EXTRACT FROM AIBook 4

INTRODUCTION to Sample Extract :

The following pages are presented as an unedited extract from an AISF Insight Book: **AIBook 4 — Governing AI: Executive and Board Accountability**.

They are included to illustrate how AIBook 4 examines the practical governance questions arising as artificial intelligence becomes embedded within organisational decision-making.

The selected chapter, **The Governance Boundary**, explores a central question facing boards, executives, and institutions:

Which decisions may be assisted by AI, and which decisions must remain under explicit human authority and accountability?

Throughout AIBook 4, governance is examined not as a technology issue alone, but as a question of responsibility, oversight, authority, and institutional legitimacy.

The chapter presented here is one component of a broader framework that also examines:

- executive accountability and stewardship
- board oversight responsibilities
- risk governance and assurance
- escalation and decision authority
- human oversight and intervention
- institutional trust and legitimacy

This extract is presented without simplification or interpretation. It illustrates the style, depth, and analytical approach applied throughout AIBook 4 and the wider AISF Insight Series.

CHAPTER 5

The Governance Boundary

Every institution that deploys AI makes an implicit decision about boundaries. The question is whether those boundaries are intentional or accidental.

Much of the debate around AI focuses on capability: what systems can do, how accurate they are, how quickly they improve. Governance, however, is concerned with a different question: what systems should be allowed to decide.

This is not a technical determination.
It is a fiduciary one.

Institutions often avoid drawing explicit boundaries because doing so feels premature. Capabilities are evolving. Use cases are expanding. Leaders worry that fixed limits may constrain future value or signal conservatism. As a result, authority boundaries are left vague, to be negotiated case by case.

When boundaries are not drawn deliberately, they are drawn by default—by vendors, by operational convenience, or by crisis. None of these actors carry board-level accountability for institutional legitimacy.

A governance boundary is not a prohibition.
It is a line of responsibility.

Some decisions can be meaningfully assisted by AI without undermining accountability. Others cannot. The distinction is rarely about complexity or data volume. It is about consequence and contestability. Decisions that materially affect rights, access, safety, or institutional trust demand a different authority posture than those that optimise internal efficiency.

This is the logic behind the **non-delegable decisions doctrine**.

The doctrine does not specify technologies or techniques. It specifies responsibility. It asserts that certain categories of judgment must remain human-owned, regardless of system performance, efficiency gains, or competitive pressure. Accuracy does not dissolve responsibility. Speed does not replace accountability.

Non-delegable decisions typically share three characteristics:

1. They carry irreversible or disproportionate harm if wrong.
2. They involve normative judgment, not just factual inference.
3. They shape trust in the institution beyond the individual case.

These characteristics are often present together, but not always recognised explicitly. When institutions fail to articulate them, delegation expands by default. Systems are introduced incrementally, each framed as limited, controlled, and supportive. Over time, the boundary

moves—not because a decision was taken to move it, but because no decision was taken to hold it.

This is boundary erosion through incrementalism.

AI is rarely deployed as a single authority-shifting intervention. It arrives in layers: scoring tools, ranking engines, recommendation systems, monitoring logic, escalation triggers. Each layer is approved independently. Each appears bounded. Each is justified on narrow grounds.

Collectively, however, these layers reshape decision authority.

One system narrows options. Another prioritises risks. A third recommends action. A fourth flags deviation. No single system "decides," yet the space for judgment steadily compresses. By the time a human intervenes, discretion exists in theory but not in practice.

Governance boundaries must therefore be set at the level of **decision classes**, not tools.

Instead of asking, "Is this system acceptable?" boards should ask:

- What category of decision does this system materially influence?
- How much discretion remains with humans in practice, not on paper?
- At what point does assistance become substitution?
- What happens when human judgment conflicts with system output?
- Who is accountable if override is not exercised?

These questions surface authority boundaries that tool-level discussions obscure.

A common failure mode is confusing **override availability** with **override realism**. Many AI systems are technically over-rideable. In practice, override may be culturally discouraged, operationally costly, or reputationally risky. Leaders may fear that overriding the system signals subjectivity, error, or non-compliance with best practice.

When override carries friction, it becomes symbolic rather than substantive.

"Human in the loop" then functions as reassurance language rather than a genuine authority safeguard. The loop exists, but it is rarely exercised. Over time, this condition normalises deference. Authority shifts without explicit acknowledgement, and responsibility weakens accordingly.

Effective governance requires testing boundaries under realistic conditions. Boards should expect management to demonstrate not only that override exists, but that it is usable under pressure: when time is limited, when outcomes are contested, and when system confidence appears high. If override is theoretically possible but practically unlikely, the boundary is already breached.

Governance boundaries also serve an identity function. They clarify what kind of institution this is. Is it one that uses systems to support judgment, or one that allows systems to substitute for it? This identity becomes visible during controversy. Institutions that cannot articulate their boundaries appear evasive, even when their systems perform well.

Boundaries therefore operate as legitimacy infrastructure.

They allow leaders to say, consistently and credibly, “This is not a decision we outsource.” They also allow leaders to explain why systems are used elsewhere without implying abandonment of responsibility. Without such boundaries, institutions struggle to explain themselves coherently when challenged.

Importantly, boundaries protect internal governance as much as external trust. They give executives and staff clarity about where authority truly resides. They reduce the quiet pressure to defer. They preserve space for judgment rather than leaving it to erode through habit.

Boards must accept a fundamental responsibility: deciding not only what the organisation can do, but what it will not do—even when technology makes delegation tempting.

That is not conservatism.
It is stewardship.

In an AI-enabled world, governance boundaries are not constraints on progress. They are conditions for legitimacy. They ensure that adoption remains a choice rather than an inevitability, and that responsibility remains visible rather than diffuse.

Institutions that do not set boundaries will eventually be forced to—after harm, scrutiny, or loss of trust. At that point, boundaries are drawn defensively, under pressure, and with limited credibility.

The role of boards and senior executives is to ensure that boundaries are set first, not last.