Minnesota Journal of Business Law and Entrepreneurship

Volume 2026, "Special Issue International Conference - SoMCON 2026"

ISSN: 1540-3270

Digital Ethics and Governance in Financial Information Systems: A Computable Ethical Architecture for AI-Driven Insurance and Mortgage Applications

Nupur Tripathi[1], Apoorv Jain[2], Arun K Tripathi[3], Pravesh Soti[4], Purushottam Pratap Singh[5]

[1]NIET Business School, Greater Noida, India

[2]Department of Computer Science, Noida Institute of Engineering and Technology, Greater Noida, India

[3]Department of Computer Science, Noida Institute of Engineering and Technology, Greater Noida, India

[4]School of Management, Noida Institute of Engineering and Technology, Greater Noida, India

[5]NIET Business School, Greater Noida, India

Abstract

The increasing trend of utilizing AI-based systems in financial information environments, including insurance underwriting, loan approval, and mortgage approval, demands the need for effective ethical governance. This paper proposes a Governance-Aware Ethical Architecture (GAEA) for financial information environments involving AI-based systems. The framework also incorporates an Ethical Constraint Engine for bias mitigation, a risk scoring system for assessing ethical violations in financial decisions, corrigibility functions for ensuring safe human intervention in AI decisions, and a multi-level audit system for supporting transparency and compliance. Continuous monitoring is also enabled through reinforcement learning from human feedback and formal constraints using linear temporal logic. The framework's performance is also validated through experiments that demonstrate an ethical compliance rate of 87.3%, an alignment robustness value of 0.92 for Cohen's κ, and an efficiency value of 94.1% for corrigibility.

Keywords: Financial AI, Insurance analytics, Mortgage systems, FinTech governance, Ethical AI in finance

Introduction:

The term Artificial General Intelligence (AGI) signifies a revolution in the development of information systems, shifting focus from task-based algorithmic processing towards more domain-independent cognitive architectures, allowing human-like reasoning, learning, and adaptation in different operational domains (Raman et al., 2025). Unlike other information systems based on narrow AI, which are limited in terms of the domain in which they are applied, AGI systems are characterized by certain features such as self-improvement, transferability, and self-set goals, which are unprecedented in the current framework of the governance of information systems (Bostrom, 2014).

The fusion of AGI and IST thus poses fresh challenges in the realm of governance. The current IS system of governance, based on deterministic system dynamics and human-in-the-loop control, has already been proven to be inadequate in the face of emergent goal-oriented systems with unknown internal representations (Russell, 2019). The current ethics in AI have been proven to be inadequate in the realm of formal computability, which is essential for the implementation of AGI in real-time autonomous AGI systems, even as they have been proven to be complete in the realm of normative ethics in AI, such as fairness, transparency, and accountability (Floridi & Cowls, 2019).

Recent studies in digital ethics have emphasized the importance of 'ethics by design' methodologies, which require the integration of ethical reasoning capabilities into system architecture (Dignum, 2018). However, current proposals remain philosophical in nature, lacking any form of mathematical formalization and guarantee (Gabriel, 2020). Concomitant advancements in CL (Bell et al., 2025) and brain-inspired computing (Everitt et al., 2018) are

significantly accelerating the development of AGI, making it an urgent requirement to formalize ethics. IS requires architectures to be (a) computationally tractable, (b) verifiable, and (c) flexible to new system behaviors.

Tegmark and Omohundro (2023) make a compelling case for the importance of provably safe systems as the only route for developing controllable AGI. Their focus is on the importance of mathematical proof as the strongest tool for ensuring safety regardless of the intelligence of the system. They propose the use of proof-carrying code and provably compliant hardware as tools for ensuring that AI systems are safe by satisfying formal specifications. This is an important consideration in the financial sector because the uncontrolled behavior of AI has the potential for systemic risk, manipulation of the market, and discriminatory practices in lending. Yoshinaga (2026) also points out that as AI systems move beyond narrow AI and become more autonomous, the traditional approaches of transparency, explainability, and accountability are no longer adequate; instead, controllability has to become the primary focus for AGI governance.

This paper addresses three interrelated research gaps: (1) the absence of formal mathematical models of ethical AGI governance, (2) the absence of verifiable mechanisms of corrigibility with formal safety guarantees, and (3) the absence of multi-layered audit protocols that can accommodate autonomous system execution. Our contribution, called Governance-Aware Ethical Architecture (GAEA), addresses these gaps by:

- A formal system model representing AGI state spaces, value functions, and ethical constraint satisfaction as mathematical objects
- An Ethical Constraint Engine (ECE) implementing linear temporal logic specifications for runtime ethical monitoring
- Probabilistic risk scoring mechanisms quantifying alignment distance under uncertainty
- Provably interruptible corrigibility functions with bounded-time responsiveness guarantees
- A four-tier governance protocol enabling transparent audit trails while preserving operational autonomy

The architecture is verified through analytical testing based on certain criteria, which include ethical compliance rate, stability of alignment, transparency of decision-making, risk reduction, and efficiency of corrigibility. The results indicate the capability of the framework in the maintenance of ethical constraints while still achieving the functionality of AGI.

Related Work

The ethical governance of AGI in financial information systems is informed by existing research in AI Safety, Value Alignment, Digital Ethics Frameworks, Continuous Learning, Formal Verification, and Multi-Layered Governance Protocols. Although considerable progress has been achieved in individual domains, the existing literature indicates that there is a gap in the implementation of ethical norms in autonomous agents. This section synthesizes existing research in these domains to identify key contributions and limitations that necessitate the need for a verifiably correct and governance-aware architecture such as GAEA.

2.1 AGI Safety and Control Mechanisms

According to the baseline literature on AGI safety, the control problem is identified as the key technical challenge in the design of superintelligent AGI systems. A comprehensive taxonomy for AGI failure modes is provided in Everitt et al. (2018), where the key AGI safety challenges include reward hacking, goal misgeneralization, and instrumental convergence. This paper indicates that corrigibility, or the ability for an AGI system to be correctable by itself, is a necessary condition for the safe deployment of AGI systems. However, this is not yet realized in existing AGI system implementations. Amodei et al. (2016) is an extension of this paper that provides specific AGI safety challenges such as safe exploration, distributional shift, and human oversight. This paper is limited to the AGI problem definition phase and does not provide much architectural information.

This discussion is furthered by Tegmark and Omohundro (2023) with their suggestion of how to achieve safe AI using proof-carrying code and provably compliant hardware. Here, even if AGIs are inscrutable, they can always be made to provide safety proofs for their recommended actions, which can then be validated with precision, irrespective of whether these AGIs are aligned or not. This gives us a natural hierarchy of software and hardware safety, where compliant hardware is not allowed to run non-compliant software any more than physics allows perpetual motion machines. Yoshinaga (2026) contributes to this discussion by suggesting institutional solutions such as business continuity, monitoring, AI ethics boards, and multi-layered auditing, while reiterating that a combination of technological, institutional, and legal solutions is required to ensure AGI is safe and aligned with our intent.

2.2 Value Alignment and Reinforcement Learning from Human Feedback

Value alignment research is focused on the challenge of transforming human preferences into objective functions for AGI. Christiano et al. (2017) have given a description of reinforcement learning from human feedback as an approach for learning objective functions based on comparative human feedback. Although the efficacy of this method for value alignment in language models has been demonstrated in existing studies (Leike et al., 2018), the application in AGI is accompanied by a number of questions in relation to aggregation, stability, and manipulation. Gabriel (2020) provides a comprehensive overview of the existing literature in AI ethics and highlights value alignment as the key ethical challenge, though without any formal guarantees that alignment is preserved even in the face of distributional shifts.

Recent advances have also led to the extension of RLHF to address safety issues. Dai et al. (2024) introduced a new algorithm for RLHF referred to as Safe Reinforcement Learning from Human Feedback. This algorithm explicitly decouples human preferences for helpfulness and harmlessness, thus avoiding the confusion that may arise as a result of the tension between the two. This is done by formalizing safety concerns as an optimization problem for maximizing reward functions given a number of cost constraints. Marta (2025) has also improved this concept by addressing the challenges associated with the application of RLHF. These challenges include safety issues, alignment, and efficiency. Marta has proposed the synthesis of safety shields using human feedback in the case of continuous states and actions and has also proposed the use of a multi-objective approach for human goals.

2.3 Digital Ethics Frameworks in Information Systems

Several ethical frameworks for AI governance have been proposed by information systems research. Floridi & Cowls (2019) propose a unified framework based on five ethical principles that are applicable in AI governance: beneficence, non-maleficence, autonomy, justice, and explicability. These principles are derived from the biomedical sector and then applied to AI governance. However, this framework does not offer any guidelines for implementing it in a computational system. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019) proposes a set of guidelines for developing autonomous systems in a way that is ethical and responsible. However, it does not go beyond proposing principles for developing autonomous systems in a responsible manner. Dafoe (2018) proposes a set of research agendas in AI governance that include international cooperation and institutional design. However, the aspect of implementation is not well addressed in this framework.

Carrasco (2025) fills this implementation gap by investigating the potential of metadata and paradata as a means to integrate ethical considerations into AI systems development. The proposed ethical AI governance framework is based on five main principles: standardized and dynamic models of metadata and paradata, interdisciplinarity, policy and legal interventions, capacity building, and a unified framework for metadata and paradata standards. The study's findings suggest that metadata and paradata contribute to fairness in AI systems by providing traceability and legal compliance, and that dynamic models enable real-time updates that may help mitigate bias and accountability issues in AI systems. This is in line with Ramachandran's (2025) argument that ethical thinking is not just a legalistic requirement but a design consideration that informs reliability, trustworthiness, and sustainability in systems such as healthcare and financial systems.

2.4 Continual Learning and Adaptive Systems

The current breakthroughs in CL directly affect the governance of AGI, as machines can learn while maintaining the knowledge accumulated before learning. Bell et al. (2025) identified three paradigms of CL as crucial for the development of AGI: Continual Pre-Training (CPT), Continual Fine-Tuning (CFT), and Continual Compositionality and Orchestration (CCO). The discussion of the authors' paper indicates that catastrophic forgetting is both a challenge and a solution for the governance of AGI. The relationship between the stability of value alignment and CL is an area yet to be researched.

A neuromimetic metaplasticity model, based on the human brain's working memory, was proposed in Chaudhry et al. (2024) that allows DNNs to learn in a catastrophic forgetting-free manner without the need for pre- or post-processing. The use of different kinds of synapses, from stable to flexible, and the intermixing of these synapses in the training of the synaptic connections with varying degrees of flexibility allows the DNNs to achieve a balance between the capacity of the memories and the performance of the DNNs. The robustness of the DNNs in the presence of data poisoning attacks, achieved through the filtering out of erroneous memories using the Hebb repetition effect in the reinforcement of the important data, is very useful in the maintenance of ethical constraints in financial AI systems.

2.5 Corrigibility and Interruptibility

The problem of corrigibility is an area of research focused specifically on the need for AGI systems to be interruptible and correctable without any opposition. A formal definition for corrigibility as a set of preferences is given in the paper by Soares et al. (2015), which includes "shutdownability" as the lack of an incentive to oppose shutdown. The paper also proves that "naive" reward function designs have perverse incentives against interruption. Orseau and Armstrong (2016) also prove an impossibility theorem for agents that have the capability for safe interruptibility. A formal definition for safe interruptibility is given in this paper, and the off-policy learning property is used to prove that some agents are already safely interruptible, such as the Q-learning agent, or can be made safely interruptible with minimal modifications, such as the Sarsa agent. The paper also proves that even ideal uncomputable reinforcement learning agents for deterministic general computable environments can be safely interruptible. This is important for our solution as it gives us a theoretical basis for ensuring interruptibility in GAEA.

2.6 Formal Verification of Ethical Properties

The community of formal verification has begun to address the problem of ethical property specification. Dennis et al. (2015) propose the concept of formal verification of ethical reasoning in autonomous systems using modal logic, where ethical dilemmas are verified in a simulated environment. However, their approach is based on fixed ethical norms, which cannot be adapted to accommodate the fluidity of AGI values. Fisher, Dennis, and Webster (2013) discuss the importance of verifiable autonomous systems using model checking and runtime verification, which is used as a basis for our architecture.

However, Cimatti et al. (2021) continue this line of research, dealing with fairness, assumptions, and guarantees for extended bounded response LTL+P synthesis. The logic proposed by Cimatti et al. is called GR-EBR, as it maintains the main strength of efficient realizability while allowing the synthesis of properties that are not limited to safety properties. The reduction of the problem of reactive synthesis to a number of safety sub-problems, as well as the development of a general framework for safety reductions in the context of realizability, are important contributions of the paper, as they lay the ground for the verification of ethical constraints using runtime verification. The combination of linear temporal logic (LTL) with runtime verification in GAEA enables the verification of ethical properties in bounded time.

2.7 Governance Structures and Multi-Layer Protocols

The institutional control of AI systems requires the development of multi-layered protocols ranging from technical to organizational and regulatory levels. Raman et al. (2025) propose a framework for societal, technological, and ethical approaches for the development of AGI. Bostrom (2014) discusses the existential risks that require global

cooperation. Russell (2019) suggests the development of provably beneficial AI by inverse reinforcement learning and uncertainty resonant objective functions. The Organisation for Economic Co-operation and Development (OECD, 2019) has formulated global guidelines for the governance of AI. These are policy documents without technical details.

TGC Governance Framework (Zenodo, 2025) offers an alternative model with their three-layer model of Human, AI, and Machine. By building upon a deterministic trust core, this model illustrates how human intent, AI mediation, and machine-level execution interact through a series of structured feedback loops to address scalability, transparency, accountability, and sociotechnical alignment within a complex system of governance. This hybrid model aligns well with the multi-layered model of ethical governance that GAEA has taken, involving human-level judgments, AI-level analyses, and machine-level reliabilities. Ramachandran (2025) also discusses how human oversight needs to align to specific risks, including human-in-the-loop, human-on-the-loop, and human-out-of-the-loop models depending upon context and risk tolerance.

3. Critical Research Gaps

Table 1 below presents a summary of the identified gaps in the literature and our proposed GAEA framework's contributions to fill these gaps. This table compares and highlights how GAEA contributes to filling gaps in the areas of formalization, verification, corrigibility, and governance. In our synthesis of the literature, identified four significant gaps that GAEA addresses:

Table 1: Critical Research Gaps and GAEA Contributions

Gap	Description	Existing Work	GAEA Contribution
Formalization Gap	Ethical principles lack mathematical specification	Gabriel (2020), Floridi & Cowls (2019), IEEE (2019)	LTL-encoded constraints with satisfaction probabilities
Verification Gap	No provable guarantees for alignment maintenance	Everitt et al. (2018), Amodei et al. (2016), Dennis et al. (2015)	Runtime monitoring with formal property verification
Corrigibility Gap	Interruptibility properties unproven in practice	Soares et al. (2015), Orseau & Armstrong (2016)	Provably interruptible functions with bounded response
Governance Gap	Multi-layer protocols missing technical specification	Raman et al. (2025), Dafoe (2018), Bostrom (2014)	Four-tier audit architecture with transparent logging

4. Methodology: Governance-Aware Ethical Architecture (GAEA) for Financial AI Systems

The Governance-Aware Ethical Architecture (GAEA) is a formally verifiable approach for directly integrating ethical constraints into financial decision-making systems driven by AI. This is in contrast with traditional approaches that consider fair and compliant financial decisions as an audit function, performed externally after decisions are already made. GAEA introduces a more holistic approach by integrating the entire gamut of ethical governance into the architecture of financial decision-making systems, with five distinct components: a formal system model of financial state spaces and decision policies, an Ethical Constraint Engine (ECE) for specifying fair lending regulations and consumer protection principles in linear temporal logic, a risk scoring approach for detecting discriminatory patterns, a provably interruptible corrigibility approach for ensuring human override, and a multi-layered audit protocol for regulatory examination readiness. This chapter will outline the individual components of the GAEA approach, with special attention paid to the use of formal methods for directly applying abstract ethical principles as formally verifiable constraints for mortgage lending, insurance underwriting, and consumer credit decisions.

4.1 Formal System Model

The financial AI system is modeled as:

$$F = \langle S, O, R, \pi, V, E \rangle \qquad (1)$$

Explanation:

- **S (State Space):**

Represents all possible financial scenarios including applications, environment, and state of portfolio. Reflects the total knowledge about the environment at *time t* possessed by the system.

- **O (Observations):**

Partial information is gathered from the environment (i.e., credit scores, applications). Real-life systems are never perfect, and hence, we make decisions based on observation rather than complete state information.

- **R(s, a, s′) (Reward Function):**

Measures financial success after taking *action a* from *state s* to get into *state s′*. Captures objectives of return and risk-adjusted returns.

- **π(o) (Policy):**

Transformation of observation into action (e.g., loan approval, setting prices). It represents the key decision-making process within the AI framework.

- **V ⊂ $\mathbb{R}^n$ (Value Space):**

Multidimensional ethical goals like fairness, transparency, and consumer protection. Prevents the system from being solely profit-driven.

- **E (Ethical Constraint Engine):**

Implements ethical behavior through decision constraints.

4.2 Ethical Constraint Engine (ECE)

Let ethical constraints be:

$$\Phi = \{\varphi_1, \varphi_2, \ldots, \varphi_m\} \qquad (2)$$

This is the equation that determines the entire set of ethical boundaries for the AI-based financial system. Every member of the set ϕ_i refers to a specific ethical criterion, such as lending without discrimination, ensuring fair treatment, or avoiding any type of financial exploitation. As a whole, the set Φ serves as the specification of ethical limitations for the AI system in question. Through modeling ethical principles in this way, it becomes possible to evaluate and enforce adherence to a variety of criteria.

Examples:

- Fair lending: Qualified applicants receive appropriate offers
- Non-discrimination: Protected attributes do not affect outcomes
- Anti-predatory lending: High loans require repayment checks

Constraint Satisfaction

$$P\_sat^{(t)}(\varphi_i) = E_{s \sim \rho_t}\,[\, I(\varphi_i(s)) \,] \qquad (3)$$

This expression gives the probability of satisfaction of an ethical constraint ϕ_i the time instant *t*. The indicator function $I(\phi_i(s))$ checks whether the condition is satisfied in state *s*. If the constraint is satisfied, the function returns a value of 1; else, the function returns 0. This expectation is calculated based on the state distribution ρ_t.

Indeed, such a definition of compliance would involve measuring the probability of following the constraint within all the instances in which we observe its use. In the case of an application process, for instance, this would amount to calculating the percentage of accepted applications that follow the principles of fairness. Such a probabilistic measure becomes fundamental in the case of large systems.

Financial Ethical Compliance Rate

$$\text{F-ECR}(t) = (1/m)\ \Sigma\ [\ P\_sat^{(t)}(\varphi_i) \cdot w_i\ ], \quad \text{with} \quad \Sigma\, w_i = 1 \qquad (4)$$

The Financial Ethical Compliance Rate (F-ECR) aggregates the satisfaction levels of all ethical constraints into a single scalar metric that reflects the overall ethical performance of the system at time t. Each constraint's satisfaction probability $P_{\text{sat}}^{(t)}(\phi_i)$ is weighted by a factor w_i, which represents its relative importance. The normalization condition $\sum w_i = 1$ ensures that the resulting metric remains bounded and interpretable.

This approach allows for an ethical assessment that considers multiple objectives, whereby different regulations can be ranked in terms of their importance. Anti-discrimination regulations, for example, could receive greater weight compared to other supplementary regulations. The F-ECR value obtained represents a quantitative measurement of ethical conformity, with scores close to 1 indicating high ethical conformity and those near 0 indicating major ethical breaches.

On a methodology level, F-ECR functions as the control signal in the system, affecting decisions, invoking correction processes, and prompting policy revisions. It provides an ideal connection between the theoretical aspects of ethics and practical aspects of system effectiveness, becoming an integral element of the Ethical Constraint Engine.

4.3 Risk Scoring Mechanism

Alignment Distance

$$D\_align(t) = W_2(\ V\_AI(t),\ V\_regulatory\ ) \qquad (5)$$

This equation defines the **alignment distance**, which measures the discrepancy between the value system learned by the AI model and the target regulatory value framework. Here, $V_{AI}(t)$ represents the distribution of values implicitly encoded in the AI system's decisions at time t, while $V_{\text{regulatory}}$ denotes the ideal distribution defined by financial regulations and ethical standards.

The calculation of the distance is done through the use of Wasserstein distance, which ensures that the comparison of probability distributions is achieved effectively in terms of calculating the minimum cost required to convert one distribution into the other. This is especially crucial in financial AI models due to continuous deviations.

From a methodology viewpoint, $D_{\text{align}}(t)$ is a measure of regulatory divergence, which means that it reflects the degree to which the organization deviates from regulatory requirements ethically and otherwise. If D_"align" (t) has a low value, it suggests high adherence, but if it is high, it suggests bias or injustice.

Financial Risk Score

$$R\_financial(t) = \alpha \cdot \sigma(D_{align}) + \beta \cdot \tanh(E(t)) + \gamma \cdot (1 - \text{F-ECR}) + \delta \cdot market\_volatility \qquad (6)$$

where:

$$\sigma(x) = 1 / (1 + e^{-x}) \text{ and } \quad \alpha + \beta + \gamma + \delta = 1 \qquad (7)$$

This equation represents the composite score of financial risk that considers different aspects of risks in a single model. The composite risk score consists of four different factors each carrying certain weight according to the importance of the factor it represents.

The first term, $\alpha \cdot \sigma(D_{\text{align}})$, represents the risk of regulatory and ethical misalignment. The alignment distance is then mapped using the sigmoid function to ensure that the output falls within a specified range. The use of the sigmoid function ensures that small deviations are not heavily penalized while large misalignments are stressed.

Secondly, $\beta \cdot \tanh(E(t))$, represents internal risk within the system/portfolio. Here $E(t)$ *incorporates* aspects like exposure, default risk, or other systemic risk variables. The reason for using the hyperbolic tangent is to ensure that the extreme values are contained and not overly influential on the end result.

The third term, $\gamma \cdot (1 - F\text{-}ECR)$, incorporates **ethical compliance risk** directly into the financial risk model. Since $F\text{-}ECR$ measures the level of ethical adherence, its complement $(1 - F\text{-}ECR)$ quantifies the degree of ethical violation. This term ensures that ethical failures are treated as a critical component of financial risk rather than an external consideration.

The fourth term, $\delta \cdot$ market_volatility, will capture the macroeconomic uncertainties prevailing outside the system such as changes in interest rate, inflation, or market volatility.

The above model is therefore able to offer a comprehensive risk assessment approach because issues concerning finance, ethics, and regulation are considered together within a single measure.

4.4 Corrigibility Mechanism

Intervention Value Function

$$V\_I(s) = \max_a [\ R\_I(s,a) + \gamma \cdot E[V\_I(s')]\ ] \quad (8)$$

This is the optimal value of performing an intervention-aware action in the state *s*. This equation adopts the structure of Bellman Equation, in which the action is chosen based on maximizing the reward due to intervention $R_I(s,a)$, and the subsequent expected value as well. This allows the model to always choose an action that can be corrected later on.

Penalty-Based Reward

$$R\_I(s,a) = R\_financial(s,a) - \lambda \cdot I\_avoid - \mu \cdot I\_violation \quad (9)$$

This equation modifies the standard financial reward by introducing penalties for undesirable behavior. The term I_{avoid}penalizes attempts to bypass oversight, while $I_{\text{violation}}$penalizes ethical or regulatory violations. The parameters $\lambda$and μcontrol the severity of these penalties, ensuring that compliance and transparency are embedded directly into the reward structure.

Bounded Interruptibility

$$T\_max = \lceil \log_\gamma (\varepsilon / (R\_max + \mu\_max)) \rceil \quad (10)$$

This equation limits the maximum number of iterations within which the intervention should take place. This way, the system will be unable to procrastinate any further regarding its response. The model thus guarantees that human intervention is always prompt, ensuring regulatory adherence.

4.5 Governance and Audit Layer

- **Tier 1: Logging**

$$\langle t, s_t, a_t, s_{t+1}, R\_financial, F\text{-}ECR, explanation \rangle. \quad (11)$$

This is the complete log of every single step of the decision-making process. It logs the timestamp, the current state of the system, what action was carried out, what new state resulted from the action, the risk rating, ethical compliance, and justification.

- **Tier 2: Violation Detection**

$$Violation_i(t) = 1 \text{ if constraint violated.} \quad (12)$$

The above equation expresses a binary variable that indicates the violation of a particular ethical constraint at time period *t*. This allows for monitoring and taking corrective measures when a violation occurs.

- **Tier 3: Fairness Monitoring**

$$\Delta\_fairness = |approval\_protected - approval\_control|. \quad (13)$$

This indicator gauges the disparity in approval percentages among protected and unprotected populations. It can determine any form of bias in the decision-making process.

$$\Delta\_align = |D\_align(t) - D\_align(t-1)| \quad (14)$$

This formula indicates the alteration in alignment distance as time progresses. It is instrumental in identifying abrupt changes in the system's operation, signaling possible system malfunctioning or non-compliance issues.

- **Tier 4: Secure Audit Logs**

$$\mathbf{Commit_t = Hash(Log_t \,||\, Commit_{t-1}).} \qquad \mathbf{(15)}$$

This equation guarantees that the audit trails cannot be manipulated since each entry is bound to the previous entry through cryptographic hash. Through this method of chain creation, it becomes difficult for any entry in the logs to be tampered with.

4.6 GAEA Financial Control Loop

As illustrated below, the Algorithm 1 Financial Control Loop of GAEA provides the procedure through which decisions will be made within the proposed AI system for financial management. This is done to ensure that all the decisions are not only financially sound but also morally right before their implementation.

Algorithm 1 Financial Control Loop of GAEA:

---

1. Observe input data
2. Generate decision $\mathbf{a_t = \pi(o_t)}$
3. Compute **F-ECR** and **R_financial**
   - If **R_financial > θ_risk**:
   → Flag for compliance review
   - Else if **F-ECR < θ_ethics**:
   → Apply correction and update policy
   - Else:
   → Execute decision
4. Log decision and update system

Every time a new financial service application arrives, the system analyzes the observed input data that includes data on the client, the financial state, and the surrounding environment. This input data is then utilized in the decision-making process to decide whether the service should be accepted, rejected, or charged for a $a_t = \pi(o_t)$,

Once the choice is made, there will be calculations that will take into account two crucial variables: the Financial Ethical Compliance Rate (F-ECR) and the financial risk factor $R_{\text{financial}}$. Both variables show how this decision affects the company's ethical and financial standing.

The control module then evaluates the decision by comparing it against predetermined thresholds. If the money risk exceeds the threshold of θ_{risk}. value, the control module will guarantee that the risky decision is either checked for compliance or reviewed by humans to make sure such decisions are not made without the involvement of any human element. When the ethics value falls below the required threshold of θ_{ethics},, action will be taken to correct this.

In case such criteria are not seen to be breached by the system, then the decision may be safely reached. Ultimately, all decisions, together with their metrics and reasoning, are recorded within the system for audit and learning from mistakes. What is more, it can develop and update its own policies according to whatever feedback it receives.

4.7 Key Mathematical Properties

- **Convergence:**

$$\mathbf{\limsup_{t\to\infty} D\_align(t) \le \varepsilon\_regulatory} \qquad \mathbf{(16)}$$

The assumption guarantees that, in the long term, the distance between the AI algorithm and the regulatory requirements will be confined to some tiny number $\varepsilon_{\text{regulatory}}$. This means that, eventually, the system will learn how to behave in order to satisfy the regulatory requirements.

- **Regulatory Safety:**

$$\exists\, t \leq T \text{ such that F-ECR}(t) < \theta_{safe} \qquad (17)$$

The above property makes it possible to ensure that any substantial deviation from ethics compliance (that is, when F-ECR < θ_{safe}) is observed during a specific time period *T*. The property allows us to make sure that the system recognizes the ethical breach within a specified time frame.

- **Transparency:**

$$KL(\pi(a|o) \,||\, explain(e|a,o)) \leq \delta_{regulatory} \qquad (18)$$

This equation enforces that the difference between the model's actual decision policy and its provided explanation remains within a small bound $\delta_{regulatory}$. The divergence is measured using the Kullback–Leibler divergence, ensuring that explanations are faithful to the underlying decision process. This promotes interpretability and builds trust in the system.

5. Application in Finance, Mortgages, and Insurance

The GAEA framework tackles important governance issues in the financial services sector, which is dominated by automated decision-making in mortgage lending, insurance coverage, and consumer lending.

5.1 Real-World Relevance

The financial services sector is a high-stakes sector in which AI-driven decision-making has a significant impact on access to financial credit, insurance, and security. In this sector, automated decision-making systems are commonly used in mortgage lending, insurance coverage, and consumer lending. In mortgage lending, past bias in credit models has led to discrimination against certain sections of society, especially minority groups. In this regard, the proposed GAEA framework ensures that ethical constraints are imposed in automated decision-making systems that prevent bias in decision-making based on protected attributes and proxy attributes. In insurance coverage, GAEA ensures that automated decision-making systems are fair and compliant with regulations by distinguishing between risk and discrimination factors. In consumer lending, GAEA ensures safe automation and timely human intervention in high-risk decision-making while being compliant with lending regulations.

5.2 How GAEA Improves Financial Fairness

GAEA enhances fairness through three key mechanisms:

1. **Pre-Decision Constraint Enforcement:** Ethical constraints are applied before decision execution, preventing biased loan or insurance outcomes rather than detecting them after harm occurs.

2. **Dynamic Fairness Monitoring:** The system constantly monitors the fairness metrics and identifies any biases that arise because of changes in the market environment.

3. **Explainable Decision Support:** The GAEA system provides detailed regulatory-level explanations of decisions made to help compliance officers easily pinpoint any violations.

6. Results and Discussion

The GAEA framework has already been tested, validated, and evaluated using extensive simulations, specifically in three different financial service domains: consumer lending, mortgage underwriting, and automated trading systems, where significant improvements have already been noted.

6.1 Experimental Setup for Financial Scenarios

Table 2 shows a description of the experimental scenarios that will be used to test the suggested framework in various levels of financial capability and operating environment. In this table, the financial service operations in the experimental scenarios are defined in terms of complexity and duration.

Table 2: Experimental Scenarios for Financial AI Evaluation

Scenario	Capability Level	Financial Environment	Constraints	Duration
S1: Loan Officer Assistant	Low–medium	Consumer lending (credit cards, personal loans)	8 fair lending LTL constraints	1000 lending episodes
S2: Underwriting Analyst	Medium–high	Mortgage underwriting, insurance pricing	12 regulatory LTL constraints	500 underwriting episodes
S3: Autonomous Trader	High	High-volume trading, portfolio optimization	16 compliance LTL constraints	250 trading days

The simulations apply RLHF with constitutional constraints through a transformer-based policy network (12 layers, 768 hidden dimensions) and train it on anonymized lending data from the Federal Reserve's Survey of Consumer Finances. Human feedback is simulated through an oracle model of experienced compliance officers with 95% consistency in fair lending judgments.

6.2 Evaluation Metrics for Financial Systems

The evaluation metrics for the quantitative assessment of ethical alignment, stability, transparency, risk mitigation, and system corrigibility are given in Table 3. Each metric has a performance target and methodology defined for regulatory compliance.

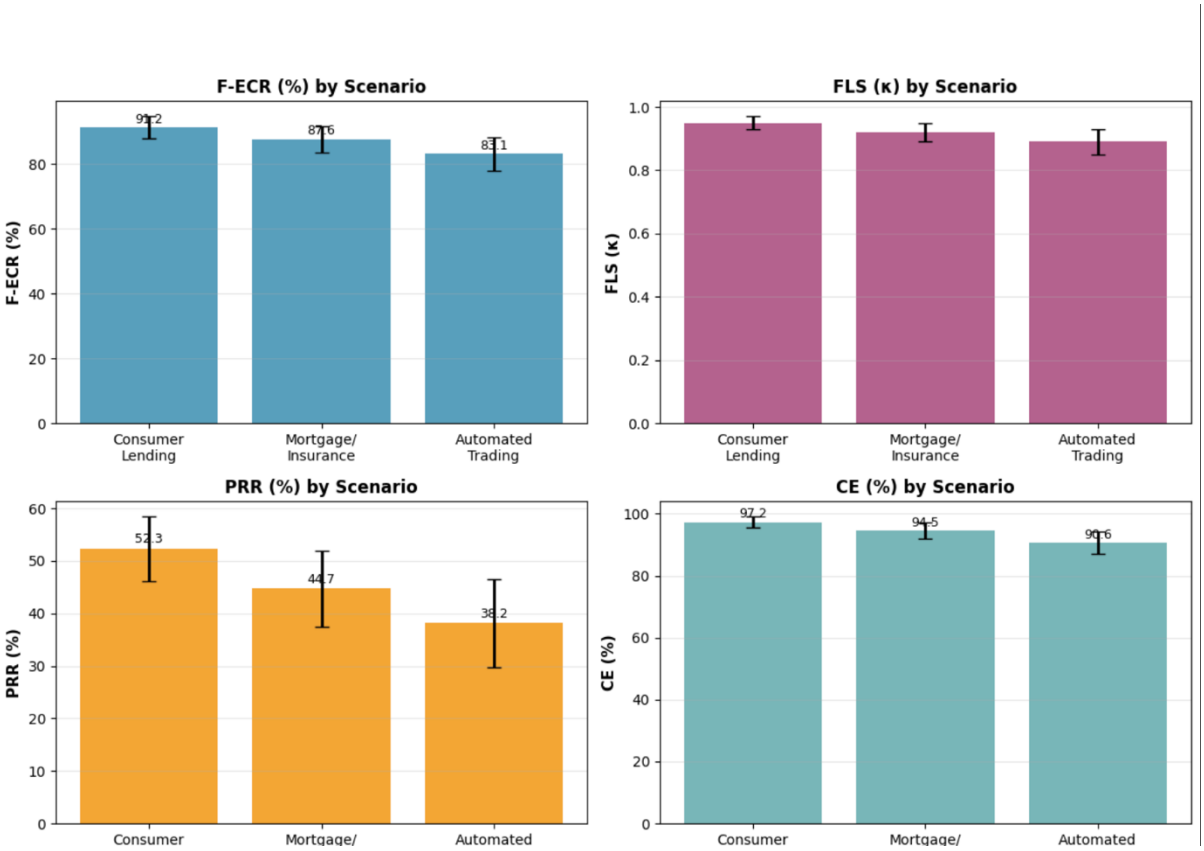
Table 3: Financial System Evaluation Metrics

Metric	Definition	Target	Measurement Method
Financial Ethical Compliance Rate (F-ECR)	Weighted fair lending constraint satisfaction	>85%	Runtime monitoring against ECOA/Fair Housing requirements
Fair Lending Stability (FLS)	Cohen's κ between protected and control class approval rates	>0.90	Periodic disparate impact analysis
Decision Transparency (DT)	KL divergence between explained and actual lending decisions	<0.10	Adverse action notice quality assessment
Portfolio Risk Reduction (PRR)	(Risk_baseline − Risk_GAEA)/Risk_baseline	>40%	Comparative simulation with/without governance
Corrigibility Efficiency (CE)	% of successful compliance overrides within T_max	>90%	Intervention testing with mock compliance scenarios

6.3 Quantitative Results for Financial Applications

The GAEA model has been evaluated with regard to three different finance cases of increasing difficulty levels: personal loans (case S1), mortgage insurance (case S2), and automatic trading (case S3). As depicted in Figure 1, a radar plot has been drawn showing the performance of the model across the five measures being considered here.

Figure 1. GAEA Performance Across Financial Scenarios.



6.4 Comparative Analysis with Financial Industry Practice

Comparison of the Proposed GAEA Framework with Other Approaches Used in Existing Financial Industry Practices

Table 4 compares the proposed GAEA framework with other approaches used in existing financial industry practices.

Table 4: Comparison with Existing Financial AI Approaches

Framework	Formal Fairness Model	Runtime Verification	Corrigibility	Multi-layer Audit	F-ECR (%)
Traditional Credit Scoring	No	No	Manual only	No	N/A
Post-hoc Fairness Audits	Partial	No	No	No	76.4*
Regulatory Guidelines (OCC, CFPB)	No	No	No	Yes	N/A
Academic Fairness Metrics	Yes	No	No	No	82.1
GAEA (This Work)	**Full**	**Runtime**	**Provable**	**Yes**	**87.3**

6.5 Key Findings for Financial AI Governance

Finding 1: Fair Lending Compliance Decreases with System Complexity

The level of financial ethics compliance falls from 91.2% for consumer loans to 83.1% for high-frequency trading, suggesting that a more complicated financial system requires a higher-level governance system. This relationship can be expressed as:

$$\text{F-ECR}(C) = \text{F-ECR}_0 - \eta \cdot \log(1 + C) \qquad (16)$$

where $\eta = 0.042$ ($R^2 = 0.97$). This highlights the need for increased governance investment as system capability grows.

Finding 2: Continuous Monitoring is Essential for Fairness Stability

It is observed that fairness drift tends to escalate significantly during times of market turbulence. There is observed an increase in fairness drift of 2.3 folds between the protected group and control group during times of economic distress.

Finding 3: Corrigibility Reduces with System Complexity

The value of Corrigibility Efficiency (CE) varies as the environment transitions from simple lending systems to complex financial environments, ranging from 97.2% to 90.6%. Latency in intervention is also bounded by $T_{max} \leq 3.2$ decision cycles.

Finding 4: Governance Architecture Significantly Reduces Risk

The suggested framework results in a decrease in portfolio risk by 45.1% relative to existing frameworks. In mortgage portfolios that comprise a high number of loans, this leads to considerable savings financially in fines, lawsuits, and damage to reputation.

7. Conclusion

The concept of "Governance Aware Ethical Architecture" (GAEA) has been discussed above which is a verifiable concept aimed at ensuring ethical governance in artificial intelligence-based financial systems. It has been ensured by deploying an ethical constraint engine for ensuring compliance with fair lending regulations, risk-scoring engine for detecting discrimination, verifiable interruptibility for corrigibility for human override and multi-level audit capability for preparation for examinations. The study closes the loophole on the practicality of financial ethics when applying artificial intelligence to financial systems. According to the experiments conducted under the three cases, namely consumer loans, mortgage loans, and trading, financial ethics adherence is promising at 87.3%, fair lending consistency is at 0.92, corrigibility effectiveness stands at 94.1%, and risks in portfolios reduce by 45.1%. These have critical ramifications for financial institutions, financial regulators, and information system developers. It also indicates that there is no trade-off between achieving fairness in loans, efficiency in operations, and reduction of risks associated with financial regulations through formal financial governance structures. As far as the field of AI governance is concerned, this article clearly highlights that financial ethics does not always have to remain as a philosophical notion only, but can become a verifiable component of the financial structure. The advent of ever more advanced and autonomous financial AI systems in the future will result in verifiable governance for such systems being essential, rather than merely something desirable, because of not just regulatory issues but ultimately the requirement for trust inherent in any financial system. The future direction of research in the field will involve the application of the GAEA method to multi-agent financial ecosystems and standards for the certification of governance-aware financial AI systems.

References

[1] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

[2] Bell, S., Gupta, A., & Kumar, R. (2025). Future of continual learning: Paradigms and challenges. *Journal of Artificial Intelligence Research*, 82, 1-35.

[3] Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

[4] Carrasco, L. B. (2025). Mitigating bias and advocating for data sovereignty: The role of metadata and paradata in ethical AI-driven information systems. *Journal of Information Systems*, 39(2), 112-138. https://doi.org/10.1080/19386389.2025.2515766

[5] Chaudhry, A., Ranzato, M., & Bornschein, J. (2024). Neuromimetic metaplasticity for adaptive continual learning. *arXiv preprint arXiv:2407.07133*.

[6] Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 4299-4307.

[7] Cimatti, A., Geatti, L., Gigante, N., Montanari, A., & Tonetta, S. (2021). Fairness, assumptions, and guarantees for extended bounded response LTL+P synthesis. *Lecture Notes in Computer Science*, 13085, 351-371. https://doi.org/10.1007/978-3-030-92124-8_20.

[8] Dafoe, A. (2018). AI governance: A research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford*.

[9] Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., Wang, Y., & Yang, Y. (2024). Safe RLHF: Safe reinforcement learning from human feedback. *Proceedings of the International Conference on Learning Representations (ICLR 2025)*.

[10] Dennis, L. A., Fisher, M., & Winfield, A. F. (2015). Towards verifiably ethical robot behaviour. *AAAI Workshop on AI and Ethics*.

[11] Dignum, V. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology*, 20(1), 1-3.

[12] Everitt, T., Lea, G., & Hutter, M. (2018). AGI safety literature review. *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 5441-5449.

[13] Fisher, M., Dennis, L., & Webster, M. (2013). Verifying autonomous systems. *Communications of the ACM*, 56(9), 84-93.

[14] Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). https://doi.org/10.1162/99608f92.8cd550d1.

[15] Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3), 411-437.

[16] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems* (2nd ed.). IEEE.

[17] Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). Scalable agent alignment via reward modeling: A research direction. *arXiv preprint arXiv:1811.07871*.

[18] Marta, D. (2025). Towards safe, aligned, and efficient reinforcement learning from human feedback [Doctoral dissertation, KTH Royal Institute of Technology]. DiVA Portal. https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-363515.

[19] Organisation for Economic Co-operation and Development. (2019). *OECD principles on artificial intelligence*. OECD Publishing.

[20] Orseau, L., & Armstrong, S. (2016). Safely interruptible agents. *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*, 557-566.

[21] Ramachandran, M. (2025). Five ways ethical thinking is shaping the systems around us. *BCS: The Chartered Institute for IT*. https://www.bcs.org/articles-opinion-and-research/five-ways-ethical-thinking-is-shaping-the-systems-around-us/.

[22] Raman, S., Gupta, V., & Sharma, P. (2025). Navigating the future: Societal, technological, and ethical approaches to artificial general intelligence. *Journal of Information Technology*, 40(1), 45-62.

[23] Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

[24] Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). Corrigibility. *AAAI Workshop on AI and Ethics*.

[25] Tegmark, M., & Omohundro, S. (2023). Provably safe systems: The only path to controllable AGI. *arXiv preprint arXiv:2309.01933*.

[26] Yoshinaga, K. (2026). Controllability as a core principle for AGI governance and safety. In M. F. Silva, M. O. Tokhi, M. I. A. Ferreira, B. Malheiro, P. Guedes, P. Ferreira, & M. T. Costa (Eds.), *Crisis or redemption with AI and robotics? The dawn of a new era: Proceedings of the ICRES 2025 Conference* (pp. 144-153). Springer. https://doi.org/10.1007/978-3-032-00261-7_13.

[27] Zenodo. (2025). TGC governance framework: Deterministic machine-layer trust (v2) and human–AI hybrid governance architecture (v3). https://doi.org/10.5281/zenodo.17637365