AI Supervision for Oversight and Control

October 21, 2025

Abstract

Oversight and control, which we collectively call supervision, are often discussed as ways to ensure that AI systems are accountable, reliable, and able to fulfill governance and management requirements. However, the requirements for "human oversight" risk codifying vague or inconsistent interpretations of key concepts like oversight and control. This ambiguous terminology could undermine efforts to design or evaluate systems that must operate under meaningful human supervision. This is important given the term is used by regulatory texts, such as the EU AI Act, the NIST AI Risk Management Framework, and the OECD AI Principles.

This paper undertakes a targeted critical review of literature on supervision outside of AI, along with a brief summary of past work on the topic related to AI. We then differentiate control as being ex-ante or real-time, and operational rather than policy or governance. In contrast, we view oversight as either performed ex-post, or a policy and governance function. We suggest that control aims to prevent failures, while oversight focuses on detection, remediation, or incentives for future prevention, and we note that preventative oversight strategies necessitate control.

Building on this foundation, we make three contributions. First, we propose a framework to align regulatory expectations with what is technically and organizationally plausible, articulating the conditions under which each mechanism is possible, where they fall short, and what is required to make them meaningful in practice. Second, we outline how supervision methods should be documented and integrated into risk management, and drawing on the Microsoft Responsible AI Maturity Model, we outline a maturity model for AI supervision. Third, we explicitly highlight some boundaries of these mechanisms, including where they apply, where they fail, and where it is clear that no existing methods suffice. This foregrounds the question of whether meaningful supervision is possible in a given deployment context, and can support regulators, auditors, and practitioners in identifying both present and future limitations.

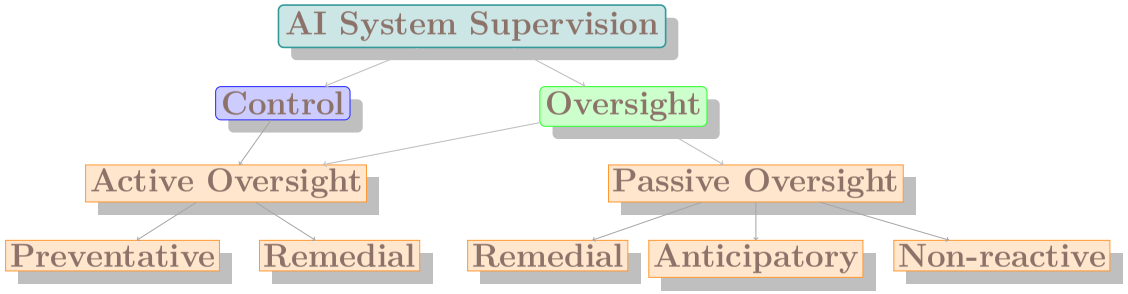
AI Supervision for Oversight and Control

David Manheim[1,2] and Aidan Homewood[3]

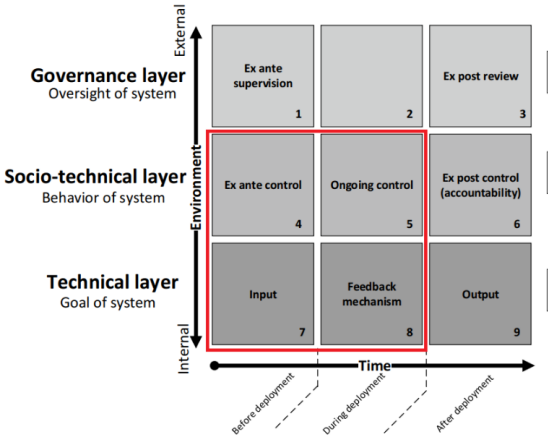
[1] Association for Long-Term Existence and Resilience, Rehovot [2] Technion Israel Institute of Technology, Haifa [3] Centre for the Governance of AI, London

AI governance and regulatory discourse often blurs real-time operational control with ex-post or supervisory oversight. Both are complex, and they are not always possible.

To enable proper supervision of AI systems, we supply a detailed reporting framework to ensure that risks are addressed, and present a 5-level maturity model.



Oversight ≠ Control



Maturity Model

5 Risks and Mitigations Public Before Training - Public Records of Failures and Updates

4 Risk Identification Matches Oversight - All Publicly Documented Before Deployment

3 Full Risk Map - Clear Oversight Documentation - Unaddressed or Unmapped Risks

2 Patchy Risk Register - Partial or Nonpublic Documentation - Minimal Mapping

1 No Risk Definitions - No Documentation - Claimed Oversight is Safetywashing