HARM66+ A Taxonomy for the Harm-Entity-Aware Post-AI Security

Javed I Khan[1]; Sharmila Rahman Prithula[1]
[1]Kent State University

Abstract

Traditional cybersecurity analysis, framed by the CIA triad, is increasingly strained by Generative AI systems that operate as open-ended socio-technical actors. These classical models often focus on proximate, individual-level impacts, systematically overlooking diffuse downstream effects such as the erosion of trust or loss of human agency.

We introduce HARM66+, a multi-level, "Harm-Entity-Aware" taxonomy comprising 66+ distinct harm classes organized into Exo-Human and Endo-Human domains [1]. By moving beyond narrow technical metrics, HARM66+ surfaces latent harms, such as Epistemic Degradation (the inability to distinguish fact from AI fiction) and institutional legitimacy erosion, that existing frameworks miss. To address the inherent subjectivity of socio-technical risk, the taxonomy maps these harms directly to specific "Harm Entities" (individuals, institutions, and ecosystems), ensuring that protections are prioritized based on the normative values of the affected community. Empirically validated against 1,500+ AI incidents and 11 ethical theories [2], this framework enables engineers to design auditable, proportionate security controls for a post-AI world.

Introduction

- The Problem: Traditional cybersecurity (CIA triad) is strained by Generative AI systems, which are open-ended socio-technical actors.
- The Gap: Many AI-mediated harms (e.g., loss of agency, trust erosion) are weakly represented or absent in existing taxonomies.
- The Consequence: Systems can satisfy formal security criteria while still producing significant real-world harm because harms cannot be explicitly named or reasoned about.

Table 1. Comparative Incident Analysis: HARM66+ vs. Traditional

Case Study	Traditional Analysis (No Taxonomy)	HARM66+ Analysis (With Taxonomy)	Outcome of Using HARM66+
AI-Generated Accusation (AIID-675)	Reports only proximate "Reputational Harm" to a principal.	Surfaces latent Institutional Legitimacy Erosion and Community Psychological Harm.	Enables safeguards for staff and students, not just the individual.
Political Disinformation (AIID-676)	Reports "Political Disinformation" as a general risk.	Identifies Collective Epistemic Harm and Normalization of Synthetic Falsity.	Triggers controls for civic institutions and voter trust.
Navigation system (AIID-857)	Reports "Navigation system safety failure" as a general risk.	Identifies Long-term familial harm; infrastructure trust erosion; regulatory legitimacy harm.	Triggers controls for Families, local community and riders

Methods and Materials

- HARM66+ Taxonomy: 66+ distinct harm classes organized into Exo-Human and Endo-Human domains.
- ENTITY7 Taxonomy: A structured victim taxonomy spanning individuals, institutions, ecosystems, and abstract societal goods.
- HV-CARD Workflow: A design chain (Technology →Affordance →Harm →Victim) used to synthesize auditable security controls.
- Ethical Grounding: Validated against 11 major ethical theories (MTEST11) to ensure the taxonomy is pluralistically defensible and not arbitrary.

Operationalizing HARM66+

1. **Case A:** Comparative Incident Analysis: HARM66+ vs. Traditional: The practical utility of the HARM66+ framework is demonstrated through a comparative analysis of real-world AI incidents, as presented in Table 1. By applying the HARM66+ taxonomy, researchers can surface "latent" harms that are systematically overlooked, including the erosion of institutional legitimacy, community-scale psychological distress, and the degradation of epistemic trust.
2. **Case B:** Standardized Reporting Workflow: Suppose a platform safety officer needs to document an incident where an AI-driven "afterlife" bot misrepresents a deceased individual's legacy. Instead of a generic privacy tag, the officer uses the hierarchical identifier for H7 Privacy & Surveillance × E1b Deceased Individuals. This creates a shared cross-domain vocabulary that ensures the incident is reported with technical and ethical precision across different jurisdictions.
3. **Case C:** Control Synthesis Workflow: Suppose a developer is building a student-facing automated grading system and needs to design safety features. By identifying the victim as E1a Living Individuals (Students) and assessing the harm's Durance (Medium) and Reversibility (Low-Medium), and Victim-level remedies (appeals process) rather than just technical model fixes.

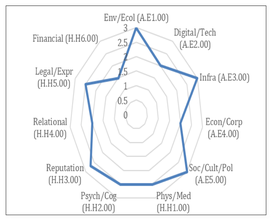

Figure 1. Durance Scores of Various Harms (1=short, 2=medium, 3=long term)

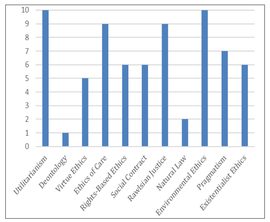

Figure 2. Relative Importance of Durance to Various Ethical Theories

Scan for Full Paper

Impact of Duration

- The duration of harm, defined as how long a negative impact persists, is a significant factor in its ethical evaluation.
- Longer-lasting harms carry greater ethical weight because their effects accumulate, increasing the moral urgency for prevention.
- The ethical severity rises sharply when harm is both long-lasting and irreversible, as this undermines future moral agency, justice, and ecological integrity.
- As shown in Figure 1, harms are categorized based on their persistence: long-term (generational/permanent), medium-term (1–10 years), and short-term (temporary with high recovery potential).

The significance of durance varies across moral frameworks, reflecting diverse priorities:

- High Importance: Utilitarianism (T1), Ethics of Care (T4), Rawlsian Justice (T7), and Environmental Ethics (T9) assign high weight because long-term harms maximize cumulative suffering or threaten generational ecological/social systems.
- Moderate Importance: Virtue Ethics (T3), Rights-Based Ethics (T5), Social Contract Theory (T6), Pragmatism (T10), and Existentialist Ethics (T11) consider durance a factor in assessing character, the intensity of a breach, or moral responsibility.
- Low Importance: Deontological Ethics (T2) and Natural Law Theory (T8) focus on the inherent wrongness of an act or its alignment with natural purpose, often regardless of how long the consequences last.

Discussion

- Surfaces Latent Harms: Reduces assessment blind spots by naming downstream socio-technical impacts.
- Proportionate Mitigation: Enables specific tools and processes for prioritized monitoring based on harm severity.
- Shared Vocabulary: Establishes a cross-domain language for consistent reporting across different sectors and jurisdictions.
- Traceability: Makes protection design traceable and auditable by linking controls directly to specific harm-victim pairs.

Conclusions

- Traditional CIA-style approaches systematically miss diffuse, cross-domain harms produced by modern AI.
- HARM66+ reduces under and over-regulation by making severity attributes (Gravity, Scale, Durance) explicit in the design process.
- This framework provides the necessary foundation for harm-aware security engineering in a post-AI world.

Contact

Javed I Khan
Kent State University
Email: javed@kent.edu

References

1. Khan, J. I., & Prithula, S. R. (2026). In Quest of an Extensible Multi-Level Harm Taxonomy for Adversarial AI: Heart of Security, Ethical Risk Scoring and Resilience Analytics. arXiv preprint arXiv:2601.16930.