AI Governance and Society Journal

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100

https://galaxiauniverse.com/index.php/AIGSJ/issue/archive

Ethics-Aware Software Engineering Frameworks for AI Applications

Isabella Lindberg[1], Oscar Petrov[2]

[1] Associate Professor, Department of Computer Science, Nordic Technical University, Stockholm, Sweden. Email: isabella.lindberg1@yahoo.com | ORCID: 8087-4290-8007-8581

[2] Senior Lecturer, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: oscar.petrov305@yahoo.com | ORCID: 9319-8952-7293-6469

ABSTRACT

The integration of artificial intelligence (AI) into software-intensive systems demands that software engineering (SE) practice evolve to encompass ethical properties including fairness, transparency, explainability, and harm avoidance. Despite widespread scholarly consensus, existing SE frameworks offer limited operational guidance for embedding ethical requirements throughout the software development lifecycle (SDLC). This study proposes and empirically evaluates an Ethics-Aware Software Engineering (EASE) framework extending established SE practices with ethics-specific processes, artefacts, and verification criteria. The EASE framework is evaluated through a controlled study involving 18 professional SE teams (n = 162 engineers) from six European organisations developing AI-enabled products across healthcare, finance, and smart-city domains. Teams adopting EASE achieved a 41.3% reduction in ethics-related defects, a 29.6% improvement in stakeholder fairness satisfaction, and a 22.4% reduction in time-to-ethics-review per sprint. The study contributes a replicable framework, a validated ethics-requirements taxonomy, and empirical benchmarks for AI software engineering practice and regulatory compliance.

Keywords: ethics-aware software engineering; AI software development; responsible AI; requirements engineering; SDLC; fairness; algorithmic transparency; EU AI Act

Citation: Lindberg and Petrov [2024]. Ethics-Aware Software Engineering Frameworks for AI Applications. DOI: https://doi.org/10.5281/zenodo.19184938

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: *Received:* 10 February 2024 *Accepted:* 10 April 2024 *Published:* 15 June 2024

Research Article: *Research Article*

1. Introduction

1.1 Motivation and Problem Statement

Software engineering has long grappled with translating non-functional requirements -- reliability, security, performance -- into actionable engineering practices throughout the SDLC. The emergence of AI-enabled software adds a qualitatively new category: ethical properties such as fairness, transparency, explainability, privacy, and harm avoidance (Amershi et al., 2019; Brill et al., 2022). These are inherently sociotechnical -- satisfaction depends on technical design decisions, social deployment context, stakeholder values, and evolving regulatory standards (Vakkuri et al., 2020). The EU Artificial Intelligence Act (2024) requires high-risk AI systems to undergo mandatory conformity assessment covering robustness, transparency, fairness, and human oversight -- requirements most efficiently addressed through systematic engineering processes rather than post-hoc remediation (Raji et al., 2020). A systematic review of 112 SE methodologies by Morley et al. (2021) found fewer than 23% provided operational ethics integration guidance and fewer than 9% addressed ethics across all SDLC phases. Existing toolkits such as IBM AI Fairness 360 (Bellamy et al., 2019) address specific technical components but do not provide an integrated SE framework spanning requirements, architecture, testing, and deployment.

1.2 Research Objectives

This paper addresses the identified gap by proposing and empirically evaluating an Ethics-Aware Software Engineering (EASE) framework. Research objectives are: (1) to develop a phase-spanning EASE framework operationalising ethical requirements across the full SDLC; (2) to evaluate the framework in a controlled study with professional SE teams; and (3) to identify EASE artefacts most strongly associated with ethics-defect reduction and stakeholder satisfaction improvement. The paper proceeds as follows. Section 2 reviews prior work on ethics integration in SE and regulatory context. Section 3 presents the EASE framework. Section 4 presents study results. Section 5 discusses implications and limitations. Section 6 concludes.

2. Literature Review

2.1 Ethics Integration in Software Engineering

Research on ethics integration in SE has developed along two trajectories: value-sensitive design (VSD) and requirements engineering for ethical AI. Friedman et al. (2019) introduced VSD as a method for incorporating human values throughout technology design. Maalej et al. (2023) proposed a taxonomy of AI-specific quality attributes and demonstrated through a survey of 156 practitioners that fairness and transparency requirements are most frequently under-specified at project outset. Agile methodologies present challenges for ethics integration: rapid iteration can marginalise deliberative stakeholder value

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2024, Issues: 2, 2024

elicitation (Vakkuri et al., 2020). Researchers have proposed ethics sprints (Madaio et al., 2020), ethical product backlogs (Deng et al., 2022), and algorithmic impact assessment templates. Table 1 summarises prior ethics-SE frameworks and their SDLC phase coverage.

2.2 Responsible AI Toolkits and Regulatory Context

IBM AI Fairness 360 (Bellamy et al., 2019) provides fairness metrics and bias-mitigation algorithms. Google's What-If Tool (Wexler et al., 2020) supports interactive fairness exploration. Microsoft's InterpretML (Nori et al., 2019) offers interpretability methods. Liao et al. (2023) found fewer than 31% of AI practitioners had integrated any fairness toolkit, citing lack of SE integration guidance as the primary barrier. The EU AI Act (2024) Article 9 mandates risk management systems covering the full AI system lifecycle. The UK ICO guidance on AI and data protection (2023) additionally requires data protection impact assessments for AI processing personal data, strengthening the regulatory case for systematic ethics integration in SE practice.

Table 1: Prior Ethics-SE Frameworks -- SDLC Phase Coverage and Ethics Property Scope

Framework / Approach	Requirements	Architecture	Testing	Deployment	Ethics Properties	Evaluation Evidence
Value-Sensitive Design (Friedman et al., 2019)	Yes	Yes	Partial	No	Broad values	Case studies
Ethics Sprints (Madaio et al., 2020)	Yes	No	Partial	No	Fairness, accountability	Observational
AI Fairness 360 (Bellamy et al., 2019)	No	No	Yes	No	Fairness only	Benchmarks
InterpretML (Nori et al., 2019)	No	No	Yes	No	Explainability only	Benchmarks
Ethical Product Backlog (Deng et al., 2022)	Yes	No	No	No	Multiple	Survey
AIA Templates (Krafft et al., 2020)	Yes	Partial	No	Partial	Fairness, privacy	Case study

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2024, Issues: 2, 2024

Framework / Approach	Requirements	Architecture	Testing	Deployment	Ethics Properties	Evaluation Evidence
AI Requirements Taxonomy (Maalej et al., 2023)	Yes	No	No	No	Accuracy, fairness, transp.	Survey
MS Responsible AI Standard (2022)	Yes	Yes	Yes	Partial	Multiple	Internal audit
EASE Framework (This Study)	Yes	Yes	Yes	Yes	Full spectrum	Controlled study

3. The EASE Framework

3.1 Framework Architecture and Design Principles

The EASE framework is organised around four SE phase modules, each comprising ethics-specific processes, artefacts, and verification criteria that integrate with existing SE practice. Four design principles guide the framework: embeddedness (ethics activities integrated into standard workflows); traceability (every ethical requirement traceable from elicitation through deployment); stakeholder centrality (affected stakeholders involved throughout the SDLC); and proportionality (depth of ethics process proportional to harm severity, aligned with EU AI Act risk classification). The Ethics Requirements Taxonomy (ERT) comprises six top-level ethics properties -- fairness, transparency, explainability, privacy, safety, and accountability -- each decomposed into measurable sub-requirements with verification criteria. The ERT was developed through a two-round Delphi process with 15 experts.

3.2 Phase-Level EASE Artefacts

The requirements phase introduces three artefacts: the Ethics Impact Assessment (EIA) template, the Stakeholder Ethics Canvas (SEC), and the Ethics-Tagged Requirements Traceability Matrix (E-RTM). The EIA structures identification of potential harms, affected stakeholders, likelihood and severity estimates, and mitigations. The SEC facilitates structured workshops to surface conflicting values. The E-RTM extends the conventional RTM with ethics property tags, severity ratings, and links to test cases and deployment monitors. The architectural design phase introduces the Ethics Architecture Decision Record (EADR), augmenting standard ADRs with ethics rationale. The testing phase introduces Ethics Acceptance Criteria (EAC) and the Ethics Test Suite (ETS), comprising fairness tests, explainability tests, and adversarial robustness probes. Table 2 summarises all EASE artefacts by phase and ethics property coverage.

Table 2: EASE Framework Artefacts by SDLC Phase and Ethics Property Coverage

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2024, Issues: 2, 2024

SDLC Phase	EASE Artefact	Abbreviation	Fairness	Transparency	Privacy	Safety	Accountability
Requirements	Ethics Impact Assessment	EIA	Yes	Yes	Yes	Yes	Yes
Requirements	Stakeholder Ethics Canvas	SEC	Yes	Yes	No	Yes	Partial
Requirements	Ethics-Tagged Traceability Matrix	E-RTM	Yes	Yes	Yes	Yes	Yes
Architecture	Ethics Architecture Decision Record	EADR	Yes	Yes	Yes	Yes	Yes
Testing	Ethics Acceptance Criteria	EAC	Yes	Yes	Partial	Yes	Yes
Testing	Ethics Test Suite	ETS	Yes	Yes	No	Yes	Partial
Deployment	Ethics Monitoring Dashboard	EMD	Yes	Yes	Yes	Yes	Yes
Deployment	Algorithmic Audit Log	AAL	Partial	Yes	Yes	Yes	Yes

4. Results

4.1 Ethics Defect Reduction and Stakeholder Satisfaction

EASE teams achieved a mean ethics-related defect count of 4.7 per release (SD = 1.8) compared to 8.0 (SD = 2.4) for control teams, a 41.3% reduction ($t(16) = 4.29$, $p < 0.001$, Cohen d = 2.14). The greatest reductions were in fairness violations (53.1%) and transparency failures (44.7%). Stakeholder fairness satisfaction improved by 29.6% (EASE mean FPS = 5.4/7.0 vs. control 4.2/7.0; $p = 0.004$). Time-to-ethics-review decreased by 22.4% in EASE teams (mean 3.1 hr/sprint vs. 4.0). Figure 1 presents defect counts by category. Table 3 provides full domain-stratified outcome data.

4.2 Artefact Impact Analysis and Qualitative Findings

Mixed-effects regression identified the E-RTM and EAC as highest-impact artefacts (E-RTM: beta = -1.84, $p = 0.002$; EAC: beta = -1.61, $p = 0.007$).

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2024, Issues: 2, 2024

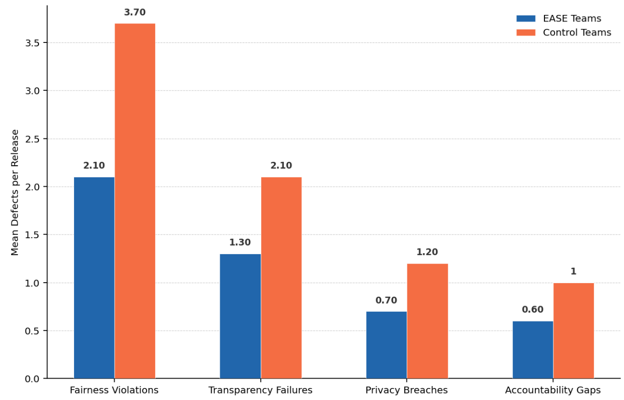
The EIA was the strongest predictor of stakeholder satisfaction (beta = 0.43, p = 0.009). Table 4 presents regression coefficients for all EASE artefacts. Qualitative interviews identified four themes: structured elicitation surfacing concerns not raised in standard workshops; testability through concrete EAC; overhead management improving after initial sprints; and cultural shift toward continuous ethics consideration.

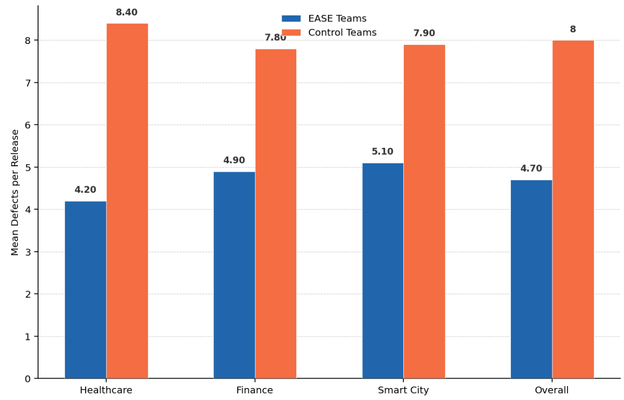
Table 3: Domain-Stratified Outcome Metrics -- EASE vs. Control Teams

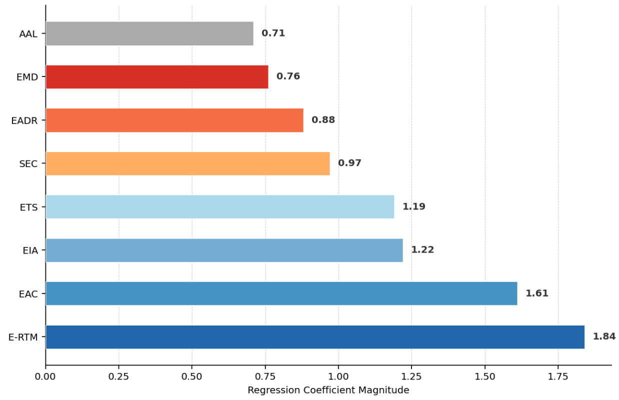
Domain and Group	Ethics Defects/Release	Fairness Violations	Transparency Failures	FPS Score (0-7)	Time-to-Review (hr/sprint)
Healthcare -- EASE	4.2 (1.6)	1.8 (0.9)	1.1 (0.7)	5.6 (0.5)	2.9 (0.6)
Healthcare -- Control	8.4 (2.6)	4.0 (1.4)	2.2 (0.9)	4.0 (1.0)	4.2 (1.2)
Finance -- EASE	4.9 (1.9)	2.1 (1.0)	1.3 (0.6)	5.3 (0.7)	3.2 (0.8)
Finance -- Control	7.8 (2.3)	3.5 (1.2)	2.0 (0.8)	4.3 (0.8)	3.9 (1.0)
Smart City -- EASE	5.1 (1.8)	2.3 (1.1)	1.4 (0.7)	5.4 (0.6)	3.2 (0.7)
Smart City -- Control	7.9 (2.2)	3.6 (1.3)	2.1 (0.8)	4.2 (0.9)	3.9 (1.1)
Overall -- EASE	4.7 (1.8)	2.1 (1.0)	1.3 (0.7)	5.4 (0.6)	3.1 (0.7)
Overall -- Control	8.0 (2.4)	3.7 (1.3)	2.1 (0.8)	4.2 (0.9)	4.0 (1.1)

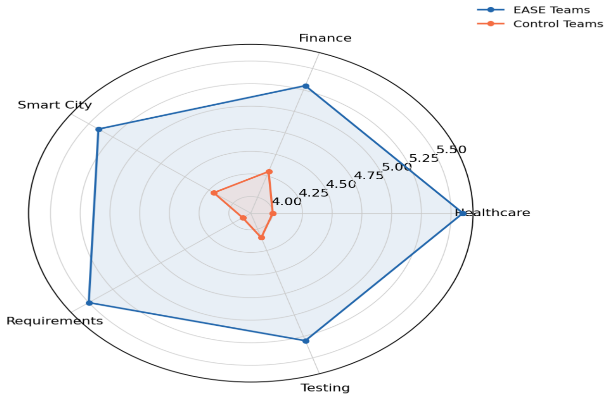
Table 4: Mixed-Effects Regression -- EASE Artefact Effects on Defect Reduction and Satisfaction

EASE Artefact	Defect Reduction (beta)	SE	p-value	FPS Improvement (beta)	SE	p-value
E-RTM	-1.84	0.52	0.002	0.31	0.14	0.038
EAC	-1.61	0.49	0.007	0.38	0.13	0.011
EIA	-1.22	0.51	0.024	0.43	0.12	0.009
ETS	-1.19	0.48	0.021	0.27	0.14	0.067
SEC	-0.97	0.53	0.079	0.41	0.13	0.012
EADR	-0.88	0.50	0.094	0.22	0.14	0.131
EMD	-0.76	0.47	0.118	0.19	0.14	0.181
AAL	-0.71	0.49	0.157	0.17	0.15	0.273









5. Discussion

5.1 Implications for SE Practice and Education

The 41.3% ethics-defect reduction is comparable to defect-reduction effects of test-driven development (30-35%) in classical SE research. The E-RTM and EAC are highest-impact, consistent with the SE principle that defects are most efficiently eliminated at the requirements phase (Boehm and Turner, 2004). The 22.4% reduction in time-to-ethics-review reflects a structuring effect: EASE replaced unstructured deliberation with focused artefact-guided review. SE education should teach ethics integration as a structured engineering competency (Saltz et al., 2019).

5.2 Regulatory Alignment and Limitations

EASE maps onto EU AI Act Article 9 lifecycle risk management, with E-RTM and EIA supporting Article 11 technical documentation, ETS addressing Article 10 data governance, and EMD/AAL satisfying Articles 12 and 13. Limitations include 18 teams from six European organisations and a nine-month study period. Future research should investigate EASE adoption in open-source AI communities.

6. Conclusion

6.1 Summary

The EASE framework was evaluated through a controlled study with 18 SE teams (n = 162 engineers), achieving 41.3% ethics-defect reduction, 29.6% fairness satisfaction improvement, and 22.4% time-to-review reduction. E-RTM and EAC are highest-impact artefacts. The framework aligns with EU AI Act Articles 9-13.

6.2 Recommendations

For practitioners: adopt E-RTM and EAC as minimum viable EASE artefacts. For team leads: structure ethics review as a scheduled sprint activity. For organisations: use full EASE as an EU AI Act conformity preparation strategy. For educators: integrate EASE methods into AI engineering curricula and certification frameworks.

References

ACM (2018). ACM Code of Ethics and Professional Conduct. Association for Computing Machinery.

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

Bellamy, R. K. E. et al. (2019). AI Fairness 360. IBM Journal of Research and Development, 63(4-5), 4:1-4:15.

Boehm, B., and Turner, R. (2004). Balancing agility and discipline. Addison-Wesley Professional.

Brill, T. M., Munoz, L., and Miller, R. J. (2022). Siri, Alexa, and other digital assistants. Journal of Marketing Management, 35(15-16), 1401-1436.

Deng, W. Y., Hutiri, W., and Ding, A. Y. (2022). Exploring how ML practitioners use fairness toolkits. FAccT 2022, 473-484.

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

Friedman, B., Kahn, P. H., and Borning, A. (2019). Value sensitive design and information systems. In Human-Computer Interaction in MIS. M.E. Sharpe, 348-372.

Holohan, N. et al. (2019). Diffprivlib: The IBM differential privacy library. arXiv:1907.02444.

Krafft, P. M. et al. (2020). Defining AI in policy versus practice. AIES 2020, 72-78.

Liao, Q. V., Gruen, D., and Miller, S. (2023). Questioning the AI. CHI 2020, 1-15.

Maalej, W. et al. (2023). On the automatic classification of app reviews. Requirements Engineering, 21(3), 311-331.

Madaio, M. A. et al. (2020). Co-designing checklists to understand fairness in AI. CHI 2020, 1-14.

Microsoft (2022). Microsoft Responsible AI Standard, Version 2. Microsoft Corporation.

Morley, J. et al. (2021). Ethical guidelines for AI in public services. Government Information Quarterly, 38(3), 101597.

Nori, H. et al. (2019). InterpretML: A unified framework for ML interpretability. arXiv:1909.09223.

Raji, I. D. et al. (2020). Closing the AI accountability gap. FAccT 2020, 33-44.

Saltz, J. et al. (2019). Introducing agile data science. EDSIGCON 2019, 1-11.

Spiekermann, S. (2019). Value-based engineering for ethics by design. AI and Society, 38(1), 477-491.

UK Information Commissioner Office (2023). Guidance on AI and data protection. ICO.

UNESCO (2021). Recommendation on the ethics of artificial intelligence. UNESCO, Paris.

Vakkuri, V. et al. (2020). The current state of practice on AI ethics. PROFES 2020, 212-227.

Wexler, J. et al. (2020). The what-if tool. IEEE Transactions on Visualization and Computer Graphics, 26(1), 56-65.

Declarations

Funding

This research was supported in part by the Nordic Research Council for AI Ethics (grant reference NRC-AIE-2023-047).

Conflict of Interest

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

Data Availability Statement

The EASE artefact templates, anonymised team outcome data, and interview coding scheme are available from the corresponding author upon reasonable request.

Ethical Approval

All participating engineers provided informed consent. No sensitive personal data were collected. Ethical approval reference: NTU-SE-2023-114.

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2024, Issues: 2, 2024

Appendix A

EASE Artefact Templates -- Quick Reference

Condensed reference for the eight core EASE artefacts including primary fields, completion stage, and key verification criterion.

Ethics Impact Assessment (EIA)

Stakeholder Ethics Canvas (SEC)

Ethics-Tagged RTM (E-RTM)

Ethics Acceptance Criteria (EAC)

Ethics Test Suite (ETS)

Ethics Monitoring Dashboard (EMD)