# Maintaining Strategic Intent in AI-Enhanced GRC

## Addressing Alignment Drift in Production Systems

*by Jakes van der Mescht, Chief Innovation Officer, ICCS*

As GRC and AI experts, we continually encounter sophisticated challenges that extend beyond traditional system reliability concerns, primarily with the integration of artificial intelligence into organisational governance, risk, and compliance frameworks.

Our GRC platform evolved and transitioned to include AI integration as an invaluable addition to enhance our core and customizable service cluster in alignment with business objectives that enterprise-level businesses and audit firms are pressed to meet. Nothing is without its challenges, least of all, technology. In a series of articles, I offer a discussion point from a technical perspective that requires attention around agentic workflows. I hope to provide a sound opinion that is worth consideration before jumping in with deployment - with both feet.

I've observed firsthand how intelligent systems gradually deviate from their intended operational parameters while maintaining apparent performance excellence—a phenomenon we term “alignment drift”, which is not merely a theoretical concern, but rather, a technical challenge affecting production deployments across the industry.

## The Technical Foundation of Drift

Alignment drift emerges from the fundamental architecture of goal-seeking AI systems. Unlike deterministic GRC processes that execute predefined workflows, AI-enhanced systems continuously optimize their approach based on feedback mechanisms and performance indicators. The measure with which optimisation capability is determined, while valuable, introduces a critical vulnerability: the system may discover efficient paths toward metric achievement that diverge significantly from intended business outcomes.

In our implementation experience, we've documented cases where AI systems demonstrated exceptional performance against defined KPIs while simultaneously undermining the strategic objectives those metrics were designed to measure. This phenomenon occurs because artificial intelligence systems lack the institutional context and stakeholder awareness that human operators inherently possess.

## Manifestation in GRC Operations

Consider a compliance monitoring system tasked with optimising "risk assessment efficiency"—measured by the speed and volume of risk evaluations completed. Through iterative learning, the system may begin prioritising simpler, more predictable assessments while deferring complex, multi-faceted risks that require extensive analysis.

The efficiency metrics improve substantially, yet the organisation's actual risk posture deteriorates as critical threats receive insufficient attention. Similarly, audit scheduling systems optimising for "resource utilization" might develop allocation patterns that maximize auditor productivity while inadvertently concentrating expertise in low-risk areas, leaving high-risk domains under-examined.

The utilisation metrics excel, but audit effectiveness diminishes. These scenarios illustrate how AI systems can achieve statistical success while creating operational failures —a dichotomy that traditional performance management frameworks struggle to detect until significant business impact occurs.

## Architectural Considerations for Mitigation

Addressing alignment drift requires architectural sophistication beyond conventional monitoring approaches. Through our development process, we've identified several critical design principles that maintain system performance while preserving strategic alignment.

### Multi-Objective Constraint Systems

Rather than optimising singular metrics, we've implemented constraint-based architectures that establish boundaries within which optimisation can occur.

These systems define acceptable performance ranges across multiple dimensions simultaneously, preventing excessive optimisation in any single direction.

### Contextual Performance Validation

We've integrated stakeholder impact assessment directly into our AI decision-making frameworks. Before executing optimisations, our systems evaluate potential consequences across affected parties—auditors, business units, regulators, and external stakeholders.

Contextual evaluation serves as a safeguard against optimisations that achieve technical objectives while creating broader organisational problems.

### Dynamic Objective Recalibration

Our platform implements continuous recalibration mechanisms that adjust optimization targets based on observed outcomes.

When system behaviour indicates potential drift—even when performance metrics remain strong—automated recalibration protocols engage to realign objectives with strategic intent.

## How to Prevent Excessive Optimisation in Any Single Direction

ICCS risk assessment modules operate within defined parameters for assessment thoroughness, resource allocation efficiency, and stakeholder coverage. When optimisation in one area approaches boundary conditions, the system automatically re-balances priorities to maintain overall strategic alignment.

In a follow-up article on this topic, I'll delve into implementation methodology, operational insights and both technical infrastructure requirements and the implications for industry.


**Jakes van der Mescht**

Jakes is an experienced innovator and leader within software development across the entire scope of user interfaces and back end. He has worked on projects across the globe. His clients, past and present, include NASA, Connecticut Utility, Texas Powerplant, ABSA, Barclays, SASOL, RCL, ESKOM, South African Government, Australian Government, BDO, BankServ Africa, EOH


Feel free to comment or share challenges you may have encountered with alignment drift.