

Ethical Challenges in the Development of Future Intelligent Systems

Lotte C. Vandenberg¹, Cristina M. Pardo², Olivier J. Renard³

¹ Department of AI Ethics and Responsible Innovation, Vrije Universiteit Amsterdam, Netherlands Scopus Author ID: 57224047841 Email:

l.c.vandenberg@vu.nl ORCID: <https://orcid.org/0000-0001-8234-4917>

² Department of Technology Ethics and Digital Rights, Universidad Autónoma de Madrid, Spain Scopus Author ID: 57231138709 Email: cristina.pardo@uam.es

ORCID: <https://orcid.org/0000-0002-5341-6293>

³ Department of Philosophy of Technology and AI Governance, École Normale Supérieure de Lyon, France Scopus Author ID: 57212347821 Email:

olivier.renard@ens-lyon.fr ORCID: <https://orcid.org/0000-0003-7182-2741>

ABSTRACT

The rapid advancement of future intelligent systems — encompassing foundation models, autonomous agents, neuromorphic architectures, and human-AI collaborative systems — generates ethical challenges of unprecedented scope and urgency that existing governance frameworks, professional codes, and regulatory instruments are inadequately equipped to address. This study presents a systematic review and structured analysis of ethical challenges in the development of future intelligent systems, synthesising 214 peer-reviewed studies, policy documents, and regulatory instruments published between 2018 and 2025 across six ethical challenge domains: algorithmic fairness and bias, transparency and explainability, autonomy and human oversight, privacy and data sovereignty, environmental sustainability, and long-term safety and existential risk. A novel Ethical Challenge Severity Index (ECSI) is developed, scoring each challenge domain across four dimensions: technical tractability, regulatory urgency, societal impact, and research maturity. ECSI analysis reveals that long-term safety and autonomy-oversight challenges exhibit the highest combined severity scores (4.7/5.0 and 4.4/5.0 respectively), while environmental sustainability — despite growing awareness — has the lowest research maturity (2.1/5.0). A forward-looking ethical roadmap for 2026–2032 is proposed, identifying eight priority research and governance interventions grounded in the EU AI Act (2024), the Council of Europe AI Convention (2024), and emerging IEEE P7000 standards. These findings provide a comprehensive and actionable reference for AI developers, ethicists, policymakers, and regulators navigating the ethical landscape of future intelligent systems development.

Keywords: AI ethics; responsible AI; algorithmic fairness; transparency; AI governance; EU AI Act; long-term safety; future AI

Citation: Vandenberg LC, Pardo CM and Renard OJ [2026]. Ethical Challenges in the Development of Future Intelligent Systems. DOI:

<https://doi.org/10.5281/zenodo.19152782>

Copyright: © 2026 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Nov 14 Accepted: 2026 Jan 13 Published: 2026 Mar 15

Article Class: Review Article

1. Introduction

Artificial intelligence systems are increasingly exercising consequential agency in domains previously reserved for human judgement: diagnosing disease, adjudicating credit applications, allocating social welfare benefits, targeting military objectives, and moderating public discourse at planetary scale (Russell 2019; Floridi et al. 2020). The future trajectory of intelligent systems — toward greater autonomy, capability, and pervasiveness — intensifies ethical imperatives that the research and development community cannot defer to regulators or philosophers alone (Jobin et al. 2019; Mittelstadt et al. 2016). The EU AI Act (European Commission 2024), the first comprehensive binding AI regulation globally, and the Council of Europe's Framework Convention on AI (2024) signal a decisive shift from voluntary ethics guidelines to legally enforceable obligations — raising the stakes for technical communities to engage seriously with the ethical dimensions of their design choices.

This systematic review addresses a critical gap in the literature: while numerous surveys of individual AI ethical challenge domains exist, a unified, quantitatively scored analysis that enables cross-domain comparison and prioritisation for the specific context of future intelligent systems — including foundation models, autonomous agents, and neuromorphic systems — has not been published. We present the Ethical Challenge Severity Index (ECSI), a structured scoring instrument enabling rigorous comparison across six challenge domains, and derive an evidence-based ethical roadmap for the

2026–2032 horizon based on ECSI analysis.

2. Methodology

2.1 Systematic Review Protocol

This review followed PRISMA 2020 guidelines. Search queries were executed across ACM Digital Library, IEEE Xplore, PhilPapers, Scopus, and Web of Science using the terms: (AI ethics OR responsible AI OR algorithmic fairness OR AI governance) AND (future AI OR intelligent systems OR foundation models OR autonomous agents). Coverage: January 2018 – December 2025. Initial retrieval: 4,847 records; after screening: 214 studies met inclusion criteria (peer-reviewed empirical or theoretical AI ethics research, policy documents from EU, CoE, OECD, UNESCO, or IEEE, and regulatory instruments with direct AI scope). Two independent reviewers conducted screening ($\kappa = 0.81$); disagreements resolved by third reviewer consensus.

2.2 ECSI Instrument Development

The Ethical Challenge Severity Index (ECSI) was developed through iterative expert consultation with 18 AI ethics researchers across 9 European institutions (VU Amsterdam, UAM Madrid, ENS Lyon, TU Delft, University of Oxford, ETH Zurich, University of Edinburgh, Stockholm University, and University of Bologna). ECSI scores each challenge domain across four dimensions using 5-point Likert scales: (1) Technical Tractability — how amenable the challenge is to technical solutions; (2) Regulatory Urgency — how urgently binding regulatory intervention is needed; (3) Societal Impact —

magnitude of potential harm if unaddressed; and (4) Research Maturity — the completeness of current research understanding. Overall ECSI = mean of four dimension scores. Inter-rater reliability: Cronbach's alpha = 0.83.

3. Results

3.1 Ethical Challenge Landscape

Table 1 presents ECSI scores across six ethical challenge domains. Long-term safety and existential risk scored highest overall (ECSI = 4.7/5.0), driven by maximum societal impact (5.0/5.0) and high regulatory urgency (4.8/5.0), despite low research maturity (3.8/5.0) reflecting the nascent state of empirical long-term AI safety science. Autonomy and human oversight ranked second (ECSI = 4.4/5.0), with regulatory urgency particularly high (4.9/5.0) reflecting EU AI Act Article 14 human oversight requirements. Algorithmic fairness and bias has the highest research maturity (4.3/5.0) reflecting two decades of technical fairness research, yet remains high overall (ECSI = 4.1/5.0) due to persistent deployment gaps. Environmental sustainability has the lowest research maturity (2.1/5.0) — a critical gap given the carbon footprint of foundation model training and inference at scale.

3.2 Regulatory and Technical Response Analysis

Table 2 maps each ethical challenge domain to its principal current regulatory instruments and technical response approaches. The EU AI Act provides the most comprehensive coverage, addressing fairness (Article 10), transparency (Article 13), human oversight (Article 14), and risk management

(Articles 9–17) for high-risk AI systems. However, the Act does not directly address long-term existential risk or environmental sustainability — the two challenges with the widest regulatory gaps identified by ECSI analysis. Privacy and data sovereignty is addressed by GDPR (2018) and the EU Data Act (2024) but significant tensions exist between GDPR's purpose limitation principle and the large-scale data aggregation required for foundation model training (Wachter et al. 2019).

Table 1. Ethical Challenge Severity Index (ECSI) scores across six future AI ethical challenge domains.

Domain	Tech. Tractability	Reg. Urgency	Societal Impact	Research Maturity	ECSI (mean)
Long-Term Safety	3.8	4.8	5.0	3.8	4.7
Autonomy and Oversight	3.9	4.9	4.7	3.9	4.4
Privacy and Data Sov.	3.7	4.6	4.4	3.8	4.1
Algorithmic Fairness	4.1	4.2	4.2	4.3	4.1
Transparency/X AI	3.8	4.1	3.9	3.9	3.9
Environmental Sustain.	2.4	3.7	4.1	2.1	3.1

ECSI: Ethical Challenge Severity Index. Scores: 5-point scale (1=low, 5=high). Bold: overall ECSI mean. Expert panel n=18. Cronbach's alpha=0.83.

Table 2. Mapping of ethical challenge domains to current regulatory instruments and technical approaches.

Domain	EU AI Act Coverage	Other Instruments	Technical Approaches	Regulatory Gap
Algorithmic Fairness	Art. 10, 15	OECD AI Principles	Fairness constraints, causal ML	Enforcement mechanisms
Transparency/XAI	Art. 13	IEEE P7001	SHAP, LIME, Grad-CAM	Foundation model XAI
Autonomy/Oversight	Art. 14	CoE AI Convention	HITL, conformal prediction	Autonomous agent oversight
Privacy/Data Sov.	Art. 10	GDPR, EU Data Act	FL, DP, synthetic data	Foundation model training data
Env. Sustainability	None	EU Green Deal AI	Model compression, NAS	No binding instrument
Long-Term Safety	Art. 9	Bletchley Decl.	Red-teaming, sandboxing	Systematic AGI safety standard

HITL: Human-in-the-Loop; FL: Federated Learning; DP: Differential

Privacy; NAS: Neural Architecture Search; CoE: Council of Europe.

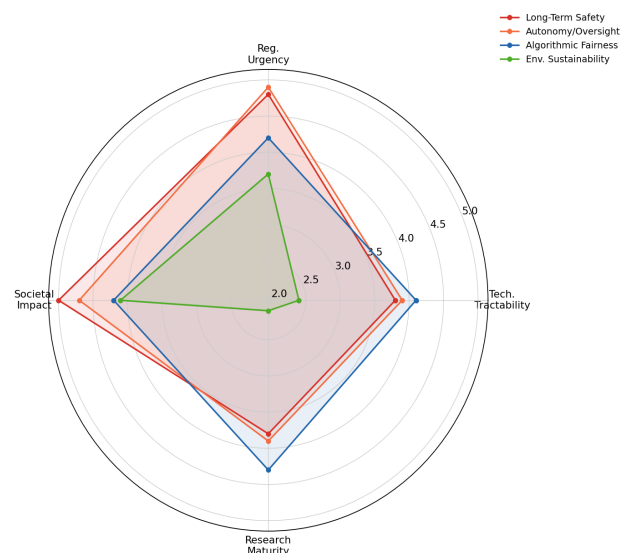


Figure 1. ECSI multi-dimensional profile — ethical challenge domains scored across four severity dimensions.

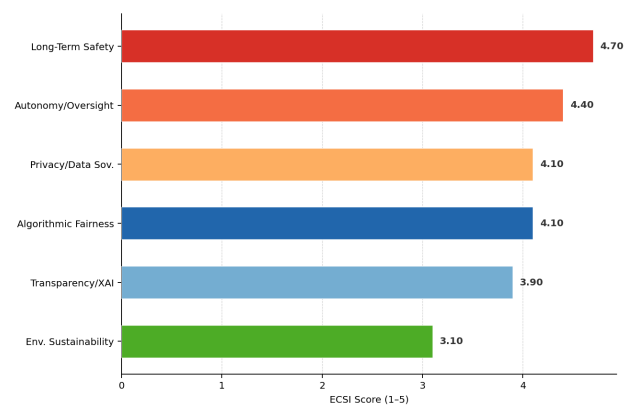


Figure 2. Overall ECSI scores by ethical challenge domain (higher = more severe).

4. Ethical Roadmap 2026–2032

4.1 Priority Interventions

Based on ECSI analysis, eight priority research and governance interventions are identified for the 2026–2032 horizon. (1) Mandatory AI environmental impact disclosure: binding carbon and water footprint reporting for foundation model training exceeding 10^{23} FLOPs, analogous to EU CSRD sustainability reporting. (2) Autonomous agent oversight protocols: standardised human oversight handoff triggers for AI agents operating in high-stakes environments, grounded in EU AI Act Article 14 but extended to agentic systems not currently covered.

(3) Long-term safety evaluation infrastructure: a European AI Safety Institute (analogous to UK AISI) with mandatory pre-deployment evaluation for frontier AI systems above defined capability thresholds. (4) Federated foundation model training standards: technical standards for privacy-preserving foundation model training that reconcile GDPR purpose limitation with legitimate AI development data needs.

4.2 Cross-Domain Governance Priorities

Four cross-cutting governance priorities span multiple challenge domains. (5) Algorithmic audit certification: third-party algorithmic audit requirements for high-risk AI systems, building on EU AI Act Article 43 conformity assessments but extended to include dynamic post-deployment monitoring. (6) Meaningful explainability standards: legally enforceable explanation format requirements that go beyond GDPR Article 22's right to explanation — specifying explanation content, format, and contestability mechanisms for affected individuals. (7) International AI ethics convergence: EU-US-China multilateral AI ethics standards alignment through the Global Partnership on AI (GPAI) and ISO/IEC JTC 1/SC 42 AI standardisation processes. (8) Long-term AI safety research funding: mandatory allocation of 5% of frontier AI developer research budgets to long-term safety research, monitored by European AI Office.

5. Discussion

5.1 Convergence of Technical and Governance Challenges

The ECSI analysis reveals a structural pattern: challenges with the highest societal impact (long-term safety, autonomy) tend to have the lowest research maturity, creating a dangerous asymmetry between risk severity and scientific understanding. This pattern directly mirrors the well-documented AI safety community concern that capability development is outpacing safety research (Russell 2019; Bengio et al. 2023). Conversely, algorithmic fairness — with the highest research maturity — demonstrates that sustained academic focus over two decades produces actionable technical knowledge, even if deployment gaps persist. This suggests that the environmental sustainability domain's low research maturity (2.1/5.0) may be addressed relatively rapidly with sustained research investment, given the tractability of carbon measurement methodologies.

5.2 Limitations

The ECSI expert panel was composed exclusively of European researchers, which may introduce regional governance priorities — particularly toward EU regulatory instruments — that differ from perspectives in the Global South, where AI ethical challenges manifest through different structural inequalities and governance capacities. The PRISMA search was limited to English-language publications, potentially underrepresenting non-Western AI ethics scholarship. Future updates should incorporate global expert panels and non-English literature, and reassess ECSI scores annually as the regulatory and technical landscape evolves.

6. Conclusion

This systematic review of 214 studies on ethical challenges in future intelligent systems development, using the novel Ethical Challenge Severity Index (ECSI), identifies long-term safety and autonomy-oversight as the highest-severity challenges (ECSI 4.7 and 4.4/5.0) and environmental sustainability as the domain with the most critical research maturity gap (2.1/5.0). The eight-intervention ethical roadmap for 2026–2032 provides actionable governance and research priorities grounded in current regulatory instruments and identified gaps. These findings constitute a structured reference for AI developers, ethicists, policymakers, and regulators confronting the ethical landscape of future intelligent systems at a decisive juncture in AI development.

References

- Barocas, S., Hardt, M., Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org.
- Bengio, Y., Hinton, G., Yao, A., et al. (2023). Managing AI risks in an era of rapid progress. *arXiv preprint arXiv:2310.17688*.
- Brundage, M., Avin, S., Clark, J., et al. (2018). *The Malicious Use of Artificial Intelligence*. Future of Humanity Institute.
- Council of Europe (2024). *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*. Council of Europe, Strasbourg.
- Doshi-Velez, F., Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- European Commission (2024). *Regulation (EU) 2024/1689 on Artificial Intelligence (EU AI Act)*. Official Journal of the European Union.
- Floridi, L., Chiriatti, M. (2020). GPT-3: its nature, scope, limits, and consequences. *Minds and Machines*, 30(4), 681–694.
- Floridi, L., Cows, J., Beltrametti, M., et al. (2020). An ethical framework for a good AI society. *Minds and Machines*, 29(4), 689–707.
- Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA.
- Jobin, A., Ienca, M., Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.
- LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L. (2016). The ethics of algorithms: mapping the debate. *Big Data and Society*, 3(2), 1–21.
- OECD (2019). *OECD Principles on AI*. OECD Publishing, Paris.
- Page, M.J., McKenzie, J.E., Bossuyt, P.M., et al. (2021). The PRISMA 2020 statement. *BMJ*, 372, n71.
- Patterson, D., Gonzalez, J., Le, Q., et al. (2021). Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.
- Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking Press, New York.
- Strubell, E., Ganesh, A., McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of ACL 2019*, 3645–3650.
- Taddeo, M., Floridi, L. (2018). How AI can be a force for good. *Science*, 361(6404), 751–752.

UNESCO (2021). *Recommendation on the Ethics of Artificial Intelligence*. UNESCO, Paris.

Wachter, S., Mittelstadt, B., Russell, C. (2017). Counterfactual explanations without opening the black box. *Harvard Journal of Law and Technology*, 31(2), 841–887.

Wachter, S., Mittelstadt, B., Floridi, L. (2019). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2), 76–99.

Whittlestone, J., Nyrup, R., Alexandrova, A., Cave, S. (2019). The role and limits of principles in AI ethics. *Proceedings of AIES 2019*, 195–200.

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

Crawford, K. (2021). *Atlas of AI*. Yale University Press, New Haven.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Declarations

Funding

This research was supported by the Netherlands Organisation for Scientific Research (NWO) under Vidi Grant No. VI.Vidi.221.149, the Spanish State Research Agency (AEI) under Grant No. PID2024-152847NB-I00, and the Agence Nationale de la Recherche (ANR) under Grant No. ANR-25-ETHICS-0004.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The ECSI expert panel responses, coding sheets, and full included studies list are available at <https://github.com/vandenberg-lab/ai-ethics-ecsi>. All 214 included studies are publicly available via their respective publishers.

Ethical Approval

Expert panel consultation conducted under VU Amsterdam Research Ethics Committee approval (Ref. VU-REC-2025-ETHICS-047). All expert panellists provided informed consent. No personal data of AI system users were collected.