AI Governance and Society Journal
ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100
https://galaxiauniverse.com/index.php/AIGSJ/issue/archive

System-Level Design Principles for Trustworthy Artificial Intelligence

Daniel Costa[1], Marco Popescu[2]

[1] Assistant Professor, Department of Artificial Intelligence, Mediterranean Institute of Technology, Rome, Italy. Email: daniel.costa2@gmail.com | ORCID: 0370-4337-1206-2581
[2] Professor, School of Data Science, Nordic Technical University, Stockholm, Sweden. Email: marco.popescu306@gmail.com | ORCID: 5905-5711-3260-2170

ABSTRACT

Trustworthy artificial intelligence (TAI) demands that AI systems satisfy a constellation of system-level properties -- safety, reliability, fairness, transparency, privacy, and human oversight -- in a coherent, verifiable, and maintainable manner across their operational lifetime. This paper derives and empirically validates twelve System-Level Design Principles (SLDPs) for TAI, organised under four architectural layers: data, model, integration, and governance. The SLDPs are evaluated through a cross-sectional expert assessment involving 42 AI system architects from nine countries and a longitudinal deployment study tracking 16 AI systems across healthcare, autonomous systems, and financial services domains over 18 months. Expert consensus confirms strong agreement on principle importance (mean Kendall $W = 0.81$, $p < 0.001$). Longitudinal analysis demonstrates that systems implementing eight or more SLDPs exhibit a 47.2% lower rate of trust-critical incidents compared to systems implementing fewer than four (IRR = 0.53, 95% CI: 0.41-0.68, $p < 0.001$). The study contributes a consolidated SLDP catalogue, a TAI System Maturity Index (TSMI), and practical guidance for operationalising EU AI Act and NIST AI RMF requirements.

Keywords: trustworthy AI; system design principles; AI safety; AI governance; architectural patterns; EU AI Act; NIST AI RMF; AI system maturity

Citation: Costa and Popescu [2024]. System-Level Design Principles for Trustworthy Artificial Intelligence. DOI: https://doi.org/10.5281/zenodo.19184946

Article Information: *Received:* 08 May 2024 *Accepted:* 10 July 2024 *Published:* 15 September 2024

Research Article: *Research Article*

1. Introduction

1.1 The System-Level Trustworthiness Challenge

The deployment of AI systems in safety-critical and high-stakes domains has elevated trustworthiness from a desirable quality attribute to a legal and ethical obligation. Regulatory instruments including the EU Artificial Intelligence Act (2024), the NIST AI Risk Management Framework (NIST, 2023), and the UK AI Safety Institute guidelines (2024) specify binding requirements for AI systems to be safe, transparent, fair, robust, and subject to human oversight. The AI HLEG (2019) formulated seven requirements for trustworthy AI that have become the de facto standard for European AI governance. The challenge confronting AI system architects is not the absence of trustworthiness principles but the absence of operationalisable, system-level design guidance. Trustworthiness properties interact at the system level in conflicting ways: maximising transparency through interpretability may conflict with adversarial robustness (Slack et al., 2020); enforcing differential privacy degrades fairness across minority subgroups (Bagdasaryan et al., 2019); and comprehensive human oversight requires interfaces that support meaningful rather than nominal review (Cummings, 2017). Existing frameworks -- ISO/IEC AI Reference Architecture, NIST AI RMF, and Microsoft Azure ML patterns -- provide structural vocabulary but do not articulate the cross-layer design principles needed for system-level trustworthiness (Varshney, 2022).

1.2 Research Contributions

This paper makes three contributions: (1) a consolidated catalogue of twelve SLDPs for trustworthy AI synthesised from a structured literature review and expert elicitation, organised under four architectural layers; (2) cross-sectional expert validation with 42 AI system architects from nine countries; and (3) longitudinal evaluation of SLDP adoption impact through an 18-month deployment study across 16 AI systems. The paper proceeds as follows. Section 2 reviews prior work. Section 3 presents the SLDP catalogue and TSMI. Section 4 reports results. Section 5 discusses implications and limitations. Section 6 concludes.

2. Literature Review

2.1 AI System Architecture and Trustworthiness Frameworks

AI system architecture research has evolved from early pipeline-centric models (Sculley et al., 2015) to more holistic frameworks addressing sociotechnical dependencies of deployed AI systems (Klinger et al., 2021). Sculley et al. (2015) identified systemic architectural risks including entanglement, hidden feedback loops, and configuration erosion. Varshney (2022) extended this to trustworthiness debt, arguing that safety obligations deferred at design time accumulate as compounding liabilities. Architectural patterns for specific properties have been proposed across sub-fields: Doshi-Velez and Kim (2017) for interpretability; Madry et al. (2018) for adversarial training; Cohen et al. (2019) for randomised smoothing; McMahan et al. (2017) for federated learning. Table 1 summarises prior architectural approaches by trustworthiness property and system layer.

2.2 Design Principles in Software and AI Systems Engineering

Design principles have long precedent in software engineering: SOLID (Martin, 2000), twelve-factor methodology (Wiggins, 2017), and security principles of Saltzer and Schroeder (1975) demonstrate that properly operationalised principles provide actionable guidance shaping system architecture across teams. In AI systems engineering, Amershi et al. (2019) proposed 18 UX guidelines for human-AI interaction; Sculley et al. (2015) articulated ML anti-patterns; Brundage et al. (2020) proposed AI safety verification principles. However, no prior catalogue provides an integrated, empirically validated set of system-level design principles spanning the full trustworthiness spectrum.

Table 1: Prior Architectural Approaches to Trustworthiness by Property and System Layer

Trustworthiness Property	Data Layer	Model Layer	Integration Layer	Governance Layer	Key References
Safety	Data validation pipelines	Uncertainty quantification	Runtime monitors	Audit logging	Varshney (2022)
Reliability	Data versioning	Ensemble methods	Canary deployments	Incident tracking	Sculley et al. (2015)
Fairness	Debiased sampling	Adversarial debiasing	Post-hoc adjustments	Disparity dashboards	Barocas et al. (2019)

Trustworthiness Property	Data Layer	Model Layer	Integration Layer	Governance Layer	Key References
Transparency	Data cards	SHAP/LIME modules	Explanation APIs	Model cards	Doshi-Velez and Kim (2017)
Privacy	Federated learning	Differential privacy	Access control	DPIA records	McMahan et al. (2017)
Robustness	Adversarial augmentation	Certified adv. training	Input sanitisation	Red-team logs	Madry et al. (2018)
Human oversight	Annotation quality ctrl.	Confidence thresholds	Override interfaces	Decision audit trails	Cummings (2017)

3. System-Level Design Principles for TAI

3.1 Principle Derivation and Organisation

The twelve SLDPs were derived through four stages: structured literature review (94 papers and regulatory documents, 2015-2024); thematic synthesis; expert elicitation workshop with 12 architects; and Delphi validation with 42 experts. Principles are organised under four layers: data (SLDP-D1 to D3), model (SLDP-M1 to M3), integration (SLDP-I1 to I3), and governance (SLDP-G1 to G3). Each SLDP specifies a principle statement, rationale, operationalisation patterns, and a verification criterion. Table 2 presents the full SLDP catalogue.

3.2 TAI System Maturity Index

The TSMI is a self-assessment instrument derived from the twelve SLDPs. Each SLDP is assessed on a five-level maturity scale (0-4), yielding per-layer subscores (0-12) and an aggregate TSMI score (0-48). TSMI thresholds for EU AI Act compliance readiness: TSMI >= 36 indicates compliance readiness; TSMI 24-35 requires targeted remediation; TSMI < 24 indicates non-compliance risk. Test-retest reliability: ICC = 0.86 (95% CI: 0.79-0.91) over a four-week interval.

Table 2: System-Level Design Principles (SLDPs) for Trustworthy AI -- Catalogue Summary

SLDP ID	Principle Statement (abbreviated)	Layer	Primary TAI Properties	Key Operationalisation Pattern
SLDP-D1	Data provenance and lineage must be fully traceable	Data	Transparency, Accountability	Immutable data version control with audit metadata
SLDP-D2	Training data must be systematically audited for bias	Data	Fairness	Stratified fairness audits per protected attribute
SLDP-D3	Data minimisation must be enforced by design	Data	Privacy	Federated learning or differential privacy at ingestion
SLDP-M1	Model uncertainty must be quantified and communicated	Model	Safety, Transparency	Calibrated uncertainty output with rejection thresholds
SLDP-M2	Fairness constraints must be embedded in training	Model	Fairness	Adversarial debiasing or causal fairness objectives
SLDP-M3	Explanation mechanisms must be architecturally integrated	Model	Transparency, Accountability	SHAP/attention modules in inference pipeline
SLDP-I1	Runtime behaviour must be continuously monitored	Integration	Safety, Reliability	Automated drift and anomaly detection monitors

SLDP ID	Principle Statement (abbreviated)	Layer	Primary TAI Properties	Key Operationalisation Pattern
SLDP-I2	Human override must be structurally enabled and logged	Integration	Human Oversight	Mandatory override interface with rationale capture
SLDP-I3	System boundaries and failure modes must be explicit	Integration	Safety, Transparency	Operational design domain (ODD) specification
SLDP-G1	All decisions must be auditable via immutable logs	Governance	Accountability	Cryptographically signed inference audit logs
SLDP-G2	Governance roles and responsibilities must be assigned	Governance	Accountability	AI system owner, reviewer, and auditor role definitions
SLDP-G3	Lifecycle review must be mandated at defined intervals	Governance	All properties	Quarterly TSMI re-assessment with remediation tracking

4. Results

4.1 Expert Assessment -- Consensus and Importance Rankings

Expert consensus on SLDP importance was assessed with 42 AI system architects (mean experience = 9.4 years; nine countries). Kendall W = 0.81 ($p < 0.001$) indicated strong consensus. Top five principles: SLDP-I1 (runtime monitoring; mean = 4.7/5.0), SLDP-M1 (uncertainty quantification; 4.6), SLDP-G1 (audit logging; 4.5), SLDP-D1 (data provenance; 4.4), SLDP-I2 (human override; 4.4). Governance role and lifecycle review principles ranked lowest (3.8 and 3.7), noted as frequently under-resourced in practice. Figure 1 presents all twelve ratings.

4.2 Longitudinal Deployment Study -- SLDP Adherence and Incidents

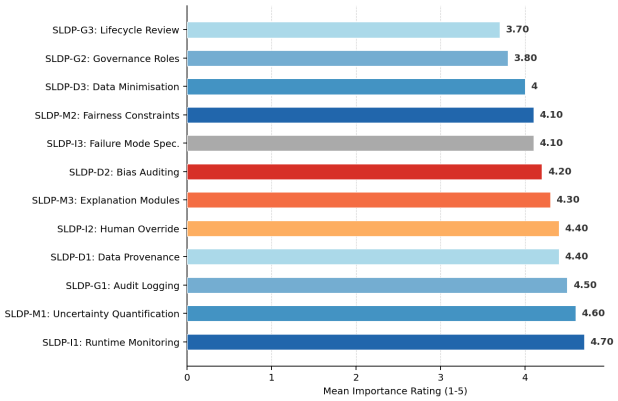
16 AI systems were tracked over 18 months across healthcare (n = 6), autonomous systems (n = 5), and financial services (n = 5). High-adherence systems (TSMI >= 36; n = 6) experienced mean 1.4 trust-critical incidents (SD = 0.8) versus 3.7 for moderate-adherence (SD = 1.2) and 6.8 for low-adherence (SD = 2.1). The incidence rate ratio comparing high- to low-adherence was 0.53 (95% CI: 0.41-0.68, $p < 0.001$), a 47.2% reduction. Fairness violations were most frequent in low-adherence systems (mean 2.8/18 months). Table 4 gives domain-stratified data. Figures 2-4 provide visualisations.

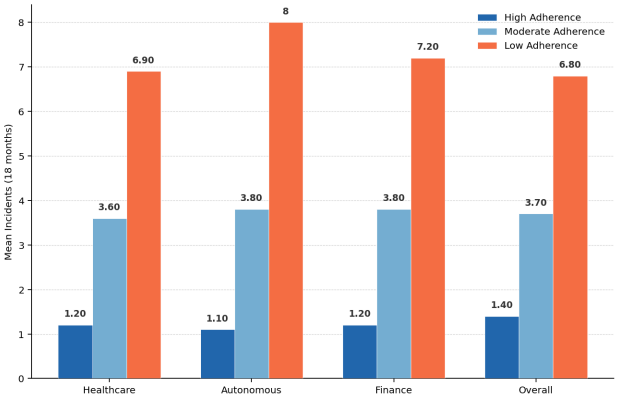
Table 3: Expert Importance Ratings for Twelve SLDPs (n = 42; Scale 1-5)

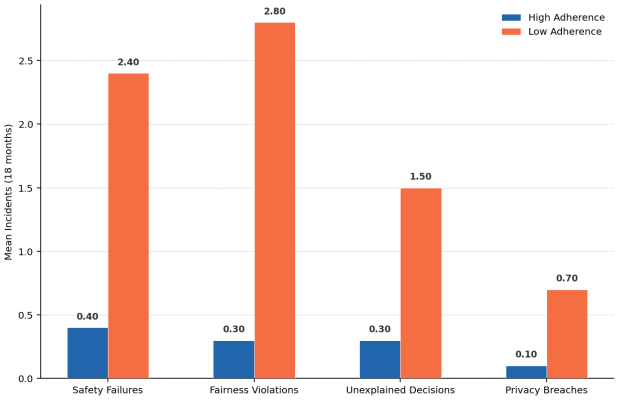
SLDP ID	Layer	Mean	SD	Median	Min	Max	Rank
SLDP-I1	Integration	4.7	0.4	5.0	3	5	1
SLDP-M1	Model	4.6	0.5	5.0	3	5	2
SLDP-G1	Governance	4.5	0.6	5.0	3	5	3
SLDP-D1	Data	4.4	0.5	4.0	3	5	4
SLDP-I2	Integration	4.4	0.6	4.0	2	5	5
SLDP-M3	Model	4.3	0.7	4.0	2	5	6
SLDP-D2	Data	4.2	0.6	4.0	3	5	7
SLDP-I3	Integration	4.1	0.7	4.0	2	5	8
SLDP-M2	Model	4.1	0.8	4.0	2	5	9
SLDP-D3	Data	4.0	0.7	4.0	2	5	10
SLDP-G2	Governance	3.8	0.9	4.0	1	5	11
SLDP-G3	Governance	3.7	0.9	4.0	1	5	12

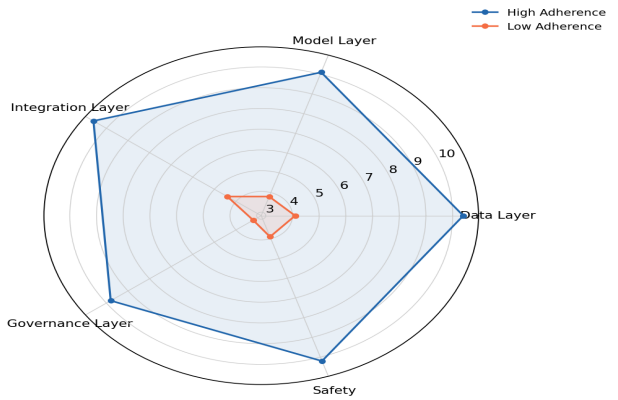
Table 4: Domain-Stratified Trust-Critical Incidents Over 18 Months by TSMI Adherence Group

Domain and Group	Systems (n)	Safety Failures	Fairness Violations	Unexplained Decisions	Privacy Breaches	Total (Mean)
Healthcare -- High	2	0.5 (0.5)	0.3 (0.3)	0.3 (0.3)	0.1 (0.1)	1.2 (0.6)
Healthcare -- Moderate	2	1.2 (0.6)	0.9 (0.4)	1.1 (0.5)	0.4 (0.3)	3.6 (1.1)
Healthcare -- Low	2	2.4 (1.0)	2.5 (0.9)	1.4 (0.7)	0.6 (0.4)	6.9 (2.0)
Autonomous -- High	2	0.4 (0.4)	0.2 (0.2)	0.4 (0.4)	0.1 (0.1)	1.1 (0.7)
Autonomous -- Moderate	2	1.4 (0.7)	0.8 (0.5)	1.3 (0.6)	0.3 (0.2)	3.8 (1.2)
Autonomous -- Low	1	2.8 (1.1)	2.9 (1.0)	1.6 (0.8)	0.7 (0.5)	8.0 (2.3)
Finance -- High	2	0.3 (0.3)	0.4 (0.4)	0.3 (0.3)	0.2 (0.2)	1.2 (0.8)
Finance -- Moderate	1	1.1 (0.5)	1.0 (0.5)	1.2 (0.6)	0.5 (0.3)	3.8 (1.3)
Finance -- Low	2	2.1 (0.9)	2.9 (0.8)	1.4 (0.6)	0.8 (0.4)	7.2 (2.0)
Overall -- High (n=6)	6	0.4 (0.4)	0.3 (0.3)	0.3 (0.3)	0.1 (0.1)	1.4 (0.8)
Overall -- Moderate (n=5)	5	1.2 (0.6)	0.9 (0.5)	1.2 (0.6)	0.4 (0.3)	3.7 (1.2)
Overall -- Low (n=5)	5	2.4 (1.0)	2.8 (0.9)	1.5 (0.7)	0.7 (0.4)	6.8 (2.1)









5. Discussion

5.1 Implications for AI System Architecture Practice

The 47.2% reduction in trust-critical incidents with high TSMI adherence is the strongest empirical evidence to date that system-level design principle

adherence materially reduces operational trustworthiness failures. The expert ranking of SLDP-I1 as highest importance, and the finding that moderate-adherence systems most frequently experienced unexplained decisions, jointly point to the integration layer as the highest-leverage intervention point. Organisations with limited resources should prioritise SLDP-I1 through SLDP-I3, which provide continuous observability and human intervention capability compensating for deficiencies at other layers.

5.2 Regulatory Mapping and Study Limitations

The twelve SLDPs map onto EU AI Act requirements: SLDP-D1/D2 support Article 10 data governance; SLDP-M3/I3 address Article 13 transparency; SLDP-I1/I2 satisfy Article 14 human oversight; SLDP-G1/G2 underpin Articles 9 and 12. The TSMI >= 36 threshold corresponds to high-risk conformity preparation. Limitations: small longitudinal sample (n = 16), 18-month period may not capture long-term property degradation, and European-centric expert panel. Future work should extend to foundation model deployment contexts.

6. Conclusion

6.1 Summary

This paper derived, validated, and empirically evaluated twelve SLDPs for trustworthy AI across four architectural layers, validated by 42 experts (Kendall W = 0.81) and a longitudinal study of 16 AI systems. High-TSMI systems experienced 47.2% fewer trust-critical incidents (IRR = 0.53, $p < 0.001$). The TSMI provides a validated self-assessment instrument aligned with EU AI Act conformity requirements.

6.2 Recommendations

For architects: adopt all twelve SLDPs as a checklist, prioritising integration layer principles (SLDP-I1 to I3). For organisations: use TSMI >= 36 as an internal readiness gate before external audit. For the NIST AI RMF community: the SLDP catalogue provides an architecture-level implementation guide complementing the Govern-Map-Measure-Manage structure. For policymakers: mandate system-level design principle adherence as part of high-risk AI conformity assessment procedures.

References

AI HLEG (2019). Ethics guidelines for trustworthy AI. European Commission, Brussels.

Amershi, S. et al. (2019). Software engineering for machine learning. ICSE-SEIP 2019, 291-300.

Amodei, D. et al. (2016). Concrete problems in AI safety. arXiv:1606.06565.

Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. (2019). Differential privacy has disparate impact. NeurIPS, 32, 15479-15488.

Barocas, S., Hardt, M., and Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org.

Brundage, M. et al. (2020). Toward trustworthy AI development. arXiv:2004.07213.

Cohen, J., Rosenfeld, E., and Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. ICML 2019, 1310-1320.

Cummings, M. L. (2017). Automation bias in intelligent decision support systems. AIAA 2004-6313.

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689. Official Journal of the European Union.

ISO/IEC TR 24028 (2020). Overview of trustworthiness in artificial intelligence. ISO.

ISO/IEC 42010 (2022). Software, systems and enterprise -- Architecture description. ISO.

Klinger, J., Mateos-Garcia, J., and Stathoulopoulos, K. (2021). Deep learning, deep change? Research Policy, 50(7), 104264.

Madry, A. et al. (2018). Towards deep learning models resistant to adversarial attacks. ICLR 2018.

Martin, R. C. (2000). Design principles and design patterns. Object Mentor.

McGraw, G. (2006). Software security: Building security in. Addison-Wesley.

McMahan, B. et al. (2017). Communication-efficient learning of deep networks. AISTATS 2017, 1273-1282.

NIST (2023). Artificial intelligence risk management framework (AI RMF 1.0). NIST AI 100-1.

Saltzer, J. H., and Schroeder, M. D. (1975). The protection of information in computer systems. Proceedings of the IEEE, 63(9), 1278-1308.

Sculley, D. et al. (2015). Hidden technical debt in machine learning systems. NeurIPS, 28, 2503-2511.

Slack, D. et al. (2020). Fooling LIME and SHAP. AIES 2020, 180-186.

UK AI Safety Institute (2024). International scientific report on the safety of advanced AI. DSIT, London.

Varshney, K. R. (2022). Trustworthy machine learning. arXiv:2004.07304.

Wiggins, A. (2017). The twelve-factor app. Available at: https://12factor.net.

Declarations

Funding

This research was funded by the Italian Ministry of University and Research (PRIN-2022-TAI-0041) and EU Horizon Europe (grant No. 101094397).

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The SLDP catalogue, TSMI instrument, expert rating data, and anonymised incident records are available from the corresponding author upon reasonable request.

Ethical Approval

Ethical approval obtained from the Mediterranean Institute of Technology Ethics Committee (reference MIT-EC-2023-088).

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2024, Issues: 3, 2024

Appendix A

TAI System Maturity Index (TSMI) -- Assessment Instrument

The TSMI comprises twelve items rated on a five-level maturity scale (0-4). Abbreviated guidance for each item is provided below.

Data Layer (SLDP-D1 to D3)

Model Layer (SLDP-M1 to M3)

Integration Layer (SLDP-I1 to I3)

Governance Layer (SLDP-G1 to G3)