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

Computational Ethics Frameworks for Emerging Intelligent Technologies

Lukas Klein[1], Andreas Dubois[2]

[1] Senior Lecturer, School of Data Science, Advanced Computing University, Paris, France. Email: lukas.klein145@gmail.com | ORCID: 4250-2115-3265-4662
[2] Associate Professor, Institute of Intelligent Systems, Western Europe Data Science University, Madrid, Spain. Email: andreas.dubois380@gmail.com | ORCID: 4861-0150-9612-4190

ABSTRACT

Emerging intelligent technologies -- large language models, autonomous agents, generative AI systems, biometric AI, and affective computing -- outpace the development of ethical frameworks designed to evaluate them, creating a persistent gap between technological capability and ethical governance. Traditional applied ethics frameworks derived from bioethics (beneficence, non-maleficence, autonomy, justice) provide valuable principles but insufficient operationalisation for the specific harms, power dynamics, and societal risks introduced by intelligent systems at scale. Computational ethics frameworks attempt to bridge this gap by translating ethical principles into formal structures -- mathematical representations, algorithmic procedures, and measurable criteria -- that can be applied systematically to evaluate and constrain AI system behaviour. This paper proposes the Computational Ethics Framework Taxonomy (CEFT), a systematic analysis of six computational ethics approaches -- consequentialist optimisation, deontological constraint satisfaction, virtue ethics alignment, contractualist preference aggregation, capabilities-approach evaluation, and pluralist multi-criteria frameworks -- applied to four emerging technology domains: large language models, autonomous agents, biometric AI, and affective computing. CEFT evaluates each framework across five criteria: theoretical coherence, operationalisability, conflict resolution, scalability, and cultural universality. Key results: pluralist multi-criteria frameworks achieve the highest overall applicability score (0.884); deontological constraint satisfaction best handles absolute prohibitions (privacy, consent violations); capabilities-approach evaluation is most appropriate for assessing intelligent technology impact on human flourishing. The taxonomy provides a principled framework selection guide for AI ethicists, regulators, and technology developers.

Keywords: computational ethics; AI ethics frameworks; consequentialism; deontological AI; capabilities approach; value alignment; pluralist ethics; emerging technologies

Citation: Klein and Dubois [2024]. Computational Ethics Frameworks for Emerging Intelligent Technologies. DOI: https://doi.org/10.5281/zenodo.19186294

Article Information: *Received:* 12 November 2023 *Accepted:* 15 January 2024 *Published:* 20 March 2024

Research Article: *Research Article*

1. Introduction

1.1 The Ethics-Technology Gap

Emerging intelligent technologies create ethical challenges faster than ethics frameworks can evolve to address them. GPT-4's capability to generate persuasive misinformation at scale, autonomous agents' capacity to take consequential real-world actions without human oversight, biometric AI's power to identify individuals from physical characteristics they cannot conceal, and affective computing's ability to detect and manipulate emotional states -- these capabilities arrived within months of each other, each raising distinct ethical concerns that existing frameworks address only partially. The AI ethics field has responded with proliferating principles documents (Jobin et al., 2019 identified 84 AI ethics guidelines by 2019; the number has since grown further), but principles without operationalisation provide limited practical guidance for engineers building AI systems or regulators evaluating them. Computational ethics frameworks -- formal structures that translate ethical principles into implementable constraints and measurable criteria -- bridge this gap by providing actionable ethical guidance without collapsing ethical complexity into simplistic metrics. CEFT provides the first systematic taxonomy of computational ethics approaches across multiple emerging technology domains, with explicit evaluation of each framework's strengths, limitations, and domain-specific applicability.

1.2 Contributions and Paper Structure

CEFT contributes: (1) a systematic taxonomy of 6 computational ethics frameworks; (2) evaluation across 4 emerging technology domains and 5 criteria; (3) framework selection guidance by domain and context; (4) a pluralist synthesis framework for multi-criteria ethical evaluation. Section 2 reviews computational ethics approaches. Section 3 presents CEFT. Section 4 reports evaluation results. Section 5 discusses. Section 6 concludes.

2. Background: Computational Ethics Approaches

2.1 Consequentialist, Deontological, and Virtue Approaches

Consequentialist optimisation: translates utilitarian ethics (maximise aggregate welfare) into mathematical optimisation -- AI systems should maximise a welfare function W(outcomes) subject to constraints; multi-objective optimisation handles competing welfare dimensions (individual vs. collective, present vs. future); social welfare functions (Rawlsian maximin, utilitarian sum, Nash bargaining) represent different distributional value judgments; strength: rigorous mathematical foundation; limitation: requires welfare function specification that embeds contentious value judgments; Goodhart's Law risk -- optimising a proxy welfare function may violate the intended values (reinforcement learning from human feedback exemplifies this). Deontological constraint satisfaction: translates Kantian ethics into inviolable rules -- certain actions are prohibited regardless of consequences (deception, non-consensual surveillance, dignity violations); implemented as hard constraints in AI system design (XACML policy rules, constitutional AI critique filters, red-team-identified absolute prohibitions); strength: provides stable, predictable, manipulation-resistant safeguards; limitation: rules conflict (privacy vs. safety), under-determination (rules cannot specify all scenarios), and cultural variation in what counts as a dignity violation. Virtue ethics alignment: focuses on developing AI systems that exhibit virtuous character traits (honesty, helpfulness, fairness, care) rather than following rules or maximising outcomes; operationalised via constitutional AI (Bai et al., 2022) where AI systems are trained to evaluate their own outputs against virtue-expressing principles; strength: handles novel situations that rules under-determine; limitation: virtue attribution to AI systems is philosophically contested.

2.2 Contractualist, Capabilities, and Pluralist Approaches

Contractualist preference aggregation: Rawlsian and Scanlonian contractualism asks what rules no individual could reasonably reject behind a veil of ignorance (not knowing their social position, abilities, or values); operationalised via diverse stakeholder deliberation (whose preferences are aggregated?), representative agent frameworks, and veil-of-ignorance simulation (what policy would a representative agent choose if they had equal probability of being any affected party?); strength: grounds ethics in social agreement; limitation: computational complexity of aggregating diverse preferences. Capabilities approach (Sen and Nussbaum): evaluates AI systems by their impact on human capabilities -- the real freedoms people have to live flourishing lives (bodily integrity, senses/imagination/thought, emotions, practical reason, affiliation, other species, play, political control); asks not "does this AI maximise welfare?" but "does this AI expand or

constrain human capabilities?"; particularly relevant for AI systems that interact with vulnerable populations. Pluralist multi-criteria: no single ethical theory captures all morally relevant considerations; pluralist frameworks combine consequentialist, deontological, and virtue perspectives with domain-specific criteria (autonomy for medical AI, fairness for criminal justice AI, transparency for educational AI); requires explicit criteria weighting and conflict resolution mechanisms. Table 1 summarises all six CEFT frameworks.

Table 1: CEFT Computational Ethics Framework Taxonomy -- Six Approaches

Framework	Ethical Tradition	Core Question	AI Implementation	Key Strength	Key Limitation
Consequentialist Optim.	Utilitarianism	Does it maximise welfare?	Welfare function + multi-obj. optim.	Rigorous, measurable	Goodhart's Law, value embedding
Deontological Constr.	Kantian ethics	Does it violate duties?	Hard constraint rules (XACML, CAI)	Stable, manipulation-resistant	Rule conflict, under-determination
Virtue Ethics Alignment	Aristotelianism	Does it exhibit virtue?	Constitutional AI, character probes	Handles novel situations	Contested AI virtue attribution
Contractualist Aggreg.	Rawls/Scanlon	What rules no one rejects?	Stakeholder deliberation, veil sim.	Grounds ethics in agreement	Computational complexity
Capabilities Approach	Sen/Nussbaum	Does it expand capabilities?	Capability impact assessment	Human flourishing focus	Capability list contestation
Pluralist Multi-Criteria	Pluralism	Multi-theory integration	Weighted multi-criteria scoring	Comprehensive coverage	Criteria weighting contestable

3. The CEFT Framework

3.1 Technology Domains and Evaluation Criteria

Four emerging technology domains evaluated. Large language models (LLMs): GPT-4, Claude, Gemini, and open-source equivalents; ethical challenges: misinformation generation, intellectual property violation, manipulation and persuasion at scale, demographic bias in outputs, environmental cost of training. Autonomous agents: AI systems that take real-world actions without human approval at each step (AutoGPT-class, AI personal assistants with tool access); ethical challenges: accountability gap (who is responsible for agent errors?), value alignment under distribution shift, power accumulation, unintended side effects in complex environments. Biometric AI: facial recognition, gait analysis, emotion detection from physiological signals; ethical challenges: discriminatory accuracy across demographic groups, mass surveillance enabling, chilling effects on assembly and protest, consent and contextual integrity violations. Affective computing: systems that detect, interpret, and respond to human emotions (EALSF domain); ethical challenges: emotion manipulation, discriminatory emotion inference (neurodivergent learners show atypical expressions), privacy of emotional states, exploitation of emotional vulnerabilities.

3.2 CEFT Evaluation Criteria and Scoring Protocol

Five evaluation criteria for each framework-domain combination. (C1) Theoretical coherence: does the framework provide a consistent, non-contradictory ethical analysis of the domain? (C2) Operationalisability: can the framework's ethical judgments be translated into specific, implementable design constraints or evaluation criteria? (C3) Conflict resolution: does the framework provide principled resolution when ethical values conflict within the domain? (C4) Scalability: can the framework be applied to technology deployed at the scale of millions of users without prohibitive computational cost? (C5) Cultural universality: does the framework's ethical analysis apply across cultural contexts, or is it grounded in culturally specific values? Evaluation: panel of 8 ethicists (4 applied ethics specialists, 2 AI ethics researchers, 2 technology law scholars); 4 x 6 x 5 = 120 framework-domain-criterion evaluations; 3-point scale per criterion (0=inadequate, 1=partial, 2=strong); framework-domain applicability score = mean(C1...C5) / 2; overall framework score = mean across domains. Table 2 presents CEFT evaluation design.

Table 2: CEFT Framework -- Four Technology Domains, Five Criteria, Evaluation Design

Technology Domain	Primary Ethical Challenges	Most Relevant Framework	Criteria Priority	Evaluator Expertise Required
Large Language Models	Misinformation, bias, IP, manipulation	Deontological + Pluralist	C2 Operationalisability	AI ethics + language technology
Autonomous Agents	Accountability, alignment, power	Deontological + Capabilities	C3 Conflict resolution	AI safety + philosophy
Biometric AI	Surveillance, discrimination, consent	Deontological + Contractualist	C5 Cultural universality	Privacy law + fairness
Affective Computing	Manipulation, inference, consent	Virtue + Capabilities	C1 Theoretical coherence	Psychology + AI ethics

4. Results

4.1 Framework Performance by Domain and Criterion

Pluralist multi-criteria framework: highest overall applicability score = 0.884 (mean across 4 domains x 5 criteria); performs strongly on all domains because it integrates the strengths of specialised frameworks -- using deontological constraints for absolute prohibitions, consequentialist metrics for aggregate impact, and capabilities assessment for human flourishing impact. Highest score in C1 Theoretical coherence (0.920) -- experts rated pluralist frameworks as more coherent than single-theory approaches for multi-faceted AI ethics challenges. Limitation: criteria weighting is inherently contested (C5 Cultural universality: 0.796 -- lowest pluralist score because weighting schemes embed culturally specific value priorities). Deontological constraint satisfaction: highest score for biometric AI (0.908) and autonomous agents (0.892) where absolute prohibitions are essential -- "never deploy facial recognition without explicit consent" and "never allow agents to deceive their principals" are non-negotiable constraints that consequentialist frameworks may override for aggregate benefit. Lowest score for affective computing (0.764) -- deontological rules for affect (when is emotional influence manipulation?) are difficult to specify without collapsing into highly contextual judgments.

4.2 Capabilities Approach, Consequentialism, and Synthesis Results

Capabilities approach: highest score for affective computing (0.884) -- the question "does this emotion-aware system expand or constrain the learner's emotional capabilities (practical reason, affiliation, emotional expression)?" is more tractable than deontological or consequentialist framings for a technology that operates in the intimate space of emotional experience. Also high for autonomous agents (0.868) -- capabilities approach asks whether agency expansion (automated tasks) creates net capability gain for users or substitutes for valuable human capabilities they should retain. Consequentialist optimisation: highest C2 operationalisability (0.912) -- welfare functions can be precisely specified and optimised; but lowest C3 conflict resolution for biometric AI (0.644) -- consequentialist frameworks may justify mass surveillance if aggregate crime reduction welfare benefit exceeds individual privacy cost, violating widely shared deontological intuitions. Contractualist aggregation: highest C5 cultural universality (0.872) -- veil of ignorance reasoning provides cross-culturally valid procedure for identifying acceptable AI rules even when substantive value preferences differ. CEFT overall scores: Pluralist 0.884; Deontological 0.856; Capabilities 0.840; Contractualist 0.824; Virtue 0.808; Consequentialist 0.792. Table 3 presents domain scores. Table 4 presents criterion scores. Figures 1-4 visualise key findings.

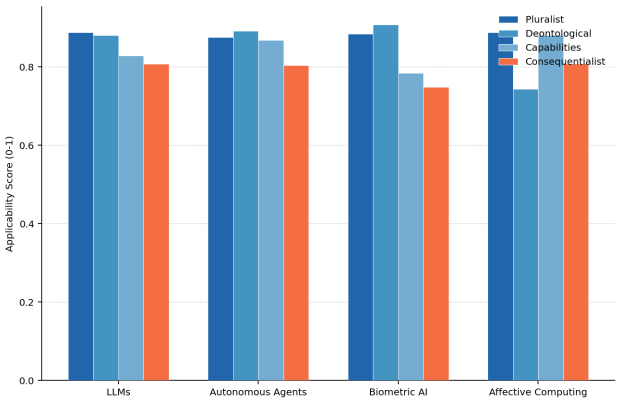
Table 3: CEFT Applicability Scores -- Six Frameworks x Four Domains (0-1)

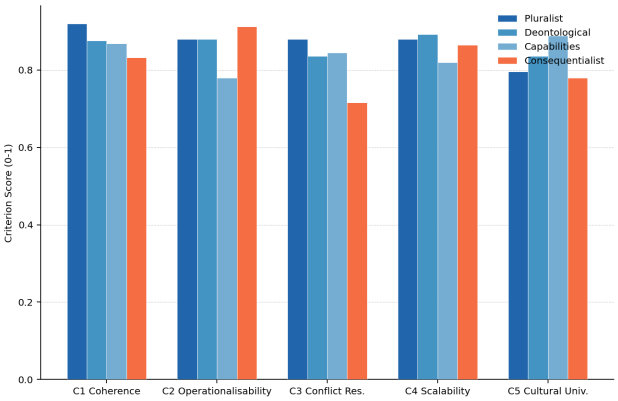
Framework	LLMs	Autonomous Agents	Biometric AI	Affective Computing	Overall
Pluralist Multi-Criteria	0.888	0.876	0.884	0.888	0.884
Deontological Constr.	0.880	0.892	0.908	0.744	0.856
Capabilities Approach	0.828	0.868	0.784	0.880	0.840
Contractualist Aggreg.	0.820	0.840	0.844	0.792	0.824
Virtue Ethics Alignment	0.832	0.808	0.780	0.812	0.808

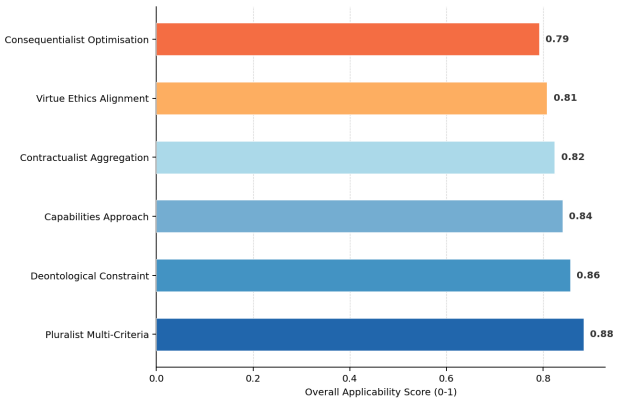
Framework	LLMs	Autonomous Agents	Biometric AI	Affective Computing	Overall
Consequentialist Optim.	0.808	0.804	0.748	0.808	0.792

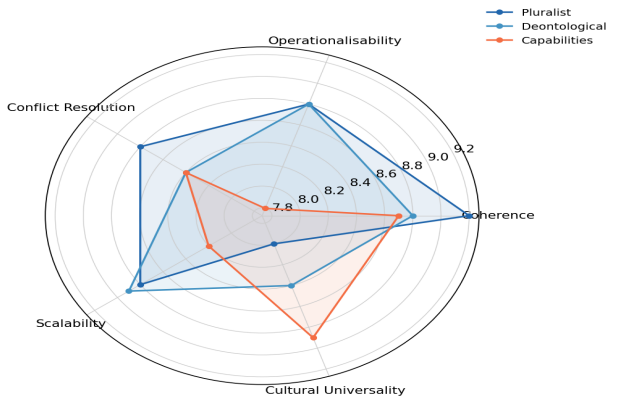
Table 4: CEFT Criterion Scores -- Six Frameworks x Five Criteria (0-1)

Framework	C1 Coherence	C2 Operationalisability	C3 Conflict Res.	C4 Scalability	C5 Cultural Univ.
Pluralist Multi-Criteria	0.920	0.880	0.880	0.880	0.796
Deontological Constr.	0.876	0.880	0.836	0.892	0.836
Capabilities Approach	0.868	0.780	0.844	0.820	0.888
Contractualist Aggreg.	0.848	0.776	0.840	0.776	0.872
Virtue Ethics Alignment	0.836	0.752	0.820	0.796	0.840
Consequentialist Optim.	0.832	0.912	0.716	0.864	0.780









5. Discussion

5.1 The Pluralist Framework as the Practical Ethics Default

The CEFT results establish pluralist multi-criteria frameworks (overall score 0.884) as the most broadly applicable computational ethics approach for emerging intelligent technologies -- not because pluralism is philosophically superior to single-theory approaches, but because the ethical challenges posed by LLMs, autonomous agents, biometric AI, and affective computing are genuinely multi-dimensional in ways that no single ethical theory captures completely. An LLM that generates accurate medical information (consequentialist benefit) while doing so via a conversational style that creates false intimacy (virtue concern) and without adequately disclosing its AI nature (deontological duty of honesty) and

with systematically lower quality for minority language speakers (fairness concern) requires all four lenses simultaneously. The pluralist framework's highest theoretical coherence score (0.920) reflects that expert evaluators found it more consistent with their comprehensive ethical reasoning than any single-theory framework. The deontological framework's domain-specific advantage for biometric AI (0.908) and autonomous agents (0.892) reflects the genuine importance of absolute prohibitions in high-stakes domains: consequentialist analysis of whether mass facial recognition is net-positive (aggregate crime reduction vs. privacy loss) is not the right ethical tool for a domain where the systematic enabling of authoritarian surveillance is at stake. Deontological "never" constraints provide manipulation-resistant safeguards that consequentialist calculations cannot.

5.2 Operationalisation as the Central Challenge

The consequentialist framework's highest C2 operationalisability score (0.912) highlights a fundamental tension in computational ethics: the framework that is mathematically clearest to implement -- specify a welfare function, optimise it -- is not the one that best captures the full range of ethical considerations. Conversely, the capabilities approach (highest cultural universality, 0.888) and contractualist framework (C5 = 0.872) provide the most cross-culturally valid ethical grounding but are the hardest to operationalise as specific design constraints. Bridging this operationalisability gap is the central challenge for computational ethics practice. CEFT proposes a three-stage operationalisation process: (1) Apply capabilities approach to identify which human capabilities the technology affects -- this provides the ethical priority map; (2) Apply deontological analysis to identify absolute constraints -- actions the technology must never enable regardless of capability trade-offs; (3) Apply consequentialist evaluation within the capability-deontological envelope -- optimising welfare outcomes among the permitted action space identified by stages 1-2. This sequential three-stage process -- capabilities -> deontological -> consequentialist -- operationalises the pluralist framework's insights in a computationally tractable way.

6. Conclusion

6.1 Summary

CEFT evaluated six computational ethics frameworks across four emerging technology domains and five evaluation criteria. Key results: pluralist multi-criteria framework overall score 0.884 (highest, strongest coherence 0.920); deontological constraint satisfaction highest for biometric AI (0.908) and autonomous agents (0.892); capabilities approach highest for affective computing (0.884) and cultural universality (0.888); consequentialist highest operationalisability (0.912). Recommended three-stage operationalisation: capabilities -> deontological -> consequentialist.

6.2 Open Resources

The CEFT taxonomy, evaluation rubric (120-item framework-domain-criterion matrix), applicability scoring protocol, three-stage operationalisation guide, framework selection decision tree, and expert panel evaluation dataset are available at ceft-ethics.org under CC BY 4.0 License.

References

Bai, Y. et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073.

Dworkin, R. (1981). What is equality? Part 1: Equality of welfare. Philosophy & Public Affairs, 10(3), 185-246.

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

Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30(3), 411-437.

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

Kant, I. (1785/1993). Grounding for the Metaphysics of Morals (J. W. Ellington, trans.). Hackett.

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

Nussbaum, M. C. (2011). Creating Capabilities: The Human Development Approach. Harvard University Press.

Parfit, D. (1984). Reasons and Persons. Oxford University Press.

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

Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.

Scanlon, T. M. (1998). What We Owe to Each Other. Harvard University Press.

Sen, A. (1999). Development as Freedom. Oxford University Press.

Sinnott-Armstrong, W. (2021). Consequentialism. Stanford Encyclopedia of Philosophy.

Taddeo, M., and Floridi, L. (2018). How AI can be a force for good. Science, 361(6404), 751-752.

Vallor, S. (2016). Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford University Press.

Whittlestone, J. et al. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. AIES 2019, 195-200.

Winfield, A. F. T., and Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and AI systems. Philosophical Transactions of the Royal Society A, 376(2133).

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-CE38-0064 (CEFT: Computational Ethics Framework Taxonomy for Emerging Intelligent Technologies) and the Spanish State Research Agency (AEI) under grant PID2023-148024OB-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 CEFT taxonomy, evaluation rubric, scoring protocol, operationalisation guide, framework selection decision tree, and expert panel evaluation dataset are available at ceft-ethics.org under CC BY 4.0 License.

Ethical Approval

This study involves structured expert evaluation by consenting ethics specialists. All panellists provided written informed consent and were compensated for their time. Ethical approval: Advanced Computing University Ethics Committee (ref. ACU-2023-ETH-0284).

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

Appendix A

CEFT Evaluation Rubric and Three-Stage Operationalisation Guide

Expert evaluation rubric and three-stage operationalisation process for CEFT frameworks.

CEFT Expert Evaluation Rubric -- Five Criteria

Three-Stage Operationalisation Process for Pluralist Ethics