Chapter 6
Architecting Trustworthy and Resilient AI Systems: Adversarial Threats, Forensic Intelligence, and Governance in Cloud-Native Environments

Madhu Babu Amarappalli
http://orcid.org/0009-0009-2723-0256
Deloitte Consulting LLC, USA

ABSTRACT

Artificial intelligence (AI) and large language models (LLMs) increasingly underpin cyber defense workflows triaging alerts, correlating telemetry, classifying malware, generating incident narratives, and accelerating analyst decision-making. Yet the same properties that make AI operationally valuable (learning from data, generalizing across contexts, and automating at scale) also expand the attack surface. Adversaries can poison training data, implant hidden backdoors, steal model behavior through APIs, induce privacy leakage, or exploit prompt-manipulation weaknesses in LLM applications. When these models are deployed cloud-natively via containerized inference services, retrieval-augmented generation (RAG), agentic toolchains, and continuous delivery pipelines attack vectors multiply across datasets, MLOps supply chains, identity layers, orchestration planes, and third-party dependencies. This chapter presents a unified framework for architecting trustworthy and resilient AI systems under adversarial pressure. We synthesize adversarial AI threat models.

DOI: 10.4018/979-8-3373-8252-4.ch006

1. INTRODUCTION

Econometric analysis is increasingly defined by data abundance and data fragility. Administrative microdata, high-frequency financial records, platform logs, and sensor streams now let researchers study behavior at unprecedented scale and granularity often in near real time. Yet these gains arrive with a structural cost: the empirical environment is routinely shaped by heterogeneous measurement systems, nonstationary behavior, and rare but consequential events. In such settings, the assumptions that made many classical estimators reliable in smaller, cleaner datasets thin tails, stable regimes, well-behaved errors, and "mostly correct" measurement are regularly strained or violated. What looks like a single "outlier" in a textbook example may, in modern pipelines, reflect any of the following: a data pipeline fault, a coding change in reporting, a sudden policy shock, a coordinated response to incentives, or an emergent economic disruption. The practical implication is that anomalous observations are no longer peripheral: they are recurring features of the data-generating process and can meaningfully shape both estimation and interpretation.

Robust statistics provides conceptual vocabulary for designing analysis procedures that remain stable under contamination. Influence functions formalize local sensitivity: they describe how an estimator reacts to a small amount of contamination placed at a particular point in the sample space, thereby making "fragility" a measurable design property rather than a vague concern (Hampel et al., 1986). Breakdown points, in turn, capture resistance to large-scale corruption by quantifying the largest fraction of contaminated observations that can drive an estimator toward arbitrarily misleading results (Hampel et al., 1986; Maronna et al., 2006). These ideas are not abstract embellishments: they directly motivate operational choices such as bounded-influence loss functions, robust scale estimation, and outlier-resistant multivariate procedures. Huber's foundational M-estimation framework illustrates the central tradeoff clearly: by replacing pure quadratic loss with a loss that grows more slowly in the tails, one can sharply reduce the impact of extreme residuals while retaining high efficiency under benign conditions (Huber, 1964). In other words, robustness is not about ignoring anomalies, it is about ensuring that their presence does not silently dominate empirical conclusions.