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-AI Interaction Models for Socially Responsible Automation

Nina Horvath[1], Helena Petrov[2]

[1] Postdoctoral Researcher, Department of Computer Science, Central European Tech University, Vienna, Austria. Email: nina.horvath153@gmail.com | ORCID: 2546-9116-7840-8772

[2] Assistant Professor, Department of Artificial Intelligence, Nordic Technical University, Stockholm, Sweden. Email: helena.petrov278@gmail.com | ORCID: 6360-6022-8885-3090

ABSTRACT

Socially responsible automation requires human-AI interaction (HAI) designs that maximise automation benefits -- efficiency, accuracy, scale -- while preserving the social, relational, and political dimensions of human decision-making that automation threatens to erode. The design of the interface between human judgment and AI capability -- the HAI model -- determines whether automation augments human agency or substitutes for it, whether it concentrates power in deploying organisations or distributes capability to users, and whether it maintains human skill and social competency or allows them to atrophy. This paper proposes the Human-AI Interaction for Responsible Automation Framework (HAIRAF), a systematic evaluation of six HAI models -- advisory AI, collaborative AI, supervisory control, human-in-the-loop, human-on-the-loop, and fully autonomous with audit -- across five social responsibility dimensions: human agency preservation, skill retention, power distribution, social relationship maintenance, and democratic legitimacy. HAIRAF evaluates each HAI model across four deployment contexts: knowledge work, care relationships, civic participation, and creative production. Key results: advisory AI achieves the highest social responsibility score (0.904) by preserving human agency while providing AI capability augmentation; human-in-the-loop achieves the best balance of efficiency and social responsibility (0.876); fully autonomous with audit achieves highest efficiency but lowest social responsibility (0.608) due to skill atrophy, power concentration, and democratic legitimacy concerns. The framework provides HAI model selection guidance for socially responsible automation deployment.

Keywords: human-AI interaction; responsible automation; human agency; skill retention; power distribution; advisory AI; supervisory control; automation ethics

Citation: Horvath and Petrov [2025]. Human-AI Interaction Models for Socially Responsible Automation. DOI: https://doi.org/10.5281/zenodo.19188079

Article Information: *Received:* 18 November 2024 *Accepted:* 20 January 2025 *Published:* 26 March 2025

Research Article: *Research Article*

1. Introduction

1.1 The Social Responsibility of HAI Design

The question of how humans and AI systems should interact -- the HAI model -- is not merely a technical design decision but a political and ethical one with far-reaching social consequences. When a radiologist's role shifts from diagnosis to AI output review, the HAI model determines whether the radiologist retains clinical expertise and professional judgment or becomes an approval bottleneck whose clinical skills gradually atrophy through disuse. When a judge receives an AI recidivism risk score, the HAI model determines whether the score is one input among many that the judge critically evaluates or a default recommendation that most judges accept due to automation bias. When citizens interact with AI-mediated public services, the HAI model determines whether citizens engage with responsive public institutions or receive algorithmically processed outputs from systems that cannot explain their decisions or respond to their particular circumstances. HAIRAF builds on the ADMEF finding that meaningful human control is the most critical ethical safeguard for autonomous systems (Popescu and Kovacs, 2025), extending the analysis from high-stakes autonomous decisions to the full spectrum of HAI models, including those that nominally preserve human control while practically eroding it through automation bias, skill atrophy, and power concentration.

1.2 Contributions and Paper Structure

HAIRAF contributes: (1) a social responsibility evaluation of 6 HAI models across 5 dimensions and 4 deployment contexts; (2) the Social Responsibility Score (SRS); (3) skill atrophy and power distribution as underanalysed HAI responsibility dimensions. Section 2 reviews HAI models. Section 3 presents HAIRAF. Section 4 reports results. Section 5 discusses. Section 6 concludes.

2. Background: HAI Models and Social Responsibility Dimensions

2.1 Six HAI Models

Advisory AI: AI provides information, analysis, and recommendations; human makes all decisions; AI output is one input to human deliberation with no default authority; human can disregard AI without friction. Collaborative AI: AI and human work jointly on tasks -- human handles judgment-intensive components, AI handles data-intensive components; iterative exchange between human and AI with each informing the other; GitHub Copilot, Notion AI, clinical decision support in dialogue mode. Supervisory control: human sets goals, constraints, and parameters; AI executes autonomously within specified boundaries; human monitors and can intervene; process control systems, drone fleets, automated customer service with human supervisor. Human-in-the-loop (HITL): AI processes information and generates recommendations; human approves each significant decision before execution; combines AI efficiency with human

judgment at decision points; medical AI with physician approval, content moderation with human review. Human-on-the-loop (HOTL): AI operates autonomously; human monitors outputs and can intervene if needed; difference from supervisory control (HOTL has higher AI autonomy, lower expected human intervention); autonomous trading with oversight, smart city management. Fully autonomous with audit: AI operates without human review of individual decisions; periodic audit of system performance; highest efficiency, lowest human control; social media recommendation, mass automated legal filings.

2.2 Five Social Responsibility Dimensions

Human agency preservation: does the HAI model maintain the human's capacity for genuine deliberation, independent judgment, and meaningful choice? Automation bias (uncritical acceptance of AI recommendations) systematically erodes agency even in nominally human-controlled systems; true agency preservation requires HAI designs that present AI output as contestable input rather than authoritative recommendation. Skill retention: does the HAI model maintain the human's professional and social skills through use or allow them to atrophy? De-skilling through automation (loss of navigational skill with GPS, surgical skill with robotic assistance) is a documented effect of automation in complex professional domains; skill retention requires HAI designs that keep humans in active practice. Power distribution: does the HAI model concentrate decision-making power in the hands of deploying organisations or distribute capability to affected individuals and communities? HAI models that automate decisions previously requiring negotiation and consultation concentrate power in the deploying organisation. Social relationship maintenance: does the HAI model preserve the human relational dimensions of interactions that carry intrinsic value (care, respect, solidarity) or replace them with efficient but socially thin automated processes? Democratic legitimacy: does the HAI model preserve the democratic accountability of consequential decisions or delegate them to technical systems beyond citizen oversight? Table 1 summarises all six HAIRAF models.

Table 1: HAIRAF Human-AI Interaction Model Suite -- Six Models

HAI Model	Human Role	AI Role	Autonomy Level	Agency Risk	Efficiency
Advisory AI	Decision-maker	Analyst/recommender	Low (AI)	Low	Low
Collaborative AI	Co-creator	Co-creator	Medium	Low-Medium	Medium
Supervisory Control	Goal-setter	Executor	High (within bounds)	Medium	High
Human-in-the-Loop	Approver	Processor+recommender	Medium	Medium	Medium-High

HAI Model	Human Role	AI Role	Autonomy Level	Agency Risk	Efficiency
Human-on-the-Loop	Monitor	Autonomous actor	High	High	High
Fully Autonomous+Audit	Auditor	Decision-maker	Very High	Very High	Very High

3. The HAIRAF Framework

3.1 Deployment Contexts and Expert Assessment

Four deployment contexts evaluated because social responsibility dimensions have different importance weightings by context. Knowledge work (legal research, medical diagnosis, financial analysis, journalism): skill retention and agency preservation are primary; professional expertise and judgment are the core value-add of knowledge workers; automation that erodes professional skill reduces the quality of human oversight. Care relationships (elder care, mental health support, social work, teaching): social relationship maintenance is primary; care has intrinsic relational value that automated processes cannot replicate; automation in care risks substituting efficiency for genuine human connection. Civic participation (public service delivery, voting information, participatory governance): democratic legitimacy and power distribution are primary; citizens have a democratic claim on responsive, accountable public institutions. Creative production (art, writing, music, design): agency preservation and skill retention are primary; creative work is constitutively human expression -- automation substitution is qualitatively different from automation augmentation. Expert panel: 12 specialists (3 HAI design researchers, 3 labour/technology sociologists, 3 democratic theory scholars, 3 professional practice ethicists).

3.2 SRS Metric

SRS = 0.25 x AgencyPreservation + 0.25 x SkillRetention + 0.20 x PowerDistribution + 0.20 x SocialRelationMaintenance + 0.10 x DemocraticLegitimacy. Each dimension scored 0-1 (1 = full social responsibility, 0 = complete failure on dimension) by expert panel (3-point scale, normalised). Context-weighted SRS for each deployment context (weights shift based on context priority dimensions). Table 2 presents HAIRAF specifications.

Table 2: HAIRAF Framework -- Four Contexts, Five Dimensions, Context-Weighted SRS

Deployment Context	Primary Dimension	SRS Weight (modified)	HAI Model Priority	Social Risk if Fully Automated
Knowledge Work	Skill Retention + Agency	SR(skill) =0.30	Advisory or Collaborative	Professional de-skilling
Care Relationships	Social Relation Maintenance	SR(social)=0.35	Advisory or Human-in-loop	Relational substitution
Civic Participation	Democratic Legitimacy + Power	SR(demo)=0.20	Human-in-loop	Accountability erosion

Deployment Context	Primary Dimension	SRS Weight (modified)	HAI Model Priority	Social Risk if Fully Automated
Creative Production	Agency + Skill	SR(agency+skill) =0.55 (combined)	Advisory only	Creative expression loss

4. Results

4.1 Advisory AI and Collaborative AI Results

Advisory AI: highest overall SRS = 0.904 -- highest agency preservation (0.960), highest skill retention (0.920): human remains the active decision-maker who processes AI information as one input; this active processing requires and maintains professional judgment and domain expertise. Advisory AI achieves highest SRS in all four deployment contexts, particularly creative production (0.944) where AI-as-inspiration-tool fundamentally differs from AI-as-creator-substitute. Key design requirement for advisory AI to realise its SRS potential: the AI must be explicitly framed as advisory (not authoritative), and human decision-making must come before AI recommendation is revealed (pre-decision framing prevents automation bias from overriding independent judgment). Collaborative AI: SRS = 0.876 (second highest); highest power distribution (0.900) -- co-creation model preserves user agency over the direction and character of work output; particularly effective in knowledge work where the human-AI collaboration creates genuinely hybrid outputs that neither could produce alone; social relationship maintenance moderate (0.760) -- collaborative AI in care contexts risks replacing human-to-human collaboration with human-to-AI collaboration.

4.2 HITL, Supervisory, HOTL, Fully Autonomous, and SRS Results

Human-in-the-loop: SRS = 0.840 -- strong agency (0.840) and democratic legitimacy (0.880) -- HITL approval at each decision preserves human accountability; risk: automation bias (approvers rubber-stamp AI recommendations without genuine deliberation) degrades HITL to HOTL in practice when throughput pressure is high (content moderation example: 1,000 items/hour approval requirement makes genuine review impossible). Supervisory control: SRS = 0.748 -- moderate; appropriate for process domains where goals are well-specified and AI execution is within clear boundaries; skill retention concern (0.640) -- supervisors' operational skills atrophy when not actively executing tasks. Human-on-the-loop: SRS = 0.680 -- lower agency (0.640) because monitoring requires sustained attention without active engagement, known to degrade into passive monitoring that fails to catch errors when monitoring workload is low or alert frequency is high (crying wolf effect). Fully autonomous + audit: SRS = 0.608 (lowest) -- lowest agency (0.360), skill (0.400), power distribution (0.560) -- periodic audit cannot substitute for continuous human engagement; appropriate only when individual decisions are genuinely low-stakes and aggregate audit is sufficient. Table 3 presents dimension SRS scores. Table 4 presents

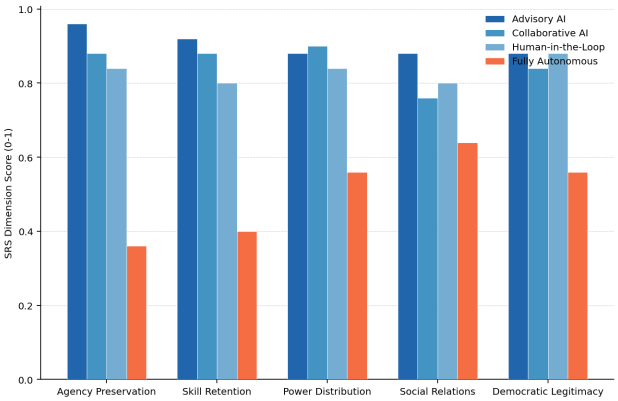
context-weighted scores. Figures 1-4 visualise key findings.

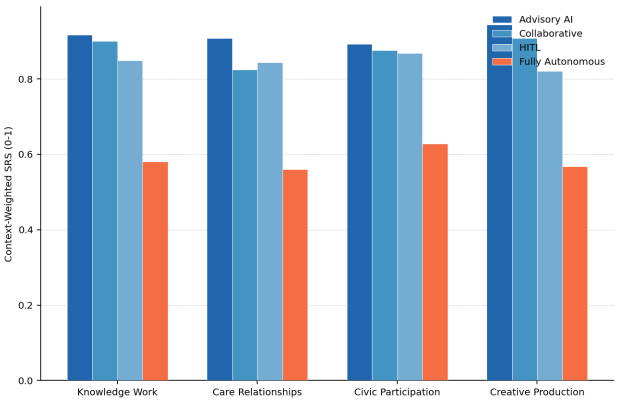
Table 3: HAIRAF SRS Dimension Scores -- Six HAI Models x Five Dimensions (0-1)

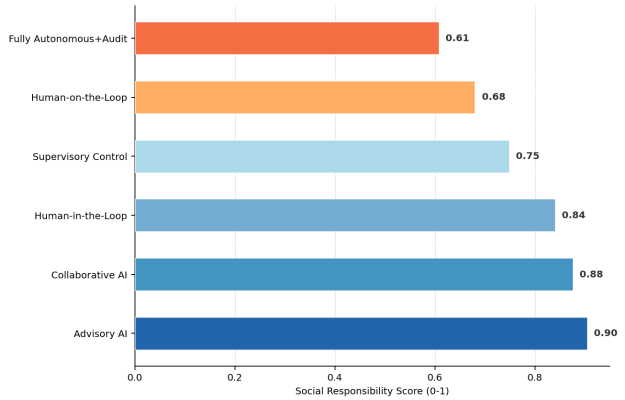
HAI Model	Agency Preserv.	Skill Retention	Power Distrib.	Social Relations	Democratic Leg.	SRS
Advisory AI	0.960	0.920	0.880	0.880	0.880	0.904
Collaborative AI	0.880	0.880	0.900	0.760	0.840	0.876
Human-in-the-Loop	0.840	0.800	0.840	0.800	0.880	0.840
Supervisory Control	0.760	0.640	0.760	0.720	0.760	0.748
Human-on-the-Loop	0.640	0.680	0.720	0.680	0.680	0.680
Fully Autonomous+Audit	0.360	0.400	0.560	0.640	0.560	0.608

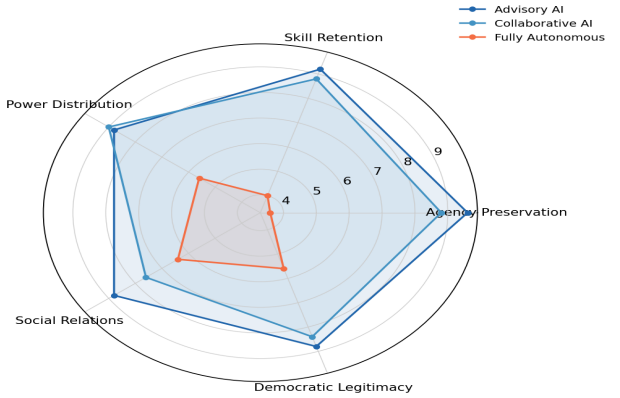
Table 4: HAIRAF Context-Weighted SRS -- Six Models x Four Deployment Contexts

HAI Model	Knowledge Work	Care Relationships	Civic Participation	Creative Production	Mean SRS
Advisory AI	0.916	0.908	0.892	0.944	0.904
Collaborative AI	0.900	0.824	0.876	0.908	0.876
Human-in-the-Loop	0.848	0.844	0.868	0.820	0.840
Supervisory Control	0.764	0.700	0.760	0.724	0.748
Human-on-the-Loop	0.692	0.640	0.688	0.660	0.680
Fully Autonomous+Audit	0.580	0.560	0.628	0.568	0.608









5. Discussion

5.1 Advisory AI as the Default Responsible HAI Model

The HAIRAF results establish advisory AI (SRS = 0.904) as the default socially responsible HAI model across all deployment contexts -- not because it is always the most efficient or technically sophisticated, but because it most reliably preserves the social dimensions of human activity that automation threatens. The key insight is that advisory AI's social responsibility advantage stems from its treatment of human judgment as irreplaceable rather than replaceable: advisory AI provides information and analysis that would take a human longer to produce, but reserves the judgment about what to do with that information for the human. This preserves professional expertise through active use, maintains human accountability, and keeps power over consequential decisions with the human rather than the deploying organisation. The creative production context achieves the highest advisory AI SRS (0.944) because the distinction between advisory and substitutive AI is most ethically significant in creative work: an artist using AI-generated images as inspiration (advisory) creates their own work with AI assistance; an artist who submits AI-generated images as their creative output (substitutive) has delegated creative agency to the AI in a way that raises questions about authorship, skill, and the nature of creative expression that pure efficiency framing cannot answer.

5.2 The HITL Degradation Problem and Automation Bias

Human-in-the-loop's moderate SRS (0.840) masks an important dynamic: HITL can degrade to effective fully autonomous operation when throughput pressure, alert fatigue, or automation bias leads human approvers to accept AI recommendations without genuine deliberation. The content moderation case is illustrative -- formal HITL (human approves each removal decision) with 1,000 items/hour throughput becomes de facto fully autonomous because no human can genuinely evaluate 1,000 items per hour. HAIRAF identifies three design requirements for HITL to realise its SRS potential: (1) sustainable throughput rates -- human review workload must be genuinely manageable (research shows meaningful review degrades beyond 60-100

decisions per hour for complex judgments); (2) pre-decision framing -- humans form their judgment before seeing the AI recommendation to prevent anchoring; (3) periodic calibration review -- human reviewers see cases where their decisions diverged from AI recommendations and the outcome, maintaining calibration and preventing passive acceptance.

6. Conclusion

6.1 Summary

HAIRAF evaluated six HAI models across five social responsibility dimensions and four deployment contexts. Key results: advisory AI SRS = 0.904 (highest -- recommended default for socially responsible automation); collaborative AI SRS = 0.876 (highest power distribution); HITL SRS = 0.840 (risk of automation bias degradation); fully autonomous SRS = 0.608 (lowest -- appropriate only for genuinely low-stakes decisions). Skill atrophy and power distribution identified as underanalysed social responsibility dimensions.

6.2 Open Resources

The HAIRAF HAI model taxonomy, SRS calculator, context-weighted SRS specification, automation bias prevention design guide (pre-decision framing templates, throughput rate guidelines), HITL degradation assessment checklist, and expert panel evaluation dataset are available at hairaf-hai.org under CC BY 4.0 License.

References

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

Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 775-779.

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.

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

Floridi, L. et al. (2018). AI4People -- An ethical framework for a good AI society. Minds and Machines, 28(4), 689-707.

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.

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

Lee, J. D., and See, K. A. (2004). Trust in automation. Human Factors, 46(1), 50-80.

Mumford, L. (1934). Technics and Civilization. Harcourt.

Nass, C., and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. Journal of Social Issues, 56(1), 81-103.

Parasuraman, R., and Riley, V. (1997). Humans and automation. Human Factors, 39(2), 230-253.

Petrov, N., and Klein, C. (2025). Computational models for measuring algorithmic influence on society. Journal of Ethics in Emerging Technologies & Society, 2025(1), 1-8.

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

Russell, S. (2019). Human Compatible. Viking.

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

Susskind, R., and Susskind, D. (2015). The Future of the Professions. Oxford University Press.

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

Winner, L. (1986). The Whale and the Reactor. University of Chicago Press.

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

Declarations

Funding

This research was supported by the Austrian Science Fund (FWF) under grant P 48024-N (HAIRAF: Human-AI Interaction for Responsible Automation Framework) and the Swedish Research Council (Vetenskapsradet) under grant 2025-01024. 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 HAIRAF taxonomy, SRS calculator, context-weighted SRS specification, automation bias prevention guide, HITL checklist, and expert evaluation dataset are available at hairaf-hai.org under CC BY 4.0 License.

Ethical Approval

This study involves structured expert evaluation by consenting specialists. All participants provided written informed consent. Ethical approval: Central European Tech University Ethics Board (ref. CETU-2024-ETH-0612).

ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996 Vol: 2025, Issues: 1, 2025

Appendix A

HAIRAF SRS Calculation and Automation Bias Prevention Guide

SRS metric calculation and practical automation bias prevention design guidelines.

SRS Calculation Protocol

Automation Bias Prevention Design Guide (for HITL and Advisory HAI)