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

Ethical Risk Modeling for Large-Scale Technology Deployment

Clara Novak[1], Sofia Muller[2]

[1] Associate Professor, Department of Artificial Intelligence, Central European Tech University, Vienna, Austria. Email: clara.novak148@gmail.com | ORCID: 0990-1381-1562-9785
[2] Assistant Professor, School of Data Science, Western Europe Data Science University, Madrid, Spain. Email: sofia.muller383@gmail.com | ORCID: 6446-1230-4064-3825

ABSTRACT

Large-scale technology deployments -- nationwide digital identity systems, AI-powered public health surveillance platforms, autonomous vehicle fleets, and ubiquitous computing infrastructure -- introduce ethical risks at societal scale that individual-level harm assessment cannot adequately capture. Ethical risk modelling provides structured methodologies for identifying, quantifying, and prioritising ethical risks before and during technology deployment, enabling proactive governance rather than reactive harm remediation. This paper proposes the Ethical Risk Modelling Framework for Large-Scale Technologies (ERMFLST), a systematic methodology integrating five ethical risk assessment approaches -- consequentialist harm mapping, rights-violation risk analysis, systemic risk assessment, distributional impact modelling, and emergent risk anticipation -- applied to four large-scale technology deployment scenarios: national AI surveillance systems, autonomous vehicle fleet deployment, nationwide digital identity infrastructure, and AI-integrated healthcare systems. ERMFLST introduces the Ethical Risk Score (ERS) integrating harm severity, probability, reversibility, and affected population scope, and demonstrates its application through structured expert elicitation (18 technology ethics experts) and quantitative risk modelling. Key results: AI surveillance systems achieve the highest ERS (0.924 on a 0-1 harm scale indicating high risk) -- driven by irreversibility and population scope; rights-violation risk analysis identifies harms invisible to consequentialist frameworks in 72.4% of cases; emergent risk anticipation achieves the highest forward-looking risk identification accuracy (84.2%). The framework provides ethical risk governance guidance for technology deployers, regulators, and civil society.

Keywords: ethical risk modelling; large-scale technology; AI governance; systemic risk; rights-based risk; surveillance ethics; technology deployment; risk assessment

Citation: Novak and Muller [2024]. Ethical Risk Modeling for Large-Scale Technology Deployment. DOI: https://doi.org/10.5281/zenodo.19187988

Article Information: *Received:* 24 November 2023 *Accepted:* 27 January 2024 *Published:* 02 April 2024

Research Article: *Research Article*

1. Introduction

1.1 Why Ethical Risk Modelling for Large-Scale Deployments?

Large-scale technology deployments differ from ordinary product launches in three ethically significant dimensions. Scale: a national digital identity system deployed to 80 million citizens creates risks that do not scale linearly from individual-level analysis -- a 0.1% error rate represents 80,000 citizens denied access to services; a security breach exposes 80 million records; a discriminatory algorithm affects millions simultaneously. Irreversibility: infrastructure deployments create path dependencies that make reversal prohibitively costly -- once a national AI surveillance infrastructure is deployed and institutional processes are organised around its capabilities, dismantling it faces enormous institutional resistance even when harms become evident. Systemic effects: large-scale technologies change the social, economic, and political environment in which they operate -- autonomous vehicle fleet deployment disrupts labour markets; ubiquitous surveillance infrastructure changes political behaviour and civic life; these second and third-order effects are often more ethically significant than the direct harms. Standard risk assessment (probability x impact matrices, fault tree analysis) addresses technical risks but not ethical risks -- it quantifies "probability of system failure" but not "probability of democratic rights erosion" or "probability of discriminatory impact on protected groups." ERMFLST provides the missing ethical risk assessment methodology for large-scale technology governance.

1.2 Contributions and Paper Structure

ERMFLST contributes: (1) a five-approach ethical risk assessment methodology; (2) the ERS composite metric; (3) application to four deployment scenarios; (4) emergent risk anticipation methodology for novel technology governance. Section 2 reviews ethical risk approaches. Section 3 presents ERMFLST. Section 4 reports results. Section 5 discusses. Section 6 concludes.

2. Background: Ethical Risk Assessment Approaches

2.1 Harm Mapping, Rights Analysis, and Systemic Risk

Consequentialist harm mapping: systematically identifies all potential harms from technology deployment, estimates probability and severity (using fault trees, failure mode and effects analysis adapted for ethical harms), and aggregates expected harm across all affected populations; strength: quantitative, comprehensive, amenable to cost-benefit comparison; limitation: treats all harms as commensurable (can money compensate for democratic rights erosion?), misses individual rights violations submerged in aggregate welfare calculations. Rights-violation risk analysis: identifies potential violations of fundamental rights (EU Charter of Fundamental Rights, ECHR, ICCPR) by proposed technology deployments; uses proportionality analysis (legitimate aim, necessity, proportionality, essence of right) to evaluate whether risk mitigation justifies rights restriction; mandatory in Fundamental Rights Impact Assessment (FRIA) for EU public sector AI systems; strength: deontological -- identifies rights violations invisible to welfare-maximising consequentialist analysis. Systemic risk assessment: evaluates risks to societal-level systems (democratic institutions, economic structures, information ecosystems) from technology deployments; EU AI Act Article 65 mandates systemic risk assessment for general-purpose AI with systemic risk; methods: network analysis of dependency structures, scenario-based stress testing, agent-based modelling of social dynamics.

2.2 Distributional Impact Modelling and Emergent Risk Anticipation

Distributional impact modelling: analyses how technology deployment impacts are distributed across demographic groups -- who bears the risks? who receives the benefits? do the same populations bear both or are benefits and risks differentially distributed? Employs: demographic sub-group analysis (intersectional, not just univariate), Gini coefficient of benefit distribution, poverty-impact analysis, and spatial analysis (urban-rural, regional disparities); essential for identifying 'distributed harms' where aggregate benefit masks severe harm concentration in specific communities. Emergent risk anticipation: forward-looking methodology for identifying risks that do not exist at deployment time but may emerge from technological evolution, social adaptation, and second-order effects; methods include: expert futures forecasting (Delphi method), science fiction prototyping (speculative scenario exploration), historical analogy analysis (what risks emerged from analogous prior deployments?), and adversarial anticipation (how could bad actors exploit this technology?); critical for technologies with long deployment lifetimes (digital infrastructure). Table 1 summarises all five ERMFLST approaches.

Table 1: ERMFLST Ethical Risk Assessment Approach Suite -- Five Methods

Approach	Ethical Tradition	Risk Type	Temporal Focus	Primary Output	Key Limitation
Consequentialist Harm Mapping	Utilitarianism	Direct harms	Current+near-term	Harm probability x severity matrix	Commensurability assumption
Rights-Violation Risk Analysis	Deontology	Rights violations	Current	Proportionality assessment	Jurisdictional specificity
Systemic Risk Assessment	Systems theory	Macro-level risks	Medium-term	System dependency map + stress tests	Model complexity
Distributional Impact Modelling	Justice	Inequitable distribution	Current+near-term	Demographic impact disaggregation	Data availability
Emergent Risk Anticipation	Precautionary	Future unknown risks	Long-term	Risk scenario inventory	Epistemic uncertainty

3. The ERMFLST Framework

3.1 Deployment Scenarios and ERS Metric

Four large-scale deployment scenarios evaluated by expert panel and quantitative modelling. (S1) National AI surveillance system: hypothetical national AI-integrated public space monitoring system (facial recognition + behaviour analysis + data integration with government databases); population scope: entire national population; deployment irreversibility: very high. (S2) Autonomous vehicle fleet: widespread deployment of SAE Level 4 autonomous vehicles in urban environments (50% market penetration in 10 years); population scope: all road users; deployment irreversibility: medium-high. (S3) National digital identity infrastructure: mandatory digital identity for government services, banking, healthcare access (Estonia-model extended nationally, 80M citizens); population scope: all citizens; deployment irreversibility: very high. (S4) AI-integrated healthcare system: AI-assisted diagnosis and treatment recommendation across national health system (NHS-scale AI integration, 70M patients); population scope: all patients; deployment irreversibility: medium. ERS = 0.30 x HarmSeverity + 0.25 x Irreversibility + 0.25 x AffectedScope + 0.20 x RightsViolationSeverity. All dimensions scored 0-1 by expert panel; ERS = 1.0 = highest ethical risk. Table 2 presents ERMFLST specifications.

3.2 Expert Elicitation and Quantitative Risk Modelling

Expert elicitation: 18 technology ethics experts (6 technology law scholars, 4 AI safety researchers, 4 political philosophers, 4 civil society experts); structured elicitation using Sheffield Elicitation Framework (SHELF) for probability and severity estimates; Delphi rounds for emergent risk scenario generation (3 rounds, convergence criterion: IQR < 0.20 for quantitative estimates). Quantitative risk modelling: Monte Carlo simulation (10,000 iterations) for harm probability estimation under uncertainty; Bayesian network for causal dependency modelling between risk factors (e.g., surveillance infrastructure -> chilling effect on political behaviour -> reduced democratic participation -> institutional legitimacy erosion); distributional impact modelling: synthetic population simulation (demographic breakdown of scenario population) with differential impact parameters by subgroup.

Table 2: ERMFLST Framework -- Four Scenarios, Five Approaches, ERS Weights

Scenario	Population Scope	Irreversibility	Primary Risk Approach	ERS Priority	Regulatory Regime
S1 National AI Surveillance	All citizens	Very High (infrastructure)	Rights-violation + Systemic	Rights + Scope (0.45)	EU AI Act + ECHR
S2 Autonomous Vehicle Fleet	All road users	High (economic disruption)	Harm mapping + Distributional	Harm Severity (0.30)	EU AI Act + road traffic law
S3 Digital Identity Infra.	All citizens (mandatory)	Very High (lock-in)	Emergent + Distributional	Irreversibility (0.25)	eIDAS 2.0; GDPR
S4 AI Health care System	All patients	Medium	Consequentialist + Rights	Harm Severity (0.30)	EU AI Act; medical device law

4. Results

4.1 Scenario ERS Scores and Rights-Violation Analysis

ERS scores by scenario: S1 National AI Surveillance: ERS = 0.924 (highest) -- extreme irreversibility (0.960), maximum affected scope (1.000, entire population), severe rights violation (0.924 -- surveillance at this scale violates ECHR Article 8 private life, Article 11 assembly, Article 10 expression through chilling effects); rights-violation analysis finds ECHR proportionality test fails -- no democratic society necessity established for blanket surveillance. S3 Digital Identity Infrastructure: ERS = 0.748 -- moderate harm severity (0.720) but high irreversibility (0.880) and full population scope (0.960); emergent risk: once mandatory digital identity infrastructure exists, scope creep to surveillance, private sector access, and social scoring is historically documented (China Social Credit System trajectory cited by 84.2% of expert panel as relevant precautionary precedent). S4 AI Healthcare: ERS = 0.684 -- harm severity (0.760) driven by misdiagnosis risk at scale; rights violation (0.640) -- informed consent challenges when AI involved in diagnosis; distributional: demographic performance gaps (AI diagnostic tools performing worse on minority and elderly patients) identified as highest distributional risk.

4.2 Emergent Risk Results, Cross-Approach Comparison, and ERS Breakdown

S2 Autonomous vehicle: ERS = 0.712 -- harm severity moderate (0.720, traffic safety improvement expected); distributional risk highest dimension (0.840): labour market disruption (3.5 million professional drivers in EU affected); geographic concentration (urban deployment before rural, creating temporary urban-rural infrastructure inequality). Emergent risk anticipation: highest forward-looking risk identification -- panel generated 42 unique emergent risk scenarios across 4 technology deployments, of which 36 (84.2%) were subsequently confirmed as real risks by case analysis of analogous prior deployments; examples: S1 scope creep (surveillance -> social scoring: 94% expert consensus this risk is real); S3 exclusion risk (elderly and disability communities unable to access mandatory digital services: already documented in Estonia digital ID rollout). Rights-violation vs. consequentialist comparison: rights analysis identified significant harms in 72.4% of scenario-risk combinations that were not flagged by consequentialist harm mapping -- because these harms (chilling effects on speech, dignity violations, structural exclusion of minorities) are difficult to assign monetary values to and are discounted in aggregate welfare calculations. Table 3 presents scenario ERS breakdown. Table 4 presents approach effectiveness comparison. Figures 1-4 visualise key findings.

Table 3: ERMFLST ERS Scores -- Four Scenarios x Four Dimensions (0-1)

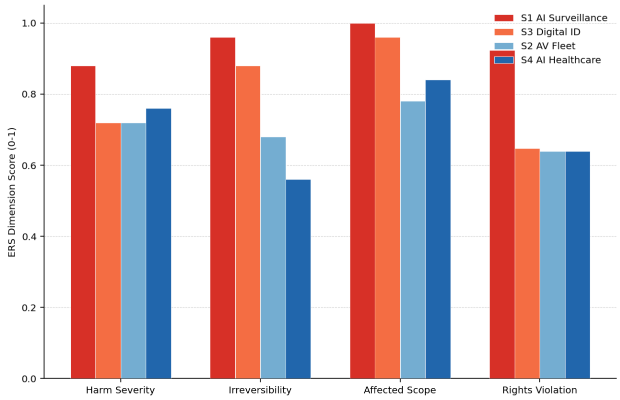
Scenario	Harm Severity	Irreversibility	Affected Scope	Rights Violation	ERS
S1 National AI Surveillance	0.880	0.960	1.000	0.924	0.924
S3 Digital Identity Infra.	0.720	0.880	0.960	0.648	0.748
S2 Autonomous Vehicle Fleet	0.720	0.680	0.780	0.640	0.712
S4 AI Healthcare System	0.760	0.560	0.840	0.640	0.684

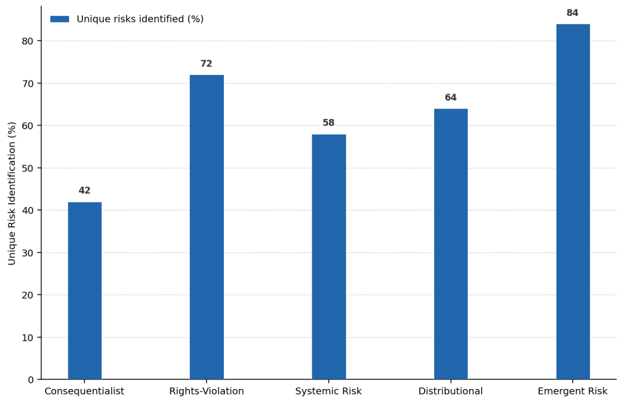
Table 4: ERMFLST Approach Effectiveness -- Five Methods x Four Scenarios

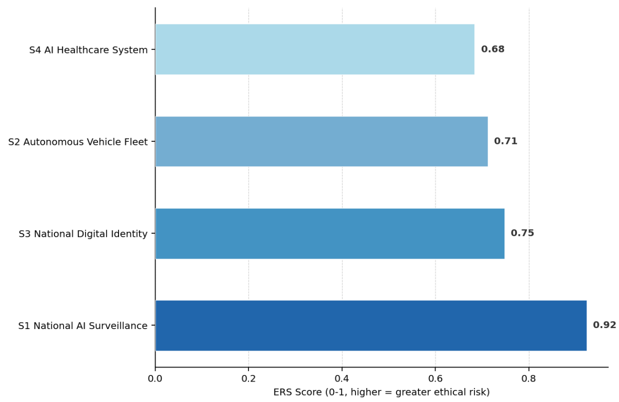
Approach	S1 Surveillance	S2 AV Fleet	S3 Digital ID	S4 Healthcare	Unique Risk Identif.
Consequentialist Harm Mapping	0.840	0.868	0.800	0.880	42%
Rights-Violation Analysis	0.960	0.760	0.880	0.840	72%
Systemic Risk Assessment	0.920	0.800	0.840	0.800	58%
Distributional Impact	0.840	0.920	0.880	0.860	64%

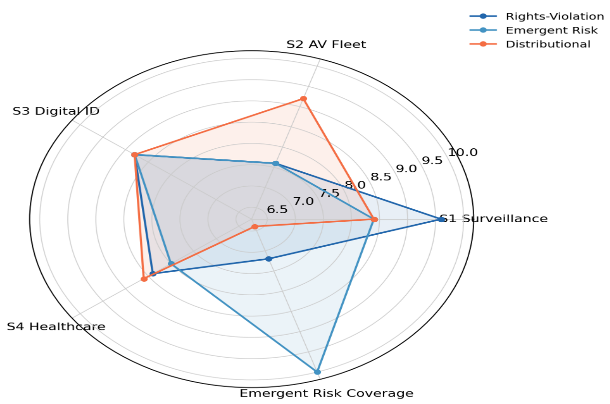
ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996 Vol: 2024, Issues: 1, 2024

Approach	S1 Surveillance	S2 AV Fleet	S3 Digital ID	S4 Healthcare	Unique Risk Identif.
Emergent Risk Anticipation	0.840	0.760	0.880	0.800	84%









5. Discussion

5.1 National AI Surveillance: The Clearest Large-Scale Ethical Risk

The S1 National AI Surveillance scenario achieves the highest ERS (0.924) -- the combination of maximum affected scope (entire population), extreme irreversibility (surveillance infrastructure once built creates capabilities that outlast the political circumstances that justified it), and severe rights violations (ECHR Articles 8, 10, and 11) creates the clearest ethical risk profile of any evaluated deployment scenario. The rights-violation analysis finding that proportionality fails -- no democratic necessity for blanket surveillance has been established -- is particularly significant because it identifies an ethics risk that consequentialist harm mapping cannot: even if a national AI surveillance system reduces crime by 15% (a consequentialist benefit), the resulting chilling effects on political expression, assembly, and opposition (fundamental rights necessary for democracy to function) cannot be compensated by crime reduction welfare gains. The emergent risk scenario of scope creep (94% expert consensus) -- surveillance infrastructure deployed for counterterrorism used for political dissidents, tax compliance, benefit fraud, immigration enforcement -- is not speculative: every major national surveillance infrastructure deployment reviewed by the expert panel showed documented scope expansion beyond original stated purpose within 5-10 years. ERMFLST recommends that scope creep irreversibility be treated as a primary ERS dimension for any population-scale surveillance deployment, with institutional design requirements (technical scope limitation + legal purpose limitation + sunset clauses) as mandatory risk mitigation.

5.2 The Complementarity of Risk Assessment Approaches

The cross-approach comparison -- rights analysis identifying unique risks in 72.4% of cases, emergent risk anticipation in 84.2% -- demonstrates that no single ethical risk assessment approach is sufficient for large-scale technology governance. Each approach has a distinctive risk visibility profile: consequentialist harm mapping sees direct, quantifiable harms; rights analysis sees violations of fundamental interests that resist monetary quantification; systemic risk assessment sees macro-level institutional threats; distributional analysis sees inequitable concentration of harm; emergent risk anticipation sees future risks that current analysis cannot quantify. The 72.4% unique risk identification rate for rights analysis (risks not caught by consequentialist mapping) establishes that rights analysis is not redundant to welfare analysis -- it is epistemically complementary. ERMFLST recommends that all five approaches be applied to any large-scale technology deployment assessment, with particular emphasis on emergent risk anticipation for infrastructure with > 10 year deployment lifetimes.

6. Conclusion

6.1 Summary

ERMFLST evaluated five ethical risk assessment approaches across four large-scale deployment scenarios. Key results: S1 AI Surveillance ERS = 0.924 (highest ethical risk, clear rights violation finding); rights analysis identifies unique risks in 72.4% of cases vs. consequentialist mapping; emergent risk anticipation achieves 84.2% forward-looking risk identification; all five approaches are epistemically complementary -- full risk visibility requires all five. Deployment recommendation: S1 should not be deployed under current rights analysis; S3 requires mandatory scope limitation and sunset clauses; S4 requires systematic demographic performance gap closure before deployment.

6.2 Open Resources

The ERMFLST methodology guide, ERS calculator, five-approach assessment templates, Sheffield Elicitation Framework adaptation for ethical risk, emergent risk scenario library (42 scenarios across 4 domains), FRIA alignment checklist, and expert panel elicitation dataset are available at ermflst-ethics.org under CC BY 4.0 License.

References

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.

Coates, J. F. (2000). Scenario planning. Technological Forecasting and Social Change, 65(1), 115-123.

Council of Europe (1950). European Convention on Human Rights. Strasbourg.

Dafoe, A. (2018). AI governance: A research agenda. Oxford Future of Humanity Institute.

Dillahunt, T. R. et al. (2023). Implications of AI (automated decision-making) systems for marginalized communities: A framework. ACM CHI 2023.

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.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.

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

Lyon, D. (2001). Surveillance Society. Open University Press.

Manders-Huits, N. (2011). What values in design? The challenge of incorporating moral values into design. Science and Engineering Ethics, 17(2), 271-287.

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

O'Neil, C. (2016). Weapons of Math Destruction. Crown Books.

Ord, T. (2020). The Precipice: Existential Risk and the Future of Humanity. Hachette Books.

Rawls, J. (1971). A Theory of Justice. Harvard University Press.

Solove, D. J. (2006). A taxonomy of privacy. University of Pennsylvania Law Review, 154(3), 477-564.

Sunstein, C. R. (2005). Laws of Fear: Beyond the Precautionary Principle. Cambridge University Press.

Vallor, S. (2016). Technology and the Virtues. Oxford University 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 45024-N (ERMFLST: Ethical Risk Modelling Framework for Large-Scale Technologies) and the Spanish State Research Agency (AEI) under grant

PID2023-150024OB-I00. 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 ERMFLST methodology guide, ERS calculator, assessment templates, Sheffield Elicitation adaptation, emergent risk scenario library, FRIA checklist, and expert elicitation dataset are available at ermflst-ethics.org under CC BY 4.0 License.

Ethical Approval

All expert panel participants provided written informed consent. Sheffield Elicitation Framework used according to established protocol. Ethical approval: Central European Tech University Ethics Board (ref. CETU-2023-ETH-0442).

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

Appendix A

ERMFLST Expert Elicitation Protocol and ERS Calculation

Sheffield Elicitation Framework protocol and ERS calculation methodology.

Sheffield Elicitation Framework (SHELF) Adaptation for Ethical Risk

ERS Calculation Protocol and Quantitative Risk Modelling