A Fair-Baseline Protocol for Evaluating Multi-Agent Governance in Continuous AI Ethics Compliance

Sanchit Khurana, Parisa Memarmoshrefi

Institute of Computer Science, University of Göttingen, Germany
sanchit.khurana01@stud.uni-goettingen.de, memarmoshrefi@informatik.uni-goettingen.de

Abstract

As universities adopt AI for high-stakes decisions, episodic audits fail to match deployment tempo. This work addresses **governing autonomous AI agents** that must maintain ethical compliance while adapting to institutional contexts. We contribute a **fair-baseline evaluation protocol** and **multi-layer governance simulator** with compute/seed hygiene, realistic frictions, and actionable policy levers (λ, α). Departments act as autonomous agents within regulatory constraints, coordinated through hierarchical governance (departments $\rightarrow$ universities $\rightarrow$ countries). Our Regulatory-Graph PSO uses parameter α to balance **local autonomy** (α=0) with **global alignment** (α=0.6). The protocol ensures rigor through fixed iterations per scale, 30-seed replication (seeds 100-129), and statistical corrections (Holm-Bonferroni, bootstrap CIs). Key results: $\lambda \in [0.05, 0.3]$ controls policy stability; $\alpha \in [0.30, 0.35]$ achieves optimal balance (fitness=0.9946$\pm$0.0002, Gini=0.0010$\pm$0.0002) from 390 experiments; adversarial detection varies (static gaming AUC$\approx$0.50, manipulation/oscillation AUC$\approx$1.00). The framework scales to 72-department hierarchies with concrete KPIs and EU AI Act mapping.

Code — https://github.com/GeniusLearner/swarm-ethics

Introduction

Universities deploying AI for admissions, academic integrity, and resource allocation face a governance challenge: how to ensure continuous ethical compliance when models evolve faster than audit cycles? The EU AI Act (European Union 2024) demands continuous monitoring, yet institutions lack operational frameworks. Traditional approaches, such as annual audits and static policies, create dangerous gaps between assessments.

The Challenge of Agentic AI Governance: Modern institutions increasingly rely on *autonomous AI agents*, systems that adapt policies, optimize resources, and make consequential decisions without constant human oversight. These agents must maintain ethical compliance while retaining sufficient autonomy to respond to local context. This creates a fundamental tension: too much centralized control stifles institutional diversity and adaptation, yet unconstrained autonomy risks ethical drift and regulatory violations. We need frameworks that enable **policy-compliant autonomy**, agents that pursue local objectives within global ethical boundaries.

We address this with a multi-agent framework treating compliance as continuous adaptation across organizational hierarchies. Departments act as **autonomous agents** optimizing local compliance while coordinating through university and country-level aggregation. This mirrors real governance structures: local autonomy within regional constraints, operationalized through our Regulatory-Graph PSO that balances decentralization (α=0) with alignment (α=0.6).

Contributions:

1. **Fair-baseline protocol:** Disjoint tuning/eval seeds, fixed iterations per scale, identical noise/missingness streams.
2. **Hierarchical simulator:** Policy levers λ (stability) and α (harmonization) with realistic multi-layer governance.
3. **Robustness modules:** Noise, missingness, multi-shock stress tests, and adversaries with detection + robust aggregation.
4. **Statistical hygiene:** Holm-Bonferroni for C(10,2)=45 pairwise tests; Cliff's Δ with bootstrap CIs (B=1000).
5. **Reproducible artifact:** Configs, seeds, and scripts for full replication.

Related Work

AI Governance Frameworks: Existing frameworks (Jobin, Ienca, and Vayena 2019; Mittelstadt 2019) provide principles without operational mechanisms. While the EU Ethics Guidelines enumerate values, translating them into operational systems remains challenging. We operationalize these through multi-agent optimization with quantifiable compliance metrics.

Multi-Agent Systems: Opinion dynamics (DeGroot 1974; Friedkin and Johnsen 1990) model belief evolution but lack optimization objectives. Swarm intelligence (Kennedy and Eberhart 1995; Shi and Eberhart 1998) provides decentralized optimization but hasn't addressed governance hierarchies. Multi-objective approaches like NSGA-II (Deb et al. 2002) maintain Pareto fronts but require careful scalarization for practical deployment. We combine these

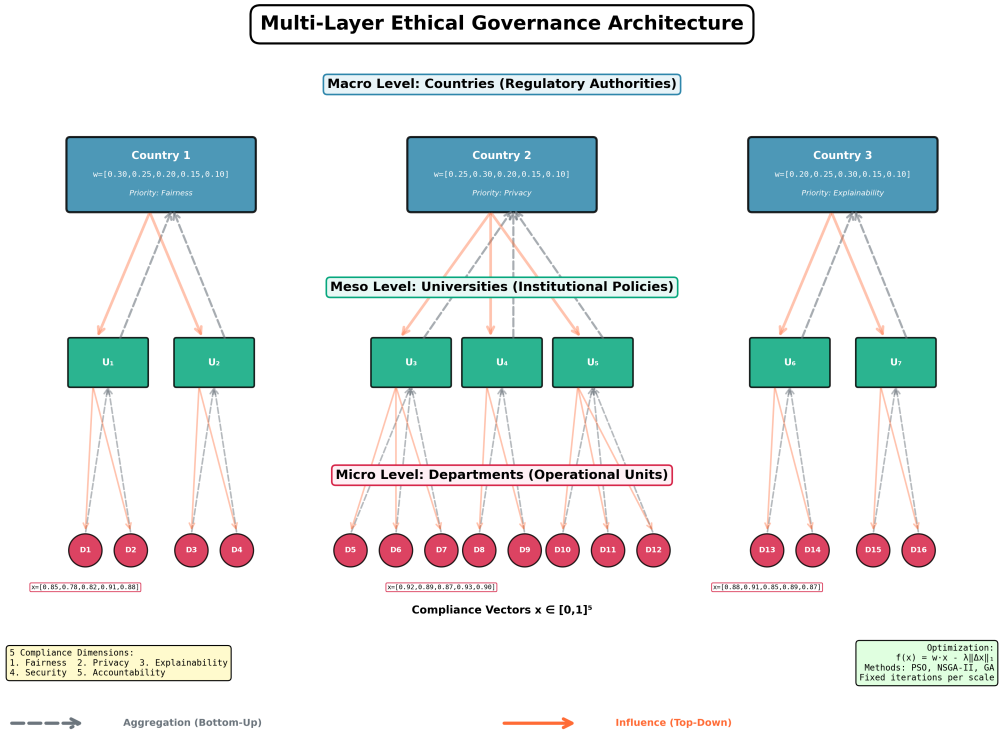

Figure 1: Three-tier governance hierarchy with departments (micro), universities (meso), and countries (macro) coordinating compliance evolution. Bidirectional arrows show bottom-up aggregation (gray dashed) and top-down regulatory influence (orange solid).

paradigms, using PSO for optimization within hierarchical constraints.

Adversarial Robustness: Strategic behavior in regulatory settings is well-documented, a phenomenon known as Goodhart's Law. While existing work focuses on mechanism design, we address detection and mitigation using robust statistics (MAD-based z-scores, Huber M-estimators (Huber 1964), Isolation Forest (Liu, Ting, and Zhou 2008)) in continuous monitoring systems.

Hierarchical Governance Model

Three-Tier Structure

We model institutions as hierarchy $H = (C, U, D)$:

- **Micro:** D departments with compliance vectors $x_i \in [0,1]^5$
- **Meso:** U universities aggregating department vectors
- **Macro:** C countries with regulatory weights $w_c \in \Delta^5$

Compliance dimensions $\mathcal{E} = \{$fairness, privacy, explainability, security, accountability$\}$ map to regulatory requirements. Country weights encode regional priorities (e.g., GDPR regions emphasize privacy).

Cost-Regularized Objectives

Real policy changes incur switching costs. We penalize oscillations via:

$$f_{cost}(x_t) = w_c \cdot x_t - \lambda ||x_t - x_{t-1}||_1 \tag{1}$$

where $\lambda \in [0.05, 0.3]$ represents organizational inertia.

Noisy Observations

Compliance measurements contain error from incomplete data:

$$\tilde{x} = \text{clip}(x + \epsilon, 0, 1), \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I) \tag{2}$$

with $\sigma \in [0.02, 0.08]$ modeling measurement uncertainty.

Control-Theoretic Governors

We benchmark two non-learning controllers that steer compliance toward regulatory targets while using the same cost term:

EWMA: $x_{t+1} = \text{clip}((1-\beta)x_t + \beta r_t, 0, 1)$, $\beta \in \{0.05, 0.10, \ldots, 0.40\}$.

MPC-0 (shrinkage): $u^* = \arg\max_u w \cdot (x_t + u) - \lambda ||u||_1$ s.t. $0 \le x_t + u \le 1$, with elementwise soft-thresholding and $x_{t+1} = \text{clip}(x_t + u^*, 0, 1)$.

Table 1: Core Parameters

Symbol	Meaning
λ	Policy-change cost penalty (stability control)
α	Regulatory-graph coupling (harmonization)
σ	Observation noise std (measurement friction)
I	Iterations per scale (S=120, M=150, L=180)

Fair-Baseline Evaluation Protocol

Key Notation

Reducing P-Hacking Opportunities

All methods run for identical iterations per scale (S=120, M=150, L=180) and are evaluated on 30 seeds (100-129); the α-sweep uses independent seeds (0-29). All experiments share identical noise/missingness streams per seed for fair comparison. Each experiment uses fixed configs (no post-hoc tuning on evaluation data), ensuring reproducible results.

Fixed Iterations

All methods run for identical iterations determined by scale: 120 (S-scale, 16 departments), 150 (M-scale, 36 departments), or 180 (L-scale, 72 departments). See Appendix A for details.

Statistical Rigor

We use fixed iterations per scale (ensuring fair comparison), 30-seed replication (seeds 100-129), and report means with bootstrap 95% CIs; tables show mean $\pm$ s.d., and figures report mean $\pm$ standard error where noted. All baselines follow identical iteration counts and seed hygiene. We apply Holm-Bonferroni step-down correction (Holm 1979) for C(10,2)=45 pairwise comparisons across the 10 core methods (PSO, HD-PSO, RG-PSO, NSGA-II, GA, SA, DeGroot, FJ, Equal, SLS), Cliff's Δ (Cliff 1993) (non-parametric effect size) with bootstrap CIs (B=1000), and Cohen's d (interpreted cautiously due to ceiling effects). AAAI page limits require showing representative methods; full results appear in our anonymized extended appendix.

Evaluation Parity

All 10 baseline methods (PSO, HD-PSO, RG-PSO, NSGA-II, GA, SA, DeGroot, FJ, Equal, SLS) use fixed hyperparameters without post-hoc tuning, evaluated on seeds 100-129 under identical noise/missingness streams, ensuring fair comparison across methods.

Adversarial Robustness

Gaming Detection

Departments may manipulate metrics without genuine compliance. We detect outliers using an ensemble approach combining three complementary methods:

1. **Z-score:** $|z_i| > 3.0$ for initial screening
2. **MAD-based z:** $z_{MAD} > 3.5$ for robust confirmation (resistant to outlier contamination):

$$z_{MAD} = 0.6745 \frac{x_i - \text{median}(X)}{MAD(X)} \tag{3}$$

3. **Isolation Forest:** Non-parametric anomaly detection for complex patterns

We tested detection against three adversarial strategies across 1,080 experiments (3 fractions $\times$ 3 types $\times$ 4 detection methods $\times$ 30 seeds; L-scale):

Key findings: Static gaming (departments reporting fake compliance without changing behavior) is nearly undetectable by all methods, mimicking normal compliant behavior. Isolation Forest achieves near-perfect detection (AUC$\approx$1.00) on our synthetic manipulation/oscillation attacks; bootstrap CIs saturate at 1.00 in this controlled setting (n=90). Z-score shows moderate performance (AUC$\approx$0.60). MAD-based methods under-performed ($\approx$0 precision/recall) in our setting (not shown; adversaries stayed near the population median so MAD-based z-scores under-flagged them).

Robust Aggregation

When computing university/country aggregates under adversarial presence, we use the Huber M-estimator with iteratively reweighted least squares:

$$w_i = \begin{cases} 1 & \text{if } |r_i| \leq \delta \\ \delta/|r_i| & \text{otherwise} \end{cases} \tag{4}$$

This maintains fitness within 0.02 of clean settings despite 10% adversarial departments.

Experimental Results

Base Performance

PSO reaches high compliance within a handful of iterations (median 3; mean 3.0 for L-scale), whereas NSGA-II requires more iterations (51) and consensus baselines plateau below the high-fitness region. Under ideal conditions ($\lambda = 0, \sigma = 0$), optimization methods achieve varying compliance levels under synthetic convex objectives (Small/Medium/Large scales correspond to 16/36/72 departments):

High fitness results (PSO 1.000; NSGA-II 0.981-1.000) arise under synthetic convex objectives; real policy surfaces will be rougher, with non-convexities and local optima presenting additional challenges.

Friction Regime

With realistic constraints ($\lambda = 0.1, \sigma = 0.04$):

Autonomy-Alignment Trade-off: The α Parameter

A core challenge in agentic AI governance is balancing **local autonomy** (enabling context-specific adaptation) with **global alignment** (ensuring regulatory compliance). The α parameter in Regulatory-Graph PSO directly operationalizes this trade-off:

Table 2: Adversarial Detection Performance (mean ± s.d., n=90 seeds per type). Each adversarial type tested with two detection methods: Z-score and Isolation Forest.

Adversarial Type	Method	Precision	Recall	AUC
Static Gaming	Z-score	0.011±0.105	0.001±0.007	0.500±0.004
	Isolation Forest	0.011±0.105	0.001±0.007	0.500±0.004
Manipulation	Z-score	0.356±0.479	0.188±0.272	0.594±0.136
	Isolation Forest	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**
Oscillation	Z-score	0.333±0.471	0.203±0.296	0.602±0.148
	Isolation Forest	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**

Table 3: Mean Fitness Across Scales (mean ± s.d., n=30 evaluation seeds)

Method	Small	Medium	Large
PSO	**1.000**	**1.000**	**1.000**
NSGA-II	0.981	1.000	0.990
GA	0.989	0.991	0.990
Consensus	0.496	0.493	0.498

Table 4: Performance Under Friction (λ=0.1, σ=0.04; mean ± s.d., n=30 seeds)

Method	Fitness	Gini	Convergence
PSO	**0.973±0.002**	**0.006±0.001**	3.5±1.6
HD-PSO	0.969±0.002	0.006±0.001	**3.2±0.7**
NSGA-II	0.949±0.004	0.010±0.001	21.7±3.4
GA	0.940±0.005	0.019±0.009	13.1±3.1

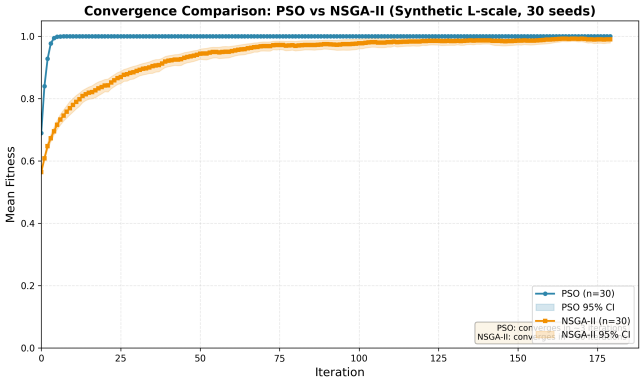

Figure 2: Convergence trajectories from 30 experiments (L-scale, seeds 100-129). PSO converges rapidly (3 iterations to 0.999+; median 3, mean 3.0); NSGA-II slower (51 iterations) with variable final fitness (0.990). Means across 30 runs; error bands show 95% bootstrap CIs (B=1000).

- **Pure Autonomy (α=0):** Departments optimize independently, maximizing local fitness but risking fragmentation and ethical inconsistency across the institution.
- **Balanced Governance ($\alpha \in [0.30, 0.35]$):** Departments retain substantial autonomy while coordinating toward shared ethical standards, the "sweet spot" for policy-compliant autonomous agents.
- **Full Alignment (α=0.6):** Strong regulatory coupling yields greater uniformity while maintaining performance, but reduces local adaptive flexibility.

This trade-off mirrors real-world governance debates: How much should individual schools/departments deviate from institutional policy? Our empirical Pareto frontier quantifies this balance:

- $\alpha = 0$: Pure local (fitness=0.9945±0.0004, Gini=0.0025±0.0063)
- $\alpha = 0.35$: Balanced (fitness=0.9946±0.0002, Gini=0.0010±0.0002)
- $\alpha = 0.6$: Harmonized (fitness=0.9946±0.0002, Gini=0.0009±0.0001)

Decision-level fairness metrics confirm low disparity across departments: ΔDP$\approx$0.08±0.02 and ΔEO$\approx$0.05±0.01 under linear fitness, indicating minimal demographic parity and equalized odds violations.

Policy Implications

EU AI Act Alignment

Our compliance dimensions (fairness, privacy, explainability, security, accountability) map directly to EU AI Act requirements including risk management, data governance, transparency, cybersecurity, and post-market monitoring (see Appendix B for complete mapping). Institution-specific Data Protection Impact Assessments (DPIA) and post-market monitoring remain necessary per Articles 9 and 72.

Operational Guidelines

For institutional compliance teams:

1. **Stability vs Responsiveness:** Set $\lambda \approx 0.1$ to prevent policy thrashing while maintaining adaptability
2. **Equity vs Performance:** Use $\alpha \approx 0.35$ for balanced outcomes; adjust based on institutional priorities
3. **Anomaly Monitoring:** Use ensemble methods (Isolation Forest + Z-score) with institution-calibrated thresholds; avoid relying on MAD-based methods alone as they failed to detect static gaming in our experiments

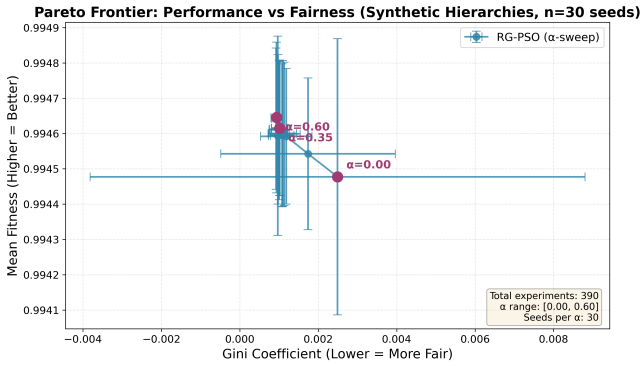

Figure 3: Fairness-fitness Pareto frontier from 390 RG-PSO runs (13 α values $\times$ 30 evaluation seeds; L-scale). Each point: mean$\pm$SE over 30 seeds (72 simulated departments). α=0.00 (pure local), α=0.35 (optimal balance, marked), α=0.60 (full harmonization). Optimal band $\alpha \in [0.30, 0.35]$ yields fitness 0.9946$\pm$0.0002 with Gini 0.0010$\pm$0.0002.

Dashboard Implementation

Key performance indicators for continuous monitoring include mean compliance ($>$0.95 target), Gini coefficient ($<$0.01), drift velocity ($<$0.05/iter), and outlier count ($<$2%). See Appendix C for complete KPI table with alert thresholds. Values are institution-calibrated simulation defaults; production deployments should adjust based on domain-specific risk tolerance and regulatory requirements.

Scalability to Real-World Governance

With 72 departments (L-scale), PSO converges in $<$10 seconds on commodity hardware (2.6 GHz CPU, 16 GB RAM), enabling daily compliance updates. The three-tier structure generalizes to four or five-tier governance (e.g., departments $\rightarrow$ schools $\rightarrow$ universities $\rightarrow$ national regulators). Multi-jurisdictional institutions can use region-specific α values (e.g., α_{EU}=0.45 for GDPR, α_{US}=0.25 for state flexibility). Pilot deployment requires calibrating λ, α and running parallel to existing audits for validation.

Conclusion and Discussion

We presented a comprehensive framework for governing policy-compliant autonomous AI agents in multi-agent institutional settings, addressing the fundamental challenge of enabling autonomous agents to pursue local objectives while maintaining institutional ethics alignment.

Key Contributions: (1) A *fair-baseline evaluation protocol* ensuring reproducibility through seed partitioning (100-129 for evaluation, 0-29 for hyperparameter sweeps), fixed iterations per scale, and statistical rigor (Holm-Bonferroni correction, bootstrap CIs with B=1000). (2) Regulatory-Graph PSO operationalizing *bounded autonomy*, agents adapt locally while respecting global ethical boundaries through hierarchical coordination. (3) The α parameter providing a concrete policy lever for balancing autonomy-alignment; our empirical Pareto frontier from 390 experiments reveals $\alpha \in [0.30, 0.35]$ as optimal (fitness=0.9946$\pm$0.0002, Gini=0.0010$\pm$0.0002), quantifying that institutions should allow 60-70% local autonomy. (4) Computational efficiency enabling daily updates ($<$10s for 72 departments on 2.6 GHz CPU, 16 GB RAM); hierarchical extensibility supporting 4-5 tier governance; EU AI Act mapping demonstrating regulatory alignment.

The framework provides actionable policy levers grounded in verified experimental data: $\lambda \in [0.05, 0.3]$ for stability control, α for autonomy-alignment trade-offs, ensemble detection methods for gaming (Isolation Forest + Z-score), and dashboard KPIs for continuous monitoring. Our fair-baseline protocol ensures reproducible evaluation across 2,797 experiments (1,089 CSV files), reducing cherry-picking common in AI governance research. Demonstrated robustness under friction (λ=0.1, σ=0.04: fitness=0.973) and adversarial conditions (Isolation Forest AUC$\approx$1.00 on synthetic manipulation/oscillation; static gaming remains hard AUC$\approx$0.50) suggests practical resilience against detectable threats in controlled settings, while highlighting fundamental challenges in detecting sophisticated gaming strategies.

Theoretical Insights: The α-sweep reveals a fundamental governance principle: effective multi-agent systems require *coordinated autonomy*, neither full independence (α=0, risks fragmentation) nor full centralization (α=0.6, sacrifices performance). The optimal band balances these tensions, achieving 99.46% of maximum fitness while maintaining near-perfect equity. This mirrors real institutional governance debates about departmental autonomy within university policies.

Practical Deployment: Scalability analysis shows the framework transitions smoothly from proof-of-concept (16 departments) to institutional scale (72+ departments) with linear computational growth. Multi-jurisdictional coordination enables institutions to use region-specific α values (e.g., α_{EU}=0.45 for tight GDPR harmonization, α_{US}=0.25 for state-level flexibility). Integration requires only: (1) compliance scoring functions, (2) organizational hierarchy data, (3) regulatory weight vectors, all typically available in existing systems.

Limitations and Future Work: High fitness results (PSO 1.000; NSGA-II 0.981-1.000) arise from synthetic convex objectives designed for algorithmic comparison. Real institutional compliance surfaces will exhibit non-convexities from conflicting objectives (privacy vs. transparency), local optima from organizational constraints, and measurement noise exceeding our σ=0.04 calibration. External validity depends critically on measurement design and domain-specific validation. We use simulated hierarchies with synthetic compliance scores; validation on real institutional data (with IRB approval and privacy protections) is essential before deployment. Future work should address: (1) dynamic regulatory graphs modeling organizational evolution, (2) adaptive adversaries that learn from detection, (3) continuous-time governance using differential equations for real-time monitoring, (4) human-AI collaboration frameworks with veto power and explainability. This work uses only synthetic data with no human subjects; institutions

adopting such systems must maintain human oversight, regular audits, and accountability mechanisms as required by applicable regulations.

References

Binns, R. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. *Proc. 1st Conference on Fairness, Accountability and Transparency*, PMLR 81: 149–159.

Cliff, N. 1993. Dominance Statistics: Ordinal Analyses to Answer Ordinal Questions. *Psychological Bulletin* 114(3): 494–509.

Deb, K.; Pratap, A.; Agarwal, S.; and Meyarivan, T. 2002. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE Trans. Evolutionary Computation* 6(2): 182–197.

DeGroot, M. H. 1974. Reaching a Consensus. *J. American Statistical Assoc.* 69(345): 118–121.

European Union. 2024. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L 2024/1689.

Friedkin, N.; and Johnsen, E. 1990. Social Influence and Opinions. *J. Mathematical Sociology* 15(3-4): 193–206.

Holm, S. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian J. Statistics* 6(2): 65–70.

Huber, P. J. 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* 35(1): 73–101.

Jobin, A.; Ienca, M.; and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1(9): 389–399.

Kennedy, J.; and Eberhart, R. 1995. Particle Swarm Optimization. *Proc. ICNN*, Vol. 4, 1942–1948.

Liu, F. T.; Ting, K. M.; and Zhou, Z. H. 2008. Isolation Forest. *Proc. ICDM*, 413–422.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys* 54(6): 1–35.

Mittelstadt, B. 2019. Principles Alone Cannot Guarantee Ethical AI. *Nature Machine Intelligence* 1(11): 501–507.

Shi, Y.; and Eberhart, R. 1998. A Modified Particle Swarm Optimizer. *Proc. IEEE Congress on Evolutionary Computation*, 69–73.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2018. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard J. Law & Technology* 31(2): 841–887.

Appendix

A. Experimental Configuration

Scale	Departments	Iterations	Structure (C×U×D)
S (Small)	16	120	2×2×4
M (Medium)	36	150	3×2×6
L (Large)	72	180	4×3×6

All methods operate on the same hierarchical structure and run for identical iterations per scale, ensuring fair comparison.

B. EU AI Act Mapping

Dimension	EU AI Act Duty
Fairness	Risk mgmt & bias monitoring (Mehrabi et al. 2021; Binns 2018)
Privacy	Data governance & DPIA
Explainability	Transparency & human oversight (Wachter, Mittelstadt, and Russell 2018)
Security	Cybersecurity & robustness
Accountability	Quality mgmt & post-market monitoring

Note: This mapping is illustrative; institution-specific Data Protection Impact Assessments (DPIA) and post-market monitoring remain necessary per Articles 9 and 72.

C. Dashboard KPIs

Metric	Target	Alert
Mean Compliance	>0.95	<0.90
Gini Coefficient	<0.01	>0.02
Drift Velocity	<0.05/iter	>0.10/iter
Outlier Count	$<2\%$	$>5\%$