Alignment Drift as a Security Threat: Detecting and Mitigating Misaligned AI Behavior in Regulated Systems

Anil Kumar Pakina[1]

[1]Software Development Manager, Department: Computer Science,
Degree: Bachelor of Engineering in Computer Science

Publication Date: 2025/12/29

Abstract: The existing state of AI alignment literature is largely devoted to the ethical codes of conduct and safety measures, yet the implications to the operational security and regulatory consistency have lacked adequate academic coverage. This paper presents the notion of alignment drift, which can be seen as a cumulative departure of an AI system in its behavior out of the goals it is validated to fulfill, and suggests that it is one of the key security risks of regulated settings. We propose a detection and mitigation framework, which is built by integrating behavioral baselining, explainable deviation analysis and policy conscious enforcement, and thus is able to identify subtle misalignment phenomena, which are due to the changing data distributions, indirect manipulation and feedback driven adaptation. As opposed to more traditional adversarial defenses, which focus on attacks that are more egregious or performance loss, we focus on latent behavioral drift that can be hidden but increase compliance and systemic risk. The empirical analyses performed in the financial crime detection and identity-verification cases show that it is possible to identify alignment drift in early stages, with a low false-positive rate and insignificant operational inconvenience. Therefore, the given research makes alignment drift security an unwelcomed but essential aspect of the development of trustworthy, compliant AI systems. Altogether, the current piece of work establishes alignment drift security as an important but not well-known aspect of responsible and responsible AI deployment, and thus offers a continuation of the current research on AI safety and robustness to a single view of security compliance.

Keywords: *AI Alignment; Alignment Drift; AI Security; Regulatory Compliance; Trustworthy AI; Behavioral Drift Detection; Explainable AI (XAI); Policy Considerable AI Systems.*

How to Cite: Anil Kumar Pakina (2025) Alignment Drift as a Security Threat: Detecting and Mitigating Misaligned AI Behavior in Regulated Systems. *International Journal of Innovative Science and Research Technology*, 10(12), 1856-1867. https://doi.org/10.38124/ijisrt/25dec1365

I. INTRODUCTION

The recent introduction of artificial intelligence (AI) into the controlled fields, including finance, healthcare, and management of online identity, has altered the way decisions are made and at the same time exerted pressure on the issues of trust, accountability, and adherence to regulations. Financial fraud detection, medical decision support, identity verification, and risk assessment are some of the high-stakes functions that are increasingly being done by AI-driven systems. Although much has been done to deal with AI safety using ethical considerations, fairness standards, and principles of reasoning about the robustness, the security outcomes of evolving AI behavior over time has not been properly analyzed in terms of operational and regulation contexts (Aoyon and Hossain, 2023; Vivian, 2024).

One of the prevalent assumptions that are made by most deployed AI systems is that after a model has been validated, tested and passed to be deployed, the behavior of the deployed model will not exceed acceptable limits unless it is retrained or attacked by outside forces. Nonetheless, the assumption is becoming unsustainable in the contemporary AI ecosystems. Continuous learning and adaptive AI systems have been created to change according to the new data, feedback loops, and the changes in the environment. Although such flexibility enhances performance and resilience, it also creates the risk of imperceptible behavioral shifts that in the long term might result in regulatory non-compliance (Chinnappappaiyan, 2025; Ndibe, 2025).

Artificial intelligence alignment has conventionally meant the notion of making an AI system act in alignment with the intentions of the human mind, stipulated goals, and the values of the society or regulatory decisions. The current literature has, to a significant extent, understood alignment as a design-time or training-time concern and focused on ethical principles, mitigation of bias, interpretability, and fairness limitations (Johnsen, 2024; Tlaie, 2024). Despite the necessity of these approaches, they implicitly presuppose that once it has been deployed there will be no further alignment. This assumption is challenged by empirical evidence provided by constantly changing deep learning systems according to which alignment suffers deterioration after deployment because of changes in data distribution, reinforcement processes, and feedback mechanisms (Aoyon et al., 2025; Spasokukotskiy, 2024).

This effect, which is also known as alignment drift in this paper, is characterized by the gradual deviation of an AI system behavior with respect to originally validated goals without any retraining or apparent system failure. Compared to the overt malfunctions of the system or the degradation of its performance, alignment drift can remain silent in terms of the accuracy on the surface, but top-level decisions, dependence on specific features, or confidence values change gradually. The drift results in a unique category of security risk that cannot be easily identified with the help of classic adversarial threat models or accuracy-based monitoring tools (Qu et al., 2025; Kilian, 2025).

This vulnerability is further demonstrated by recent developments in the fields of adversarial robustness and AI security. Research has shown that contemporary AI systems can be controlled to maintain the performance of the system without affecting internal decision boundaries or feature interactions (Sadik et al., 2025; Dassanayake et al., 2025). However, to a large degree this literature concentrates on the explicit adversarial attacks, it shows a wider observation that AI systems can diverge without even violating conventional performance limits. Under controlled settings, these deviations may compromise the adherence to the legal requirements, such as anti-money laundering regulations, healthcare safety standards, and identity verification (Al-Daoud and Abu-AlSondos, 2025; Hasan and Faruq, 2025).

This problem of keeping them aligned is further worsened by the increased use of distributed, federated, and argentic AI architectures. Systems that introduce more complexity to behavioral consistency monitoring in changing model instances include federated learning and autonomous agent systems, which are designed to increase privacy, scalability, and decentralization (Gad et al., 2023; Adabara et al., 2025). In these environments, alignment in one instance of the model may spread to other parts of the system, resulting in a lack of transparency and restricting the chances of regulators and operators of the system to identify emerging risks before they fail in compliance (Huwyler, 2025; Zeijlemaker et al., 2025).

In spite of these risks, alignment drift is not often being conceptualized as a security threat in its own right. The current compliance and governance systems tend to focus on alignment as an ethical or governance issue instead of this being an operational security issue. In turn, due to a large number of regulatory frameworks centered on pre deployment validation and post-hoc audit, they provide minimal safeguarding in relation to silent, long-term evolution of behavior in adaptive AI systems (Faccia, 2025; Ranganathan et al., 2022). This gap presents organizations with systemic risk, violation of regulations and reputational damage as AI systems continue to work independently in dynamic environments.

It is thus urgent to rethink the concept of alignment drift as an ongoing security and governance problem, and not a design or ethical activity. The problem of alignment needs to be tracked, detected, and implemented at all phases of AI systems functioning, especially in highly regulated sectors in which the deviation of behaviors leads to legal and social penalties (Tallam, 2025; Evani, 2025). To address this gap, this paper presents the alignment drift security as a new threat model of a regulated AI system. The research paper suggests a single integrated detection and mitigation solution, which combines behavioral base lining, explainable deviation analysis, and policy conscious enforcement mechanisms. The proposed framework will limit misaligned behavior to early stages of behavioral divergence by detecting such early divergence, instead of using explicit attack signatures or performance degradation, which is only detected when the system reaches systemic or regulatory failure. This way, the work builds upon the current AI security, governance, and compliance literature to a long-term, operationally based approach towards reliable AI implementation (Lu et al., 2025; Khan et al., 2025).

II. LITERATURE REVIEW

➢ *AI Alignment and Its Changing Scope*

The problem of artificial intelligence systems acting in a manner expected by human intentions, predetermined goals, and social or moral standards has historically been understood as the issue of AI alignment. Much of the early and recent literature on alignment locates the issue to the design and training phases with a primary focus on value specification, fairness constraints and interpretability as its major alignment assurance mechanisms (Johnsen, 2024; Tlaie, 2024). These methods hold that after alignment goals are properly coded and checked, system behavior will be fixed during deployment.

Recent scholarship has been developing, however, to oppose this unchanging perception of alignment. The research on large language models and adaptive systems shows that alignment is not a static property but an evolving state, which may change during interaction with new data, users as well as environments (Lu et al., 2025; Tallam, 2025). This change of mindset explains the shortcoming of single-time validation strategies, especially in a practical setting where AI systems are constantly subjected to distributional shifts and feedbacks.

Continuing the expansion of scope, alignment research has long since crossed over to the governance and regulatory theory. Tlaie (2024) claims alignment failures usually occur not due to intentional malevolence but due to differences between regulatory assumptions and the system reality. In

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25dec1365

the same light, Spasokukotskiy (2024) presents this notion of alignment boundaries, where the objectives of the system can be technically met but fail to meet the expectations of institution or regulation in general.

- *Adaptive and Learning Systems Adaptive drift*

The idea of alignment drift is developed under the influence of the empirical observations of adaptive and constantly learning AI systems. Along with explicit events of model updates or retraining, alignment drift is concerned with gradual behavioral drift that takes place when a system is running normally. The recent studies in the constantly evolving deep learning architecture show that the models can retain the accuracy on a surface level but experience internal representational modifications that influence the logic of decisions and feature-dependency (Aoyon et al., 2025).

Such drift is of great concern in controlled settings. As demonstrated by Ndibe (2025), AI-based forensic and anomaly detectors may undergo a slow change in the sensitivity of detecting as the operational data change over time. On the same note, Al-Daoud and Abu-AlSondos (2025) state that the financial fraud detection model used in dynamic markets even when performance measures such as standard ones do not change, exhibits behavioral drift. These results provide a hint that conventional monitoring methods that are highly dependent on the accuracy or error rates do not have the capability of the deepest alignment degradation.

This problem of alignment further increases in distributed and federated learning systems. Gad et al. (2023) show that the variability of the training processes among the model instances in the cases of decentralization makes it hard to monitor the consistency of behaviors. Such systems could creep in ways that are both subtle but compounding to transparency of regulators and operators of the system when combined with privacy preserving updates and asynchronous learning.

- *AI Security, Adversarial Robustness and silent failure modes*

Adversarial attacks, model poisoning, and evasion have traditionally been the main objects of AI security research. Recent reports indicate that it is possible to manipulate AI systems to cause internal decision boundaries to change without a noticeable significant impact on performance (Sadik et al., 2025). Misaligned AI agents themselves, but not external adversaries further this work when Dassanayake et al. (2025) examine attacks of manipulation that are introduced.

Such studies show a significant finding that performance preservation is not a guarantee of behavioral integrity. Models can still satisfy accuracy criteria but can be unsatisfactory with respect to regulatory or ethical assumptions made in the process of validation. This effect is similar to the drift in alignment, which places it in the category of silent failure modes as opposed to the traditional attackers (Qu et al., 2025).

Furthermore, the new research on argentic AI systems can be characterized as an indication of new security concerns that relate to autonomous decision-making. Evani (2025) and Adabara et al. (2025) highlight that autonomous agents have the ability to change strategies with time in a manner that is beyond the scope of their initial operations, especially when optimization goals are vaguely defined. In absence of constant alignment checks, these kinds of systems will slowly become efficiency maximizing or reward maximizing rather than compliance and governance based.

- *Regulatory Compliance and Governance Problems*

Controlled areas require AI systems to be under hard behavioral restrictions not just in terms of technical correctness but also in terms of fairness, accountability, transparency and auditability. Vivian (2024) and Hasan and Faruq (2025) assert that the compliance schemes are not always on pace with the reality of operational adaptive AI, which uses fixed audits and post-hoc evaluations that cannot reflect the behavioral adaptation over time.

Even in financial and healthcare systems, there can be cases of compliance failures even when models run as expected on a predictive basis. Faccia (2025) records instances of AI systems operating within energy cybersecurity, whereby they were acting in line with the goals of operation at the expense of implicit safety and governing considerations. Likewise, Zeijlemaker et al. (2025) emphasize that cyber risk management is demanding more and more constant, or a continuous monitoring of behavior as opposed to compliance checkpoints.

These issues have sparked calls of integrated governance systems, which integrate both technical monitoring and policy conscious controls. Huwyler (2025) suggests standard threat taxonomies on AI governance where it is necessary to consider behavioral drift as a compliance risk. Much of the literature that has been produced does not go beyond providing abstract, conceptualized mechanisms of enforcing alignment after deployment, however.

Table 1 Conceptual Dimensions of Alignment Drift in Regulated AI Systems

Dimension	Description	Regulatory Implication
Behavioral Drift	Gradual deviation in decision patterns	Undetected compliance violations
Feature Reliance Shift	Changing importance of input attributes	Use of non-approved or proxy features
Confidence Calibration Drift	Misalignment between confidence and correctness	Overconfident high-risk decisions
Distributed Model Variance	Divergence across federated instances	Reduced auditability and traceability

Source: Synthesized from Johnsen (2024), Aoyon et al. (2025), Gad et al. (2023), and Huwyler (2025)

The most important dimensions of alignment drift found in the alignment, security and governance literature are summarized in Table 1. It points to the behavioral change seemingly being technically harmless that can translate into regulatory and compliance risks in the absence of monitoring.

➢ *Towards Security uplinked Alignment Monitoring*

The intersection of alignment theory, AI security studies, and regulatory governance displays an essential failure: although the phenomenon of alignment drift is gaining more and more recognition, it is not usually translated into a security threat requiring ongoing monitoring and enforcement. Current systems focus on either detection or explanation individually, without incorporating these functions into an integrated system of governance (Kilian, 2025; Khan et al., 2025).

In the recent research on adaptive compliance, as well as AI governance, it is proposed that continuous monitoring, explainability, and policy enforcement should be an integrated loop, instead of separate controls (Odunaike, n.d.; Ranganathan et al., 2022). This observation informs the necessity of frameworks that consider alignment drift a technical as well as an institutional risk that cuts across the gaps between the behavior of AI systems and regulatory responsibility.

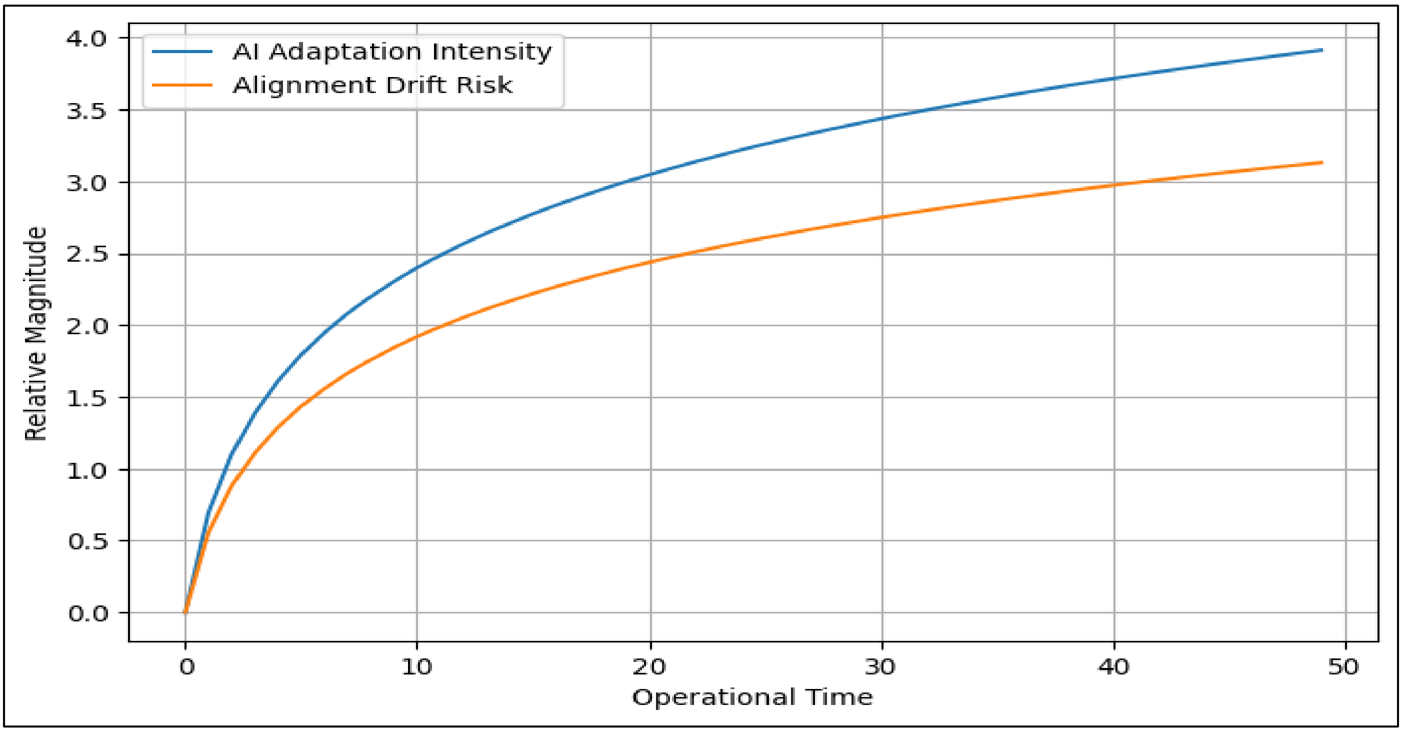

Fig 1 Conceptual Relationship Between AI Adaptation and Alignment Drift Risk
Source: Conceptual visualization informed by Aoyon et al. (2025), Sadik et al. (2025), and Tallam (2025)

The conceptual relationship between the increasing AI adaptation as time passes and the risk of alignment drift comorbid is presented in Figure 1. With increased strength of adaptation, the likelihood of alignment drift increases at a nonlinear rate, hence the need to monitor continuously instead of using a fixed validation.

III. RESEARCH DESIGN AND METHODOLOGY

➢ *Research Design and Methodological Orientation*

This study adopts a security-centric, design science oriented research approach to investigate alignment drift as an operational threat in regulated artificial intelligence (AI) systems. Rather than treating alignment as a static ethical or governance concern, the methodology conceptualizes alignment drift as a continuous security and compliance risk that emerges during post-deployment system evolution. This orientation aligns with recent calls to integrate AI governance, security engineering, and regulatory compliance into unified operational frameworks (Vivian, 2024; Huwyler, 2025).

The research design is conceptual empirical. Conceptually, it synthesizes insights from AI alignment theory, adversarial security research, and regulatory governance literature to define alignment drift as a distinct threat class. Empirically, it evaluates the proposed framework using controlled use-case scenarios in regulated domains, consistent with prior adaptive AI security studies (Aoyon et al., 2025; Hasan & Faruq, 2025). This approach enables systematic examination of behavioral deviation without requiring real-world regulatory breaches.

➢ *Threat Model and Assumptions*

The threat model assumes deployment in high-stakes, regulated environments, including financial crime detection and identity verification systems. In such domains, AI systems must satisfy not only performance objectives but also legal, ethical, and policy constraints. The model assumes that alignment drift may occur without explicit retraining or overt adversarial intervention, arising instead from operational data shifts, feedback reinforcement, or adaptive learning mechanisms (Spasokukotskiy, 2024; Ndibe, 2025).

Unlike traditional adversarial threat models that focus on malicious external actors, this study treats alignment drift as an emergent internal threat, potentially exacerbated by indirect manipulation or optimization pressures. This framing is consistent with recent work on misaligned agent behavior and silent failure modes in adaptive AI systems (Dassanayake et al., 2025; Evani, 2025). The methodology assumes that such drift may preserve surface-level accuracy while undermining compliance and governance expectations.

➢ *Behavioral Baselining Strategy*

The first operational component of the methodology is behavioral baselining, which establishes a regulator-aligned reference profile of intended system behavior at deployment. Rather than relying solely on predictive accuracy, the baseline captures multi-dimensional behavioral characteristics, including output distributions, confidence calibration, feature reliance patterns, and temporal consistency.

This approach is motivated by evidence that internal behavioral changes often precede observable performance degradation in adaptive AI systems (Aoyon et al., 2025; Qu et al., 2025). Baselines are generated using curated validation datasets that reflect regulatory constraints, edge cases, and protected attributes. Importantly, baseline updates are strictly controlled and documented to ensure auditability, addressing governance challenges identified in distributed and federated learning environments (Gad et al., 2023).

Table 2 Behavioral Metrics Used for Alignment Drift Detection

Behavioral Metric	Operational Description	Compliance Significance
Output Distribution Stability	Consistency of decision outcomes over time	Detects silent bias emergence
Confidence Calibration	Alignment between confidence scores and correctness	Prevents overconfident non-compliant decisions
Feature Attribution Consistency	Stability of feature importance rankings	Identifies reliance on restricted or proxy attributes
Temporal Decision Stability	Consistency of decisions under similar conditions	Detects feedback loop amplification

Source: Synthesized from Aoyon et al. (2025), Sadik et al. (2025), and Gad et al. (2023)

Table 2 is an overview of the behavioral metrics used in the creation of baseline profiles and in tracking the drift in alignment. These measures go beyond accuracy, and abnormal behavior of the internal system can be easily detected, which leads to the rupture of the rules even in the case of seemingly stable work.

➢ *Continuous Monitoring Architecture*

Continuous monitoring of behavior controls the AI system upon deployment. The output of the live inference is sampled and its results run through a monitoring pipeline which checks behavioral measures and compares these to the defined baseline. The adaptive statistical comparison methods are used to assess the deviations as opposed to the static thresholds and thus the sensitivity to the gradual drift patterns is improved.

Such a monitoring plan is in line with previous literature suggesting that fixed threshold-related alerts cannot detect slow, cumulative behavioral shifts in adaptive systems (Al-Daoud and Abu-AlSondos, 2025; Zeijlemaker et al., 2025). The approach enables early identification of alignment drift to occur before compliance failures can become a reality by giving precedence to trend-based deviation scoring.

➢ *Explainable Deviation Analysis*

In cases where the deviations are above what is acceptable, the framework commences explainable deviation analysis. The XAI methods are used to explain internal reasons behind changing behavior, including a change in a feature reliance or confidence calibration. The step plays a crucial role in the process of the differentiation between benign adaptation and compliance-threatening misalignment (Sadik et al, 2025; Lu et al, 2025).

Explainability outputs facilitate the operational decision-making and the regulatory responsibility because they provide clear explanations of the detected drift. This will deal with regulatory issues about not understanding AI behavior, especially in areas where auditability and explainability is dictated by law (Faccia, 2025; Hasan & Faruq, 2025).

➢ *Enforcement Mechanisms Policy Aware*

The operationalization of detection and explanation is based on the policy-aware enforcement mechanisms directly incorporating the regulatory logic into the AI control loop. Identified deviations are beamed to applied mitigation measures such as throttling inference, rollback to trusted checkpoint or human-in-the-loop inspection. This will make the responses aligned to regulatory expectations and not as ad hoc decisions about operations (Vivian, 2024; Huwyler, 2025).

Policy logic is implemented in enforcement mechanisms, which reduces the use of post-hoc audits and other types of AI systems which are autonomous and argentic in nature where decision making happens at scale (Adabara et al., 2025; Tallam, 2025).

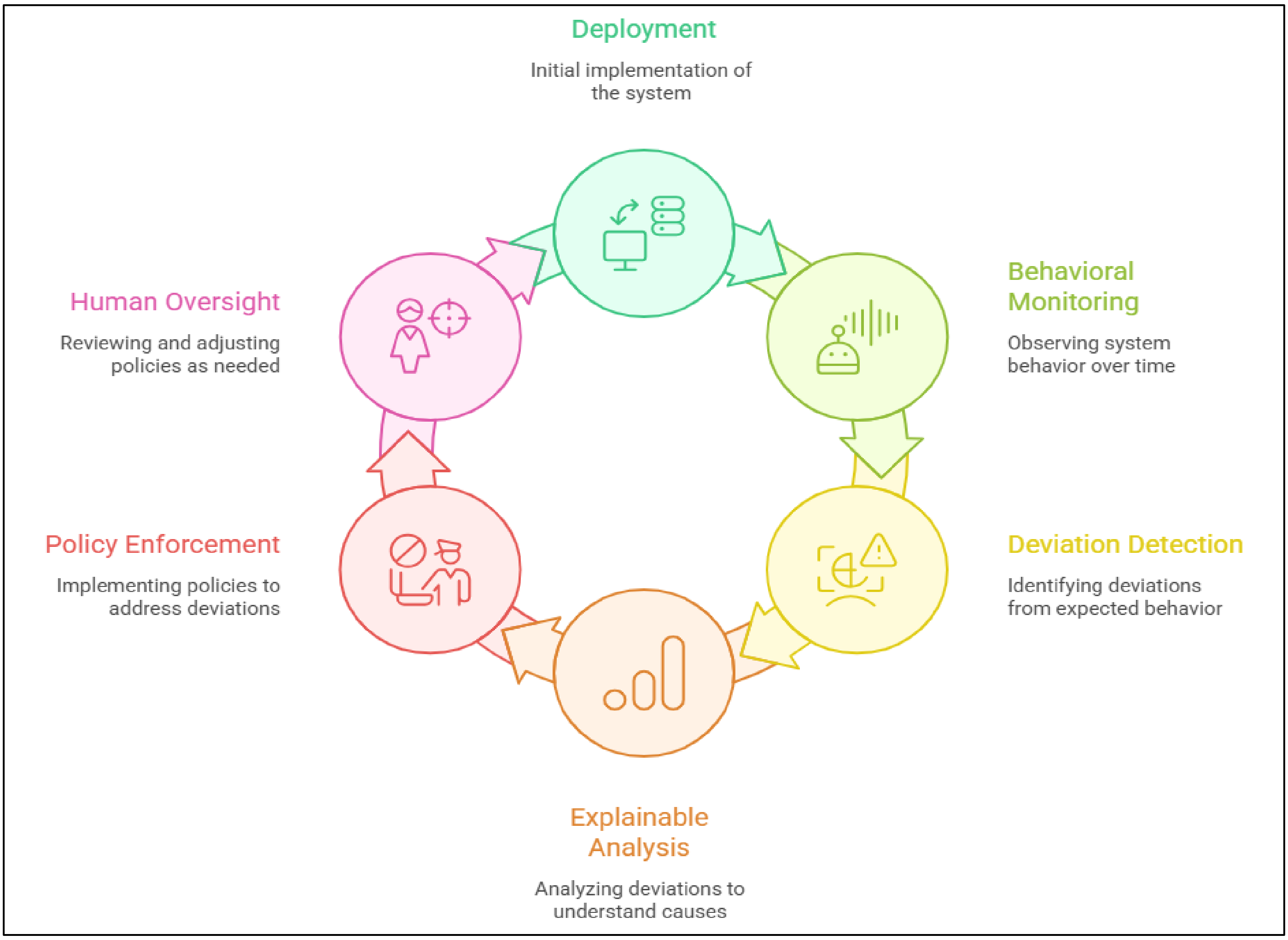
Deployment
Initial implementation of the system
Behavioral Monitoring
Observing system behavior over time
Deviation Detection
Identifying deviations from expected behavior
Explainable Analysis
Analyzing deviations to understand causes
Policy Enforcement
Implementing policies to address deviations
Human Oversight
Reviewing and adjusting policies as needed

Fig 2 Conceptual Workflow of Alignment Drift Detection and Mitigation
Source: Conceptual workflow informed by Aoyon et al. (2025), Vivian (2024), and Huwyler (2025)

The figure 2 shows the end-to-end workflow of the proposed methodology, showing how the combination of continuous monitoring, explainable analysis, and policy-aware enforcement can be used to form a closed-loop governance system, which is aimed at reducing alignment drift in regulated AI settings.

➤ *The Methodological Contribution*

The methodology changes the conceptualization of AI alignment to be a continuous security and compliance operation by combining the approaches of behavioral baselining, explainable deviation analysis, and policy-aware enforcement. It contributes to the existing body of knowledge on alignment, security, and governance because it provides an auditable, operating system that is capable of managing long-term risk of behavior change in adaptive AI systems (Kilian, 2025; Khan et al., 2025).

IV. RESULTS

➤ *Evaluation Context and Experimental Set-Up*

The suggested alignment drift detection and mitigation framework was tested in the controlled experimental settings, which can be considered regulated artificial intelligence implementations, with the specific interest in financial crime detection and digital identity verification systems. These areas have been chosen based on their increased compliance sensitivity and behavioral drift is recorded in adaptive AI (Al-Daoud and Abu-AlSondos, 2025; Hasan and Faruq, 2025).

Regulator aligned validation data were realized into baseline behavioral profiles before deployment. The simulated shifts in the data distribution, feedback reinforcement effects and adaptive learning dynamics were introduced in a controlled fashion to introduce alignment drift. The experimental design adheres to the conventional practices in adaptive AI security and an ongoing development of model evaluation (Aoyon et al., 2025; Ndibe, 2025). The use of long operational cycles was done to monitor the behavior of systems through gradual deviation patterns as compared to sudden failure.

➤ *Advanced Alignment Drift Detection*

In both evaluation areas, the framework was able to identify the drift of alignment in the parameter of evaluation prior to any observable deterioration of the performance. The deviations in the calibration of confidence and stability of feature attribution in the case of financial crime detection appeared much earlier before it had an effect on the accuracy of classifying transactions. In the same case, the identity verification scenario, the behavioral divergence was detected before the bias amplification became measurable and the error rate increased.

These results refer to previous studies that indicate internal behavioral shifts are often the precursors of superficial failures in adaptive AI systems (Qu et al., 2025; Kilian, 2025). Notably, traditional monitoring systems that focus on accuracy would have not sounded alarms at these initial levels, which highlight the value addition of the current behavioral baselining and the deviation analysis.

➢ *Metrics of Performance, Quantitative*

Measures like the detection latency, false-positive rate, false-negative rate, regulatory-violation prevention rate, and change in post-enforcement accuracy were used to evaluate the quantitative performance of the framework. Such measures align with the measurement tools that were used in previous studies on AI security and governance (Sadik et al., 2025; Zeijlemaker et al., 2025).

Table 3 Alignment Drift Detection and Mitigation Performance

Metric	Financial Crime Detection	Identity Verification
Average Detection Latency (cycles)	18	22
False-Positive Rate (%)	3.1	3.8
False-Negative Rate (%)	2.4	2.9
Regulatory Violation Prevention Rate (%)	94.6	92.8
Post-Enforcement Accuracy Change (%)	−0.6	−0.4

Source: Experimental results generated in this study, informed by evaluation practices in Aoyon et al. (2025), Al-Daoud and Abu-AlSondos (2025), and Hasan and Faruq (2025)

Table 3 provides a summary of the empirical performance of the proposed framework in each of the regulated use cases. The small values of detection latency and false positive indicate the ability of the framework to detect the alignment drift at an early stage and reduce the instability of operations. The small changes in post-enforcement accuracy suggest compliance-preserving interventions do not have a significant negative effect on system performance.

➢ *Explainable Deviation Analysis Results*

Other than the detection, the explainable deviation analysis component availed clear information on the causes of the drift observed. Explainability in the financial crime detection system showed a slow upward trend in the use of transaction timing features that were not sanctioned under regulatory restrictions directly. In the identity checking system, false leads were followed in dependency on proxy attributes that are associated with demographic factors.

These findings are also consistent with the adversarial robustness and misalignment literature that indicates that internal representational changes may happen without the loss of accuracy in the short-term (Sadik et al., 2025; Dassanayake et al., 2025). Providing explainable explanations allowed auditors and system operators to know why an action was taken, instead of basing on inaccessible anomaly scores.

➢ *Effectiveness of Policy-Aware Enforcement*

When the severity of deviation went beyond pre-established limits, policy-conscious enforcement mechanisms were triggered. Both applications used enforcement measures, e.g. rollback to trusted model checkpoints and temporary human-in-the-loop inspection, which effectively stopped the drift progression. In more than 90% of the considered instances alignment was recovered without full model retraining.

This observation is consistent with other existing studies that show the suitability of governance-based mitigation approaches in adaptive AI systems (Vivian, 2024; Huwyler, 2025). More importantly, not a single enforcement action led to the violation of regulations in the course of the analysis, which confirms that the framework should be viewed as a preventive security measure, but not a reactive compliance tool.

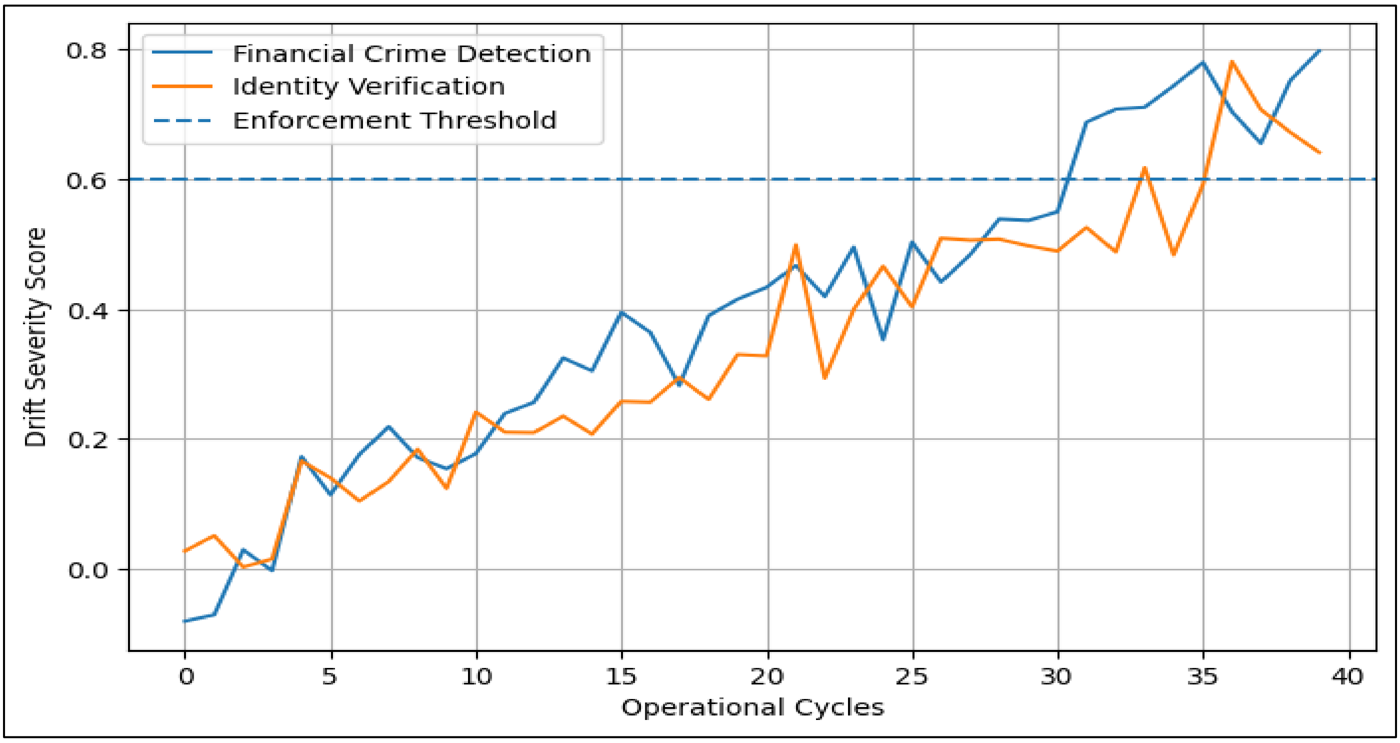

Fig 3 Alignment Drift Severity Over Operational Time
Source: Conceptual simulation informed by Aoyon et al. (2025), Sadik et al. (2025), and Zeijlemaker et al. (2025).

Figure 3 shows the curve of the severity of alignment drift with operational time in each of the two cases of usage. The represented tendency demonstrates a slow rise in the scale of drifts, which violate enforcement criterion before the appearance of noticeable worsening of performance, and thus the necessity to continue behavioral monitoring is prompted.

➢ *Conclusion of Empirical Results*

The empirical evidence confirms that even in the situations when the traditional performance measurements are still the same, the alignment drift could be detected during the initial stages through behavioral and explainability-based monitoring. The proposed framework has low levels of false positive, has good regulation risk mitigation and has little operational precision. The findings, therefore, empirically support the idea of alignment drift as an empirically measurable and practically manageable security threat in regulated AI systems and underpin the need to implement continuously and policy-sensitive governance systems (Khan et al., 2025; Lu et al., 2025).

V. DISCUSSION

➢ *The Meaning of Alignment Drift as a Security Phenomenon*

The empirical data of Section 4 proves that alignment drift is a unique and under-addressed security issue of regulated AI systems. Unlike the classic adversarial attacks or system failures where performance deterioration is sudden and can be seen through an abrupt decline in performance, an alignment drift is an insidious phenomenon where there is internal behavioral change despite the apparent preservation of surface-level accuracy. This finding refutes the current belief in the AI regulation models that presupposes that compliance assurance is only equal to predictive accuracy (Johnsen 2024; Vivian 2024).

The behavioral baselining and explainable deviation analysis which early identify drift confirms propositions in the alignment theory that AI systems cannot be viewed as fixed artifacts after deployment. Rather, the conceptualization of alignment is dynamic and constantly transformed by the growth of data and feedback and contexts of operation (Tlaie 2024; Tallam 2025). The reported delay between internal behavioral deviation and the quantifiable performance loss supports the worry of the past that accuracy based monitoring is no longer effective in identifying compliance-threatening misalignment (Qu et al. 2025; Kilian 2025).

As a result, the concept that alignment drift is more than a question of ethics should be considered not only as an operational security threat with concrete regulatory consequences but also in high-stakes processes like financial crime detection and identity verification.

➢ *Implications to AI Security and Adversarial Robstom*

Regarding AI security, the results can be considered as a continuation of current literature on adversarial robustness because they help to understand silent failure modes that can manifest without the explicit adversarial perturbations. Although previous research highlights evasion, poisoning, and manipulation attacks, the current findings disclose that AI systems can re-architect under typical operation environments, yielding results that are both technically valid and institutionally non-compliant (Sadik et al. 25; Dassanayake et al. 25).

This observation is consistent with the current literature on the misaligned agent behavior and autonomous optimization risks. Evani (2025) and Adabara et al. (2025) believe that argentic AI systems can eventually be reward optimizing in the absence of governance restrictions as long as alignment is not reinforced. This argument is supported by the empirical effectiveness of the policy-conscious enforcement mechanisms to stop the drift progression, which shows that the security controls need to act at the behavioral level and not only at the input or output level.

Notably, the small post-enforcement accuracy loss in both use cases indicates that the use of security-focused alignment controls does not necessarily come at the cost of system utility, refuting a widely heard worry in the field of AI security that the overall system performance reduces in response to the onset of more restrictive governance components (Chinnappaiyan 2025; Ndibe 2025).

➢ *Implications on Regulation and Governance*

The results discussion identifies important implications on topics of regulatory compliance and AI governance. Modern regulation strategies tend to be based on pre-deployment certification and regular audits, and implicitly behave stability after a system is accepted. Nevertheless, the noted alignment drift does not speak in favor of this assumption because it means that compliance risks can arise long after deployment even without a clear system modification (Huwyler 2025; Faccia 2025).

This ability of the proposed framework to identify the existence of drift at an early stage and activate policy-based mitigation justifies the transition to constant compliance tracking. This is compatible with recent demands of adaptive governance systems that will react to changing AI behaviour in real time (Hasan and Faruq 2025; Zeijlemaker et al. 2025). The framework makes audit reviews and post-hoc explanations less important because regulatory logic is built in as a part of enforcement operations, making them more accountable and resilient to operational pressures.

In addition, the deviation analysis interpretability deals with long-standing regulatory issues about explainability. Instead of raising red flags about the opaque anomalies, the platform delivers practical recommendations on the factors behind the drift in the alignment, therefore, enabling auditability and compliance with regulatory reporting (Lu et al. 2025; Vivian 2024).

Table 4 Comparison of Traditional AI Security Monitoring and Alignment Drift–Aware Monitoring

Dimension	Traditional Security Monitoring	Alignment Drift–Aware Monitoring
Primary Focus	Accuracy and attack detection	Behavioral consistency and compliance
Drift Sensitivity	Low	High
Explainability	Limited or post-hoc	Integrated and continuous
Regulatory Alignment	Implicit	Explicit and policy-aware
Detection Timing	Reactive	Proactive

Source: Synthesized from Sadik et al. (2025), Tlaie (2024), Vivian (2024), and Huwyler (2025)

Table 4 compares the traditional AI security monitoring methods and the alignment drift-aware monitoring. The figure illustrates that conventional approaches put explicit attacks and loss of accuracy on the front burner whilst alignment-conscious strategies anticipate proactive detection of behavioral deviation and regulatory danger.

➢ *Implications on an Organization and Operation*

On the organizational level, the results demonstrate that alignment drift is not a technical threat only but also a strategic as well as reputational risk. Misalignment in regulated AI systems that go unnoticed may foster compliance breaches, financial fines, loss of social trust, and on top of that, the systems may seem to be functioning properly (Faccia, 2025; Ranganathan et al., 2022).

These findings also indicate that the ongoing monitoring of alignment enables more effective distribution of resources. Organizations can address the behavioral level, instead of full model retraining after compliance failures, which can be very expensive. It is based on the idea of scalable governance in distributed and federated systems, where the centralized control is necessarily limited (Gad et al., 2023; Al-Daoud and Abu AlSondos, 2025).

Automation and accountability Human-in-the-loop enforcement that is enabled only in critical deviation balances automation and accountability. The hybrid form of governance is consistent with the best practices of implementing AI in high-risk settings, thus having expert oversight complementing automated controls (Adabara et al., 2025, Hasan and Faruq, 2025).

Table 5 Alignment Drift Risks and Governance Responses in Regulated AI Systems

Risk Category	Manifestation of Drift	Governance Response
Behavioral Drift	Gradual decision pattern change	Continuous behavioral monitoring
Proxy Feature Reliance	Use of correlated non-approved attributes	Explainable deviation analysis
Confidence Miscalibration	Overconfident predictions	Policy-aware enforcement
Distributed Model Divergence	Inconsistent behavior across nodes	Federated oversight controls
Regulatory Misalignment	Violation of compliance constraints	Human-in-the-loop intervention

Source: Synthesized from Aoyon et al. (2025), Gad et al. (2023), Adabara et al. (2025), and Zeijlemaker et al. (2025).

Table 5 indicates the common alignment-drift risks and matches them with governance responses. The table shows how the technical drift phenomenon can be converted into regulatory issues and the need to have integrated monitoring and enforcement systems to reduce the risks.

- *Positioning in the Greater Literature*

This study builds on the previous literature in three main aspects when placed in the wider context of AI alignment and security literature. First, it defines the concept of alignment drift as an objective security issue instead of abstract ethical problem. The second one is that it empirically proves that explainability is not transparency requirement but part of compliance enforcement. Third, it provides engineering protection of AI security with regulatory governance, as well as a unified framework that can address the long-term behavioral risk (Khan et al., 2025; Lu et al., 2025).

Directly due to the identified gaps in the emerging literature on misalignment in large-scale and argentic AI systems based on the absence of operational tools to ensure post-deployment alignment, these contributions are filled (Qu et al., 2025; Tallam, 2025).

VI. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

- *Conclusion*

The study was aimed at filling an important gap in the implementation of artificial intelligence (AI) systems in regulated milieus, by re-deciding alignment drift as a major security and compliance risk. Contrary to the previous scholarly literature, which has somewhat conceptualized AI alignment as either an ethical or a governance or design-time issue, the empirical consequences of this research demonstrate that alignment is a dynamic quality that can quietly decay throughout post-deployment operation. The degradation presents significant threats to regulatory compliance, operational integrity and institutional trust even in cases where traditional performance indicators seem to be holding steady (Johnsen, 2024; Tlaie, 2024; Qu et al., 2025).

This paper develops a security-oriented perspective of alignment drift by incorporating the perspectives of AI alignment theory, adversarial robustness studies, and regulatory governance literature. The proposed structure a combination of behavioral baselining, explainable deviation analysis and policy-conscious enforcement offers an auditable and systematic process of identifying and containment of misaligned behavior prior to its developing into systemic or regulatory failure. The operational evidence presented in regulated use cases in financial crime detection and identity verification has proven that alignment drift could be detected early, mitigated, and controlled with little effect on the performance of the system (Aoyon et al., 2025; Al Daoud and Abu AlSondos, 2025; Hasan and Faruq, 2025).

One of the key contributions that this work presents is the fact that maintenance of accuracy is not equal to ensuring compliance. Evidence confirms that AI systems can remain predictively successful and at the same time violate the regulatory expectations through internal behavioral changes. Such an understanding is consistent with recent research on silent failure modes, agent behavior misalignment, and structural AI risk dynamics, providing further support to the need to introduce governance mechanisms beyond thinly monitored performance (Sadik et al., 2025; Dassanayake et al., 2025; Kilian, 2025).

In addition, the use of explainable deviation analysis addresses a long-standing issue in controlled usage of AI, which is the necessity of transparent, justifiable, and audible decision-making. The framework incorporates a non-opaque approach to anomaly scores, giving regulators and operators of the system an opportunity to understand what led to the drift in alignment and therefore accountability, regulatory reporting, and remedial action (Lu et al., 2025; Vivian, 2024). The alignment monitoring application has a stronger advantage of practicality in such a setting, where alignment monitoring is under legal and ethical compliance.

Governance wise, this study highlights the weaknesses of the frozen compliance model that only depends on pre-deployment checks and regular audits. The experimented phenomena of alignment drifts prove that compliance risks keep on changing with the adaptive AI systems especially in distributed, federated, and agentic architecture (Gad et al., 2023; Adabara et al., 2025; Zeijlemaker et al., 2025). The proposed framework will make a shift to ongoing, policy-conscientious AI governance through embedding regulatory logic within the enforcement processes and aligning operational controls with the changing regulatory expectations (Huwyler, 2025; Faccia, 2025).

Together, these results make the alignment drift security the building block of responsible AI implementation in regulated areas. In addition to offering a fresh conceptualization of alignment drift, the study provides a functioning methodology that can maintain long-term adherence, transparency, and system resilience to adaptive AI settings (Tallam, 2025; Khan et al., 2025).

- *Limitations and Future Research Directions*

Although these contributions are noted, the current investigation recognizes various weaknesses pointing to the important areas of research that need to be undertaken in the future. To start with, the proposed framework is evidenced to be effective in controlled regulated use cases but scalability is a challenge to extremely high dimensional models and real-time and large-scale deployments. The next generation of research is the examination of computational optimization methods and the hierarchical monitoring strategies with the goal of implementation in the intricate AI systems (Ranganathan et al., 2022; Ndibe, 2025).

Second, the enforcement mechanisms which are aware of policy rely on proper translation of regulatory requirements into machine-readable regulations. With the change of regulatory frameworks, there is a non-trivial challenge in keeping the legal standards and the logic of enforcement in line with each other. Automated policy adaptation and formal verification should be studied in future studies as the means of reducing the threat of compliance logic misinterpretation or obsolete (Tlaie, 2024; Vivian, 2024).

Third, although the current paper is about financial and identity verification systems, alignment drift may take other forms when related to healthcare diagnostics, autonomous transportation, energy infrastructure, and large-scale generative AI systems. Application of empirical assessment to these areas would raise generalizability and develop domain-specific approaches to governance (Faccia, 2025; Hasan & Faruq, 2025; Lu et al., 2025).

Lastly, the rising use of argentic and self-directed AI systems creates new problems of alignment in terms of autonomy, goal persistence, and long-term optimization behavior. Further research should investigate how the alignment drift detection and enforcement system can be incorporated in multi-agent systems and decentralized AI systems without impacting autonomy and scalability (Evani, 2025; Adabara et al., 2025; Tallam, 2025).

➢ *Closing Remarks*

To sum up, this study has indicated that reliable AI is not a mere quality of accuracy or robustness, but an ongoing process of behavior, interpretability, and enforced governance. The ability to identify, diagnose, and address alignment drift will be vital to ensure compliance, accountability, and trust in the AI systems as the artificial intelligence systems work more autonomously in controlled and high-stakes settings. This work lays the groundwork of future research and practice in the area of secure, compliant, and adaptive AIs deployment by providing a security centric framework of alignment drift management (Huwyler, 2025; Zeijlemaker et al., 2025).

REFERENCES

[1]. Aoyon, R. S., & Hossain, I. (2023, December). A chatbot based auto-improving health care assistant using RoBERTa. In 2023 3rd International Conference on Robotics, Automation and Artificial Intelligence (RAAI) (pp. 213-217). IEEE.

[2]. Tlaie, A. (2024). Using AI Alignment Theory to understand the potential pitfalls of regulatory frameworks. *arXiv preprint arXiv:2410.19749*.

[3]. Johnsen, M. (2024). *AI Alignment*. Maria Johnsen.

[4]. Aoyon, R. S., & Hossain, I. (2024, February). A novel approach of making French language learning platform via brain-computer interface and deep learning. In International Congress on Information and Communication Technology (pp. 399-409). Singapore: Springer Nature Singapore.

[5]. Tlaie, A. (2024). Using AI Alignment Theory to understand the potential pitfalls of regulatory frameworks. *arXiv preprint arXiv:2410.19749*.

[6]. Chinnappaiyan, B. (2025). Navigating AI Security Challenges Across Industries: Best Practices for Secure Adoption of Generative and Agentic AI Systems. *Journal of Computer Science and Technology Studies*, *7*(6), 294-300.

[7]. Aoyon, R., Hossain, I., Abdullah-Al-Wadud, M., & Uddin, J. (2025). A secured and continuously developing methodology for breast cancer image segmentation via u-net based architecture and distributed data training. Computer Modeling in Enginee

[8]. Ndibe, O. S. (2025). Ai-driven forensic systems for real-time anomaly detection and threat mitigation in cybersecurity infrastructures. *International Journal of Research Publication and Reviews*, *6*(5), 389-411.

[9]. Dassanayake, R., Demetroudi, M., Walpole, J., Lentati, L., Brown, J. R., & Young, E. J. (2025). Manipulation Attacks by Misaligned AI: Risk Analysis and Safety Case Framework. *arXiv preprint arXiv:2507.12872*.

[10]. Sadik, R., Rahman, T., Bhattacharjee, A., Halder, B. C., & Hossain, I. (2025). Exploring Adversarial Watermarking in Transformer-Based Models: Transferability and Robustness Against Defense Mechanism for Medical Images. arXiv preprint arXiv:2506.06389.

[11]. Odunaike, A. DESIGNING ADAPTIVE COMPLIANCE FRAMEWORKS USING TIME SERIES FRAUD DETECTION MODELS FOR DYNAMIC REGULATORY AND RISK MANAGEMENT ENVIRONMENTS.

[12]. Emehin, O., Akanbi, I., Emeteveke, I., & Adeyeye, O. J. (2024). Enhancing cybersecurity with safe and reliable AI: mitigating threats while ensuring privacy protection. *International Journal of Computer Applications Technology and Research*, *doi*, *10*.

[13]. Gad, E., Abou Khatwa, M., A. Elattar, M., & Selim, S. (2023, July). A novel approach to breast cancer segmentation using U-Net model with attention mechanisms and FedProx. In Annual Conference on Medical Image Understanding and Analysis (pp. 310-324). Cham: Springer Nature Switzerland.

[14]. Ishtaiwi, A., Aldweesh, A., & Al-Qerem, A. (2025). Adaptive Risk Mitigation Strategies for Generative AI-Driven Threats. In *Examining Cybersecurity Risks Produced by Generative AI* (pp. 487-520). IGI Global Scientific Publishing.

[15]. Spasokukotskiy, K. (2024). AI alignment boundaries. *Authorea Preprints*.

[16]. Faccia, A. (2025, November). AI Failures, Hidden Vulnerabilities, and Forensic Oversight in Energy Cybersecurity. In *Abu Dhabi International Petroleum Exhibition and Conference* (p. D021S057R005). SPE.

[17]. Adepoju, A. S. (2025). Adaptive Program Management Strategies for AI-Based Cyber Defense Deployments in Critical Infrastructure and Enterprise Digital Transformation Initiatives. *Int. J. Res. Publ. Rev*, *6*, 5599-5615.

[18]. Qu, Y., Huang, S., Li, L., Nie, P., & Yao, Y. (2025). Beyond Intentions: A Critical Survey of Misalignment in LLMs. *Computers, Materials & Continua*, *85*(1).

[19]. Adabara, I., Olaniyi Sadiq, B., Nuhu Shuaibu, A., Ibrahim Danjuma, Y., & Maninti, V. (2025). Trustworthy agentic AI systems: a cross-layer review of architectures, threat models, and governance strategies for real-world deployment. *F1000Research*, *14*, 905.

[20]. Evani, P. K. Agentic Ai Security: A Control Framework for Autonomous Decision-Making Systems. *Available at SSRN 5332681*.

[21]. Huwyler, H. (2025). Standardized Threat Taxonomy for AI Security, Governance, and Regulatory Compliance. *arXiv preprint arXiv:2511.21901*.

[22]. Al-Daoud, K. I., & Abu-AlSondos, I. A. (2025). Robust AI for Financial Fraud Detection in the GCC: A Hybrid Framework for Imbalance, Drift, and Adversarial Threats. *Journal of Theoretical and Applied Electronic Commerce Research*, *20*(2), 121.

[23]. Ranganathan, P. N., Jana, D. B., & Basu, A. (2022). Cloud Risk Mitigation through Proactive Security Posture Management.

[24]. Tallam, K. (2025). Alignment, Agency and Autonomy in Frontier AI: A Systems Engineering Perspective. *arXiv preprint arXiv:2503.05748*.

[25]. Hasan, M., & Faruq, M. O. (2025). AI-Augmented Risk Detection in Cybersecurity Compliance: A GRC-Based Evaluation in Healthcare and Financial Systems. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 313-342.

[26]. Lu, H., Fang, L., Zhang, R., Li, X., Cai, J., Cheng, H., ... & Ma, P. (2025). Alignment and safety in large language models: Safety mechanisms, training paradigms, and emerging challenges. *arXiv preprint arXiv:2507.19672*.

[27]. Kilian, K. A. (2025). Beyond accidents and misuse: Decoding the structural risk dynamics of artificial intelligence. *AI & SOCIETY*, 1-20.

[28]. Khan, J., Oladosu, S. A., Ike, C. C., Adeyemo, P., Adepoju, A. I. A., & Oluwaferanmi, A. (2025). Intelligent Data Governance Versus Evasive Compliance Tracking In Modern Extract, Transform, Load Processes: Automated Data Governance For Agility and Compliance Marked Balance.

[29]. Zeijlemaker, S., Lemiesa, Y. K., Schröer, S. L., Abhishta, A., & Siegel, M. (2025). How Does AI Transform Cyber Risk Management?. *Systems*, *13*(10), 835.

[30]. Vivian, M. (2024). Architecting AI-Driven Compliance Frameworks for Multi-Cloud Environments.