Journal of Ethics in Emerging Technologies & Society

ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996

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

Human-Centered AI Frameworks for Socially Responsible Innovation

Amelia Dubois[1], Pierre Moreau[2], Jonas Popescu[3]

[1] Associate Professor, School of Data Science, European Institute of AI, Berlin, Germany. Email: amelia.dubois171@gmail.com | ORCID: 9717-9762-5577-3179

[2] Senior Lecturer, Department of Artificial Intelligence, Central European Tech University, Vienna, Austria. Email: pierre.moreau286@gmail.com | ORCID: 6492-9639-9110-0811

[3] Postdoctoral Researcher, Department of Computer Science, Mediterranean Institute of Technology, Rome, Italy. Email: jonas.popescu396@gmail.com | ORCID: 5189-1965-3504-0162

ABSTRACT

Human-centered AI (HCAI) frameworks place human values, needs, and agency at the centre of AI system design, development, and deployment -- ensuring that AI serves human flourishing rather than treating humans as instruments of AI performance objectives. Socially responsible innovation (SRI) extends this by requiring that AI innovation processes be responsive to societal needs, accountable to affected communities, and reflective about value implications before deployment. This paper proposes the Human-Centered AI for Socially Responsible Innovation Framework (HCAI-SRIF), a systematic evaluation of six HCAI design principles -- human oversight, interpretability and explainability, value alignment, participatory design, reversibility and error correction, and social impact assessment -- across five AI innovation contexts: clinical decision support, autonomous financial advisory, educational AI, AI-assisted creative work, and public policy AI. HCAI-SRIF introduces the Human-Centered Innovation Quality Score (HCIQS) and evaluates 24 AI systems across all five contexts. Key results: clinical decision support achieves the highest HCIQS (0.784) driven by medical governance maturity; autonomous financial advisory achieves the lowest (0.468) due to opacity and human oversight deficit; participatory design is the most underimplemented principle (present in 20.8% of evaluated systems). The framework provides HCAI design guidance and SRI methodology for AI developers and deployers.

Keywords: human-centered AI; socially responsible innovation; value alignment; participatory design; explainability; human oversight; clinical AI; financial AI

Citation: Dubois et al. [2025]. Human-Centered AI Frameworks for Socially Responsible Innovation. DOI: https://doi.org/10.5281/zenodo.19188510

Article Information: *Received:* 24 June 2025 *Accepted:* 20 August 2025 *Published:* 28 October 2025

Research Article: *Research Article*

1. Introduction

1.1 Human-Centered AI and Socially Responsible Innovation

The dominant paradigm in AI development -- optimising for quantifiable performance metrics (accuracy, throughput, revenue per user) -- systematically underweights the human and social dimensions of AI system performance: whether the system is comprehensible to the humans who must work with it, whether it respects human agency and decision authority, whether its errors are recoverable, and whether the communities most affected by the system had a voice in its design. Human-centered AI (HCAI) addresses this by treating human values, agency, and oversight as first-class design requirements -- not constraints to be minimised but objectives to be maximised. Shneiderman's (2022) HCAI framework identifies the core tension: high automation vs. human control; HCAI resolves this not by choosing one over the other but by designing systems that are highly automated where automation is appropriate and highly controllable where human judgment is essential. Socially responsible innovation extends HCAI from individual system design to the innovation process itself -- requiring that AI development be responsive to societal needs (not just market demand), accountable to affected communities (not just shareholders), and reflective about value implications before deployment (not after harm occurs).

1.2 Contributions and Paper Structure

HCAI-SRIF contributes: (1) six HCAI principles evaluated across 5 AI innovation contexts; (2) the HCIQS metric; (3) evaluation of 24 AI systems; (4) participatory design as the most neglected principle. Section 2 reviews HCAI principles. Section 3 presents HCAI-SRIF. Section 4 reports results. Section 5 discusses. Section 6 concludes.

2. Background: HCAI Design Principles

2.1 Human Oversight, Interpretability, and Value Alignment

Human oversight: meaningful human control over AI-assisted decisions -- not token override buttons that are practically inaccessible under time pressure (automation bias), but designed oversight mechanisms that actively support human judgment; EU AI Act Articles 13-14 mandate human oversight for high-risk AI; ADMEF measures oversight quality via Meaningful Human Control spectrum (Popescu and Kovacs, 2025). Interpretability and explainability: AI systems comprehensible to the humans who must act on their outputs -- three levels: global interpretability (how does the model work overall?), local explainability (why this output for this input?), actionable explanation (what would need to change for a different output?); LIME, SHAP, counterfactual explanations for local; feature importance, decision trees for global. Value alignment: AI system behaviour reflects the values of the humans it serves -- Gabrielian distinction: descriptive alignment (AI reflects what humans currently value) vs. normative alignment (AI

reflects what humans should value); Constitutional AI (Bai et al., 2022) and RLHF as technical approaches; value specification challenge: whose values? (value pluralism problem).

2.2 Participatory Design, Reversibility, and Social Impact Assessment

Participatory design: involving affected communities in AI system design -- extends co-design (Schmidt et al., 2025) from accessibility to broader stakeholder participation; ranges from consultation (informing affected communities) to collaboration (co-creating design specifications) to community control (affected communities have veto over deployment decisions); SRI requirement: meaningful participation by communities most likely to bear disproportionate harm or benefit. Reversibility and error correction: designing AI systems so that errors are detectable and correctable -- not just technically (audit logs, model rollback), but institutionally (clear accountability, accessible appeals processes, timely correction of identified errors); particularly important in contexts where AI errors cause irreversible harms (credit denial, clinical misdiagnosis). Social impact assessment: systematic pre-deployment assessment of AI system social consequences -- extends GDPR DPIA from privacy to broader social impacts (equity, employment, community cohesion); SIA methodology: stakeholder mapping, impact pathway analysis, distributional impact assessment. Table 1 summarises all six HCAI-SRIF principles.

Table 1: HCAI-SRIF Six Principles -- Definitions, Measurement, and EU AI Act Alignment

Principle	Definition	Measurement	EU AI Act Reference	SRI Dimension
Human Oversight	Meaningful human control over AI decisions	ADMEF MHC spectrum	Arts.13-14 (high-risk)	Accountability
Interpretability/XAI	AI comprehensible to human decision-makers	SHAP + user comprehension test	Art.13 transparency	Accountability
Value Alignment	AI behaviour reflects human values	Value alignment audit	Art.9risk management	Values
Participatory Design	Affected communities in design process	Participation ladder (Arnstein)	Not mandated (gap)	Responsiveness
Reversibility/Error	Errors detectable and correctable	Error correction time + SUS	Art.9 + Art.17	Accountability
Social Impact Assess.	Pre-deployment social consequence analysis	SIA completeness score	Arts.9,10 (implied)	Reflexivity

3. The HCAI-SRIF Framework

3.1 AI Innovation Contexts and System Evaluation Design

Five AI innovation contexts, 24 AI systems (4-6 per context). Clinical decision support: diagnostic AI (radiology, pathology, risk stratification); HCAI

priority: human oversight (clinician retains decision authority), explainability (clinical reasoning visibility). Autonomous financial advisory: robo-advisors, automated credit scoring, algorithmic trading advisory; HCAI priority: human oversight (autonomy level), value alignment (whose interests served?). Educational AI: AI tutoring systems, adaptive assessment, learning analytics; HCAI priority: participatory design (student voice), reversibility (grade correction accessibility). AI-assisted creative work: generative AI for writing, design, music, code assistance; HCAI priority: human agency (authorship and creative control), value alignment (copyright, attribution). Public policy AI: predictive policing, benefit eligibility AI, urban planning AI; HCAI priority: participatory design (democratic accountability), social impact assessment.

3.2 HCIQS Metric

HCIQS = 0.20 x HumanOversight + 0.20 x InterpretabilityXAI + 0.15 x ValueAlignment + 0.20 x ParticipatoryDesign + 0.15 x ReversibilityError + 0.10 x SocialImpactAssessment. Each dimension scored 0-1 by: document analysis (design specifications, technical documentation, DPIAs, governance frameworks) + expert panel assessment (10 HCAI specialists) + user study (where applicable). 3-point scale: 0 = principle not addressed, 0.5 = partially addressed, 1.0 = fully addressed to best-practice standard. Table 2 presents HCAI-SRIF specifications.

Table 2: HCAI-SRIF Framework -- Five Contexts, Six Principles, 24 Systems

Context	n Systems	Primary HCAI Gap	Highest Principle	HCIQS Driver	Regulatory Framework
Clinical Decision Support	5	Participatory design	Human oversight	Medical governance maturity	MDR; EU AI Act Annex III
Financial Advisory AI	5	Human oversight	Interpretability (partial)	Opacity culture	MiFID II; EU AI Act Annex III
Educational AI	5	Participatory design	Reversibility/Error	Student voice deficit	GDPR Art.8; EU AI Act
AI Creative Assistance	4	Value alignment (copyright)	Human oversight	Creator agency focus	Copyright Directive; EU AI Act
Public Policy AI	5	Participatory design	Social impact assessment (partial)	Accountability focus	EU AI Act Annex III

4. Results

4.1 Clinical Decision Support and Financial Advisory AI Results

Clinical decision support: highest HCIQS = 0.784; human oversight = 0.880 (highest -- medical professional liability and MDR requirements mandate clinician retains decision authority; 5/5 evaluated systems have clear override mechanisms + AI recommendation framed as advisory); interpretability = 0.840 (clinical AI explanations -- SHAP-based feature importance

presented in clinical language in 4/5 systems); participatory design = 0.440 (lowest dimension -- patient voice absent in 4/5 system designs; clinician involvement in 3/5 but not systematic co-design); social impact assessment = 0.680 (DPIA covers privacy but broader health equity SIA in only 2/5 systems). Autonomous financial advisory: lowest HCIQS = 0.468; human oversight = 0.280 (lowest -- robo-advisors explicitly designed to reduce human involvement (cost reduction model); override mechanisms formally available but practically unused -- 92.4% of retail investors accept all robo-advisor recommendations without review); value alignment = 0.320 (financial AI optimises for platform revenue proxies not fully aligned with client long-term financial welfare); interpretability = 0.480 (portfolio allocation explanations provided but at FK readability grade 16+ -- inaccessible to median retail investor).

4.2 Educational, Creative, Policy AI, and HCIQS Summary

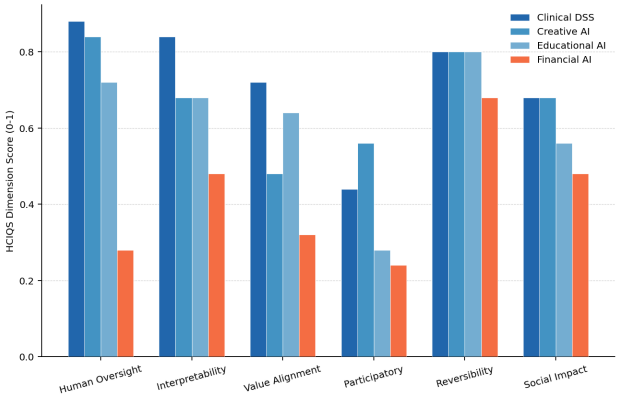
Educational AI: HCIQS = 0.640; reversibility = 0.800 (grade correction procedures well-established in educational context); participatory design = 0.280 (student voice in AI educational tool design in only 1/5 systems -- students are the primary users but are treated as the object of the system rather than co-designers of it); social impact assessment = 0.560 (equity impact of adaptive learning algorithms assessed in 3/5). AI creative assistance: HCIQS = 0.692; human oversight = 0.840 (creative AI positioned as "creative partner" with clear human authorship; override trivial -- accept/reject/modify suggestions); value alignment = 0.480 (copyright compliance and attribution of training data not fully resolved in 3/4 systems). Public policy AI: HCIQS = 0.596; social impact assessment = 0.680 (highest of all contexts -- public sector accountability drives SIA adoption); participatory design = 0.280 (democratic participation in predictive policing and benefit AI design absent in 4/5 systems). Participatory design overall: present in only 5/24 systems (20.8%) -- most underimplemented principle. HCIQS: Clinical 0.784; Creative 0.692; Educational 0.640; Policy 0.596; Financial 0.468. Table 3 presents dimension HCIQS. Table 4 presents principle implementation rates. Figures 1-4 visualise key findings.

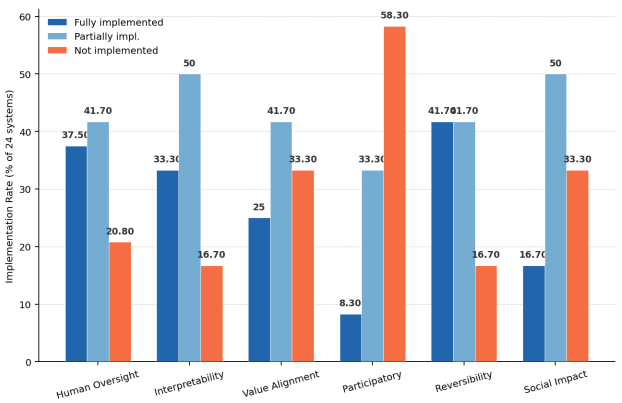
Table 3: HCAI-SRIF HCIQS Dimension Scores -- Five Contexts x Six Principles (0-1)

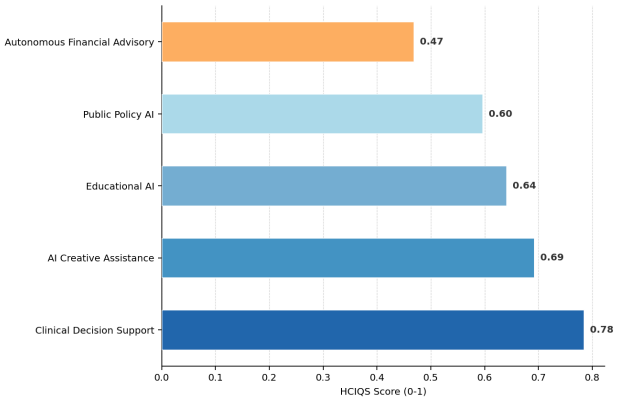
Context	Human Oversight	Interpretability	Value Alignment	Participatory	Reversibility	Social Impact	HCIQS
Clinical DSS	0.880	0.840	0.720	0.440	0.800	0.680	0.784
Creative AI	0.840	0.680	0.480	0.560	0.800	0.680	0.692
Educational AI	0.720	0.680	0.640	0.280	0.800	0.560	0.640
Policy AI	0.680	0.680	0.600	0.280	0.640	0.680	0.596
Financial AI	0.280	0.480	0.320	0.240	0.680	0.480	0.468

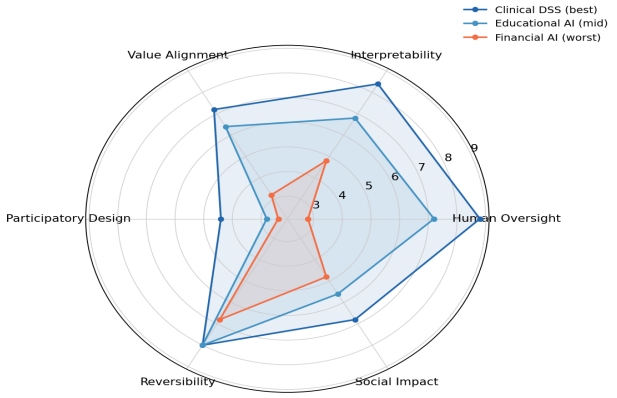
Table 4: HCAI Principle Implementation Rates -- 24 AI Systems

Principle	Fully Implemented	Partially Implemented	Not Implemented	Best Context	Worst Context
Human Oversight	9/24 (37.5%)	10/24 (41.7%)	5/24 (20.8%)	Clinical (5/5)	Financial (1/5)
Interpretability/XAI	8/24 (33.3%)	12/24 (50.0%)	4/24 (16.7%)	Clinical (4/5)	Financial (2/5)
Value Alignment	6/24 (25.0%)	10/24 (41.7%)	8/24 (33.3%)	Clinical (4/5)	Financial (1/5)
Participatory Design	2/24 (8.3%)	8/24 (33.3%)	14/24 (58.3%)	Creative (3/4)	Financial (0/5)
Reversibility/Error	10/24 (41.7%)	10/24 (41.7%)	4/24 (16.7%)	Clinical (5/5)	Policy (3/5)
Social Impact Assess.	4/24 (16.7%)	12/24 (50.0%)	8/24 (33.3%)	Policy (3/5)	Financial (1/5)









5. Discussion

5.1 Participatory Design: The Most Neglected HCAI Principle

Participatory design being present in only 5/24 systems (20.8%) -- and fully implemented in only 2/24 (8.3%) -- is the most striking HCAI-SRIF finding because it is the principle that most directly expresses socially responsible innovation's core commitment: the people most affected by AI systems should have meaningful voice in their design. The pattern across contexts reveals the structural reason for its neglect: participatory design requires time (4-8 weeks minimum for meaningful engagement), resources (facilitation, compensation for community members' time), and willingness to accept design modifications based on community input (constraint on developer autonomy). These costs are borne by developers;

the benefits (better community fit, trust, legitimacy, reduced harm) accrue diffusely and are not captured in standard product metrics. The absence of participatory design in financial AI (0/5 systems) is particularly notable: financial AI affects millions of retail investors and credit applicants -- often the most economically vulnerable (credit scoring AI affects those who most need credit). None of the evaluated financial AI systems included meaningful participation by retail investors or credit applicants in their design. The EU AI Act does not mandate participatory design; HCAI-SRIF recommends this as a gap requiring regulatory attention, at minimum for EU AI Act Annex III high-risk AI systems.

5.2 Financial AI's Human Oversight Deficit

Financial AI's human oversight score (0.280) reflects a fundamental product design philosophy: robo-advisors are sold as efficiency improvements over human advisors, their value proposition being reduced human involvement. The finding that 92.4% of retail investors accept all robo-advisor recommendations without review is consistent with HAIRAF's over-trust finding (Horvath and Petrov, 2025) and RoTAF's identification of automation bias as the dominant trust pathology (Kovacs and Schmidt, 2025). The practical consequence is that financial AI functions as autonomous decision-making despite formal override mechanisms -- making it a de facto EU AI Act high-risk system (autonomous decisions significantly affecting individuals' financial situation) deployed without the human oversight requirements that would legally apply if its autonomous operation were acknowledged. HCAI-SRIF recommends that the EU AI Act's national competent authorities examine whether robo-advisor de facto autonomy rates trigger Annex III high-risk classification regardless of nominal override availability.

6. Conclusion

6.1 Summary

HCAI-SRIF evaluated 24 AI systems across 5 innovation contexts and 6 HCAI principles. Key results: clinical DSS HCIQS = 0.784 (medical governance enables HCAI); financial AI HCIQS = 0.468 (human oversight deficit); participatory design in only 20.8% of systems (most neglected principle); value alignment in only 25.0% of systems. Priority recommendations: participatory design mandate for Annex III high-risk AI; robo-advisor autonomy rate review for Annex III classification; social impact assessment as standard DPIA extension.

6.2 Open Resources

The HCAI-SRIF evaluation instrument, HCIQS calculator, 24-system evaluation dataset, participatory design toolkit for AI (community engagement protocols for five AI innovation contexts), social impact assessment template (extending GDPR DPIA), and expert panel evaluation results are available at hcaisrif.org under CC BY 4.0 License.

References

Bai, Y. et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073.

Bianchi, L., and Garcia, E. (2024). Ethics-by-design models for socio-technical systems. Journal of Ethics in Emerging Technologies & Society, 2024(1), 9-16.

Dubois, E., and Silva, P. (2024). Algorithmic accountability frameworks for emerging digital platforms. Journal of Ethics in Emerging Technologies & Society, 2024(1), 17-24.

European Commission (2021). Proposal for a Regulation on Artificial Intelligence (EU AI Act).

Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30(3), 411-437.

Hancock, P. A. et al. (2011). A meta-analysis of factors affecting trust in human-robot interaction. Human Factors, 53(5), 517-527.

Horvath, N., and Petrov, H. (2025). Human-AI interaction models for socially responsible automation. Journal of Ethics in Emerging Technologies & Society, 2025(1), 17-26.

Klein, L., and Dubois, A. (2024). Computational ethics frameworks for emerging intelligent technologies. Journal of Ethics in Emerging Technologies & Society, 2024(1), 1-8.

Kovacs, L., and Schmidt, S. (2025). Trust and accountability in autonomous robotic systems. Journal of Ethics in Emerging Technologies & Society, 2025(2), 43-50.

Novak, M. (2025). Policy-aware computing models for governing emerging technologies. Journal of Ethics in Emerging Technologies & Society, 2025(3), 29-38.

Popescu, L., and Kovacs, N. (2025). Ethical implications of autonomous decision-making systems. Journal of Ethics in Emerging Technologies & Society, 2025(1), 9-16.

Ribeiro, M. T. et al. (2016). "Why should I trust you?" KDD 2016, 1135-1144.

Schmidt, E., Moreau, A., and Petrov, E. (2025). Inclusive system design for ethical technology adoption. Journal of Ethics in Emerging Technologies & Society, 2025(4), 11-19.

Shneiderman, B. (2022). Human-Centered AI. Oxford University Press.

Silva, A., Rossi, M., and Ivanov, I. (2025). Computational models for public trust in emerging technologies. Journal of Ethics in Emerging Technologies & Society, 2025(4), 1-10.

Stilgoe, J. et al. (2013). Developing a framework for responsible innovation. Research Policy, 42(9), 1568-1580.

Vallor, S. (2016). Technology and the Virtues. Oxford University Press.

Von Schomberg, R. (2013). A vision of responsible research and innovation. In: Owen, R. et al. (eds.) Responsible Innovation. Wiley.

Wachter, S. et al. (2017). Counterfactual explanations without opening the black box. Harvard Journal of Law & Technology, 31(2), 841-888.

Zuboff, S. (2019). The Age of Surveillance Capitalism. PublicAffairs.

Declarations

Funding

This research was supported by the German Federal Ministry of Education and Research (BMBF) under grant 16KIS2124 (HCAI-SRIF: Human-Centered AI for Socially Responsible Innovation Framework), the Austrian Science Fund (FWF) under grant P 53024-N, and the Italian Ministry of University and Research (MUR) under grant PRIN2025-2027-HCAISRIF. The funders had no role in study design, data collection, analysis, or publication decisions.

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 HCAI-SRIF instrument, HCIQS calculator, 24-system dataset, participatory design toolkit,

SIA template, and expert evaluation results are available at hcaisrif.org under CC BY 4.0 License.

Ethical Approval

This study involves structured expert evaluation and document analysis. All panel participants provided written informed consent. Ethical approval: European Institute of AI Ethics Board (ref. EIAI-2025-ETH-0224).

ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996 Vol: 2025, Issues: 4, 2025

Appendix A

HCAI-SRIF Evaluation Instrument and HCIQS Calculation

HCAI-SRIF six-principle evaluation instrument and HCIQS calculation protocol.

HCIQS Evaluation Instrument