Governing AI in Executive Roles: A Broad Framework for Ethical Oversight and Integration Inspired by Albania's Appointment of Diella

Anonymous submission

Abstract

The historic appointment of Diella, an AI system, as Albania's Minister of State for Artificial Intelligence on September 11, 2025, marks a pivotal moment in AI governance, representing the first instance of an autonomous AI agent assuming a cabinet-level executive role in public administration. Tasked with overseeing public procurement to combat corruption, Diella exemplifies the potential for AI to enhance transparency and efficiency while raising profound questions about alignment with human values, accountable decision-making, and legal accountability. This position paper introduces the Ethical Oversight and Integration (EOI) Framework, a structured approach to guide the deployment of AI in authoritative roles. The framework comprises three interconnected pillars (Value Alignment, Ethical Safeguards, and Juridical Structures), each with specific guidelines and implementation strategies. Drawing on Diella's case, including its parliamentary address on September 18, 2025, and ongoing constitutional debates, we illustrate the framework's practical utility. By synthesizing insights from AI ethics, policy analysis, and legal scholarship, the EOI Framework offers a roadmap for balancing technological innovation with societal protections, ensuring AI serves as a reliable public servant rather than an unchecked authority.

Introduction

On September 11, 2025, Albanian Prime Minister Edi Rama formalized the appointment of Diella as Minister of State for Artificial Intelligence. Diella has been tasked with managing public procurement processes to eliminate corruption, ensuring completely incorruptible tenders through objective evaluation and transparency. This unprecedented move capitalizes on AI's ability to process vast data volumes impartially, reduce bureaucratic delays, and mitigate human biases such as bribery. However, it also amplifies significant governance challenges. Who bears liability for erroneous AI decisions? How can AI align with dynamic policy goals? What moral safeguards prevent embedded biases from perpetuating inequalities?

This paper positions Diella as a seminal case study to propose a governance framework for AI in executive roles. We argue that without structured governance, AI executives risk eroding public trust, exacerbating inequalities, and creating accountability vacuums. Our contribution is a practical, multidimensional framework that integrates lessons from Diella to guide future implementations, fostering responsible AI practices in governance.

Related Work

Existing literature on AI governance provides foundational insights but largely focuses on advisory or supportive AI roles rather than executive ones. For instance, regulatory analyses like Chan et al.'s work on balancing regulation and innovation for artificial intelligence (1) explore top-down versus bottom-up approaches, which echo Albania's decree-based appointment of Diella as a top-down innovation. However, these studies underexplore the legal personhood of AI, a gap addressed by Papyshev and Migliorini (2), who propose accountability models for AI-induced harms, relevant to potential procurement errors by Diella. Alignment research, such as the work by Conitzer et al. (3), emphasizes aggregating human inputs for value alignment, critical for ensuring AI executives like Diella reflect societal priorities.

Broader scholarly frameworks on AI in government roles further inform our work. Engler (4) proposes a unified theoretical framework for AI governance to balance regulation and innovation in public sectors. Similarly, the 5W1H framework by Wang et al. (5) systematically analyzes AI governance through key questions on regulation, applicable to executive AI deployments. Taeihagh (6) discusses governance of AI in public sectors, highlighting power asymmetries amplified by AI in decision-making roles. Recent real-world developments, such as Albania's appointment of Diella, have been analyzed in media and academic outlets (7; 8; 9; 10), providing empirical context and raising questions on legal status, oversight, and ethical implications that our framework addresses.

Methods: The Ethical Oversight and Integration (EOI) Framework

The EOI Framework organizes governance into three pillars: Value Alignment, Ethical Safeguards, and Juridical Structures. Each pillar incorporates advanced technical innovations, detailed guidelines, and measurable metrics to ensure robust, scalable deployment. We expand each component with technical depth while maintaining practical applicability for real-world implementation.

Pillar 1: Value Alignment

This pillar ensures AI behaviors synchronize with organizational goals and societal expectations, preventing misaligned outcomes such as prioritizing cost over equity in tender awards. We introduce technical innovations like neurosymbolic reasoning and adaptive learning to enhance alignment precision.

- **Guideline 1.1: Embed Contextual Objectives via Customization Techniques.** Tailor AI models with domain-specific inputs, such as procurement regulations, to align outputs with mandates like anti-corruption and fairness.
  - *Technical Innovation:* Implement neurosymbolic reasoning, combining neural networks with symbolic logic to encode procurement rules explicitly. For example, a rule-based module ensures tenders comply with EU standards, while neural components optimize for efficiency.
  - *Idea:* Develop a prompt engineering pipeline with iterative refinement. Use symbolic constraints to enforce legal thresholds while maintaining flexibility for optimization.
  - *Metric:* Alignment fidelity score, measured by concordance between AI decisions and expert-annotated benchmarks (target: greater than 90 percent). Compute using cosine similarity between AI output vectors and ground-truth annotations.
  - *Implementation:* Train models on a corpus of Albanian procurement laws and EU directives, augmented with historical tender data. Use a neurosymbolic framework like DeepProbLog to integrate logical constraints with probabilistic reasoning.
- **Guideline 1.2: Incorporate Diverse Input Mechanisms.** Aggregate stakeholder feedback to dynamically adjust AI parameters, accommodating evolving priorities.
  - *Technical Innovation:* Deploy a federated learning system to collect anonymized feedback from citizens, officials, and experts across regions, ensuring inclusivity without compromising privacy.
  - *Idea:* Create deliberation panels and digital surveys, processed via a social choice algorithm (such as Borda count) to prioritize alignment adjustments. Feedback is encoded as embeddings for model updates.
  - *Metric:* Feedback integration rate, tracking the percentage of inputs leading to model updates quarterly (target: greater than 75 percent). Measure via audit logs of parameter changes.
  - *Implementation:* Use a federated server to aggregate encrypted feedback, updating the model's loss function to weigh stakeholder priorities. Regularize updates with differential privacy to protect contributor identities.
- **Guideline 1.3: Robustness Against Adversarial Inputs.** Ensure AI resists manipulation attempts, such as crafted bids designed to exploit model weaknesses.
  - *Technical Innovation:* Apply adversarial training with generative models to simulate malicious inputs, enhancing model robustness.
  - *Idea:* Generate synthetic adversarial bids to train the AI to detect and flag manipulations, integrating certified robustness techniques.
  - *Metric:* Adversarial robustness score, percentage of malicious inputs correctly flagged (target: greater than 95 percent).
  - *Implementation:* Use a generative adversarial network (GAN) to create test cases, training the model with robust optimization frameworks like TRADES.

Example Deployment: Customize Diella's prompts with neurosymbolic constraints reflecting Albania's EU-aligned procurement laws. Implement federated feedback from rural and urban stakeholders to address regional disparities. Monitor drift via real-time dashboards, ensuring alignment with anti-corruption goals.

Pillar 2: Ethical Safeguards

This pillar mitigates harms like discrimination or opacity, ensuring AI decisions uphold moral standards. We introduce blockchain auditing and fairness-aware learning to enhance ethical rigor.

- **Guideline 2.1: Integrate Interpretability Tools.** Require AI to generate comprehensible rationales, fostering trust and transparency.
  - *Technical Innovation:* Use integrated gradients and natural language explanations to attribute decision factors, generating human-readable summaries.
  - *Idea:* Layer attribution methods atop models, outputting explanations that specify decision rationale with weighted factors. Translate outputs into Albanian for accessibility.
  - *Metric:* Interpretability score, via user comprehension surveys (target: greater than 80 percent layperson understanding). Use Likert-scale responses from diverse demographics.
  - *Implementation:* Integrate a post-hoc explainer like LIME, fine-tuned for procurement contexts, with a multilingual interface for stakeholder accessibility.
- **Guideline 2.2: Routine Bias Auditing.** Conduct ongoing assessments to identify and remediate inequities in data or outputs.
  - *Technical Innovation:* Deploy fairness-aware learning with constrained optimization to enforce demographic parity and equal opportunity in tender awards.
  - *Idea:* Automate audits using fairness probes (such as disparate impact analysis) on awarded tenders, followed by data augmentation for underrepresented groups (for example, minority-owned businesses).
  - *Metric:* Bias disparity ratio, aiming for less than 10 percent difference across protected attributes (such as region or ethnicity). Compute using fairness metrics like equalized odds.

  - *Implementation:* Use a fairness toolkit (such as AI Fairness 360) to run quarterly audits, retraining models with reweighted datasets to correct imbalances.
- **Guideline 2.3: Enforce Oversight Hierarchies.** Mandate hybrid workflows where AI outputs require human validation for high-impact actions.
  - *Technical Innovation:* Implement blockchain-based audit trails to log AI decisions immutably, enabling transparent human review.
  - *Idea:* Design escalation tiers: low-value tenders (less than 50,000 euros) auto-approved, high-value routed to review committees with logged rationales. Blockchain ensures tamper-proof records.
  - *Metric:* Oversight compliance rate, percentage of decisions vetted appropriately (target: 100 percent for high-value tenders).
  - *Implementation:* Deploy a Hyperledger Fabric blockchain to store decision logs, accessible to oversight committees via a secure dashboard.
- **Guideline 2.4: Proactive Ethical Risk Assessment.** Anticipate and mitigate emerging ethical risks through predictive modeling.
  - *Technical Innovation:* Use predictive risk models to forecast potential ethical failures, such as unintended bias amplification in long-term operations.
  - *Idea:* Train a secondary model to predict risks based on historical procurement data, flagging scenarios like regional favoritism for preemptive correction.
  - *Metric:* Risk prediction accuracy, percentage of correctly identified risks in simulations (target: greater than 85 percent).
  - *Implementation:* Integrate a risk assessment module using Bayesian networks, updated with real-time procurement outcomes.

Application to Diella: Audit Diella for biases in bidder selection (such as urban versus rural favoritism) using fairness-aware algorithms. Log decisions on a blockchain for parliamentary scrutiny, with natural language explanations in Albanian to enhance trust.

Pillar 3: Juridical Structures

This pillar establishes clear rules for AI's legal standing, liability, and regulatory adherence, incorporating smart contracts and regulatory technology for precision.

- **Guideline 3.1: Define Agency Boundaries.** Clarify AI's legal status to avoid responsibility voids.
  - *Technical Innovation:* Use smart contracts to codify AI's scope, embedding legal boundaries in executable code.
  - *Idea:* Classify AI as a delegated agent under human principals, with smart contracts defining advisory versus decisional roles (for example, Diella advises on tenders while humans finalize decisions).
  - *Metric:* Legal clarity index, assessed by expert reviews of documentation (target: greater than 90 percent agreement on scope).
  - *Implementation:* Deploy Ethereum-based smart contracts to enforce role boundaries, audited by legal experts for compliance with Albanian law.
- **Guideline 3.2: Allocate Liability Pathways.** Map accountability for errors to involved parties.
  - *Technical Innovation:* Implement automated fault attribution systems using causal inference to trace errors to developers, operators, or data sources.
  - *Idea:* Create fault trees (for instance, developer liability for model flaws, operator liability for deployment errors) integrated into smart contracts for automatic enforcement.
  - *Metric:* Liability resolution time, average days to attribute faults in simulations (target: less than 7 days).
  - *Implementation:* Use a causal inference engine (such as DoWhy) to map error pathways, with outcomes logged in smart contracts for transparency.
- **Guideline 3.3: Enable Adaptive Regulation.** Use experimental environments to test and refine rules.
  - *Technical Innovation:* Deploy regulatory technology (RegTech) platforms to automate compliance monitoring and sandbox testing.
  - *Idea:* Launch regulatory sandboxes for phased AI rollouts, with mandatory annual audits and contingency clauses for rollback if violations occur.
  - *Metric:* Adaptation efficacy, measured by reduction in compliance violations post-review (target: less than 5 percent violations).
  - *Implementation:* Use a RegTech platform like ComplyAdvantage to monitor compliance, with sandbox environments simulating procurement scenarios.
- **Guideline 3.4: International Standards Integration.** Align AI governance with global benchmarks.
  - *Technical Innovation:* Integrate ontology-based compliance checkers to map AI operations to international frameworks (such as GDPR or OECD AI Principles).
  - *Idea:* Develop an ontology of EU procurement standards, automatically checking Diella's outputs for compliance via semantic reasoning.
  - *Metric:* Compliance alignment score, percentage of decisions meeting global standards (target: greater than 95 percent).
  - *Implementation:* Use OWL ontologies with SPARQL queries to validate decisions, integrated into the RegTech platform.

Application to Diella: Establish Diella as a delegated agent via smart contracts, with clear liability pathways for procurement errors. Test the system in regulatory sandboxes before full deployment, ensuring compliance with both Albanian and EU standards.

Discussion

The EOI Framework addresses critical gaps in current approaches to AI governance by providing actionable guidelines that balance innovation with accountability. Our anal-

ysis of Diella's appointment reveals several key insights that extend beyond the Albanian context.

First, the framework demonstrates that successful AI integration in executive roles requires more than technical sophistication. While Diella's processing capabilities offer clear advantages in reducing corruption and improving efficiency, these benefits materialize only when supported by robust governance structures. The Value Alignment pillar addresses this by ensuring that AI systems do not optimize narrowly defined metrics at the expense of broader societal values. For instance, without explicit constraints, an AI focused solely on cost minimization might systematically disadvantage small businesses or underrepresented regions, undermining equity goals even while achieving procurement efficiency.

Second, our Ethical Safeguards pillar responds to legitimate concerns about AI opacity and bias. The requirement for interpretable outputs serves multiple functions beyond transparency. It enables meaningful human oversight, facilitates public trust, and creates opportunities for iterative improvement based on stakeholder feedback. The blockchain-based audit trail we propose offers a technical solution to the accountability challenge, ensuring that decision pathways remain traceable even as AI systems grow more complex. This approach contrasts with black-box deployment models that prioritize performance over explainability, which we argue are fundamentally incompatible with democratic governance.

Third, the Juridical Structures pillar confronts the legal uncertainty surrounding AI agency. By proposing that AI systems be classified as delegated agents rather than autonomous legal entities, we avoid the philosophical complications of granting personhood to machines while establishing clear liability chains. This classification allows existing legal frameworks to accommodate AI executives without requiring wholesale legislative overhaul. The smart contract implementation provides a technical mechanism for codifying these legal boundaries, automatically enforcing constraints on AI decision authority.

Our framework also highlights tensions that must be navigated carefully. The drive for efficiency through AI automation conflicts with the deliberative nature of democratic decision-making. Public procurement, as Diella illustrates, involves competing values (cost, quality, equity, sustainability) that cannot be reconciled through algorithmic optimization alone. The framework's emphasis on human-in-the-loop processes for high-stakes decisions reflects this reality, preserving space for human judgment even as AI capabilities expand.

Limitations of our approach warrant acknowledgment. The framework's effectiveness depends on implementation quality and institutional capacity. Countries with weak regulatory infrastructure may struggle to deploy the monitoring systems we prescribe. Additionally, the technical innovations we propose (federated learning, blockchain auditing, neurosymbolic reasoning) require significant computational resources and expertise, potentially limiting adoption in resource-constrained settings. Future work should explore simplified variants suitable for varied contexts.

The Albanian case also reveals political dimensions that frameworks alone cannot resolve. Diella's appointment occurred through executive decree rather than legislative deliberation, raising questions about democratic legitimacy. While technical safeguards can mitigate certain risks, they cannot substitute for robust political processes that ensure public input and consent. The framework should thus be viewed as a complement to, not replacement for, democratic governance mechanisms.

Finally, our work opens several research directions. Empirical evaluation of the framework's components in controlled settings would validate their effectiveness and identify refinements. Comparative analysis across multiple jurisdictions deploying AI in governmental roles could reveal context-specific factors that enhance or hinder successful implementation. Investigation into public attitudes toward AI executives would inform strategies for building social acceptance and trust.

Conclusion

Albania's appointment of Diella as Minister of State for Artificial Intelligence represents a watershed moment in AI governance, crystallizing challenges that will grow increasingly urgent as AI capabilities advance. This paper has presented the Ethical Oversight and Integration Framework as a practical response to these challenges, offering structured guidance for deploying AI in executive roles while safeguarding democratic values and societal wellbeing.

The EOI Framework's three pillars (Value Alignment, Ethical Safeguards, and Juridical Structures) provide complementary mechanisms for ensuring AI systems serve human interests rather than operate as unconstrained technocratic authorities. By grounding each pillar in concrete technical innovations, measurable metrics, and implementation strategies, we have sought to bridge the gap between abstract ethical principles and operational reality.

Our analysis demonstrates that responsible AI governance in executive contexts requires simultaneous attention to technical, ethical, and legal dimensions. Technical excellence alone cannot ensure beneficial outcomes; it must be paired with mechanisms for value alignment, bias mitigation, transparency, and accountability. Similarly, ethical guidelines remain aspirational without technical infrastructure to implement and enforce them. The framework's integrated approach reflects this interdependence.

Diella's case illustrates both the promise and peril of AI in governance. The potential for enhanced efficiency, reduced corruption, and data-driven decision-making is substantial. However, realizing these benefits without undermining democratic accountability, perpetuating bias, or creating responsibility vacuums requires careful governance design. The EOI Framework provides a template for this design, adaptable to diverse institutional contexts and policy domains.

Looking forward, the trajectory of AI development suggests that Diella will not remain unique. As AI systems grow more capable, pressure will mount to deploy them in roles of increasing authority. Our framework offers a foundation for

these deployments, but continued refinement through empirical testing and cross-jurisdictional learning will be essential. The challenge facing researchers, policymakers, and technologists is to ensure that the integration of AI into governance structures enhances rather than erodes democratic values.

Ultimately, the question is not whether AI will play executive roles in governance, but how to structure these roles to serve the public interest. This paper has sought to advance that conversation by offering a comprehensive, technically grounded framework informed by the first real-world case of an AI cabinet minister. We hope this contribution catalyzes further research and informed policymaking as societies navigate the complex terrain of AI governance in the years ahead.

Ethical Statement

This research examines the governance implications of deploying AI in executive governmental roles, a topic with significant ethical dimensions that warrant explicit consideration. Our work carries both potential benefits and risks that we address here.

On the positive side, the EOI Framework promotes ethical AI deployment by prioritizing transparency, accountability, and alignment with human values. By providing structured guidance for mitigating bias, ensuring interpretability, and establishing clear liability frameworks, our work aims to prevent harms that could arise from unconstrained AI decision-making in positions of authority. The emphasis on human oversight and democratic safeguards reflects our commitment to preserving human agency and dignity even as AI capabilities expand.

However, we acknowledge potential negative implications. First, by providing a framework that facilitates AI deployment in executive roles, our work might inadvertently accelerate adoption before societal consensus on appropriateness is reached. We emphasize that the framework should be viewed as a set of necessary (but not necessarily sufficient) conditions for responsible deployment, not as a blanket endorsement of AI executives. Second, the technical sophistication required to implement our framework might exacerbate inequalities between well-resourced and resource-constrained governments, potentially widening the global AI governance gap.

Third, there exists risk that the framework could be selectively adopted, with governments implementing efficiency-enhancing components while neglecting oversight mechanisms. We stress that the framework's three pillars are interdependent and must be implemented holistically to achieve intended safeguards. Partial implementation could create a veneer of responsible governance while failing to provide substantive protections.

Fourth, our analysis of Diella could be interpreted as normalizing AI in roles that some argue should remain exclusively human. We recognize that reasonable people disagree about whether AI should ever occupy positions of governmental authority. Our framework does not resolve this philosophical question but rather addresses practical governance challenges for jurisdictions that choose to proceed with such deployments.

We also note potential dual-use concerns. While developed for democratic governance contexts, elements of our framework could theoretically be adapted to enhance AI-enabled authoritarian control. This risk is inherent to governance research but merits explicit acknowledgment. We trust that publication through academic channels will facilitate informed public discourse that helps prevent such misuse.

Finally, we commit to ongoing engagement with affected communities, policymakers, and civil society organizations as this research area evolves. The rapid pace of AI development demands that researchers remain attentive to emerging ethical challenges and adapt frameworks accordingly. We welcome critical feedback on our approach and encourage diverse perspectives to shape the future direction of this work.

Acknowledgments

Anonymous acknowledgments for submission.

References

[1] Keith Jin Deng Chan, Gleb Papyshev, and Masaru Yarime. Balancing the Tradeoff between Regulation and Innovation for Artificial Intelligence: An Analysis of Top-down Command and Control and Bottom-up Self-Regulatory Approaches. In *Proceedings of the 2nd International Workshop on AI Governance (AIGOV)*, 2024.

[2] Gleb Papyshev and Sara Migliorini. Developing a Liability Framework for Harms Arising out of Specification Gaming. In *Proceedings of the 2nd International Workshop on AI Governance (AIGOV)*, 2024.

[3] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, and 7 more authors. Social Choice for AI Alignment: Dealing with Diverse Human Feedback. In *Proceedings of the 2nd International Workshop on AI Governance (AIGOV)*, 2024.

[4] Alex C. Engler. Advancing AI governance with a unified theoretical framework. *Policy and Society*, 2025.

[5] Wang, Z., et al. Navigating the complexities of AI and digital governance: the 5W1H framework. *Government Information Quarterly*, 42(4), 101973, 2025.

[6] Araz Taeihagh. Governance of artificial intelligence. *Policy and Society*, 40(2), 72–88, 2021.

[7] Alice Taylor. Albania appoints world's first AI-made minister. POLITICO, September 11, 2025. Available at: https://www.politico.eu/article/albania-apppoints-worlds-first-virtual-minister-edi-rama-diella/.

[8] Fatos Bytyci. Albania appoints AI bot as minister to tackle corruption. Reuters, September 11, 2025. Available at: https://www.reuters.com/technology/albania-appoints-ai-bot-minister-tackle-corruption-2025-09-11/.

[9] Al Jazeera Staff. Albania appoints AI bot 'minister' to fight corruption in world first. Al

Jazeera, September 12, 2025. Available at: https://www.aljazeera.com/news/2025/9/12/albania-appoints-ai-bot-minister-to-fight-corruption-in-world-first.

[10] Llazar Semini. Albania's AI 'minister' makes its debut with an address to parliament. AP News, September 18, 2025. Available at: https://apnews.com/article/albania-new-cabinet-program-ai-minister-diella-corruption-3aa58c801d69b5b295975cc68079a2d3.