

Safe Enough to Try GenAI: A Multi-Level Sociotechnical Approach for Responsible Generative AI in Public-Serving Organizations

Ashley Khor

University of Pittsburgh

Abstract

This research examines how public-serving organizations can responsibly design, evaluate, and use generative AI under real-world constraints, developing a multi-level sociotechnical approach that connects institutional conditions, operational decision-making, and human-centered evaluation practices.

Background and Research Question

Public-serving organizations are under increasing pressure to adopt generative AI (GenAI) while maintaining commitments to safety, accountability, care, and public value. Existing responsible AI work has produced substantial agreement on principles (Fjeld et al. 2020), yet the movement from principles to situated practice remains under-specified. This gap is especially acute for nonprofits and public-serving organizations that work with vulnerable populations, operate under resource constraints, and rely on foundation-model systems shaped by upstream actors outside their control. My dissertation asks: how can public-serving organizations make responsible generative AI safe enough to try without treating either adoption or avoidance as the only ethical option? More specifically, I examine why implementation gaps persist, what kinds of sociotechnical infrastructure are needed to translate values into practice, and how evaluation and prompt-design methods can preserve human meaning in high-emotion settings.

Approach

I develop a practice-grounded account of responsible GenAI implementation through a multi-year research program connected to the EAAMO Living Labs working group and collaborations with public-serving organizations in survivor-support and refugee-support domains. The work combines

conceptual analysis, practice-based engagement, participatory evaluation design, and exploratory LLM experiments. I use Ostrom's (2005) Institutional Analysis and Development framework as a diagnostic lens to locate implementation failures across material conditions, community attributes, rules-in-use, action situations, interactions, and outcomes. I then specify the infrastructure needed at three levels: operational decision scaffolds for situated design and use, collective-choice arrangements for legitimate working-rule formation and capability building, and constitutional-choice arrangements through communities of practice and sectoral norm development.

Work Accomplished

I have completed a set of studies that examine responsible GenAI implementation across institutional, operational, evaluative, and interpretive levels. Together, these studies develop "Safe Enough to Try GenAI" as a middle-ground approach: public-serving organizations should not be forced to choose between uncritical adoption and blanket refusal, but need structured ways to experiment, evaluate, govern, and revise GenAI systems under conditions of uncertainty.

At the institutional level, my AIES submission on sociotechnical infrastructure uses the Institutional Analysis and Development framework to diagnose why responsible GenAI implementation gaps persist. It shows how model properties such as probabilistic generation, perceptual opacity, temporal instability, and interaction-dependent outputs combine with structural asymmetries in power, information, capability, and accountability. This analysis reframes responsible GenAI as an infrastructural problem rather than a problem of principles alone, and proposes a conceptual sociotechnical infrastructure that supports the development and adoption of responsible GenAI practices in public-serving organizations across operational, collective-choice, and constitutional-choice levels.

At the operational level, my AIES paper on sociotechnical microethics identifies the mechanisms through which public values become actionable in everyday decisions. It specifies microethical granularity, epistemic specification of values and harms, participatory authority, lifecycle linkage, and organizational capability as mechanisms that help organizations move from abstract commitments to situated design, deployment, and use-time judgment.

At the evaluation level, my AAAI 2025 Fall Symposium paper, “Safety is a Process, not a Score” (Khor 2025), develops a participatory, symbol-aware safety evaluation methodology for GenAI tools in high-emotion social good contexts. The methodology was adopted as part of the pre-launch evaluation process for a nonprofit GenAI tool for survivors in 2025. This work shows how safety evaluation can attend not only to factual accuracy or toxic content, but also to tone, implication, narrative framing, emotional resonance, and the relational cues that shape whether a system feels safe and appropriate for vulnerable users.

At the interpretive level, two studies examine how GenAI systems can support analysis of sensitive human meaning without flattening ambiguity. My APPAM paper, “From Hopes to Healing Systems,” uses LLM-assisted and qualitative mixed methods to analyze survivor-authored hopes for healing in gender-based violence recovery. My AIES prompt-design study then examines how contextualized prompt design—through role attributes, construct definitions, few-shot examples, and ambiguity-handling instructions—shapes LLM classification of survivor-authored statements into healing-domain categories. Using an ambiguity-aware evaluation approach, the study treats classification variation as evidence of how prompts resolve, preserve, or overstate meaning in sensitive language. The findings are intentionally bounded: prompt scaffolds can shape interpretive behavior, but their effects vary by model and prompt type. Collectively, these studies support a calibrated claim that contextualized and ambiguity-aware prompt design can function as interpretive framing, underscoring the need to evaluate sensitive text-analysis workflows for ambiguity handling, construct fidelity, and rationale quality.

Contributions and Significance for AIES

The contribution of this research is a middle-ground approach to responsible GenAI: neither techno-solutionist adoption nor blanket refusal, but structured experimentation under conditions of uncertainty. The work contributes to AIES in three ways. Conceptually, it explains why responsible GenAI implementation breaks down when values encounter foundation-model properties, organizational constraints, and weakly specified formal and informal rules. Methodologically, it offers approaches for participatory, symbol-aware, and ambiguity-aware evaluation in high-emotion contexts where conventional performance metrics are insufficient. Practically, it translates responsible AI into

decision scaffolds that public-serving organizations and end users can use to design, evaluate, deploy, and revise GenAI systems with greater confidence. The work is grounded in public-serving settings, but its analytic transfer is broader: it applies where organizations have high duties of care, limited control over upstream models, and a need to make situated judgments about whether GenAI is safe enough to try.

Validity, Scope, and Claim Calibration

The project is intentionally practice-grounded rather than population-generalizing. Its validity rests on analytic generalization, mechanism specification, and traceability between method, data, interpretation, and claim. The IAD analysis supports a diagnostic claim about recurring infrastructural needs under identifiable conditions. The microethics framework is supported by practice-based engagements that reveal how values become operational problems in concrete design, evaluation, and use situations. The prompt-design study provides an early empirical signal rather than definitive model-behavior claims. Its purpose is to identify how interpretive scaffolds can alter classification and rationale patterns, and to show why sensitive-language workflows require ambiguity-aware evaluation.

Ongoing Work

I am now integrating these studies into the dissertation and extending the practice-based program through Living Labs collaborations. The next phase strengthens empirical reporting, clarifies validity criteria for co-production and participatory evaluation, and develops more portable decision scaffolds for organizations facing similar constraints. The dissertation specifies conditions for mechanism portability: the approach is most relevant where organizations operate in high-stakes or high-emotion service contexts, where affected communities or trusted intermediaries can inform values and harms, and where implementation requires iterative governance rather than one-time approval.

References

- Fjeld, J.; Achten, N.; Hilligoss, H.; Nagy, A.; and Srikumar, M. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. SSRN Scholarly Paper No. 3518482. Social Science Research Network. doi:10.2139/ssrn.3518482.
- Khor, A. 2025. Safety Is a Process, Not a Score: A Symbol-Aware Safety Evaluation Methodology for GenAI for Social Good Tools in High-Emotion Contexts. Paper presented at the AAAI 2025 Fall Symposium on AI for Social Good, November 5.
- Ostrom, E. 2005. Understanding Institutional Diversity. Princeton, NJ: Princeton University Press.