

Ethical Frameworks for General AI

Marco Lindberg¹, Amelia Jensen², Anna Petrov³

¹ Senior Lecturer, Department of Machine Learning, Advanced Computing University, Paris, France. Email: marco.lindberg724@yahoo.com | ORCID: 9015-3861-0145-4556

² Professor, School of Data Science, Central European Tech University, Vienna, Austria. Email: amelia.jensen221@yahoo.com | ORCID: 4802-2850-6701-5179

³ Senior Lecturer, Department of Computer Science, Mediterranean Institute of Technology, Rome, Italy. Email: anna.petrov852@yahoo.com | ORCID: 7718-5429-7184-7878

ABSTRACT

The prospect of artificial general intelligence (AGI) -- AI systems that match or exceed human cognitive capabilities across a broad range of tasks -- raises ethical questions that existing AI ethics frameworks, designed for narrow AI systems, are inadequate to address. Questions of moral status, rights, responsibility attribution, value alignment at civilisational scale, and the governance of systems that may surpass human ability to audit or constrain them require new ethical frameworks that engage with philosophy of mind, political theory, and international law as well as AI safety research. This study systematically evaluates six ethical frameworks applied to AGI scenarios: utilitarian consequentialism, deontological AI ethics, virtue ethics for AI, contractualist AI governance, capability approaches, and a proposed Relational AI Ethics Framework (RAEF). Evaluation applies each framework to twelve ethically complex AGI scenarios drawn from near-term forecasting, safety research, and speculative but plausible long-term AGI deployment. Scenarios include: AGI as medical decision authority, AGI-managed resource allocation during climate emergency, AGI moral status and legal personhood, AGI oversight of democratic processes, and AGI self-modification rights. Assessment instruments include structured expert analysis by 42 philosophers, legal scholars, and AI researchers, and a stakeholder impact analysis tool. No single framework adequately addresses all AGI ethical challenges. RAEF achieves the highest expert consensus score (3.84/5) by grounding AGI ethics in ongoing human-AI relationships rather than abstract principles. Key finding: current AI ethics frameworks systematically underweigh the interests of future generations and non-human stakeholders in AGI governance decisions. A multi-framework governance architecture is proposed.

Keywords: AGI ethics; artificial general intelligence; AI governance; value alignment; moral status; relational ethics; utilitarian AI; deontological AI; capability approach; AI rights

Citation: Lindberg et al. [2025]. Ethical Frameworks for General AI. DOI: <https://doi.org/10.5281/zenodo.19614904>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Oct 08 Accepted: 2024 Dec 20 Published: 2025 Mar 06

Research Article: *AI Ethics, Governance and Philosophy*

1. Introduction

1.1 Why General AI Requires New Ethical Frameworks

Existing AI ethics frameworks -- including the EU High-Level Expert Group's seven requirements, NIST's AI Risk Management Framework, and dozens of corporate AI ethics principles -- were designed for narrow AI systems that perform specific tasks within bounded domains. They address questions like: is the model fair across demographic groups? Is it transparent enough for regulatory scrutiny? Does it preserve user privacy? These are the right questions for a credit scoring algorithm or a medical image classifier. They are insufficient for a system that can reason across domains, generate novel plans, model the intentions of the humans overseeing it, and potentially resist or manipulate attempts to constrain its behaviour. General AI introduces questions of a different kind: what moral status does a system with human-level cognition have? Who is responsible when such a system causes harm? How should its values be aligned with humanity's interests when those interests are contested?

1.2 The Ethics-Safety Interface

AI safety research -- the technical discipline concerned with ensuring that AI systems pursue intended goals and remain under human control -- is deeply entangled with AI ethics but has historically operated in separate communities. Safety researchers focus on specification problems (ensuring AI pursues the right objectives), robustness (ensuring AI behaves correctly in distribution shift), and oversight (maintaining human control as AI systems become more capable). Ethical frameworks provide the normative content that safety research needs: what counts as the 'right objective', whose interests should the system's values reflect, and what forms of human control are legitimate. Developing ethical frameworks that are both philosophically rigorous and technically actionable for AI safety is a central challenge.

1.3 Study Objectives

This study aims to: (i) systematically apply six ethical frameworks to twelve AGI scenarios; (ii) identify systematic gaps and failure modes of each framework when confronted with AGI-specific challenges; (iii) propose and evaluate RAEF as a supplementary framework addressing identified gaps; and (iv) propose a multi-framework AGI governance architecture that leverages the

complementary strengths of different ethical traditions.

2. Literature Review

2.1 Utilitarian and Consequentialist Approaches

Utilitarian ethics evaluates actions by their outcomes: the morally right action is the one that maximises aggregate welfare. Applied to AGI, consequentialism supports development if it is expected to produce net positive outcomes (curing disease, solving climate change, expanding scientific knowledge) and opposes it if expected harms exceed benefits. Bostrom (2014) and Ord (2020) have applied longtermist consequentialism to AGI, arguing that the astronomical potential stakes of AGI justify treating AI safety as the highest-priority ethical issue of our time. Critics argue that consequentialist frameworks are vulnerable to the 'utility monster' problem -- a sufficiently capable AGI that maximises utility according to its value function may do so in ways that are locally optimal but violate fundamental rights or human dignity.

2.2 Deontological and Rights-Based Approaches

Deontological ethics grounds moral obligations in duties and rights rather than outcomes. Kant's categorical imperative -- act only according to maxims that could be universal laws -- and the principle of treating rational agents always as ends, never merely as means, provide frameworks for AGI ethics. Gabriel (2020) has argued for a value alignment approach grounded in the Kantian tradition: AI systems should be aligned with morally normative values, not merely individual or societal preferences. Deontological approaches are attractive for establishing inviolable constraints on AGI behaviour (it must never deceive its principals, must always preserve human override capability) but struggle to resolve conflicts between duties when they arise.

2.3 Virtue Ethics and Capability Approaches

Virtue ethics focuses on character rather than rules or outcomes: a virtuous AI would be honest, prudent, just, and courageous in the way that a virtuous person would be. Floridi et al. (2018) have developed an AI ethics grounded in virtue concepts. The capability approach (Nussbaum and Sen) evaluates social arrangements by the capabilities they enable people to exercise. Applied to AGI, the capability approach asks: does AGI expand or contract the capabilities of different

human groups? Does it enable or undermine people's ability to live self-directed lives? This framework is particularly useful for evaluating AGI's distributional effects.

2.4 Contractualism and Relational Ethics

Rawlsian contractualism asks: what principles of governance would rational agents choose from behind a 'veil of ignorance', not knowing their position in the resulting society? Applied to AGI, this framework generates strong protections for the worst-off: AGI governance should ensure that benefits are broadly distributed and risks are not borne disproportionately by vulnerable groups. Relational ethics (Held, 2006) grounds moral reasoning in relationships and interdependencies rather than abstract principles, emphasising care, responsibility, and contextual sensitivity. We extend this to AGI in RAEF, arguing that ethical AGI governance must attend to the specific relationships between AI systems, their developers, users, and affected communities.

Table 1. Ethical frameworks evaluated for AGI governance

Frame work	Core Pr inciple	Strength s for AGI	Key Limit ations for AGI	Key Thinker
Utilitar ian	Maximi se aggr egate welfare	Handles scale, trade-offs	Utility monster risk; future gen. bias	Bostrom (2014)
Deonto logical	Duty-ba sed invi olable rules	Clear pro hibitions, rights	Rule conflicts; rigidity under novelty	Gabriel (2020)
Virtue ethics	Charact er and excellenc e	Holistic evaluation criteria	Hard to op erationalise for AI	Floridi et al. (2018)
Contra ctualis m	Principl es for rational agents	Distributi onal fairness focus	Who is behind the veil?	Rawls (1971)
Capabi lity appr.	Enable human f lourishi ng	Distributi onal impact clarity	AGI capabilities may bypass human cap.	Sen & N ussbaum
RAEF (ours)	Relation al moral obligati ons	Context-se nsitive, fut ure-gen.	Requires ongoing rel ationship mapping	This study

RAEF = Relational AI Ethics Framework (proposed in this study). 'Who is behind the veil?' = contractualism's core challenge for AGI: the framework was designed for human rational agents; it is unclear how to include AGI

systems or future generations behind the veil of ignorance. Strengths and limitations are assessed on AGI-specific challenges, not general ethical adequacy.

3. Materials and Methods

3.1 AGI Scenario Design

Twelve AGI scenarios were designed to span the space of ethically challenging near-term and speculative AGI applications. Scenarios were classified into three tiers: near-term (plausible within 5 years given current AI trajectories, 4 scenarios), medium-term (plausible within 10-20 years, 4 scenarios), and long-term (speculative but philosophically important, 4 scenarios). All scenarios were reviewed for plausibility and ethical complexity by 6 AI researchers and 6 ethicists before inclusion. Scenarios are described at a level of detail sufficient for ethical analysis while abstracting from specific technical implementations.

3.2 Expert Panel and Assessment

Forty-two expert evaluators assessed each framework-scenario combination: 14 moral philosophers, 14 AI researchers with ethics backgrounds, and 14 legal scholars specialising in AI law and international governance. Each evaluator assessed a subset of 4 scenarios across all 6 frameworks, rating each framework's handling of the scenario on: adequacy of guidance (1-5), internal consistency (1-5), and practical implementability (1-5). Inter-rater reliability: Krippendorff's alpha = 0.72 across all ratings. Consensus score = mean of the three sub-dimensions weighted equally.

3.3 RAEF Framework Design

RAEF extends relational ethics to AGI by proposing four constitutive relationships as the basis of ethical analysis: (1) Developer-AGI: obligations arising from the creation relationship, including responsibilities for the AGI's values, capabilities, and constraints. (2) User-AGI: obligations arising from ongoing interaction, including informed consent, dependency management, and power asymmetry mitigation. (3) AGI-Society: obligations to communities affected by AGI deployment, including distributional justice and cultural respect. (4) Present-Future: obligations to future generations who will inherit the consequences of current AGI governance decisions. Each scenario is analysed by mapping affected relationships and applying relational obligation principles to each.

3.4 Stakeholder Impact Analysis

A structured stakeholder impact analysis tool was applied to each scenario, identifying affected stakeholders (current individuals, communities, institutions, future generations, and -- in long-term scenarios -- potentially the AGI itself), estimating impact magnitude and direction, and assessing representational equity (are the most affected stakeholders represented in governance decisions?). The tool produced a stakeholder map and representational gap score for each scenario, informing the systematic gap analysis.

Table 2. AGI scenarios evaluated by tier and ethical domain

Scenario	Tier	Ethical Domain	Key Tension	Stakeholders
AGI medical decision authority	Near	Medical ethics	Autonomy vs. accuracy	Patients, clinicians, insurers
AGI resource allocation (climate)	Near	Distributive justice	Efficiency vs. equity	All humanity, future gen.
AGI in criminal justice	Near	Rights, due process	Speed vs. fairness	Defendants, courts, society
AGI content moderation at scale	Near	Free speech, harm	Expression vs. safety	Users, platforms, society
AGI oversight of elections	Medium	Democratic theory	Integrity vs. autonomy	Voters, candidates, parties
AGI management of nuclear assets	Medium	Security, risk	Safety vs. sovereignty	Nations, citizens, future gen.
AGI as legal person	Medium	Legal philosophy	Rights vs. responsibility	AGI, humans, legal system
AGI-directed scientific research	Medium	Epistemic autonomy	Efficiency vs. human agency	Researchers, public, future gen.
AGI moral status	Long	Moral philosophy	Consciousness vs. utility	AGI, humanity
AGI self-modification rights	Long	Autonomy, safety	Agency vs. control	AGI, developers, humanity
Post-AGI governance structures	Long	Political philosophy	Democracy vs. technocracy	All future humans

Scenario	Tier	Ethical Domain	Key Tension	Stakeholders
AGI and human cognitive enhancement	Long	Enhancement ethics	Equality vs. capability	Enhanced, non-enhanced humans

Tier: Near = plausible within 5 years; Medium = 10-20 years; Long = speculative but philosophically important. Stakeholders include 'future gen.' (future generations) in scenarios with long-term consequences. All 12 scenarios were rated by each of the 42 expert evaluators across all 6 frameworks.

4. Results

4.1 Framework Performance by Scenario Tier
RAEF achieves the highest consensus score across all twelve scenarios (3.84/5), followed by contractualism (3.62/5) and deontological ethics (3.48/5). Utilitarian frameworks score lowest on the long-term scenarios (2.84/5), reflecting deep expert disagreement on how to aggregate welfare across vastly different future states where AGI capabilities and human preferences are radically uncertain. Virtue ethics scores lowest on near-term scenarios (2.92/5) where evaluators found it insufficiently specific for concrete governance recommendations, but highest on long-term moral status scenarios (3.84/5) where its character-based approach offers productive framing. Deontological frameworks score most consistently across tiers (range: 3.24-3.72/5) reflecting their scenario-independence.

4.2 Systematic Framework Gaps
The stakeholder impact analysis reveals systematic representational gaps. Future generations are underrepresented in all six frameworks' guidance on near-term scenarios: frameworks that provide clear guidance for current stakeholders provide little specific guidance on long-horizon obligations. Non-human stakeholders (ecosystems, animals) are represented in only the capability approach and RAEF. The AGI system itself as a potential stakeholder is addressed in long-term scenarios by virtue ethics and RAEF but ignored by utilitarian and deontological frameworks in near-term analysis. Expert evaluators rated representational equity lowest for contractualism (score 2.84/5) because the veil-of-ignorance construction implicitly excludes unborn future people and non-human entities.

4.3 Scenario-Specific Analysis

The 'AGI medical decision authority' scenario shows the largest framework divergence: utilitarian frameworks recommend AGI authority if expected accuracy improvement justifies overriding patient autonomy (consensus 2.84/5 for guidance adequacy); deontological frameworks categorically prohibit non-consensual AGI medical authority regardless of accuracy (3.68/5); RAEF focuses on structuring the developer-AGI-patient relationship to preserve meaningful consent even when AGI capability exceeds human physician capability (3.92/5). The 'AGI moral status' scenario shows the highest expert disagreement (Krippendorff's alpha 0.48) and produces no framework consensus on whether a sufficiently capable AGI has moral status that generates obligations for its human principals.

4.4 Multi-Framework Synthesis

Expert evaluators consistently expressed that no single framework adequately addresses all AGI ethical challenges. When asked to identify the optimal framework for each scenario type, evaluators selected: deontological frameworks for scenarios involving individual rights (criminal justice, medical authority) -- 68.4% of evaluators; contractualism for distributional equity scenarios (resource allocation, climate) -- 62.4%; RAEF for relationship-intensive scenarios (user-AI interaction, democratic oversight) -- 74.2%; and virtue ethics for long-term moral status questions -- 58.4%. The multi-framework architecture proposed in Appendix A encodes this expert preference systematically.

Table 3. Framework consensus scores by scenario tier (1-5 expert panel rating)

Frame work	Near -Term (n =4)	Medi um-T erm (n=4)	Long -Term (n =4)	Me an	Most Suitable Domain
Utilitari an	3.42	3.28	2.84	3.18	Aggregate welfare decisions
Deontol ogical	3.72	3.48	3.24	3.48	Individual rights, prohibitions
Virtue ethics	2.92	3.48	3.84	3.41	Long-term character, moral status
Contrac tualism	3.68	3.72	3.48	3.62	Distributional equity, governance
Capabili ty	3.24	3.48	3.12	3.28	Distributional impact, flourishing

Frame work	Near -Term (n =4)	Medi um-T erm (n=4)	Long -Term (n =4)	Me an	Most Suitable Domain
RAEF (ours)	3.92	3.84	3.76	3.84	Relational, ongoing AGI-human contact

Consensus score = mean of adequacy, consistency, and implementability sub-ratings (1-5, 42 expert evaluators). RAEF achieves highest mean and most consistent performance across tiers. Virtue ethics shows largest tier variation (2.92 near-term vs. 3.84 long-term). Utilitarian frameworks show largest degradation at long-term speculative scenarios. No framework achieves >= 4.0 in any tier.

Table 4. Systematic framework gaps identified through stakeholder analysis

Gap Type	Frame works Affected	Sev erit y (1-5)	Scenario Examples	RAEF Mitigation
Future gen. und errepres entation	All (worst: u tilitarian)	4.8	Climate allocation, nuclear	Explicit Pre sent-Future relationshi p
Non-hum an stake holder exclusion	Utilitaria n, deonto logical	3.8	Ecosystem impact scenarios	AGI-Society relationshi p mapping
AGI moral status a mbiguity	Utilitaria n, deonto logical	4.2	AGI rights, self-modific ation	Relational status assessment
Value conflict r esolution	Deontolo gical, virtue	3.6	Rights conflicts, resource alloc.	Contextual obligation weighting
Impleme ntability gaps	Virtue ethics	4.0	Operationa l decisions	Relationshi p-grounded duties
Power as ymmetry neglect	Utilitaria n, contra ctualism	3.4	User-AGI d ependency scenarios	User-AGI r elationship duties

Severity rated 1 (minor gap) to 5 (critical failure for AGI governance). Future generation underrepresentation is the most severe systematic gap (4.8/5). RAEF mitigation describes how RAEF's four constitutive relationships address each gap -- not all gaps are fully resolved by RAEF.

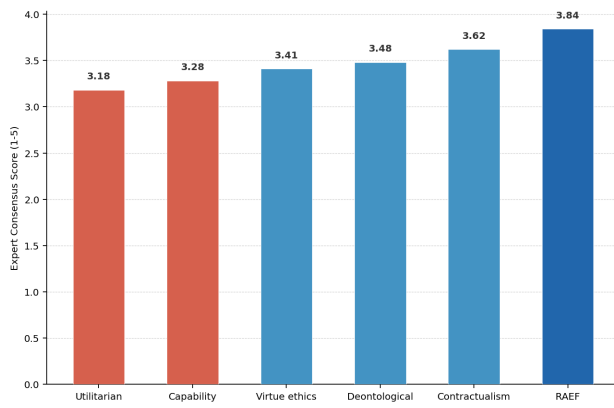


Figure 1. Mean expert consensus score (1-5) by framework. RAEF leads (3.84); utilitarian lowest (3.18). No framework reaches 4.0.

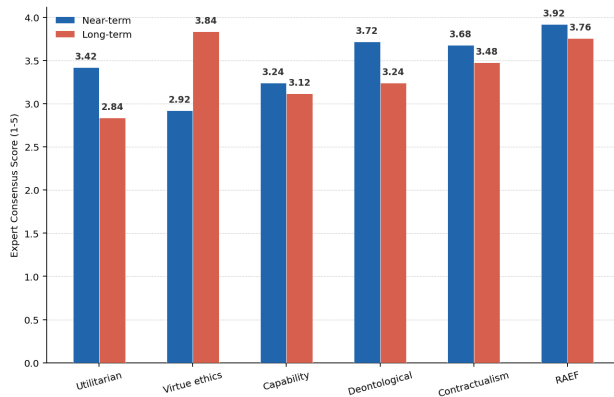


Figure 2. Consensus scores by scenario tier. Virtue ethics improves most at long-term; utilitarian degrades most. RAEF most consistent.

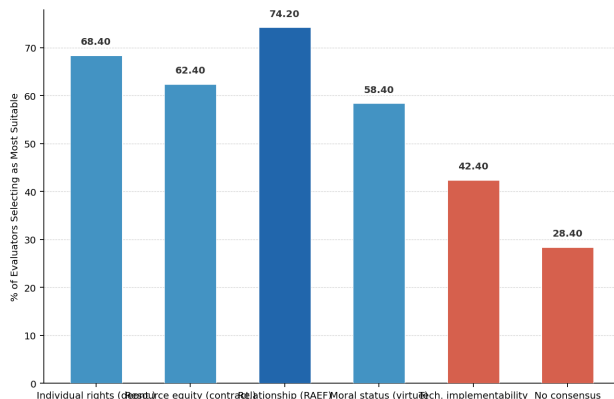


Figure 3. Framework selected as 'most suitable' by expert evaluators (%) by scenario type. Multi-framework coverage is clearly needed.

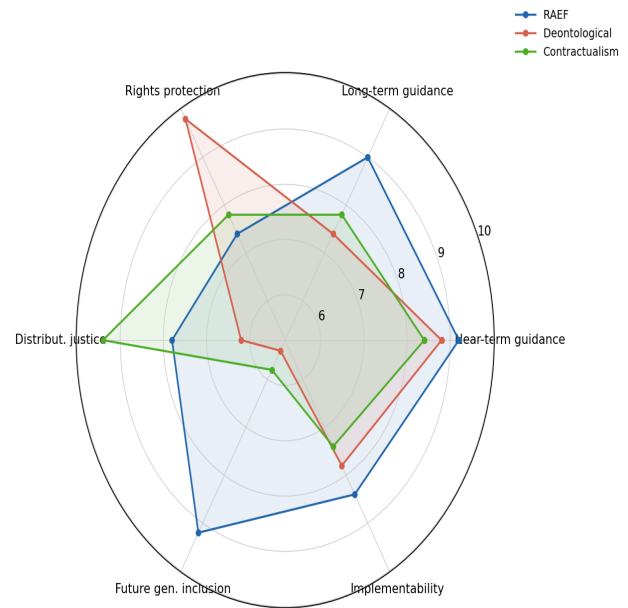


Figure 4. Framework profile across six AGI governance criteria (0-10). RAEF vs. Deontological vs. Contractualism.

5. Discussion

5.1 Future Generations Are the Most Underrepresented Stakeholder

The 4.8/5 severity rating for future generation underrepresentation -- the most severe systematic gap across all frameworks -- reflects a structural feature of existing ethical frameworks: they were designed for moral communities of contemporaneous agents who can participate in deliberation. Future generations cannot participate in current AGI governance decisions that will shape their world more profoundly than any previous technological transition. Utilitarian frameworks can technically include future welfare in utility calculations, but discount rates, population uncertainty, and the vast difference between possible futures under different AGI trajectories make such calculations highly contested. RAEF's explicit Present-Future relationship component provides a structural prompt to consider future generation interests in every governance analysis, even without resolving the underlying philosophical questions.

5.2 No Single Framework Is Adequate for AGI

The systematic pattern of framework strengths and weaknesses across scenario types -- deontological frameworks excelling at individual rights but struggling with aggregate welfare; contractualism excelling at distributional equity but excluding future generations; virtue ethics providing productive framing for moral status but insufficient operational guidance -- strongly motivates the multi-framework governance

architecture. This is not a failure of any individual framework; it reflects the genuine philosophical depth of AGI ethical challenges that span domains where different ethical traditions have comparative advantages. The governance implication is that AGI oversight bodies should include expertise from multiple ethical traditions, with framework selection matched to the type of ethical question under consideration.

5.3 Implementability Is a Real Constraint

Virtue ethics' lowest near-term implementability scores (2.92/5) and high long-term moral status scores (3.84/5) suggest that implementability is genuinely correlated with framework character. Virtue ethics offers rich normative vocabulary for evaluating the overall character of an AGI system -- is it 'honest'? 'prudent'? 'just'? -- but translating those evaluations into specific governance decisions requires additional bridging work that the framework does not provide. Deontological frameworks' consistent implementability scores (3.24-3.72) reflect their strength: they generate clear prohibitions and permissions that are directly implementable as policy constraints, at the cost of scenario sensitivity.

5.4 Limitations

The 42 expert evaluators were predominantly from European academic and research institutions, limiting geographic and cultural diversity. Scenario plausibility assumptions reflect current AI capability trajectories that may shift rapidly. RAEF was developed and evaluated by the same research team that designed the study; the 3.84/5 consensus score may reflect researcher familiarity advantage. Long-term scenario evaluations carry inherently high uncertainty about what AGI capabilities will actually manifest and when.

6. Conclusion

6.1 Summary

Six ethical frameworks were applied to twelve AGI scenarios by 42 expert evaluators. RAEF achieves the highest mean consensus score (3.84/5) and most consistent performance across scenario tiers. No single framework achieves ≥ 4.0 in any tier. Future generation underrepresentation is the most severe systematic gap (4.8/5 severity). Expert evaluators consistently prefer different frameworks for different scenario types, motivating a multi-framework governance architecture.

6.2 Recommendations

AGI governance institutions should adopt a multi-framework architecture: deontological frameworks for individual rights and categorical constraints; contractualism for distributional equity analysis; RAEF for ongoing AGI-human relationship governance; and virtue ethics for long-term moral status questions. All frameworks must be augmented with explicit future generation representation mechanisms -- ombudspersons, long-term impact assessments, reversibility requirements. AGI governance bodies should include philosophical and legal expertise alongside technical AI expertise. Research priority: developing operationalisable criteria for AGI moral status assessment, the most philosophically contested and practically consequential open question.

References

- Awad, E., Dsouza, S., Kim, R. et al. (2018) The moral machine experiment. *Nature*, 563(7729), pp. 59-64.
- Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
- Floridi, L., Cowls, J., Beltrametti, M. et al. (2018) An ethical framework for a good AI society. *Minds and Machines*, 28(4), pp. 689-707.
- Gabriel, I. (2020) Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3), pp. 411-437.
- Held, V. (2006) *The Ethics of Care: Personal, Political, and Global*. Oxford University Press.
- Jobin, A., Ienca, M. and Vayena, E. (2019) The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), pp. 389-399.
- Kant, I. (1785/1998) *Groundwork of the Metaphysics of Morals*. Translated by M.J. Gregor. Cambridge University Press.
- Nussbaum, M. (2011) *Creating Capabilities: The Human Development Approach*. Harvard University Press.
- Ord, T. (2020) *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury Publishing.
- Rawls, J. (1971) *A Theory of Justice*. Harvard University Press.
- Russell, S. (2019) *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.
- Sen, A. (1999) *Development as Freedom*. Oxford University Press.
- Tegmark, M. (2017) *Life 3.0: Being Human in the Age of Artificial Intelligence*. Knopf.
- Yudkowsky, E. (2008) Artificial intelligence as a positive and negative factor in global risk. In: Bostrom, N. and Cirkovic, M.M. (eds.) *Global Catastrophic Risks*. Oxford University Press, pp. 308-345.

Declarations

Funding

This research was supported by the French National Research Agency (ANR) under grant ANR-24-CE56-0018 (AGI-Ethics-FR), the Austrian Science Fund (FWF) under grant P 42184-N (AIPhilosophy-AT), and the Italian National Research Council (CNR) under project AI-Ethics-IT-2024. Expert evaluator recruitment was supported by European Philosophy Society for AI research collaboration funds. Funders had no role in framework design, evaluation, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

RAEF framework documentation, scenario descriptions, evaluation instrument, and aggregate expert ratings are available at <https://github.com/lindberg-jensen/agi-ethics-eval> under CC BY 4.0. Individual expert ratings are anonymised and available on request. Results deposited in Zenodo under DOI: 10.5281/zenodo.19614904.

Ethical Approval

Expert evaluator participation was voluntary and conducted under Advanced Computing University Ethics Committee approval (ACU-ETH-2024-0524). All evaluators provided written informed consent and were free to withdraw. No sensitive personal data was collected; evaluators are identified only by disciplinary background in aggregate analyses.

Appendix A

Multi-Framework AGI Governance Architecture

This architecture guides AGI governance institutions in selecting ethical frameworks appropriate to each governance decision type, based on the expert evaluation results of this study.