RESEARCH SUMMARY FOR FIELDWORK

Human oversight in AI and agentic systems in the public sector

In a nutshell	Human oversight is often described as a person reviewing or intervening in AI-driven decisions. In practice, the space for meaningful intervention is often shaped earlier, through design choices, workflow structures, role definitions, escalation rules, and assumptions about what people will be able to notice, understand, and correct. In AI and agentic systems, this becomes more difficult because tasks may be linked across multiple steps, responsibilities may be distributed, and human involvement may shift from direct review to supervision, exception handling, or prior specification. This study examines how oversight is understood, specified, and organised in public sector contexts where AI and agentic systems are being adopted, piloted, or considered. Rather than focusing only on mature deployments, it also looks at how organisations plan, define, and justify oversight in earlier stages of system design and introduction.
Background	Public sector organisations are rapidly exploring and adopting AI to improve efficiency, consistency, and service delivery. At the same time, organisational readiness remains uneven. Around 90% of organisations plan to explore or deploy AI-enabled or agentic systems within the next 2 to 3 years yet fewer than 25% report high maturity in the data capabilities required to support these systems[1]. This creates a practical tension: systems are being deployed faster than the structures needed to understand and govern them. AI is increasingly used not only to support decisions, but also to structure and execute parts of the decision process itself. This introduces a form of delegated decision authority, where aspects of judgement are embedded in technical systems, workflows, and operational rules. Recent developments in agentic systems extend this further. These systems may be designed to plan, coordinate, or act across multiple steps within organisational workflows, which raises new questions about coordination, control, and accountability when multiple systems and actors interact. AI governance frameworks often treat human oversight as a key safeguard. In practice, this is usually framed as a person reviewing outputs, monitoring system behaviour, or intervening when needed. However, this assumes that decision-makers can understand how the system reaches outcomes and that meaningful intervention remains possible at the point of use. In reality, many important decisions are already shaped earlier in the process through system design, data selection, thresholds, permissions, and workflow constraints. This suggests that oversight may be better understood not as a single action, but as a distributed organisational arrangement spanning design, procurement, monitoring, and escalation.

[1] 'Data foundations for government – From AI ambition to execution', Capgemini Research Institute report, May 2025

Research questions	How is human oversight understood, specified, and organised across organisational roles, processes, and governance mechanisms in AI and agentic systems in public sector contexts? Subquestions: • How are assumptions about human monitoring, intervention, and responsibility reflected in organisational roles and processes? • What mechanisms are used to structure oversight in practice (e.g. monitoring, escalation, review), and how are these expected to work? • How do emerging agentic systems challenge existing assumptions about where oversight sits, how it operates, and what humans are expected to do? The study takes a flexible approach to case selection, recognising that AI and agentic systems are at different stages of development across the public sector. Cases may include systems in use, pilots, or initiatives at earlier stages of design, procurement, or consideration. *N. B. The study uses the term "agentic systems" pragmatically to refer to systems described as more autonomous, multi-step, or action-oriented, recognising that terminology is still evolving.*
Methodology	This study uses an exploratory qualitative design to examine how human oversight is interpreted and operationalised in practice. The focus is on understanding how oversight works in real organisational settings, rather than measuring how often specific practices occur. The study examines AI-enabled decision systems embedded in organisational workflows, with particular attention to emerging agent-based systems that can initiate or coordinate actions across tasks. These systems are relevant because they change how decision-making is structured. Oversight mechanisms rely on assumptions about how people monitor system behaviour and recognise potential issues. They also assume that individuals can decide when and how to intervene. The study explores how these assumptions are reflected in actual organisational arrangements. The research is based on 3–5 case studies of AI-enabled systems that are adopted, piloted or currently being designed within public sector organisations. Data sources include: • Document analysis: (procurement materials, governance frameworks, public documentation of AI systems, audit or risk procedures) • Semi-structured interviews: with actors involved in design, implementation, or governance (e.g. public officials, vendors, consultants) The material is analysed using qualitative thematic analysis, with a focus on identifying patterns in: • how oversight responsibilities are distributed • how system behaviour is monitored • how escalation and intervention are structured in practice