system in both the stakes involved and the institutional accountability mechanisms available. The facilitator/inhibitor framework of Sterz et al. [1] provides a useful meta-level mapping tool, but empirical replication remains necessary.

The experiment interface, analysis scripts, and pre-processed dataset used in this study are made publicly available to support such replication (see footnote).[1]

[1] `https://github.com/Diego-Viero/Effective-Human-Oversight-Master-Thesis`

9

Conclusion

9.1. Interpretation of Main Effects

This study tested two interface design levers, information signal granularity and intervention option range, against two measures of human oversight quality: correctness of fraud classification and delegation decisions, and perceived workload. Neither lever reached the pre-registered significance threshold of $\alpha = .008$ in either dependent variable, and the interaction between them was also non-significant. Given that the sample ($N = 144$) represents approximately half the pre-registered target, all null results should be read as inconclusive rather than as evidence of absence.

Within these constraints, the one substantive pattern to carry forward is the potential main effect of information signal granularity on oversight effectiveness ($F(2, 134) = 3.29$, $p = .040$, $\hat{\eta}_p^2 = 0.047$). Its direction could be non-monotonic as per our exploratory analysis: the risk level indicator outperformed both plain information and the model confidence score, contrary to the hypothesised ordering. As discussed in Section 7.2.1, this reversal is consistent with a signal-usability account in which categorical cues are more immediately actionable for non-expert overseers than continuous probability scores, whose interpretation requires complementary domain knowledge not uniformly present in a general-population sample. The broader implication is that richer model outputs do not automatically produce better-informed oversight; the appropriate level of signal granularity depends on the expertise of the overseer population for which the interface is designed.

Regarding workload, the absence of significant differences between conditions offers limited practical reassurance that graduated intervention sets need not impose a measurable cognitive cost in short-horizon tasks, though whether this holds under sustained real-world oversight conditions remains an open question.

9.2. Interpretation of Interaction Effects

No significant interaction between information signals and intervention options was found on oversight effectiveness ($F(4, 134) = 0.47$, $p = 0.756$) or perceived workload ($F(4, 134) = 1.59$, $p = 0.181$). The complementarity hypothesis, that richer signals and deeper intervention options would jointly produce more than the sum of their individual contributions, is not supported in this preliminary data.

However, this result is difficult to interpret conclusively, given that the complementarity literature suggests human–AI synergy is moderated by expertise in ways the current non-expert sample cannot adequately test [1, 5]. Testing whether the interaction emerges in a verified-expert sample therefore remains the most informative next step, and is already accounted for in the pre-registered analysis plan.

9.3. Closing Remarks

This thesis set out to bring empirical grounding to a regulatory requirement, effective human oversight, that has so far been defined more clearly in law than in practice. By operationalising two interface-level facilitators from the framework of Sterz et al. [1], information signal granularity and intervention option

range, and testing them in a controlled fraud detection task, it contributes a first empirical estimate of their effects in a non-expert overseer population. The central finding, that a categorical risk indicator may outperform a richer continuous signal for overseers without verified domain expertise, points to a practically important constraint: interface design cannot be evaluated in isolation from the people it is designed for. The expert-screened wave, preregistered and ready to collect, remains the most direct next step towards a more complete answer. More broadly, as AI systems take on an expanding role in consequential decisions, the question of how oversight interfaces can genuinely support human judgement, rather than merely satisfy a compliance requirement, will only become more pressing, and this study represents one step towards answering it.

Generative AI Disclosure

This chapter discloses the use of generative artificial intelligence (AI) tools during the preparation of this thesis, in accordance with TU Delft's guidelines on transparency in academic work.

Generative AI tools were used throughout the thesis process as productivity and writing aids. Their use was limited to supporting tasks and did not involve delegating intellectual or analytical judgement to the tools. All content, interpretations, arguments, and conclusions presented in this thesis are the author's own. Specifically, AI assistants were used to improve the clarity, fluency, and structure of written prose across chapters, including rephrasing sentences for conciseness and ensuring consistency in tone and style. AI tools were also consulted to explore possible organisational approaches to chapters and sections, with suggestions evaluated critically and adapted to the specific requirements of the thesis. Additionally, AI assistants supported the development of Python scripts for statistical analysis, data and result processing; all generated code was reviewed, tested, and adapted by the author before being integrated into the analysis pipeline, and all analytical decisions were made independently by the researcher. AI tools further assisted in generating and refining code used to produce figures and plots, all of which were reviewed for accuracy and appropriateness prior to inclusion. Finally, AI assistants were occasionally consulted for help with LaTeX syntax, package usage, and resolving typesetting issues during document preparation.

Acknowledgements

This thesis would not have been possible without the support and guidance of many people, and I would like to take a moment to express my gratitude.

I would like to thank my supervisors, Ujwal Gadiraju and Tim Draws, for their guidance and support throughout this project. Your feedback was fundamental in shaping the direction of this work, and your availability and engagement throughout the process are greatly appreciated.

A special thanks goes to *StudioLab*, with particular thanks to Aadjan van der Helm, Martin Havranek & Jerry De Vos, who kindly welcomed me and provided a stimulating working environment for the duration of this thesis. This journey would not have been as enjoyable, nor would I have learned as much, without you.

I would also like to thank the Web Information Systems Group for providing the budget that made it possible to run the user study presented in this thesis. Finally, I would like to thank the professionals at OTTO who generously gave their time to participate in the interviews conducted as part of this research. Your insights were invaluable in grounding this work in practice.

To all of you, thank you.

Diego Viero

June, 2026

Bibliography

[1] S. Sterz et al., "On the quest for effectiveness in human oversight: Interdisciplinary perspectives", in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, pp. 2495–2507.

[2] J. Salgado-Criado, "Human oversight of artificial intelligence: An operations management perspective", *Journal of Industrial Engineering and Management*, vol. 18, no. 2, pp. 285–304, 2025.

[3] M. Langer, K. Baum, and N. Schlicker, "Effective human oversight of ai-based systems: A signal detection perspective on the detection of inaccurate and unfair outputs", *Minds and Machines*, vol. 35, no. 1, p. 1, 2024.

[4] Y. Sun, E. Jang, F. Ma, and T. Wang, "Generative ai in the wild: Prospects, challenges, and strategies", in *Proceedings of the 2024 chi conference on human factors in computing systems*, 2024, pp. 1–16.

[5] Y. Zhang, Q. V. Liao, and R. K. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making", in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 295–305.

[6] S. Gaube et al., "Keeping an eye on ai: A framework for effective human oversight of ai systems", 2026.

[7] European Parliament and Council of the European Union, *Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act)*, Official Journal of the European Union, L 1689, 12 July 2024, ELI: http://data.europa.eu/eli/reg/2024/1689/oj; Text with EEA relevance, Jul. 2024. [Online]. Available: `https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng`.

[8] J. Laux, "Institutionalised distrust and human oversight of artificial intelligence: Towards a democratic design of ai governance under the european union ai act", *AI & society*, vol. 39, no. 6, pp. 2853–2866, 2024.

[9] L. J. Skitka, K. Mosier, and M. D. Burdick, "Accountability and automation bias", *International Journal of Human-Computer Studies*, vol. 52, no. 4, pp. 701–717, 2000.

[10] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err.", *Journal of experimental psychology: General*, vol. 144, no. 1, p. 114, 2015.

[11] M. Langer, V. Lazar, and K. Baum, "How to test for compliance with human oversight requirements in ai regulation", in *CHI25 Conference on Human Factors in Computing System Workshop on Sociotechnical AI Governance, Yokohama, Japan*, 2025.

[12] L. Methnani, A. Aler Tubella, V. Dignum, and A. Theodorou, "Let me take over: Variable autonomy for meaningful human control", *Frontiers in Artificial Intelligence*, vol. 4, p. 737 072, 2021.

[13] B. Green, "The flaws of policies requiring human oversight of government algorithms", *Computer Law & Security Review*, vol. 45, p. 105 681, 2022.

[14] M. Almada, "Human intervention in automated decision-making: Toward the construction of contestable systems", in *Proceedings of the Seventeenth International Conference on artificial intelligence and law*, 2019, pp. 2–11.

[15] B. Wagner, "Liable, but not in control? ensuring meaningful human agency in automated decision-making systems", *Policy & Internet*, vol. 11, no. 1, pp. 104–122, 2019.

[16] J. V. Alves et al., "Financial Fraud Alert Review Dataset", Apr. 2025. DOI: `10.6084/m9.figshare.28351172.v1`. [Online]. Available: `https://springernature.figshare.com/articles/dataset/Financial_Fraud_Alert_Review_Dataset/28351172`.

[17] S. Jesus et al., "Turning the Tables: Biased, Imbalanced, Dynamic Tabular Datasets for ML Evaluation", *Advances in Neural Information Processing Systems*, 2022.

[18] C. Faas, S. Kerstan, R. Uth, M. Langer, and A. M. Feit, "Design considerations for human oversight of ai: Insights from co-design workshops and work design theory", in *Proceedings of the 31st International Conference on Intelligent User Interfaces*, ser. IUI 26, 2026, pp. 804–821.

[19] P. Hemmer, M. Schemmer, N. Kühl, M. Vössing, and G. Satzger, "Complementarity in human-ai collaboration: Concept, sources, and evidence", *European Journal of Information Systems*, vol. 34, no. 6, pp. 979–1002, 2025.

[20] K. Inkpen et al., "Advancing human-ai complementarity: The impact of user expertise and algorithmic tuning on joint decision making", *ACM Transactions on Computer-Human Interaction*, vol. 30, no. 5, pp. 1–29, 2023.

[21] N. Diakopoulos, "Accountability in algorithmic decision making", *Communications of the ACM*, vol. 59, no. 2, pp. 56–62, 2016.

[22] J. Goldenfein, "Algorithmic transparency and decision-making accountability: Thoughts for buying machine learning algorithms", *Jake Goldenfein,'Algorithmic Transparency and Decision-Making Accountability: Thoughts for buying machine learning algorithms' in Office of the Victorian Information Commissioner (ed), Closer to the Machine: Technical, Social, and Legal aspects of AI (2019)*, 2019.

[23] M. Yurrita, T. Draws, A. Balayn, D. Murray-Rust, N. Tintarev, and A. Bozzon, "Disentangling fairness perceptions in algorithmic decision-making: The effects of explanations, human oversight, and contestability", in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–21.

[24] K. Alfrink, I. Keller, G. Kortuem, and N. Doorn, "Contestable ai by design: Towards a framework", *Minds and Machines*, vol. 33, no. 4, pp. 613–639, 2023.

[25] N. Karusala, S. Upadhyay, R. Veeraraghavan, and K. Z. Gajos, "Understanding contestability on the margins: Implications for the design of algorithmic decision-making in public services", in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–16.

[26] C. Doctorow, "Ai companies will fail. we can salvage something from the wreckage", *The Guardian*, Jan. 2026. [Online]. Available: `https://www.theguardian.com/us-news/ng-interactive/2026/jan/18/tech-ai-bubble-burst-reverse-centaur`.

[27] L. Cavalcante Siebert et al., "Meaningful human control: Actionable properties for ai system development", *AI and Ethics*, vol. 3, no. 1, pp. 241–255, 2023.

[28] J. Kuzma, J. Paradise, G. Ramachandran, J.-A. Kim, A. Kokotovich, and S. M. Wolf, "An integrated approach to oversight assessment for emerging technologies", in *Emerging Technologies*, Routledge, 2020, pp. 249–271.

[29] J. Laux and H. Ruschemeier, "Automation bias in the ai act: On the legal implications of attempting to de-bias human oversight of ai", *ArXiv*, vol. abs/2502.10036, 2025. [Online]. Available: `https://api.semanticscholar.org/CorpusId:276395100`.

[30] S. Knapi, A. Malhi, R. Saluja, and K. Främling, "Explainable artificial intelligence for human decision support system in the medical domain", *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 740–770, 2021.

[31] Q. Zhang, M. L. Lee, and S. Carter, "You complete me: Human-ai teams and complementary expertise", in *Proceedings of the 2022 CHI conference on human factors in computing systems*, 2022, pp. 1–28.

[32] B. C. Cheong, "Transparency and accountability in ai systems: Safeguarding wellbeing in the age of algorithmic decision-making", *Frontiers in Human Dynamics*, vol. 6, p. 1 421 273, 2024.

[33] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration", *Human factors*, vol. 52, no. 3, pp. 381–410, 2010.

[34] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, "To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making", *Proceedings of the ACM on Human-computer Interaction*, vol. 5, no. CSCW1, pp. 1–21, 2021.

[35] M. Estévez-Almenzar, R. Baeza-Yates, and C. Castillo, "Human response to decision support in face matching: The influence of task difficulty and machine accuracy", *arXiv preprint arXiv:2505.13218*, 2025.

[36] D. Byrne, "A worked example of braun and clarkes approach to reflexive thematic analysis", *Quality & quantity*, vol. 56, no. 3, pp. 1391–1412, 2022.

[37] D. Viero, U. Gadiraju, and T. Draws, *Effective human oversight of ai systems*, May 2026. DOI: 10.17605/OSF.IO/8NYB6. [Online]. Available: osf.io/8nyb6.

[38] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research", in *Advances in psychology*, vol. 52, Elsevier, 1988, pp. 139–183.

[39] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge, 2013.

[40] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk", in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2008, pp. 453–456.

[41] J. L. Gastwirth, Y. R. Gel, and W. Miao, "The Impact of Levenes Test of Equality of Variances on Statistical Theory and Practice", *Statistical Science*, vol. 24, no. 3, pp. 343–360, 2009. DOI: 10.1214/09-STS301. [Online]. Available: https://doi.org/10.1214/09-STS301.

[42] J. S. Long and L. H. Ervin, "Using heteroscedasticity consistent standard errors in the linear regression model", *The American Statistician*, vol. 54, no. 3, pp. 217–224, 2000.

[43] P. A. Games and J. F. Howell, "Pairwise multiple comparison procedures with unequal ns and/or variances: A monte carlo study", *Journal of Educational Statistics*, vol. 1, no. 2, pp. 113–125, 1976.

[44] J. Cohen, *Statistical power analysis for the behavioral sciences*. routledge, 2013.

[45] M. Geissbuehler and T. Lasser, "How to display data by color schemes compatible with red-green color perception deficiencies", *Optics express*, vol. 21, no. 8, pp. 9862–9874, 2013.

[46] K. L. Mosier, L. J. Skitka, M. D. Burdick, and S. T. Heers, "Automation bias, accountability, and verification behaviors", in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, SAGE Publications Sage CA: Los Angeles, CA, vol. 40, 1996, pp. 204–208.

[47] D. Dunning, J. A. Meyerowitz, and A. D. Holzberg, "Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability.", *Journal of personality and social psychology*, vol. 57, no. 6, p. 1082, 1989.

[48] R. Jain, S. Bridgers, L. Janzer, R. Greig, T. H. Teh, and V. Mikulik, "Human-ai complementarity: A goal for amplified oversight", *arXiv preprint arXiv:2510.26518*, 2025.

A

Interview Questions

Experience and Role Context

- Can you describe your current role and main responsibilities within your department?
- How long have you been working at OTTO, more specifically in your department?
- What does a typical review session look like for you? Can you walk me through some of the steps?
- How would you describe your confidence level when identifying a PSF (product sales forecast) issue or anomaly?
- How much time do you typically spend on these review sessions?

Oversight Workflow

- How do you decide which predictions to examine more closely during a review session?
- What do you do when you identify something that seems unusual or incorrect?
- Can you describe a recent example where you found an inaccuracy? What happened next?
- What tools or resources do you use during the review process?
- What information about the predictions do you have access to during reviews?

Perceived Value and Challenges

- From your perspective, what purpose does this review process serve?
- Can you think of a time when this review process led to a meaningful change or improvement?
- What aspects of the review process do you find most straightforward? Most difficult?
- What kinds of inaccuracies or issues are easiest for you to spot? Which is the hardest?
- Is there anything that makes it harder to conduct these reviews effectively?

Training and Expertise Differences

- What background or knowledge do you draw on when reviewing model predictions?
- Was there anything you had to learn specifically for doing these reviews?
- How did you learn to do this type of review work?
- Do you feel your background/role influences what you notice or focus on during reviews?

Forward-Looking / Improvement Perspectives

- If you could change one thing about how these reviews are conducted, what would it be?
- What additional information or tools do you think would make these reviews easier or more effective?

- Are there aspects of the models' decision-making that you wish you understood better?
- If you were designing an oversight process from scratch, what would you prioritize?
- Is there anything else about your experience with this review process that you think is important but we haven't discussed?