Chapter 5
Long-Term Monitoring and Performance Evaluation of AI Systems

Dhanalakshmi Jaganathan
https://orcid.org/0000-0002-3919-4469
SRM Institute of Science and Technology, India

Prabhu Chakkaravarthy
https://orcid.org/0009-0000-3180-8051
SRM Institute of Science and Technology, India

ABSTRACT

With AI increasingly used in healthcare, finance, and autonomous systems, ongoing monitoring is vital to ensure performance, fairness, and ethical compliance. Long-term oversight helps control risks, improve systems, and maintain trust. Monitoring frameworks detect issues like model drift, bias, or degradation caused by data shifts, user behavior, or environmental changes—prompting regular retraining. Beyond accuracy, key metrics include fairness, transparency, robustness, efficiency, and compliance. Explainability ensures stakeholders understand AI decisions, while robustness testing checks performance under varied inputs. AI observability tools with logging and analytics detect abnormal behavior early. Human oversight complements automated systems to catch subtle issues and ensure social responsibility. Monitoring many complex AIs remains a challenge, demanding better tools for transparency and adaptation. Regular testing, human involvement, and governance ensure AI stays effective, ethical, and lawful. These practices support responsible AI development and foster user trust.

DOI: 10.4018/979-8-3373-2787-7.ch005

INTRODUCTION

Importance of Long-Term AI Monitoring

A significant number of industries—such as healthcare, finance, and autonomous systems—are increasingly dependent on artificial intelligence (AI) systems to manage critical operations, deliver insights, and automate complex decision-making tasks (Amershi et al., 2019). However, the deployment of an AI model is not the end of the development lifecycle. Rather, it marks the beginning of an ongoing journey requiring persistent monitoring and performance evaluation to ensure that these systems remain accurate, fair, and trustworthy over time (Sculley et al., 2015).

Long-term monitoring is essential because AI models do not operate in static environments. Real-world data distributions, user behaviors, and external conditions continuously shift, sometimes in subtle ways that may not be immediately apparent. Over time, these changes can lead to issues such as model drift, where the AI's predictions become less reliable or even biased as the input data evolves away from the data it was originally trained on (Gama et al., 2014). Early detection of such gradual problems—like emerging biases, declining accuracy, or shifts in user interaction patterns—enables organizations to intervene before these issues severely impact outcomes.

Persistent oversight also plays a crucial role in addressing evolving ethical, regulatory, and societal expectations. As AI becomes more integrated into decision-making processes that affect individuals and communities, stakeholders—including end users, regulators, and advocacy groups—demand transparency in how these systems function. Monitoring frameworks that include explainability features make it possible to trace, justify, and audit the decisions made by AI, which is vital to maintaining public confidence (Mitchell et al., 2019).

Moreover, long-term monitoring supports continuous improvement through retraining and adaptation. When performance evaluations reveal that a model's effectiveness is waning—whether due to changing data patterns, new regulatory requirements, or shifts in business objectives—organizations can proactively update the model. This process helps prevent the AI from becoming obsolete or producing unintended negative consequences, such as perpetuating outdated biases or making unsafe recommendations (Chen et al., 2020).

Trust in AI systems is fundamentally linked to their ability to operate reliably and fairly. Continuous monitoring reassures users and regulatory bodies that AI is not producing harmful, discriminatory, or erroneous outcomes. For instance, in applications like healthcare, a lapse in model accuracy or fairness could result in misdiagnosis or unequal treatment, while in finance, it could lead to unjust loan decisions. Therefore, robust monitoring practices are not merely technical safeguards