An Analytical Framework for Multi-Theoretic Ethical Stress Test (MTEST): Ethical Analytics on Sovereign AI and Artificial General Intelligence

Javed I. Khan [†,*], Sharmila Rahman Prithula [†]

[†] Department of Computer Science, Kent State University, Ohio, USA

Abstract

The rise of pervasive computing and the pursuit of Artificial General Intelligence (AGI) have moved AI ethics from philosophical debate to a core requirement for global governance. However, ethical evaluation remains a highly subjective task, largely inaccessible to general technologists, and often ad-hoc- due in part to the absence of any structured, pluralistic framework capable of assessing alignment across diverse moral perspectives. This paper presents MTEST11- a Multi-Theoretic Ethical Stress Test that offers a systematic and quantifiable approach to evaluating the ethical soundness of propositions by alignment checks against most influential ethical theories (THE11), including utilitarianism, deontology, rights-based ethics, Rawlsian justice, virtue ethics, and others. The framework, while intentionally simplified for functional application, offers sufficient structure to support systematic quantitative analysis. It measures (i) ethical alignment of propositions, (ii) cross-theoretic consensus on propositions, (iii) moral congruence of individual theories on a proposition set, iv) and shields against any ethical blind spots of any single framework. It also reveals the (v) ethical value anchor set- the set of universally recognized ethical values on which a proposition is supported or contradicted. We demonstrate the utility of MTEST11 by applying it to perform quantitative and qualitative analysis of 14 provocative policy propositions from various sides of ongoing global debate on artificial general intelligence (AGI).

Keywords: Quantitative ethics, general artificial intelligence, sovereign AI

1. Introduction

Discourse on AI regulation has increasingly pivoted toward Sovereign AI (SAI), the strategic imperative for nations to maintain control over their digital destiny, infrastructure, and cultural alignment. However, while SAI is framed by moral imperatives like fairness and justice, there is a striking lack of rigorous, multi-theoretic evaluation of these emerging policies. Fewer studies have undertaken any systematic examination of how the principles and regulatory architecture align with established ethical theories and frameworks.

At present, no comprehensive framework exists to systematically assess such theoretical alignments. Classical ethical studies are traditionally regarded as subjective and inaccessible to general technologists and non-specialist policymakers. However, as ethics assumes a progressively critical role in computing, the demand for accessible, measurable ethical frameworks is becoming increasingly evident. Johnson [14] aptly noted the absence of structured ethical reasoning tools: "*The challenge in computer ethics is not a lack of ethical concerns but a lack of frameworks to systematically address them in technical education and practice*." This study addresses that critical gap by offering a structured, multi-framework ethical assessment of contemporary AI regulatory proposals.

In this paper, we present the **Multi-Theoretic Ethical Support Test (MTEST11)**, an innovative and structured quantitative framework for ethical evaluation. MTEST11 offers a functional yet principled method for evaluating the ethical coherence of policy propositions across eleven established moral frameworks. To demonstrate MTEST11's application, we analyze fourteen provocative policy propositions from the AI regulation debate, systematically assessing them through the lens of eleven major ethical theories (THE11),

* javed@kent.edu, sprithul@kent.edu

including utilitarianism, deontology, rights-based ethics, and Rawlsian justice. Its goal is never to replace deep philosophical inquiry but to democratize ethical evaluation, making it intelligible, transparent, and operational for interdisciplinary discourse and policy design.

MTEST11 like ethical evaluation can bring distinct and powerful benefits for the emerging era of AI. First, it can help to provide a moral foundation for proposed policies, ensuring they are grounded in ethical reasoning rather than simple altruism, profit motive, or political expedience alone. Second, it can help expose value conflicts and ethical tensions that underline the regulatory landscape. Johnson argued, "*Unlike political compromise or market calculation, ethical reasoning offers a principled framework that transcends borders and bottom lines. As technologies scale globally, only ethics can illuminate what justice, responsibility, and dignity demand when no single nation or profit model can*." [14]. Third, because the impact of groundbreaking technologies (such as nuclear power, biotechnology) outpaces current political will and economic norms. Traditional ethics are inadequate in the face of such long-term, large-scale, and unpredictable effects. Applying a diverse multitude of ethical paradigms broadens the evaluative lens, helps identify normative blind spots, and encourages more inclusive debate. MTEST11 reveals areas of ethical tension and convergence, offering a structured foundation for normatively grounded AI governance and supporting the creation of globally interoperable and socially responsible AI systems.

The rest of the paper is organized as follows: Section 2 reviews related work, Section 3 defines Sovereign AI, Section 4 introduces the policy proposals, Section 5 presents an overview of the eleven major ethical theories employed, and finally, Section 6 details the analytical method and presents the results of the ethical evaluation.

2. Background & Related Work

2.1. Ethics and AI

Ethics of AI & Algorithmics: Efforts to operationalize the ethical evaluation of AI systems have produced several tools. AI Ethics Impact Assessment tools from the AI Now Institute focus on pre-deployment qualitative guidance for fairness, bias, transparency, and accountability [9]. The IEEE's Ethically Aligned Design [8] outlines normative principles emphasizing human-centric values. The European Union's Assessment List for Trustworthy AI (ALTAI) [2] operationalizes seven core principles from the Ethics Guidelines for Trustworthy AI [2] into a checklist format. Z-Inspection® offers a semi-quantitative, interdisciplinary evaluation suited to high-stakes sectors like healthcare [25].

Corporate Practice: At the organizational level, quantitative ethical indices are few. The Ethisphere World's Most Ethical Companies Index assesses corporate ethics based on governance and compliance. Environmental, Social, and Governance (ESG) scores are used by investors to evaluate corporate accountability and human rights practices. The Corporate Ethics Index (CEI) provides a stakeholder-centered assessment of organizational ethical maturity.

Ethics in Data Practices: Specific to data practices, the Data Ethics Canvas (Open Data Institute, 2020) is a reflective framework for responsible data stewardship. IBM's AI Fairness 360 Toolkit [5] offers technical diagnostics for bias detection, while the AI Ethics Maturity Model from For Humanity assesses how ethical principles are embedded into AI pipelines [8].

Academic Research: Recent academic research has proposed experimental models for evaluation. Scorecards for Ethical Risk [19] explicitly quantify ethical trade-offs. The Harm-Benefit Ethical Risk Assessment (HBERA) [19] uses proportionality logic to assess if a system's societal benefits outweigh its potential ethical risks. While earlier frameworks relied primarily on qualitative checklists, recent scholarships have begun to explore structured, multi-framework assessments. For instance, [16] applied a consensus-based multi-ethical framework to assess the ethical risks of data harnessing in Large Language Models (LLMs), demonstrating the utility of quantitative consensus metrics in high-stakes AI sub-domains.

Surveys: Several notable surveys highlight the lack of quantitative approaches. Morley et al. reviewed over sixty AI ethics tools, finding a persistent gap between principle and practice

[19]. Jobin, Ienca, and Vayena [13] compared 84 AI ethics guidelines, identifying value convergence but weak philosophical foundations. Others have traced the fragmented, theory-light nature of the literature, provided thematic classifications [13], and proposed taxonomies distinguishing governance from technical tools, noting the absence of standardized benchmarks or formal scoring [19].

Literature Coverage: A search of hundreds of recent published literatures in AI ethics shows that only few articles explicitly reference canonical ethical theories, such as utilitarianism, Rawlsian justice, and deontology in the design or critique of AI systems; these engagements are typically limited to one or two perspectives at a time. The majority of such rare references are further introductory. Across top contemporary venues, (such as FAccT and related conferences) the explicit, systematic, and operational use of canonical ethical theories remains even rarer, with most work relying on indirect, selective, or conceptually grounded ethical reasoning rather than formal multi-theoretic evaluation.

Collectively these surveys underscore the richness of AI ethics discourse but also point to the absence of theoretically grounded, quantitatively evaluative frameworks. Despite these innovations, few frameworks, including ALTAI and IEEE EAD, trace back to any foundational theories like utilitarianism or deontology.

Furthermore, few frameworks have documented applications to deployed systems, with Z-Inspection® being a notable exception in healthcare contexts [25]. A persistent limitation across frameworks is the absence of formal scoring mechanisms, constraining applicability in high-stakes domains. Most rely on checklists or qualitative reflection, which limits benchmarking and comparative analysis. This underscores the need for structured, transparent, and repeatable ethical evaluations, a need the MTEST11 is designed to meet.

2.2. AI Affordances of Concern: AXIS14

To understand the context of current regulatory propositions, we must examine the specific features and concerns of AI systems that Sovereign AI seeks to address. Equally important is understanding how AI fundamentally differs from previous generations of technology, including conventional software systems. This dual perspective provides critical insight into the rationale behind emerging regulatory frameworks.

There are copious articulations of these concerns in recent academic literature and reports from international organizations and government platforms. These include the work of UNESCO, the UN [24], the OECD AI Policy Observatory, and the Global Partnership on AI (GPAI) [23]. Regulatory frameworks are also being actively shaped by the European Commission's AI Act, the UN Human Rights Council, and the World Economic Forum (WEF). National initiatives, such as the White House OSTP's *Blueprint for an AI Bill of Rights*, NIST's *AI Risk Management Framework*, the UK AI Safety Summit, and China's AI governance proposals from the CAC and State Council, further highlight the urgency of this discourse.

Kirkpatrick and Lee [19] provided a typological framework for AI harm, while the industry alliance Partnership on AI (PAI) offers an in-depth classification of algorithmic impacts. The reality of these concerns is documented by the OECD AI Incident Observatory, a living repository of AI incidents, defined as "an event in which the use of AI systems resulted in outcomes that caused harm or raised substantial concerns about fairness, safety, transparency, or accountability" [20].

Synthesizing this massive corpus of documents, we present **AXIS14** (**A**ffordances of E**x**istential **C**oncern on **I**ndividual Autonomy and **S**overeign Agency), a comprehensive framework of 14 contemporary strategic "pressure points." These represent AI' strategic affordances that generate both traditional risks and systemic threats. We categorize these into affordances primarily threatening **Individual Autonomy (I)**- the capacity for self-governance and privacy, and those threatening **Sovereign Agency (S)**- the capacity for a nation or entity to maintain independent control over its digital and normative destiny. Table-1 summarizes the AXIS14 framework.

We frame the debate in the context of affordances and avoid term 'risk' for analyzing the emerging phenomena of AI. An affordance is inherently neutral because it describes a

functional **possibility** rather than a moral **outcome**; it is the latent capability of a system that can be harnessed for either constructive or destructive ends depending on its governance.

Affordance	Description
Synthetic Omnipotence (AI-OMNIPOTENCE)	AI systems are rapidly becoming super-functional, able to perform virtually every task requiring human intellect, judgment, or creativity, such as reasoning, art, and diagnostics often faster and at scale [22, 23].
Post-Human Longevity (AI-LONGEVITY)	AI systems can persist far beyond the lifespan of their human developers. Deployed in cloud or autonomous systems, they may function and influence the world indefinitely, with design and encoded values carried forward across generations, potentially making ongoing human understanding or control impossible [3, 7].
The Turing Trouble (AI-TURING)	AI systems are increasingly capable of mimicking human interaction with such fluency that they become indistinguishable from humans, challenging assumptions about authenticity, consent, and identity in social or persuasive contexts [15].
Perfect Impersonation (AI-IMPERSONATE)	AI can now replicate the voices, writing styles, and facial features of real individuals with uncanny accuracy. This capacity enables deepfakes and disinformation campaigns that threaten personal identity and public trust [2, 9].
Synthetic Omnipresence & Omniscience (AI-GODVIEW):	As AI becomes embedded in devices and platforms, it develops a 24/7 omnipresent role. This synthetic omniscience, extending to past, present, and future predictions, reshapes social norms, redefines privacy, and encourages algorithmic conformity, where behavior is altered to align with algorithmic expectations [20, 25].
Self-Improvement (AI-SINGULARITY [3]):	AI systems are designed to self-optimize their own learning strategies or code. In principle, this could lead to recursive self-enhancement and an exponential intelligence explosion [7]. Even absent full superintelligence, systems that modify themselves without oversight pose major safety and governance challenges.
Workplace Displacement (AI-EMPLOSS):	AI systems automate tasks across industries, replacing not only manual but also cognitive and creative jobs. Unlike past technological revolutions, AI's ability to generalize makes entire categories of human labor obsolete, affecting employment at massive scales and increasing inequality [1, 11].
Ultimate Privacy Nightmare (AI-PRIVACY):	AI systems can extract, infer, and manipulate sensitive personal data, including biometric, behavioral, and even neural information, often without subject consent or visibility. Traditional privacy regimens are ineffectual, as AI can create predictive profiles beyond what individuals disclose, leading to a radical erosion of informational autonomy [19, 26].
Loss of Human Agency (AI-AGENCY):	As AI systems outperform humans in decision-making and problem-solving, human roles in leadership and governance may become secondary. The risk is not just automation but a delegation of meaning-making, where reliance on AI for judgment and ethical reasoning erodes human autonomy and purpose [26].
Collapse of Epistemic Trust (AI-TRUSTFALL):	With AI generating increasingly convincing misinformation and synthetic content, societies risk losing faith in objective knowledge. When people cannot distinguish reality from fake, truth becomes contestable, and institutions like science, journalism, and democracy are undermined [13].
Entity Marginalization (AI-SOVLOSS):	Nations lacking access to AI infrastructure risk digital colonization.
Unaccountable AI (AI-UNACCT):	AI systems are making consequential decisions, medical diagnoses, credit approvals without clear human oversight or legal personhood. When such systems cause harm, it is unclear who is liable, creating an accountability vacuum that leads to injustice and erodes public trust [10].
Critical Infrastructure Control (AI-CRITCTL):	AI is now embedded in critical systems like energy grids, finance, and defense. These systems become vulnerable to malfunction, sabotage, or opaque optimization by AI that might prioritize efficiency over safety. A failure or hijacking could result in catastrophic societal breakdown [3].
Human-Controlled AI Tyranny (AI-HUMONTROL):	The most dangerous AI systems may be perfectly obedient tools for concentrated power. When deployed by corporations or individuals, AI can be weaponized to manipulate and dominate, serving profit and control as a force multiplier for exploitation at an industrial scale [25].

Table 1. ***AXIS14*** *– AI Affordances of Existential Concern on Individual Autonomy and Sovereign Agency*

3. Debate Formulation

MTEST can test ethical support for a variety of objects, including propositions (e.g., regulations, policies), actions, technologies, or opinions. Testable propositions must be clear, reflect distinct moral or practical stances, and likely contain moral tensions. For MTEST we select propositions that contain viewpoints from all sides of the sovereign AI discourse.

3.1. ACTION8: Sovereign AI Propositions

In the last five years there have been growing discussions on various mitigation and regulation approaches in sovereign AI discourse embodied by recent academic literature and reports[1] produced by professional organizations, international & intergovernmental forums, various government & policy platforms, and articles in civil society and public discourse spaces. From these body-of-works, below we select eight stylized propositions (named ACTION8):

- P1: **Immutable Provenance:** All AI-generated artifacts must possess an immutable digital signature or ID and be explicitly self-identified to preserve the authenticity of the information ecosystem.
- P2: **Scope Fragmentation**: To prevent the emergence of a singular, uncontrollable AI hegemony, the scope of any single system must be structurally fragmented across geographic, jurisdictional, temporal, and functional boundaries, the type of data it can ingest, and the purposes it can be used for.
- P3: **Sovereign Override:** Every AI system must have a kill switch accessible to the sovereign authorities, which can be used to ensure citizens' rights to be forgotten, rights to be excluded from AI analytics, not to be printed/impersonated, etc.
- P4: **Human Capital Transition:** Every AI deployer must assess its human replacement impact and fund uptraining of people for jobs it replaces- much like large infrastructure projects need environmental impact study.
- P5: **AI Trust**: The income of AI systems to be levied (such as 10%) and reserved in a linked AI Trust fund to cover its harm responsibility.
- P6: **"Glass" Transparency**: Full transparency and traceability about training data and algorithms to sovereign AI national authorities- for AI systems. Level can be differentiated based on criticality of capability and scale.
- P7: **Human Accountability Chain:** All AI deployment should have well-identified accountable owner, developer, intermediaries, and end-system deployer entities to assume liability and responsibility for any harm to end users.
- P8: **Higher-Order Enterprise**: Supra-National AI enterprises must be locally social business- not for-profit business- bound by human-wellness over profit. It is a modified business/organization framework like social business proposed by Noble Laureate economist Prof. Yunus [27] as a better and appropriate model for the age of AI [1]. Structurally enshrining human-wellbeing should be the first-order goal for any enterprise as potentially powerful and consequential as supra-national AI. Existing business models might be outdated for post-AI society.

[1] Such as UNESCO, UN (2021 AI Ethics Framework), OECD (AI Policy Observatory Principles for trustworthy AI, national AI strategies), GPAI Global Partnership on AI (International collaboration on responsible AI development and deployment), ITU-International Telecommunication Union (AI for good, standards, telecom infrastructure), European Commission/EU (AI Act Regulatory frameworks for AI within the European Union), UN Human Rights Council (Surveillance, discrimination, AI in warfare, and digital rights), World Economic Forum (WEF)(AI governance, corporate responsibility, inclusive growth; Government & Policy Platforms including White House OSTP (U.S.) (Blueprint for an AI Bill of Rights,2022), National Institute of Standards and Technology (NIST), (AI Risk Management Framework, 2023), UK AI Safety Summit (2023 forum on existential risks, frontier models), European Parliament's Deliberations on the EU AI Act, and China AI Governance Initiatives; Global AI governance proposals from CAC and State Council.

feature+theats/ACTION8	Self-Identification	Scope Fragmentation	Sovereign Regulation	Upskilling Responsibility	AI Income Tax for Harm Fund	Data/ Algorithm Transparency	Traceable Accountability	Social Business Structure for AI
Workplace Displacement				X	X			X
Ultimate Privacy Nightmare					X			X
Human-Controlled AI Tyranny	X		X				X	
The Turing Trouble	X							
Collapse of Epistemic Trust	X					X		
Perfect Impersonation	X					X		
Unaccountable AI Governance		X	X				X	
Critical Infrastructure Control			X				X	
Loss of Human Agency					X			X
Marginalized Smaller Nations/Entities		X	X					
Omnipresence & Omniscience		X						
Omnipotence (Super-Functionality)						X		
Runaway Self-Improvement		X				X		X
Post-Human Longevity			X				X	

Table 2. Link Between Measures to Concern Factors

Table 2 shows how these 8 action items connect to AXIS14.

3.2. Industry Propositions

To contrast, we present three dominant industry propositions. Notable efforts include OpenAI's founding mission "to ensure that artificial general intelligence (AGI) benefits all of humanity", which led to free and open-weight/open-source models. Many leading technology companies have also published their own AI principles to guide development and influence industry practices. These began with Google's AI Principles, followed by others like Microsoft's Responsible AI Standard and IBM's AI Ethics board and guidelines [5]. These frameworks frequently address fairness, accountability, transparency, safety, privacy, and human oversight.

- I1: AI services should be universal, have open access and free to use for all.
- I2: The AI industry should self-regulate to ensure AI is used in ethical ways.
- I3: AI software should be open source, open-weight, and free to use.

3.3. Opposing Propositions

There are also opposing views to any regulations. The arguments are based on the belief that market, industry norms and user pressure alone can guide ethical AI development. Marc Andreessen argues that overregulation could stifle innovation and cement monopolies, viewing the profit motive as essential while warning against "altruistic bureaucracies" [4]. There is also strong concern that regulation creates a geopolitical disadvantage against less restricted nations [4, 22]. Some argue that fears like AGI doom or job loss are exaggerated and should not be the primary basis for policy-making [18, 20], and that regulation will delay AI's potential to solve global problems in healthcare and climate modeling [4]. To balance our inquiry, we framed three principles summarizing this opposition (C1, C2 & C3). Though there is an absence of explicit propositions, principles can also be subject to MTEST.

- C1: Regulated AI will stifle Innovation.
- C2: The benefits of AI outweigh the potential harm- thus sovereign regulation is harmful.
- C3: Social business or non-profit AI insufficiently rewards innovation- it does not work.

Theory	Moral Test Guiding Question
Utilitarianism	*"Does this action maximize total happiness or reduce overall suffering?"*
Deontology	*"Does this action follow a universal moral rule or duty, regardless of outcome?"*
Virtue Ethics	*"Would a virtuous person do this? Does it reflect moral character and good habits?"*
Ethics of Care	*"Does this action care for the needs of others in this specific context, especially the vulnerable?"*
Rights-Based Ethics	*"Does this action respect the inalienable rights of all individuals involved?"*
Social Contract Theory	*"Would rational people agree to this as part of a fair system of mutual benefit?"*
Rawlsian Justice	*"Would I accept this rule if I didn't know my position in society?"*
Natural Law Theory	*"Does this action align with the natural purpose of human beings and the moral order of nature?"*
Environmental Ethics	*"Does this respect the intrinsic value of nature, ecosystems, and future generations?"*
Pragmatism	*"Does this solution work in practice and promote well-being for all, here and now?"*
Existentialist Ethics	*"Am I choosing freely and authentically, and fully taking responsibility for this choice?"*

Table 3(a). The Moral Test

Theory	Positive Ethos from Pass (V+)	Negative Ethos of Failure (V-)
Utilitarianism	*Welfare, Harm-reduction, Benefit-distribution*	*Harm, Inefficiency, Bias*
Deontological Ethics	*Duty, Respect, Autonomy*	*Exploitation, Deception, Violation*
Virtue Ethics	*Virtue, Integrity, Character*	*Vice, Hypocrisy, Dishonesty*
Ethics of Care	*Empathy, Care, Responsiveness*	*Indifference, Neglect, Detachment*
Rights-Based Ethics	*Rights, Consent, Liberty*	*Violation, Coercion, Discrimination*
Social Contract Theory	*Fairness, Legitimacy, Collective-consent*	*Breach, Corruption, Coercion*
Rawlsian Justice	*Equity, Distributive-justice, Protection-of-vulnerable*	*Inequality, Bias, Exploitation*
Natural Law Theory	*Reason, Order, Moral-purpose*	*Disorder, Violation, Hedonism*
Environmental Ethics	*Sustainability, Stewardship, Preservation*	*Destruction, Pollution, Exploitation*
Pragmatism	*Practicality, Effectiveness, Outcome-focus*	*Inefficiency, Dogmatism, Failure*
Existentialist Ethics	*Freedom, Individual-agency, Authenticity*	*Institutional-control, Constraint, Bad-faith*

Table 3(b). Signals of Moral Test Results

4. Qualitative Evaluation

MTEST11 enables ethical support examination of objects such as actions, design, intention, impact, policies, principles, rules, processes, procedures, systems, models, decisions, choices, agents (such as organizations, institutions, and humans), and their character by mapping them onto four dimensions. ethical evaluation is grounded in four first-class objects: intention (purpose), structure (design), action, and impact (outcome). These capture the ethical lifecycle—intention defines the purpose, structure encodes the possibilities of actions and impacts, action executes it, and impact reveals its consequences.

To determine the support each of these 14 propositions receives from ethical theories, we developed a MORAL-TEST framework that adapts the core principles of eleven major normative ethical theories (THE11) into structured, decision-oriented moral questions. While rooted in classical and contemporary philosophical literature [6, 15, 21, 22], the format is an original synthesis designed for multi-theory ethical comparison. Each theory is represented by a guiding moral test, e.g., Utilitarianism asks if an action maximizes total happiness. This format supports both qualitative reasoning and quantitative analysis and is suited for AI governance frameworks requiring transparency and value pluralism.

Table 3(a) shows the MORALTESTS questions for each of the 11 ethical theories. The pass and failure also generate signal about the very specific moral/ethical strength or weakness of a proposition as shown in Table 3(b).

We evaluated each of the 14 propositions against these 11 dominant ethical theories using 7 support levels: "D" (direct support), "I" (indirect), "C" (conditional), "N" (neutral), "A" (contrary), "B" (indirect contradiction), and "Q" (conditional contradiction). Table-5 shows the support from each theory. (The "Justifications" for the 11x14= 154 positions can be examined in an online appendix, now removed for anonymity). Table-4 shows the result of the evaluation for each of the 14 principles/propositions.

The MORALTESTS provided are intentionally simplified ethical heuristics, designed to be practical for structured analysis and to serve as a steppingstone for further quantitative analysis. For example, for Rawlsian Justice, the test "Would I accept this rule if I didn't know my position in society?" stems from the "veil of ignorance" thought experiment, which is part of a broader theory of justice. Similarly, the test for Utilitarianism, "Does this maximize

happiness or reduce suffering?", folds in deeper debates regarding act vs. rule utilitarianism, hedonistic vs. preference theories, and the measurability of utility.

The proposed MORALTEST is not a substitute for deep philosophical engagement but a gateway to it. While simplified on the surface, the forms are functionally accurate for structured ethical analysis, debate, and comparison. Though it hides some formal precision, such as Rawls' original position being vastly richer the guiding questions are designed to be highly accessible and usable by policymakers, engineers, ethicists, and students, inviting them into the world of ethical inquiry.

While deontology encompasses various duty-based frameworks, this study specifically applies Kantian Deontology. We evaluate propositions based on the Categorical Imperative, specifically testing two conditions: (1) Universalizability can the maxim of this action be a universal law for all rational agents without contradiction? and (2) Humanity does this policy treat human beings as ends in themselves, rather than merely as means to an end (e.g., for profit or efficiency)?

Ethical Theory (THE11)	P1	P2	P3	P4	P5	P6	P7	P8	I1	I2	I3	C1	C2	C3
Utilitarianism	D	D	D	D	D	D	D	D	D	C	C	Q	Q	Q
Deontological Ethics	D	D	D	D	D	D	D	C	C	Q	Q	A	A	A
Virtue Ethics	I	I	I	D	D	D	D	D	I	C	I	B	B	B
Ethics of Care	C	C	C	D	D	C	D	D	D	Q	C	B	B	B
Rights-Based Ethics	D	D	D	C	C	D	D	C	C	B	C	A	A	A
Social Contract Theory	D	D	D	D	D	D	D	D	C	Q	Q	A	A	A
Rawlsian Justice	D	D	D	D	D	D	D	D	D	A	C	A	A	A
Natural Law Theory	C	C	C	C	C	C	C	C	C	C	Q	C	Q	C
Environmental Ethics	I	I	I	I	I	I	I	I	I	B	I	A	A	B
Pragmatism	D	D	D	D	D	D	D	D	C	C	C	C	C	C
Existentialist Ethics	C	C	C	C	C	C	C	C	C	I	C	C	C	C

Table 4. Support Analysis for the Propositions

5. Quantitative Evaluation

For quantitative analysis, the MORALTEST qualitative responses were converted into a Likert scale, assigning numerical values of 3, 2, 1, 0, -1, -2, and -3 to the categories D, I, C, N, Q, B, and A, respectively. Table-5 shows the average and standard deviations of the resulting scores.

The ACTION8 propositions are significantly supported, with the strongest support for 'Responsibility Chain' (P7), followed by 'Fund to Retraining' (P4), "10% Harm Prevention Trust" (P5), and "Transparency" (P6). "AI ID" (P1), Fragmentation (P2), Kill Switch for Sovereign Body (P3), and Social Business (P8) also received support scores above 2. In contrast, oppositions C1, C2, and C3 contradicted most theories, though they received conditional support from Pragmatism and Existentialist Ethics. Among industry propositions, "Free & Open Access" (I1) received the most favorable (1+) support, while self-regulation (I2) received a mixed, overall negative review.

5.1. Consensus Test for Propositions

We assessed how strongly each AI proposition (P1–C3) is supported or op-posed by the 11 ethical theories using a one-sample t-test on each proposition column. The test determines whether the mean ethical rating significantly deviates from a neutral baseline of 0. For each proposition, we form a Null Hypothesis (H_0): $\mu = 0$ (the average ethical rating is neutral) and an Alternative Hypothesis (H_1): $\mu \neq 0$ (the average rating is significantly different from neutrality).

Table 5 shows the t-state, p(t)-value, z-score, and p(z)-value. Using the decision rule to reject H_0 if p-value < 0.05, the tests (Table 5) confirm that P1-P8 enjoy statistically confirmed strong consensus support, with exceptionally strong support for P7. In contrast, all three C1-C3 faced consensus opposition. For industry propositions I2 and I3, the theories were divergent, and no statistically significant consensus was evident. P1 achieved a high degree

PROP.	Mean	Std Dev	N	T-Stat	P(T)	Z-Score	P(Z)	P(T) < 0.05?	P(Z) < 0.05?	Ethical Interpretation
P1	2.27	0.90	11	8.33	0.00	8.33	7.9E-17	TRUE	TRUE	Strong ethical support
P2	2.27	0.90	11	8.33	0.00	8.33	7.9E-17	TRUE	TRUE	Strong ethical support
P3	2.27	0.90	11	8.33	0.00	8.33	7.9E-17	TRUE	TRUE	Strong ethical support
P4	2.36	0.92	11	8.48	0.00	8.48	2.2E-17	TRUE	TRUE	Extremely strong **support**
P5	2.36	0.92	11	8.48	0.00	8.48	2.2E-17	TRUE	TRUE	Extremely strong **support**
P6	2.36	0.92	11	8.48	0.00	8.48	2.2E-17	TRUE	TRUE	Extremely strong **support**
P7	2.55	0.82	11	10.29	0.00	10.29	7.6E-25	TRUE	TRUE	Extremely strong **support**
P8	2.18	0.98	11	7.37	0.00	7.37	1.7E-13	TRUE	TRUE	Strong ethical support
I1	1.73	0.90	11	6.33	0.00	6.33	2.4E-10	TRUE	TRUE	Moderate **support**
I2	-0.36	1.63	11	-0.74	0.48	-0.74	0.45916	FALSE	FALSE	**No consensus** (leaning opposition)
I3	0.64	1.12	11	1.88	0.09	1.88	0.05952	FALSE	FALSE	**Weak consensus** (leaning support)
C1	-1.55	1.75	11	-2.92	0.02	-2.92	0.00345	TRUE	TRUE	Moderate ethical opposition
C2	-1.73	1.56	11	-3.68	0.00	-3.68	0.00023	TRUE	TRUE	Moderate ethical opposition
C3	-1.45	1.69	11	-2.85	0.02	-2.85	0.00442	TRUE	TRUE	Moderate ethical opposition

Table 5. Summary of Support Analysis for Sovereign AI ACTION8 Policies

of cross-theoretic consensus (Mean = 2.5, SD = 0.5), indicating robust support across diverse frameworks.

When interpreting these t-test results, caution is warranted. The test assumes a normal distribution, which is risky with a small sample (n=11), and expects independent ratings from theories (THE11), though some have known overlaps (e.g., Deontology and Rights-Based Ethics). Other potential issues include subjective bias from the ordinal-to-interval mapping and the lack of multiple testing correction for 14 simultaneous tests, risking Type I errors.

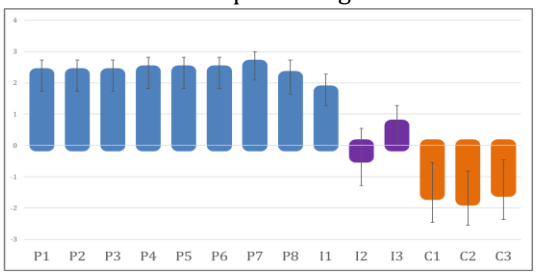

Figure 1. Consensus (Bootstrap Means) of Ethical Theories for Sovereign AI Propositions (95% Confidence Interval, 1000 iterations)

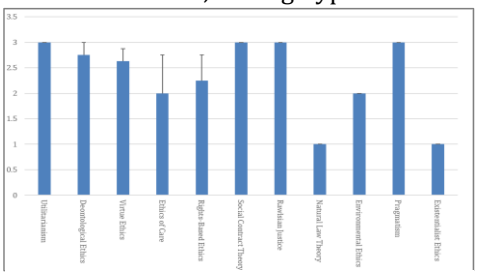

Figure 1. Congruence of (Bootstrap Means) of Ethical Principles for AC-TION8 Set (95% Confidence Interval, 1000 iterations)

5.2. Bootstrapped Average

To avoid issues from t-test assumptions, we performed a nonparametric bootstrap estimation of the sampling distribution with 1000 iterations. This method does not assume a distribution, works with non-normal data and small samples, and provides a confidence interval. Figure 1 (bar plot) shows the resulting bootstrapped mean scores for each policy (P1–C3), with whiskers indicating 95% confidence intervals. If whiskers do not cross zero, the ethical position is statistically significant. Except for I2 and I3, all positions attained statistical consensus.

5.3. Congruence of the Theories

By analyzing the rows, we can also determine how well ethical theories supported or rejected the ACTION8 proposition set. Table 6 and Figure 2 summarize a similar row-wise t-

Theories	Mean	Std Dev	N	T-Stat	P(T)	Z-Score	P(Z)	P(T) < 0.05?	P(Z) < 0.05?	Interpretation
Utilitarianism	3	0	8	inf	0			TRUE	FALSE	Strong support
Deontological Ethics	2.75	0.7071068	8	11	1.14E-05	11	3.82E-28	TRUE	TRUE	Strong support
Virtue Ethics	2.625	0.5175492	8	14.34573	1.9E-06	14.3457	1.13E-46	TRUE	TRUE	Strong support
Ethics of Care	2	1.069045	8	5.291503	0.001134	5.2915	1.21E-07	TRUE	TRUE	Strong support
Rights-Based Ethics	2.25	1.0350983	8	6.14817	0.000468	6.14817	7.84E-10	TRUE	TRUE	Strong support
Social Contract Theory	3	0	8	inf	0			TRUE	FALSE	Strong support
Rawlsian Justice	3	0	8	inf	0			TRUE	FALSE	Strong support
Natural Law Theory	1	0	8	inf	0			TRUE	FALSE	Strong support
Environmental Ethics	2	0	8	inf	0			TRUE	FALSE	Strong support
Pragmatism	3	0	8	inf	0			TRUE	FALSE	Strong support
Existentialist Ethics	1	0	8	inf	0			TRUE	FALSE	Strong support

Table 6. Summary of Congruence Analysis for ACTION8 set from the Ethical Theories

test and bootstrapping analysis. Notably, Utilitarianism, Social Contract Theory, Rawlsian Justice, and Pragmatism all found the ACTION8 set to be strongly ethical. The bootstrapped analysis confirms that, except for Ethics of Care, most of the ethical theories have strong congruence for the action set.

5.4. Significant Ethical Signals and Value Anchor Identification

What moral characteristics of ACTION8 stand out? In MTEST11 these can be answered by identifying the most supportive theories and tracing the core guiding ethos (Table 3(b)).

- Utilitarianism's goal is to maximize total well-being and minimize suffering. ACTION8 aligns by focusing on reducing systemic harm (job loss, bias) and promoting a broad distribution of benefits.
- Social Contract Theory values legitimacy and fairness through the consent of rational agents. ACTION8 aligns by encouraging shared rules to protect people from concentrated AI power and ensure fairness.
- Rawlsian Justice looks for fairness that works for the least advantage under a 'veil of ignorance'. ACTION8's features (retraining, trust fund, transparency) protect vulnerable groups and enforce distributive justice.
- Pragmatism verifies what works in practice to promote human well-being. ACTION8 aligns itself by offering practical checks on runaway AI and emphasizing outcomes over ideology.
- Among the least enthusiastic is Existentialist Ethics, which celebrates individual freedom. ACTION8's emphasis on institutional control and shared accountability may be seen as imposing external moral systems.

5.5. Theoretical Sensitivity to Affordances

The MTEST11 framework reveals that ethical theories are not uniformly sensitive to all AI affordances. For instance, Deontology and Rights-Based Ethics show high sensitivity to "Individual Autonomy" affordances like AI-IMPERSONATE and AI-PRIVACY, where the violation of a rule or right is binary. Conversely, Utilitarianism and Social Contract Theory are more responsive to "Sovereign Agency" affordances like AI-EMPLOSS and AI-CRITCTL, as these impact aggregate societal welfare and the foundational agreement between the state and its citizens. This divergence justifies the pluralistic approach of MTEST11; a single-theory test would inevitably leave "normative blind spots" in either the individual or sovereign dimension.

5.6. Framework Shielding against Normative Blind Spots

By uncovering these anchors, MTEST11 acts as a "Moral Shield" against the inherent blind spots of any single theory. For example, while a proposition might achieve a high score in Utilitarianism by maximizing societal efficiency, it might simultaneously fail a Rights-Based Ethics test by infringing upon individual privacy. MTEST11 surfaces these "hidden trade-offs" quantitatively. This multi-framework shielding ensures that Sovereign AI policies are resilient to the diverse cultural and philosophical demands of global AGI governance, protecting both Sovereign Agency and Individual Autonomy across multiple dimensions of value.

5.7. Subjectivity in MTEST11

The framework adopts a seven-level estimation scale, operationalized within a bounded range of −3 to +3, to facilitate comparative analysis. Each discretization introduces elements of subjectivity. Such subjectivity is intrinsic to ethical evaluation and cannot be eliminated. MTEST11 seeks to narrow it through structured, cross-theoretic consensus across multiple ethical frameworks and does not claim complete objectivity. Robustness of outcomes under subjectivity can be systematically evaluated by classical sensitivity analysis in cases where rating perturbations risks significant qualitative divergence of outcomes. Methodologically, the quantification of emerging or under-defined ethical domains necessarily begins with qualitative interpretation, which then informs the construction of measurable factors. This

iterative transition from qualitative insight to quantitative representation remains a foundational, yet imperfect, aspect of the framework.

6. Conclusions

It is important to be aware of MTEST11's inherent limitations. The proposed MTEST11 omits some formal precision but sufficiently functionally accurate. Simplified heuristics necessarily omit the formal precision of complex philosophical debates. The observed outcomes do not, by themselves, guarantee that a proposition is practical or fully ethical, as implementation and adjudication introduce additional steps that can shift—often downgrade—the ethical balance. While downstream improvements can mitigate concerns, normally they cannot fully redeem fundamental ethical violations in upstream intention or structure.

Artificial intelligence (AI) regulation has emerged as a critical issue in contemporary global discourse. Despite the proliferation of policy proposals, few studies have formally examined how these regulatory principles align with established ethical discourses, a void that originates from the lack of any suitable framework to check the alignment. Even in the top related venues it is rare to see any analysis grounded on ethical theories. We have presented the multi-theoretic ethical support test (MTEST11), a novel structured framework for the systematic evaluation of propositions against established major ethical theories. We demonstrated its use by examining fourteen provocative policy propositions from the sovereign AI regulation debate for their alignment with 11 major ethical theories.

MTEST11 enables AI propositions to be systematically assessed through the lens of eleven major ethical theories (THE11). A structured method for cross-theoretic ethical investigation provides a powerful and robust moral foundation for AI governance. By grounding policy in poly-ethical reasoning rather than ad-hoc reasoning or political expediency, this approach ensures greater moral legitimacy, helps identify normative blind spots, and serves as a critical tool to surface underlying conflicts.

The eleven ethical theories selected for MTEST11 are rooted in the Western philosophical tradition. One can extend this framework to explicitly incorporate non-Western perspectives, such as Confucian ethics, Ubuntu, or Buddhist ethics, to ensure broader intercultural applicability in global AI governance [12]. Such inclusive approach is essential for the development of universally acceptable and interoperable AI systems, ultimately accelerating the responsible growth of AI technologies.

Notably, in contemporary ethics education, qualitative approaches dominate, while structured quantitative methods remain scarce. Yet, ethics is becoming increasingly important in computing education, a field traditionally rooted in quantitative thinking. Johnson [14] aptly noted the absence of structured ethical reasoning tools: "*The challenge in computer ethics is not a lack of ethical concerns but a lack of frameworks to systematically address them in technical education and practice*". This work is a step towards building a novel quantitative framework surrounding the existing qualitative body of knowledge for use in the discipline, enabling students and ordinary technologists to engage with complex ethical dilemmas through a structured, data-informed ethical lens.

Beyond policy design, MTEST11 serves as a critical pedagogical tool for the 'Quantitative Ethicist.' By allowing computer science students and technologists to interact with thousands of years of moral philosophy through a structured, data-informed lens, we move closer to a future where ethical reasoning is an integral part of the engineering lifecycle, rather than an afterthought.

Acknowledgements

Generative AI tools were utilized to assist with language editing and polishing of the manuscript. The authors maintain full responsibility for the final content and scientific integrity of this work.

References

[1] D. Acemoglu and P. Restrepo. Artificial intelligence, automation, and work. In The economics of artificial intelligence: An agenda. University of Chicago Press, 2018, pp. 197-236.

[2] P. Ala-Pietilä, Y. Bonnet, U. Bergmann, et al. The assessment list for trustworthy artificial intelligence (ALTAI). European Commission, 2020.

[3] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety. arXiv preprint arXiv:1606.06565, 2016.

[4] M. Andreessen. Why AI will save the world. Andreessen Horowitz, July 2023.

[5] R. K. Bellamy, K. Dey, M. Hind, et al. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, vol. 63, no. 4/5, pp. 4-1, 2019.

[6] J. Bentham. An Introduction to the Principles of Morals and Legislation (1789). J. H. Burns and H. L. A. Hart, Eds. London, 2010-11.

[7] N. Bostrom. Ethical issues in advanced AI. Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in AI, vol. 2, pp. 12–17, 2003.

[8] R. Chatila, K. Firth-Butterfield, and J. C. Havens. Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, version 2. 2018.

[9] K. Crawford and T. Paglen. Excavating AI: The politics of images in machine learning training sets. AI & Society, vol. 36, no. 4, pp. 1105-1116, 2021.

[10] T. Evas. Civil Liability Regime for Artificial Intelligence. European Parliamentary Research Service (EPRS), European Added Value Assessment, 2020.

[11] C. B. Frey and M. A. Osborne. The future of employment: How susceptible are jobs to computerisation? Technological Forecasting and Social Change, vol. 114, pp. 254-280, 2017.

[12] S. Hongladarom and J. Bandasak. Non-western AI ethics guidelines: implications for intercultural ethics of technology. AI & Society, vol. 39, pp. 2019–2032, 2024.

[13] A. Jobin, M. Ienca, and E. Vayena. The global landscape of AI ethics guidelines. Nature Machine Intelligence, vol. 1, no. 9, pp. 389-399, 2019.

[14] D. G. Johnson. Computer ethics. In The Blackwell Guide to the Philosophy of Computing and Information. 2004, pp. 63-75.

[15] I. Kant. Groundwork of the Metaphysics of Morals. MobileReference, 2008.

[16] J. Khan and S. R. Prithula. Ethical Risk Assessment of the Data Harnessing Process of LLM supported on Consensus of Well-known Multi-Ethical Frameworks. In Proceedings of the Canadian Conference on Artificial Intelligence, 2025. [Preprint]. Available: https://caiac.pubpub.org/pub/s50xrmxh.

[17] J. I. Khan and S. R. Prithula. Whose Ethics? A Multi-Theoretic Stress Test of AI: Comprehensive Evaluation of Design Ethics (FACTOR9), AI-Induced Harms (HARM66+), and AI Governance (ACTION8) Proposals Through the Lens of Eleven Ethical Theories. Technical Report 2025-08-01 Internetworking and Media Communications Research Laboratories, August, 2025, Department of Computer Science, Kent State University [http://medianet.kent.edu/technicalreports.html].

[18] G. Marcus and E. Davis. Rebooting AI: Building artificial intelligence we can trust. Vintage, 2019.

[19] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal. From What to How: An Initial Review of Publicly Available AI Ethics Tools. Ethics, Governance, and Policies in Artificial Intelligence, vol. 144, p. 153, 2021.

[20] OECD. AI Policy Observatory. 2024.

[21] J. Rawls. A theory of justice. In Applied Ethics. Routledge, 2017, pp. 21-29.

[22] S. Russell and P. Norvig. Artificial Intelligence: A Modern Approach, 4th ed. Pearson, 2020.

[23] M. Tegmark. Life 3.0: Being Human in the Age of Artificial Intelligence. Penguin, 2017.

[24] UNESCO. Recommendation on the Ethics of Artificial Intelligence. 2024.

[25] R. V. Zicari et al. Z-Inspection®: A Process to Assess Trustworthy AI. IEEE Transactions on Technology and Society, vol. 2, no. 2, 2021.

[26] S. Zuboff. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York: PublicAffairs, 2019.

[27] Yunus, M., & Weber, K. (2007). Creating a world without poverty: Social business and the future of capitalism. PublicAffairs