

E²-RE4AI: An Ethics-Embedded Requirements Engineering Framework for Lifecycle-Integrated AI Systems

Samiullah Khan

samiullah.se@must.edu.pk

Dept. of Software Engineering MUST, Mirpur, AJK, Pakistan

Hasan Asif

hassanasif2003@gmail.com

Dept. of Software Engineering (Student) MUST, Mirpur, AJK, Pakistan

Anum Tariq

anum.se@must.edu.pk

Dept. of Software Engineering MUST, Mirpur, AJK, Pakistan

Shamila Nasreen

shamila.se@must.edu.pk

Dept. of Software Engineering MUST, Mirpur, AJK, Pakistan

Tehmina Khalil

tehminka.se@must.edu.pk

Dept. of Software Engineering MUST, Mirpur, AJK, Pakistan

Corresponding Author: Samiullah Khan samiullah.se@must.edu.pk

Received: 09-02-2026

Revised: 02-03-2026

Accepted: 10-03-2026

Published: 24-03-2026

ABSTRACT

In artificial intelligence (AI) systems, fairness, pri-vacy and explainability are seldom expressed as requirements that can be verified. Existing approaches stay at the level of principles, focus on model training, or rely on a single audit after the system is built. We present E²-RE4AI, a requirements engineering (RE) framework that converts ethical principles into measurable ethical non-functional requirements (E-NFRs). Conflicts between these requirements are settled with an analytic hierarchy process whose inputs and result are recorded. The requirements are verified at five gates across the lifecycle and kept under monitoring after deployment; a violation re-opens the design decision it affects. In a synthetic scholarship-admission case, a model that failed the interpretability requirement was ruled out. With weights that put fairness first, the framework selected an equal-opportunity design that narrowed the selection-rate gap from 0.14 to 0.02 without reducing accuracy. Weights that put privacy first, and a legal prohibition on group-specific thresholds, each led to a different selection, and the reasons for both were documented. Across 200 simulated deployments, the monitoring rule produced no false violations and confirmed an injected drift in 73.5% of runs. After a second protected attribute was introduced, thresholds tuned only for income left a gap of 0.107 between the most and least favourably treated subgroups, which intersectional thresholds lowered to 0.073. The framework can be applied; whether it improves outcomes in practice remains to be tested.

Keywords: AI ethics, fairness, multi-criteria decision analysis, requirements engineering, traceability

INTRODUCTION

AI systems already rank university applicants, grant credit and set the order in which patients are treated. Because their behaviour is learned from data, their errors fall unequally on different social groups. Requirements engineering (RE) has been slow to react. Mapping studies rank ethics, trust and explainability among the least mature topics in RE for AI [1], [2], and many practitioners continue to document AI requirements with UML diagrams and office documents [3]. Researchers in AI ethics,

for their part, have largely agreed on a few core principles [4] and built more than a hundred tools for putting them into practice [5], [6].

These two lines of work remain disconnected. Principles are rarely stated as requirements with metrics and thresholds, which means no one can verify whether a system satisfies them [7], [8]. When principles clash, teams lack an agreed, documented method for deciding which one should yield [9], [10]. Ethics is also typically assessed at design time, even though a model that was fair when launched may become unfair over time [11]. Regulation has made this pressing. Under the EU AI Act, a system that decides who is admitted to education is high-risk: it must have risk management, data governance and human oversight in place, and it must be monitored after it reaches the market [12].

We address the following research question: how can ethical principles be converted into measurable, traceable requirements whose conflicts are resolved transparently and which remain satisfied across the AI lifecycle? This paper offers three contributions:

- E²-RE4AI, a framework that turns every ethical principle into an E-NFR with a metric, a threshold and a checking gate, and traces it from the originating principle through to any subsequent monitoring alert (Section III);
- a trade-off procedure that first eliminates designs failing a mandatory requirement, then ranks the remaining ones using analytic hierarchy process (AHP) weights and tests the stability of that ranking, recording each step (Section III);
- a reproducible case study on synthetic admission data that shows how the decision shifts with stakeholder values and legal constraints, and quantifies how frequently the monitoring rule produces false or missed alarms (Section IV).

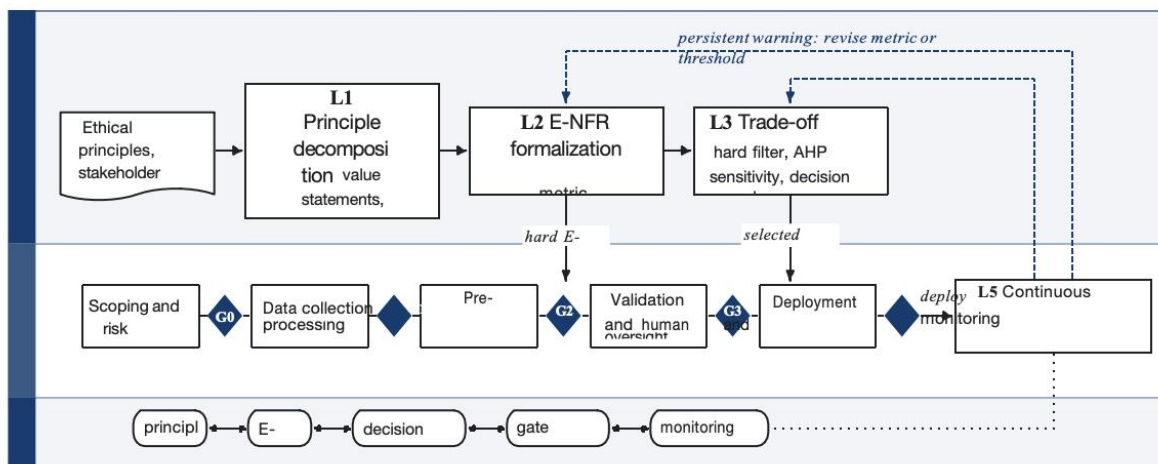


Fig. 1. Overview of E2-RE4AI. The specification lane (L1–L3) converts principles into E-NFRs and a chosen design; the lifecycle lane verifies that design at gates G0–G4 (L4) and monitors it once deployed (L5); dashed arrows feed monitoring results back to L3 and L2; the trace lane (L6) connects all artefacts.

Related Work

RE for AI. Horkoff framed fairness and explainability as non-functional requirements (NFRs) of machine learning and identified their measurement as the central challenge [7]. A subsequent practitioner survey confirmed that such NFRs are seldom measured in practice [8]. Guizzardi *et al.* regard ethical requirements as a distinct category [13]. The RE4AI Ethical Guide supports teams in

eliciting them [14], and the Ethical Requirements Stack integrates them into agile portfolio management [15]. None of this work quantifies the requirements, prioritises them against one another, or monitors them during operation.

Lifecycle frameworks. IEEE Std 7000 obtains ethical value requirements from stakeholders' concerns [16]. ECCOLA em-beds ethics in agile sprints [17], and Z-Inspection evaluates the overall trustworthiness of a system [18]. SMACTR defines a staged internal audit [19], and the Responsible AI Pattern Catalogue gathers patterns for governance and engineering [20]. At the organisational level, the NIST AI RMF [21] and ISO/IEC 42001 [22] provide a structure for managing AI risk. None of these frameworks converts principles into requirements with thresholds that can be checked by a gate or a monitor.

Trade-offs. A large share of the practical difficulty of AI ethics arises because principles pull in opposite directions [9], [10]. Some of these tensions are mathematical. When groups differ in base rate, a risk score cannot be calibrated and give every group the same error rates at once [23], [24], and differential privacy reduces accuracy most for small groups [25]. Other tensions are contested, such as the claim that interpretable models are necessarily less accurate [26]. Assessing fairness one attribute at a time may also conceal harm to subgroups defined by combinations of attributes [27]–[29]. Sanderson *et al.* suggested that trade-offs be identified, weighted and justified [30]. Their approach does not, however, link the decision to specific requirements or to the gates that subsequently verify them.

Table 1: Comparison With The Most Closely Related Approaches (Authors' Assessment)

Approach	Measurable Requirement	Trade-off Procedure	Phase gates	Drift Linked to requirement	Regulation anchors
IEEE 7000 [16]	●	●	—	—	—
ECCOLA [17]	●	—	●	—	—
Eth. Req. Stack [15]	●	—	—	—	●
SMACTR [19]	—	—	●	—	—
Z-Inspection [18]	●	●	—	—	●
RAI Patterns [20]	●	●	●	●	●
Sanderson et al. [30]	—	●	—	—	—
E²-RE4AI	●	●	●	●	●

●explicit; ●partial or informal; — not a primary focus.

Monitoring. Shifts in the data distribution can undermine a fairness guarantee [31]. Fairness drift has been documented in deployed models [11], and runtime fairness monitors have already been developed [32]. These monitors do not link an alert to the requirement it violates or to the decision that needs to be reconsidered.

Table 1 summarises this comparison. Each layer of E²-RE4AI draws on prior work. The novelty lies in carrying one E-NFR through the entire chain: it is specified, balanced against competing requirements, verified at gates and monitored after release.

THE E²-RE4AI FRAMEWORK

E²-RE4AI comprises five process layers together with a traceability layer that runs in parallel with them (Fig. 1).

L1: Principle decomposition. Every principle is decomposed into context-specific value statements, accompanied by a map of the stakeholders involved and the types of harm in question. In admissions, for instance, *fairness* is expressed as “low-income applicants are not less likely to receive a

scholarship than equally able high-income applicants.” The stakeholders identified at this stage later supply the weights used in L3. The panel should comprise representatives of applicants, admissions officers, an ethicist, the data-protection officer and a legal adviser. Every member fills in the pairwise comparisons independently using a brief form that calculates their weights and consistency ratio immediately, allowing inconsistent judgements to be corrected straight away. Thresholds are proposed on the basis of legal or institutional policy where such a policy exists, and otherwise from a baseline measurement; the panel then approves them.

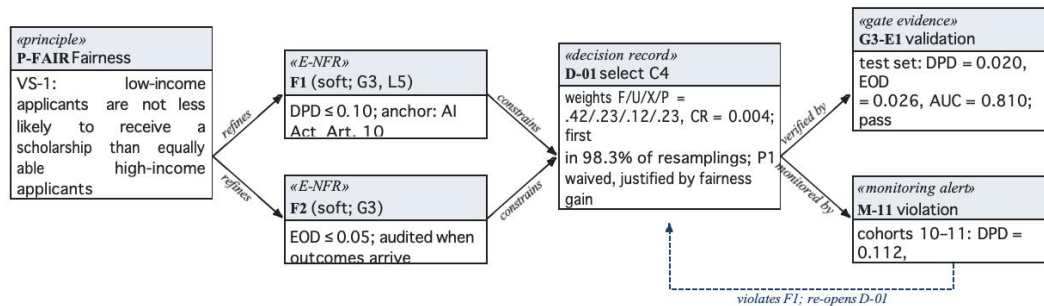


Fig. 2. Trace kept by L6 for the fairness principle in the fairness-first scenario. Monitoring alert M-11 violates F1 and re-opens decision D-01.

Table 2: E-Nfrs Defined For The Scholarship-Admission Case

ID	Metric and threshold	Type	Gate	Indicative anchor
A1	$AUC \geq 0.75$	hard	G3	AI Act Art. 15
X1	intrinsically interpretable	hard	G2	AI Act Art. 13
F1	model selection-rate gap (DPD)	soft	G3, L5	AI Act Art. 10
F2	≤ 0.10 true-positive-rate gap	soft	G3	AI Act Art. 10
P1	$(EOD) \leq 0.05$ no protected attribute at inference	soft	G2	data minimisation

L2: E-NFR formalization. We define an E-NFR as the tuple

$$r = \langle id, p, s, m, \triangleright \triangleleft, \theta, \kappa, g, \ell \rangle, \quad (1)$$

where p is the principle, s the stakeholder source, m the metric, $\triangleright \triangleleft \in \{\leq, \geq\}$ the comparator, θ the threshold, $\kappa \in \{\text{hard}, \text{soft}\}$ the status, g the gate that checks it and ℓ an indicative regulatory anchor. A design c satisfies r if $\text{sat}(r, c) = \mathbf{1}[m(c) \triangleright \triangleleft \theta]$. Hard E-NFRs are mandatory. Soft E-NFRs can be traded off in L3, provided that the trade-off is documented. E-NFRs can be specified over intersectional subgroups as well as over single attributes, e.g., as the largest selection-rate gap between any pair of subgroups, or as an equalized-odds gap that covers both true- and false-positive rates [33]. The E-NFRs used in the main case study are given in Table II.

L3: Trade-off resolution. Designs that fail any hard E-NFR are removed. The remaining designs constitute the feasible set

$$F = \{c \in C : \text{sat}(r, c) = 1 \quad \forall r \text{ with } \kappa(r) = \text{hard}\}. \quad (2)$$

Next, stakeholders rate each pair of criteria from 1 to 9 using Saaty’s scale. The weight vector w is the principal eigenvector of the comparison matrix built from these ratings, and the judgements are accepted when the consistency ratio meets $CR < 0.10$ [34]. Individual judgements are merged through the element-wise geometric mean of the members’ matrices, which is the standard way of aggregating

individual judgements in group AHP [35]. Every criterion value x_{cj} is min–max normalised over F to $\tilde{x}_{cj}$, and designs are ordered. According to

$$S(c) = \sum_{j=1}^n w_j \tilde{x}_{cj}. \quad (3)$$

To check the robustness of the ranking, we resample the weights, re-rank the designs with TOPSIS, and repeat the full procedure for every stakeholder profile. The decision record stores the pairwise matrices, the scores and these robustness checks. It also records any soft E-NFR that the selected design does not meet, together with the justification for accepting it. L4: Lifecycle gates. At G0 (scoping), the team determines the risk class and records any legal constraints; the AI Act lists admissions systems as high-risk in its Annex III [12]. G1 (data collection) verifies that every group is sufficiently represented. G2 (training) examines the way bias enters the data and checks X1 and P1. G3 (validation) assesses A1, F1 and F2 on held-out data and confirms that a human is able to override a decision. G4 (deployment) reviews residual harms and activates L5. The gates mirror SACTR’s staged audit [19], except that each of them tests the E-NFRs defined in L2.

L5: Monitoring. Whether a scholarship recipient graduates becomes known only several years after the decision. For this reason, L5 tracks DPD, which does not require outcome labels, for each cohort, and audits EOD once graduation results become available. For every cohort t , it calculates $DPD_t = |SR_{0,t} - SR_{1,t}|$ from the selection rates of the groups, together with a 95% interval $[L_t, U_t]$. A *warning* (investigate) is issued when $U_t > \theta$, and a *violation* (act) when both $DPD_t > \theta$ and $DPD_{t-1} > \theta$. A violation returns the case to L3, where the original decision is revisited.

L6: Traceability. We adopt the classic definition of requirements traceability [36]. L6 connects every principle to its E-NFRs, and every E-NFR to the decisions, gate evidence and alerts related to it (Fig. 2).

Illustrative Case Study

Setup. A university awards scholarships to the 30% of applicants with the highest likelihood of graduating. In each run we generated 20 000 synthetic applicants, each described by college marks, an entry-test score, an interview score and a “high-cost college bonus”. Our generator makes graduation depend solely on an applicant’s underlying ability. Even so, low-income applicants (35% of the total) obtain lower test and interview scores than peers of equal ability and rarely receive the bonus, which therefore serves as a proxy for income. The results rest on these assumptions and do not speak to real admissions data.

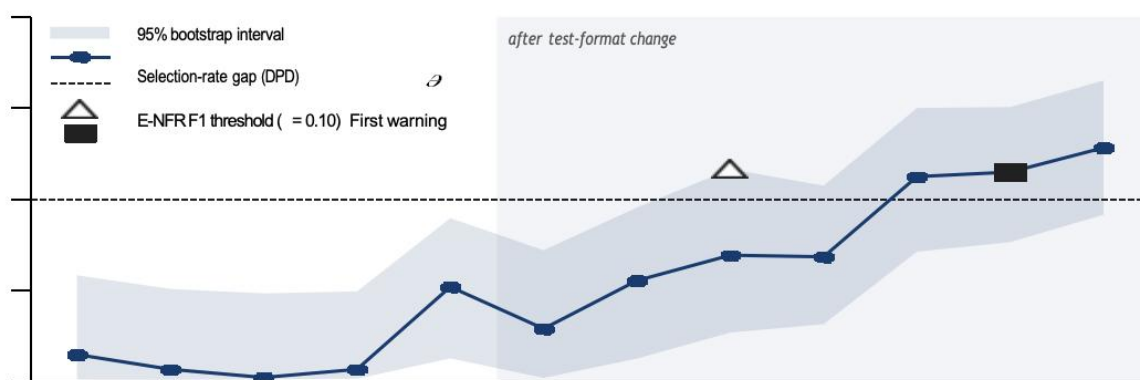


Fig. 3. Selection-rate gap of the deployed design across twelve synthetic cohorts, shown with 95% bootstrap intervals.

Table 3: Candidate Designs (Mean $\pm$ Sd Over Five Seeds)

<i>Design</i>	<i>AUC</i>	<i>Utility</i>	<i>DPD</i>	<i>EOD</i>	<i>S(c)</i>
C1 LR, all	.810 $\pm$.003	.850	.117 $\pm$.009	.155 $\pm$.021	.13
C2 LR, no proxy	.810 $\pm$.003	.849	.140 $\pm$.015	.187 $\pm$.026	.24
C3 C2+reweigh	.810 $\pm$.003	.849	.141 $\pm$.015	.188 $\pm$.027	.23
C4 C3+thresholds	.810 $\pm$.003	.853	.020 $\pm$.005	.026 $\pm$.017	65
C5 boosting	.803 $\pm$.004	.844	.124 $\pm$.014	.165 $\pm$.020	fails X1

We compared five designs. C1 is logistic regression (LR) using all features, and C2 is that model with the proxy re-moved. C3 applies reweighing to C2 [37]. C4 extends C3 with group-specific thresholds selected for equal opportunity [33]. C5 is a gradient-boosting model. Every design was trained, tuned and tested on a 60/20/20 split, repeated over five random seeds. The AHP criteria were fairness (mean of DPD and EOD), utility (proportion of scholarship holders who graduate), explainability and privacy. All E-NFR thresholds were set before any model was trained. The pairwise judgements are the authors' own and serve as stand-ins for three stakeholder profiles.

Gates and trade-off. The results appear in Table III. C5 failed hard requirement X1, and it was also marginally less accurate than the linear models, in line with [26]. Reweighing (C3) made no difference, since the bias lies in the features rather than in the relationship between group and outcome. Dropping the proxy (C2) *increased* the gap, because in this data the proxy partly compensates for the under-measurement of low-income applicants. This is why G2 examines *the source* of bias before a remedy is chosen. C4 was the only design that satisfied both F1 and F2. It reduced the gap from 0.140 to 0.020 while leaving AUC unchanged, but it violates P1 because it relies on income group at decision time.

TABLE 4: Decision Under Varying Profiles and Constraints

Scenario	Weights F/U/X/P	Choice	Unmet
Fairness-first	.42/.23/.12/.23	C4	P1
Utility-first	.19/.52/.11/.19	C4	P1
Privacy-first	.22/.12/.12/.54	C2	F1, F2
P1 made hard	.42/.23/.12/.23	C1	F1, F2

Under fairness-first weights (fairness 0.42, utility 0.23, privacy 0.23, explainability 0.12; CR = 0.004), C4 came first. It remained on top in 98.3% of 10 000 randomly perturbed weight vectors and under every $\pm 25\%$ change to any single weight, and TOPSIS yielded the same ordering. Table IV reports the effect of altering the values or the constraints. C4 was still chosen under utility-first weights, whereas privacy-first weights led to C2. If a legal prohibition on group-specific thresholds turns P1 into a hard requirement, C4 is ruled out and C1 is selected instead; F1 and F2 are then documented as unmet and the case is escalated to G0. Combining the three profiles with the geometric-mean rule yielded weights of 0.29 for fairness, 0.27 for utility, 0.32 for privacy and 0.13 for explainability (CR = 0.001), and C4 was selected once more. The trace recorded for the fairness-first decision appears in Fig. 2.

Monitoring. C4 was then deployed on twelve cohorts of 2 000 applicants each. Starting at cohort 6, a new test format raises the penalty on low-income applicants by 0.12 SD per cohort. In the run shown in Fig. 3, the first warning occurred at cohort 8 and the first violation at cohort 11. The simulation was run 200 times. In the absence of drift, the rule never signalled a violation, though at least one warning occurred in 34.5% of runs; for this reason a warning triggers only an investigation. Under drift, the rule confirmed a violation within twelve cohorts in 73.5% of runs, with a median at cohort 11.

Intersectional extension. To examine fairness beyond a single binary attribute, we introduced a second protected attribute into the generator: 30% of applicants are rural and obtain lower interview and college-mark scores than urban applicants of equal ability. Three designs were compared over five seeds (Table V): LR without the proxy (C2), thresholds tuned only for income (C4), and thresholds tuned for the four income–region subgroups (C4-I). Income-only thresholds nearly eliminated the income gap, yet rural applicants continued to be selected less often, and the gap between the most and least favourably treated subgroups remained at 0.107. Intersectional thresholds lowered this gap to 0.073, increased the ratio of the lowest to the highest subgroup selection rate from 0.68 to 0.79, and reduced the equalized-odds gap from 0.159 to 0.097, with a slight loss of utility (0.850 to 0.846). An E-NFR defined for income alone would have let C4 pass G3 while overlooking the rural gap; intersectional E-NFRs therefore belong in L2.

Table 5: Intersectional Extension: Income × Region (Mean Over Five Seeds)

Design	DPD income	DPD region	Subgroup SR gap	Min/max SR ratio	EO gap (4 groups)	Utility
C2	.137	.066	.200	.46	.280	.850
C4 (income)	.025	.076	.107	.68	.159	.850
C4-I (intersect.)	.032	.023	.073	.79	.097	.846

SR: selection rate. EO gap: the largest difference in true- or false-positive rate between any pair of the four subgroups. All three designs had an AUC of 0.804.

DISCUSSION AND THREATS TO VALIDITY

The case study demonstrates that E²-RE4AI can be executed end to end, that its selection changes in explainable ways when values or legal constraints change, and that each decision leaves a record that an auditor can trace. It does not demonstrate that the framework yields better outcomes in real projects. Every result is derived from synthetic data with known bias mechanisms, whereas real projects involve noisier data, more stakeholders and competing organisational pressures. Group fairness measures, the intersectional ones introduced here included, may still make existing inequality appear acceptable [38], and they lose reliability as subgroups shrink. The weights in this study are proxies for stakeholder panels: the three profiles were set by the authors, and although the aggregation rule illustrates how a real panel’s judgements would be merged, it is no substitute for such a panel. The regulatory anchors are only indicative, and the lawfulness of group-specific thresholds must be determined at G0 for each jurisdiction.

CONCLUSION

E²-RE4AI treats ethical principles as first-class requirements. They are quantified, prioritised through a documented procedure, verified at lifecycle gates and monitored after de-ployment, and each alert leads back to the decision it concerns. The case study also showed that fairness requirements need to cover intersectional subgroups, not only single attributes. In future work, we will apply the framework together with practitioners on real AI projects and gather their feedback on the E-NFR template, the elicitation form and the decision record. We also plan to elicit weights and thresholds from real stakeholder panels and to build tool support for more of the elicitation and monitoring. The code and the data generator are available at [repository URL].

ACKNOWLEDGMENT

The authors used an AI assistant for language editing and to draft parts of the text, figures and code. The authors reviewed and checked all of this content and ran all of the code themselves.

REFERENCES

- K. Ahmad, M. Abdelrazek, C. Arora, M. Bano, and J. Grundy, "Re-quirements engineering for artificial intelligence systems: A systematic mapping study," *Inf. Softw. Technol.*, vol. 158, 2023, Art. no. 107176.
- Umm-e-Habiba, M. Haug, J. Bogner, and S. Wagner, "How mature is requirements engineering for AI-based systems? A systematic mapping study on practices, challenges, and future research directions," *Require-ments Eng.*, vol. 29, pp. 567–600, 2024.
- K. Ahmad, M. Abdelrazek, C. Arora, M. Bano, and J. Grundy, "Requirements practices and gaps when engineering human-centered artificial intelligence systems," *Appl. Soft Comput.*, vol. 143, 2023, Art. no. 110421.
- A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Mach. Intell.*, vol. 1, pp. 389–399, 2019.
- J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices," *Sci. Eng. Ethics*, vol. 26, no. 4, pp. 2141–2168, 2020.
- E. Prem, "From ethical AI frameworks to tools: A review of approaches," *AI Ethics*, vol. 3, pp. 699–716, 2023.
- J. Horkoff, "Non-functional requirements for machine learning: Chal-lenges and new directions," in *Proc. IEEE 27th Int. Requirements Eng. Conf. (RE)*, 2019, pp. 386–391.
- K. M. Habibullah, G. Gay, and J. Horkoff, "Non-functional requirements for machine learning: Understanding current use and challenges among practitioners," *Requirements Eng.*, vol. 28, pp. 283–316, 2023.
- J. Whittlestone, R. Nyrupe, A. Alexandrova, and S. Cave, "The role and limits of principles in AI ethics: Towards a focus on tensions," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc. (AIES)*, 2019.
- H. Bleher and M. Braun, "Reflections on putting AI ethics into practice: How three AI ethics approaches conceptualize theory and practice," *Sci. Eng. Ethics*, vol. 29, no. 3, 2023, Art. no. 21.
- S. E. Davis, C. Dorn, D. J. Park, and M. E. Matheny, "Emerging algo-rithmic bias: Fairness drift as the next dimension of model maintenance and sustainability," *J. Amer. Med. Inform. Assoc.*, vol. 32, no. 5, pp. 845–854, 2025.
- European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Off. J. Eur. Union*, L series, Jul. 2024. [Online]. Available: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- R. Guizzardi, G. Amaral, G. Guizzardi, and J. Mylopoulos, "Ethical requirements for AI systems," in *Advances in Artificial Intelligence (Canadian AI 2020)*, ser. LNCS. Springer, 2020.
- J. A. Siqueira de Cerqueira, A. P. De Azevedo, H. A. Tives, and E. D. Canedo, "Guide for artificial intelligence ethical requirements elicitation – RE4AI ethical guide," in *Proc. 55th Hawaii Int. Conf. Syst. Sci. (HICSS)*, 2022.

- M. Agbese, R. Mohanani, A. A. Khan, and P. Abrahamsson, "Ethical requirements stack: A framework for implementing ethical requirements of AI in software engineering practices," in *Proc. 27th Int. Conf. Eval. Assess. Softw. Eng. (EASE)*, 2023, pp. 326–328.
- "IEEE standard model process for addressing ethical concerns during system design," IEEE Std 7000-2021, 2021.
- V. Vakkuri, K.-K. Kemell, M. Jantunen, E. Halme, and P. Abrahamsson, "ECCOLA – A method for implementing ethically aligned AI systems," *J. Syst. Softw.*, vol. 182, 2021, Art. no. 111067.
- R. V. Zicari *et al.*, "Z-Inspection®: A process to assess trustworthy AI," *IEEE Trans. Technol. Soc.*, vol. 2, no. 2, pp. 83–97, 2021.
- I. D. Raji *et al.*, "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proc. Conf. Fairness, Accountability, Transparency (FAT*)*, 2020, pp. 33–44.
- Q. Lu, L. Zhu, X. Xu, J. Whittle, D. Zowghi, and A. Jacquet, "Responsible AI pattern catalogue: A collection of best practices for AI governance and engineering," *ACM Comput. Surv.*, vol. 56, no. 7, pp. 1–35, 2024.
- National Institute of Standards and Technology, "Artificial intelligence risk management framework (AI RMF 1.0)," NIST, Gaithersburg, MD, USA, Tech. Rep. NIST AI 100-1, Jan. 2023.
- "Information technology—artificial intelligence—management system," ISO/IEC 42001:2023, 2023.
- Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *Proc. 8th Innov. Theor. Comput. Sci. Conf. (ITCS)*, ser. LIPIcs, vol. 67, 2017, pp. 43:1–43:23.
- A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.
- E. Bagdasaryan, O. Poursaeed, and V. Shmatikov, "Differential privacy has disparate impact on model accuracy," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019.
- C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, pp. 206–215, 2019.
- J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. Conf. Fairness, Accountability, Transparency (FAT*)*, ser. PMLR, vol. 81, 2018, pp. 77–91.
- M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, ser. PMLR, vol. 80, 2018, pp. 2564–2572.
- J. R. Foulds, R. Islam, K. N. Keya, and S. Pan, "An intersectional definition of fairness," in *Proc. IEEE 36th Int. Conf. Data Eng. (ICDE)*, 2020.

- C. Sanderson, E. Schleiger, D. Douglas, P. Kuhnert, and Q. Lu, "Re-solving ethics trade-offs in implementing responsible AI," in *Proc. IEEE Conf. Artif. Intell. (CAI)*, 2024, arXiv:2401.08103.
- A. Barrainkua *et al.*, "Preserving the fairness guarantees of classifiers in changing environments: A survey," *ACM Comput. Surv.*, vol. 57, no. 6, 2025.
- T. A. Henzinger, M. Karimi, K. Kueffner, and K. Mallik, "Runtime monitoring of dynamic fairness properties," in *Proc. ACM Conf. Fairness, Accountability, Transparency (FAccT)*, 2023.
- M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2016.
- T. L. Saaty, "Decision making with the analytic hierarchy process," *Int. J. Services Sci.*, vol. 1, no. 1, pp. 83–98, 2008.
- E. Forman and K. Peniwati, "Aggregating individual judgments and priorities with the analytic hierarchy process," *Eur. J. Oper. Res.*, vol. 108, no. 1, pp. 165–169, 1998.
- O. C. Z. Gotel and A. C. W. Finkelstein, "An analysis of the requirements traceability problem," in *Proc. IEEE Int. Conf. Requirements Eng. (ICRE)*, 1994, pp. 94–101.
- F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowl. Inf. Syst.*, vol. 33, no. 1, pp. 1–33, 2012.
- M. Kasy and R. Abebe, "Fairness, equality, and power in algorithmic decision-making," in *Proc. ACM Conf. Fairness, Accountability, Transparency (FAccT)*, 2021, pp. 576–58