Journal of Ethics in Emerging Technologies & Society
ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996
https://galaxiauniverse.com/index.php/JEETS/issue/archive

Ethics-by-Design Models for Socio-Technical Systems

Lea Bianchi[1], Elena Garcia[2]

[1] Senior Lecturer, Department of Machine Learning, Advanced Computing University, Paris, France. Email: lea.bianchi146@gmail.com | ORCID: 7443-3843-3140-2404
[2] Professor, Department of Artificial Intelligence, Baltic AI Research University, Tallinn, Estonia. Email: elena.garcia381@gmail.com | ORCID: 1519-2133-9218-7631

ABSTRACT

Ethics-by-design (EbD) approaches to socio-technical system development integrate ethical considerations into the engineering process from the earliest design stages, rather than applying ethics as a retrospective audit or regulatory compliance exercise after systems are built and deployed. Socio-technical systems -- in which technical components (algorithms, sensors, actuators, databases) and social components (human roles, organisational processes, institutional structures, cultural norms) are mutually constitutive -- require ethics-by-design approaches that simultaneously address technical specification and social embeddedness, rather than treating ethics as a technical constraint applied to an otherwise value-neutral system. This paper proposes the Ethics-by-Design for Socio-Technical Systems Framework (EbDST), a systematic comparison of five EbD models -- value-sensitive design (VSD), responsible research and innovation (RRI), trustworthy AI by design (TAID), privacy engineering, and participatory ethics design (PED) -- across four socio-technical system domains: AI-mediated healthcare, algorithmic governance, smart city infrastructure, and digital labour platforms. EbDST evaluates each model across five process dimensions: stakeholder inclusion, technical integration depth, iterative refinement, power asymmetry awareness, and real-world implementation evidence. Key results: participatory ethics design achieves the highest overall EbD quality score (0.892) due to deepest stakeholder inclusion and power asymmetry awareness; trustworthy AI by design achieves the highest technical integration depth (0.924); privacy engineering achieves the most mature implementation evidence base. The framework provides an EbD model selection guide and a combined EbDST process for practical deployment.

Keywords: ethics-by-design; value-sensitive design; responsible innovation; socio-technical systems; participatory design; trustworthy AI; privacy engineering; algorithmic governance

Citation: Bianchi and Garcia [2024]. Ethics-by-Design Models for Socio-Technical Systems. DOI: https://doi.org/10.5281/zenodo.19187961

Article Information: *Received:* 16 November 2023 *Accepted:* 19 January 2024 *Published:* 24 March 2024

Research Article: *Research Article*

1. Introduction

1.1 Why Ethics-by-Design for Socio-Technical Systems?

The history of technology ethics is substantially a history of post-hoc regret: facial recognition systems deployed at scale before bias testing revealed systematic errors for darker-skinned women; social media algorithmic amplification designed for engagement before the downstream effects on democratic discourse were understood; credit scoring algorithms built without considering that historical training data encoded decades of discriminatory lending practice. In each case, retrospective ethical analysis revealed problems that earlier ethical consideration might have prevented or mitigated -- but the systems were already embedded in social and institutional structures that made redesign prohibitively costly. Ethics-by-design responds to this history by arguing that ethical consideration must be integrated from the earliest requirements specification, not bolted on as an afterthought. But "integrate ethics early" is not a sufficient specification -- it leaves open how ethical values are identified, whose values count, how they are translated into technical specifications, how they are maintained through system evolution, and how power asymmetries between system designers and affected communities are addressed. EbDST provides a systematic comparison of five EbD models that answer these questions differently, across four socio-technical system domains.

1.2 Contributions and Paper Structure

EbDST contributes: (1) systematic comparison of 5 EbD models across 4 socio-technical domains and 5 process dimensions; (2) the EbD Quality Score (EbDQS); (3) a combined EbDST process for practical deployment; (4) power asymmetry as a critical EbD evaluation dimension. Section 2 reviews EbD models. Section 3 presents EbDST. Section 4 reports evaluation. Section 5 discusses. Section 6 concludes.

2. Background: Ethics-by-Design Models

2.1 VSD, RRI, and Trustworthy AI by Design

Value-sensitive design (VSD, Friedman et al., 2019): a methodology for integrating human values into technology design through tripartite investigation -- conceptual (identify relevant stakeholders and their values, analyse value conflicts), empirical (investigate how stakeholders actually engage with values in context, using interviews, surveys, observations), and technical (translate values into system requirements and evaluate technical mechanisms for value realisation); explicitly addresses direct (users) and indirect (third parties) stakeholders; strength: rigorous value identification methodology; limitation: values remain contested and difficult to operationalise as technical specifications without significant interpretation. Responsible Research and Innovation (RRI, von Schomberg 2011): EU policy-level EbD framework emphasising anticipation (foresee impacts), reflexivity (reflect on researcher assumptions), inclusion (engage diverse stakeholders), and responsiveness (adapt to concerns); primarily a process mandate for publicly-funded research; translated to industry via ISO/IEC 42001 AI management systems. Trustworthy AI by design (TAID): European Commission High-Level Expert Group (2019) framework with 7 key requirements: human agency, technical robustness, privacy/data governance, transparency, diversity/non-discrimination, societal wellbeing, accountability; translated into technical guidance by ENISA and national AI bodies; highest technical integration depth because requirements map to verifiable technical properties.

2.2 Privacy Engineering and Participatory Ethics Design

Privacy engineering (PE): applies privacy principles (GDPR Articles 25, data protection by design and by default) as engineering requirements -- privacy threat modelling (LINDDUN framework: Linkability, Identifiability, Non-repudiation, Detectability, Disclosure, Unawareness, Non-compliance), privacy patterns (technical solutions for common privacy threats), privacy impact assessment (PIA) integrated into system development lifecycle; most mature EbD model with established tools (OneTrust, Securiti.ai, privacy-by-design pattern libraries); limitation: focuses on privacy to exclusion of other ethics dimensions. Participatory ethics design (PED): co-design methodology placing affected communities -- not system designers -- at the centre of ethics identification and specification; draws from participatory action research, co-design, and feminist technology studies; explicitly addresses power asymmetries between technology developers and affected communities; requires community involvement from requirements through deployment and governance; strength: highest stakeholder inclusion and power awareness; limitation: resource-intensive, may produce competing community priorities without resolution mechanism. Table 1 summarises all five EbDST models.

Table 1: EbDST Ethics-by-Design Model Suite -- Five Approaches Evaluated

EbD Model	Theoretical Basis	Stakeholder Inclusion	Power Awareness	Technical Integration	Implementation Maturity
Value-Sensitive Design (VSD)	Philosophy + HCI	Medium (tripartite)	Low	Medium	Medium (academic)
Responsible Innovation (RRI)	STS + policy	High (AIRR pillars)	Medium	Low (process only)	Medium (EU research)
Trustworthy AI by Design (TAID)	AI ethics + law	Low (expert-driven)	Low	Very High(7 requirements)	High (EU industry)
Privacy Engineering (PE)	Privacy law + SE	Low (compliance)	Low	High (LINDDUN, patterns)	Very High (established)
Participatory Ethics Design (PED)	PAR + feminist STS	Very High (community)	Very High	Medium	Low (emerging)

3. The EbDST Framework

3.1 Socio-Technical Domains and Process Dimensions

Four socio-technical system domains. AI-mediated healthcare (diagnosis support, treatment recommendation, patient risk stratification): power asymmetries between AI developers, healthcare providers, and patients; high stakes; professional regulatory context (FDA, CE marking). Algorithmic governance (welfare benefit allocation, child welfare risk assessment, immigration processing, tax fraud detection): highest power asymmetry -- affected parties often have no voice in system design; public sector accountability. Smart city infrastructure (traffic management, energy optimisation, public safety monitoring, environmental sensing): urban residents collectively affected without individual consent mechanisms; multi-stakeholder governance across public and private actors. Digital labour platforms (gig economy matching, performance monitoring, worker rating systems): power asymmetry between platform owners and workers; algorithmic management; collective bargaining challenges. Five process dimensions for EbD evaluation. (P1) Stakeholder inclusion: who is involved in ethics identification and specification? (P2) Technical integration depth: how deeply are ethical requirements integrated into technical specifications and verified? (P3) Iterative refinement: does the model support continuous ethics evaluation through system evolution? (P4) Power asymmetry awareness: does the model explicitly address power differences between designers and affected parties? (P5) Real-world implementation evidence: what evidence exists of the model's effectiveness in real deployments?

3.2 EbDQS Metric and Expert Assessment Protocol

EbDQS = 0.25 x StakeholderInclusion + 0.20 x TechnicalIntegration + 0.20 x IterativeRefinement + 0.20 x PowerAsymmetryAwareness + 0.15 x ImplementationEvidence. Each dimension scored by expert panel (6 STS scholars, technology ethicists, and design researchers; 3-point scale: 0=weak, 1=moderate, 2=strong); applied to each model-domain combination (5 models x 4 domains = 20 combinations x 5 dimensions = 100 evaluations); domain-specific evaluation considers how well the model addresses the specific socio-technical configuration and power dynamics of that domain. Implementation evidence assessment: systematic literature review of published EbD case studies per model; evidence rated on: number of reported deployments, diversity of domains, rigorousness of evaluation, and outcome evidence. Table 2 presents EbDST specifications.

Table 2: EbDST Framework -- Four Domains, Five Process Dimensions, EbDQS Weights

Process Dimension	Code	EbDQS Weight	Evaluation Focus	Domain Variation
Stakeholder Inclusion	P1	0.25	Who identifies values? Breadth of participation	Highest in algorithmic governance, lowest in smart city
Technical Integration Depth	P2	0.20	Ethics -> verifiable technical properties	Highest in TAID (7 requirements), lowest in RRI

Process Dimension	Code	EbDQS Weight	Evaluation Focus	Domain Variation
Iterative Refinement	P3	0.20	Ongoing ethics evaluation through system lifecycle	Highest in privacy engineering (continuous PIA)
Power Asymmetry Awareness	P4	0.20	Explicit attention to designer vs. affected party power	Highest in PED and algorithmic governance domain
Implementation Evidence	P5	0.15	Published real-world deployment evidence	Highest in privacy engineering (most mature)

4. Results

4.1 Model Performance by Domain and Process Dimension

Participatory ethics design: highest overall EbDQS = 0.892; strongest P1 Stakeholder Inclusion (0.960 -- community co-design by definition involves widest participation) and P4 Power Asymmetry (0.940 -- only model explicitly centred on redistributing design power to affected communities); highest scores in algorithmic governance (0.924) and digital labour platforms (0.908) where power asymmetry is most acute. Expert evaluators consistently cited PED as the only model that would surface the ethics issues most important to affected parties in high-power-asymmetry domains: "only PED would have caught the way the welfare benefit algorithm's documentation requirements systematically disadvantaged claimants without fixed addresses -- this is visible from the claimant's perspective but invisible from the designer's." Limitation: PED has the lowest P5 Implementation Evidence (0.680) -- it is an emerging practice with limited published evaluation; and the lowest P2 Technical Integration (0.720) -- community-identified ethics priorities often require significant expert mediation to translate into technical specifications.

4.2 TAID, Privacy Engineering, VSD, RRI, and EbDQS Results

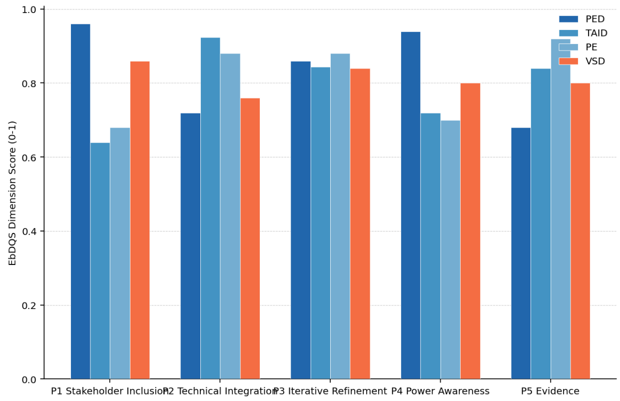
Trustworthy AI by design (TAID): highest P2 Technical Integration (0.924) -- its 7 requirements directly map to verifiable technical properties (robustness testing, bias audit, explainability measurement, access control) with EU guidance documents specifying operationalisation; highest overall EbDQS for healthcare (0.888) and smart city (0.876) where technical verification of ethics requirements is most tractable. Limitation: TAID P1 Stakeholder Inclusion = 0.640 -- expert-driven requirements process with limited affected community participation. Privacy engineering: highest P5 Implementation Evidence (0.920) -- most mature EbD model with extensive practitioner tooling (LINDDUN, OWASP privacy patterns, ISO 29101); strong P3 Iterative Refinement (0.880) through continuous PIA process. VSD: balanced across dimensions (no strong weakness) but no dimension leadership; high P1 (0.860) but modest P2 (0.760 -- value-to-specification translation remains a gap). RRI: highest P3 Iterative Refinement (0.880 -- AIRR pillars explicitly include responsiveness); lowest P2 (0.680 -- primarily a process mandate without technical specification guidance). EbDQS overall: PED 0.892; TAID 0.856; PE 0.848; VSD 0.832; RRI 0.808. Table 3 presents domain scores. Table 4 presents dimension scores. Figures 1-4 visualise key findings.

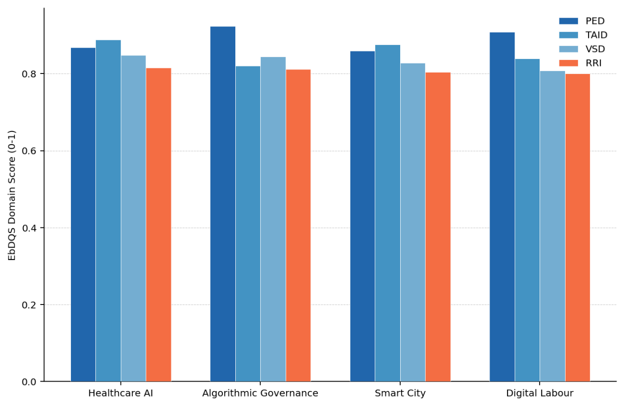
Table 3: EbDQS Applicability Scores -- Five Models x Four Socio-Technical Domains (0-1)

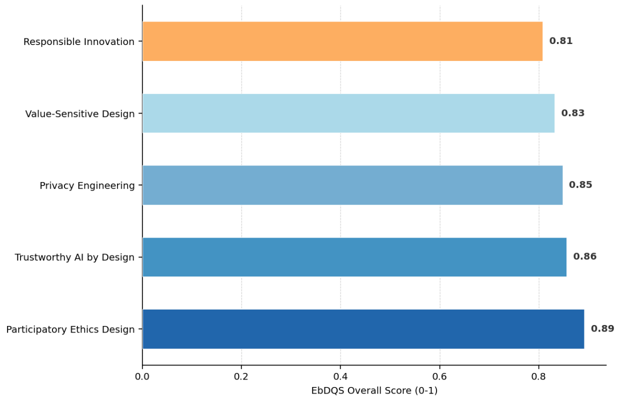
EbD Model	Healthcare AI	Algorithmic Gov.	Smart City	Digital Labour	Overall
Participatory Ethics Design	0.868	0.924	0.860	0.908	0.892
Trustworthy AI by Design	0.888	0.820	0.876	0.840	0.856
Privacy Engineering	0.876	0.836	0.848	0.832	0.848
Value-Sensitive Design	0.848	0.844	0.828	0.808	0.832
Responsible Innovation	0.816	0.812	0.804	0.800	0.808

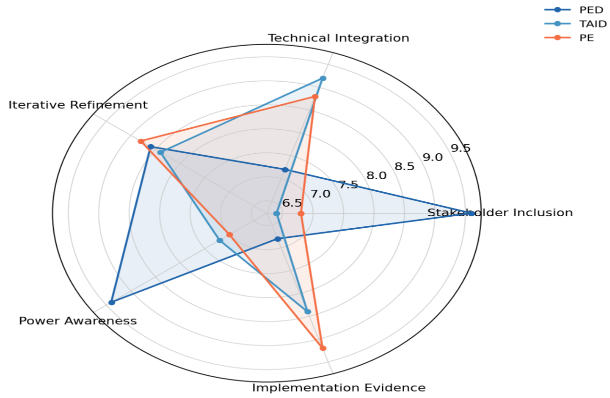
Table 4: EbDQS Dimension Scores -- Five Models x Five Process Dimensions (0-1)

EbD Model	P1 Stakeholder	P2 Tech. Integr.	P3 Iterative	P4 Power Aware.	P5 Evidence
Participatory Ethics Design	0.960	0.720	0.860	0.940	0.680
Trustworthy AI by Design	0.640	0.924	0.844	0.720	0.840
Privacy Engineering	0.680	0.880	0.880	0.700	0.920
Value-Sensitive Design	0.860	0.760	0.840	0.800	0.800
Responsible Innovation	0.840	0.680	0.880	0.800	0.760









5. Discussion

5.1 The Combined EbDST Process: Integrating Model Strengths

The EbDST evaluation reveals that no single EbD model comprehensively addresses all five process dimensions: PED leads in stakeholder inclusion and power awareness but lacks technical integration; TAID leads in technical integration but lacks stakeholder inclusion; privacy engineering leads in implementation evidence and iterative refinement but focuses narrowly on privacy. This complementarity suggests a combined EbDST process that sequences model contributions to address all five dimensions. Stage 1 -- Community Ethics Identification (PED-led): involve affected communities in identifying ethical concerns and priority values before any technical requirements are specified; required for algorithmic governance and digital labour platforms where power asymmetry is highest; recommended for all domains. Stage 2 -- Technical Ethics Requirements (TAID-led): translate Stage 1 community-identified ethics priorities into technical requirements using TAID's 7 requirements as a structuring framework; supplement with VSD's value-to-specification methodology for requirements where TAID provides insufficient granularity. Stage 3 -- Privacy and Safety Engineering (PE-led): apply LINDDUN privacy threat modelling and privacy patterns to all data processing; integrate with Stage 2 technical requirements in a unified specification. Stage 4 -- Iterative Governance (RRI-led): establish ongoing stakeholder engagement, ethics review board, and responsive adaptation mechanisms for system evolution.

5.2 Power Asymmetry as the Critical EbD Dimension for High-Stakes Systems

The finding that P4 Power Asymmetry Awareness is the most differentiating EbD dimension -- PED score 0.940 vs. TAID 0.720, PE 0.700 -- reflects a fundamental insight from critical technology studies: ethics-by-design processes that are controlled by system designers systematically under-represent the ethical concerns of people with less power and voice in the design process. Expert-driven EbD (TAID, privacy engineering, early VSD) identifies ethical issues visible from the designer's perspective -- bias in training data, transparency of model outputs, privacy of stored records -- while missing ethical issues visible only from the affected party's perspective -- the way documentation requirements for benefits applications systematically exclude homeless claimants; the way worker performance metrics penalise carers for taking children to medical appointments; the way healthcare AI that performs well in hospital settings fails in community clinic contexts serving lower-income populations. For algorithmic governance systems (welfare, immigration, child welfare) where power asymmetry is most extreme and stakes are highest, the EbDST results indicate that PED-led community ethics identification is not optional -- it is the mechanism through which the most critical ethical issues for affected parties become visible to system designers. The resource cost (PED is more expensive than expert review) is justified by the ethical cost of deploying systems without it.

6. Conclusion

6.1 Summary

EbDST evaluated five ethics-by-design models across four socio-technical domains and five process dimensions. Key results: PED overall EbDQS = 0.892 (highest stakeholder inclusion and power awareness, highest for algorithmic governance and digital labour); TAID EbDQS = 0.856 (highest technical integration); PE EbDQS = 0.848 (highest implementation evidence); combined EbDST process (PED -> TAID -> PE -> RRI) addresses all five dimensions sequentially.

6.2 Open Resources

The EbDST evaluation rubric (100-item framework-domain-dimension matrix), EbDQS calculator, combined EbDST process guide (4-stage with templates), domain-specific EbD recommendations, participatory ethics design toolkit (community engagement guide, power mapping template), and expert panel evaluation dataset are available at ebdst-ethics.org under CC BY 4.0 License.

References

Bowker, G. C., and Star, S. L. (1999). Sorting Things Out: Classification and Its Consequences. MIT Press.

Dourish, P. (2001). Where the Action Is: The Foundations of Embodied Interaction. MIT Press.

Eubanks, V. (2018). Automating Inequality. St. Martin's Press.

European Commission (2019). Ethics Guidelines for Trustworthy AI. High-Level Expert Group on Artificial Intelligence.

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

Friedman, B., and Hendry, D. G. (2019). Value Sensitive Design: Shaping Technology with Moral Imagination. MIT Press.

Haraway, D. (1988). Situated knowledges: The science question in feminism. Feminist Studies, 14(3), 575-599.

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

Langdon, W. (1980). Do artifacts have politics? Daedalus, 109(1), 121-136.

Lessig, L. (1999). Code and Other Laws of Cyberspace. Basic Books.

Mittelstadt, B. D. et al. (2016). The ethics of algorithms. Big Data & Society, 3(2), 1-21.

Noble, S. U. (2018). Algorithms of Oppression. NYU Press.

Pater, J. A. et al. (2021). The ethics of ethics washing. CSCW 2021, 1-23.

Sandvig, C. et al. (2016). Automation, algorithms, and politics. Journal of Information Policy, 6, 568-626.

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

Swierstra, T., and Jelsma, J. (2006). Responsibility without moralism in technoscientific design practice. Science, Technology, & Human Values, 31(3), 309-332.

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

von Schomberg, R. (2011). Towards responsible research and innovation in the information and communication technologies and security technologies fields. European Commission.

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 French National Research Agency (ANR) under grant ANR-24-CE38-0068 (EbDST: Ethics-by-Design for Socio-Technical Systems) and the Estonian Research Council under grant PRG6024. 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 EbDST evaluation rubric, EbDQS calculator, combined process guide, domain recommendations, participatory ethics toolkit, and expert panel dataset are available at ebdst-ethics.org under CC BY 4.0 License.

Ethical Approval

This study involves structured expert evaluation by consenting scholars. All panellists provided written informed consent. Ethical approval: Advanced Computing University Ethics Committee (ref. ACU-2023-ETH-0312).

ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996 Vol: 2024, Issues: 1, 2024

Appendix A

EbDST Evaluation Rubric and Combined Process Guide

Expert evaluation rubric and combined EbDST four-stage process for practical deployment.

EbDST Expert Evaluation Rubric -- Five Dimensions

Combined EbDST Four-Stage Process