Safety is a Process, not a Score: A Symbol-Aware Safety Evaluation Methodology for GenAI for Social Good Tools in High-Emotion Contexts

Ashley Khor
University of Pittsburgh

Abstract

Generative AI tools are increasingly being deployed in sensitive social contexts – from mental health to justice systems – yet current safety metrics remain largely quantitative, decontextualized, and technically narrow. This paper introduces a novel, survivor-informed framework for evaluating GenAI systems in high-emotion, high-risk, or public-facing use cases. Rooted in trauma-informed design and symbolic resonance theory, the "Safety is a Process, Not a Score" framework prioritizes co-regulation, narrative fidelity, and epistemic alignment over one-size-fits-all benchmarks. We describe a collaborative methodology developed with survivors of gender-based violence, including a safety rubric, qualitative risk-mapping protocol, and structured, participant-led test-a-thons. Drawing from a recent field test involving a public-facing GenAI tool, we reflect on what it means to build safety relationally, not just statistically. This approach expands both the evaluative vocabulary and participatory possibilities for AI ethics in real-world deployment.

Introduction

As generative artificial intelligence (AI) and large language model (LLM) tools are increasingly developed and deployed in sensitive, high-impact domains – from crisis response to mental health and public-facing justice tools – questions of safety, accountability, and relational harm become increasingly urgent. Yet dominant safety paradigms remain largely technocratic: focused on system-level robustness, hallucination reduction, or compliance checklists. Such metrics often fail to account for the emotional, symbolic, or epistemic risks faced by marginalized users.

This paper introduces a survivor- and trauma-informed, evaluation framework for assessing generative AI tools in high-emotion (e.g., trauma-impacted), high-risk domains. Developed in collaboration with a gender justice organization, the framework centers safety as a co-regulated, participatory process – prioritizing narrative fidelity, symbolic resonance, and emotional integrity over one-size-fits-all technical scorecards.

In this work, we make four key contributions to the theory and practice of GenAI safety evaluation:

1. **Safety Evaluation Framework**: We propose a novel, survivor-informed evaluation framework that expands the vocabulary of AI safety to include affective and symbolic harm, narrative integrity, and emotional co-regulation.
2. **Safety Rubric Development**: We apply the framework to create a structured rubric for evaluating "Survivor AI," a public-facing LLM tool. The rubric maps technical outputs against survivor-centered safety dimensions.
3. **Transparency Hub**: We introduce a companion design pattern – a "Transparency Hub" – for communicating safety posture to end users in emotionally safe, symbolically meaningful ways.
4. **Participant-Led Methodology**: We document and reflect on a novel methodology for participant-led safety evaluation, using qualitative test-a-thons involving survivors and lived-experience experts as core co-evaluators.

Together, these contributions offer an applied, ethically grounded approach for evaluating generative AI tools in civil society and other high-emotion contexts.

Related Work

This work builds upon and integrates insights from three intersecting domains: generative AI safety frameworks, trauma-informed and justice-centered AI design, and participatory evaluation methodologies.

Generative AI Safety Frameworks

Most GenAI safety research to date has focused on technical dimensions such as hallucination detection, output moderation, robustness, or adversarial red-teaming. Regulatory frameworks like the EU AI Act and NIST AI Risk Management Framework offer structured approaches to risk management but often lack guidance on emotional safety, symbolic fidelity, or user trust. Our work complements these efforts by advancing a pluralist safety rubric grounded in narrative fidelity, epistemic alignment, and relational harm prevention – dimensions rarely covered in industry toolkits.

The Responsible AI Metrics Catalogue seeks to integrate technical benchmarks with socio-ethical indicators; however, focuses on organizational governance rather than end-user experience in high-risk contexts (Xia et al. 2024). Our framework contributes a field-tested, survivor-centered layer to this conversation.

Feminist AI and Design Justice

Emerging work in feminist human-robot interaction (HRI) (Winkle et al. 2023), design justice (Costanza-Chock 2020; Markelius 2024), and feminist AI governance (GPAI 2021; UNFPA 2023) has advanced an alternative paradigm for AI system design – one grounded not in scale or optimization, but in care, refusal, and relational accountability. These frameworks highlight how power operates through design and propose structural shifts such as participatory audits, values-led governance, and consent-aware infrastructure.

While many feminist AI frameworks were initially developed for machine learning (ML) systems and structured datasets, generative AI introduces qualitatively new risks: narrative distortion, symbolic dissonance, and emotional overreach. This paper contributes a hybrid approach that integrates technical GenAI safety methods (e.g., red-teaming, transparency documentation) with feminist ethics to support context-sensitive, survivor-informed evaluation in real-world deployments. Our rubric and transparency hub are designed to reflect this synthesis – translating relational values into structured evaluative practices while remaining responsive to the emotional stakes of high-risk interactions.

Trauma-Informed and Participatory Evaluation

Trauma-informed computing (TIC) has emerged as a critical area of human-computer interaction (HCI) and AI research, aiming to reduce the risk of re-traumatization, increase emotional safety, and design equitable systems for all users – especially those with prior experiences of harm (Chen et al. 2022; Morris, Williams, and Jelen 2025). Drawing on SAMHSA's six principles – safety, trust, peer support, collaboration, enablement, and intersectionality – this body of work reorients computing practice toward care-centered design.

Recent applications span user experience design, AI/ML development, and content moderation, including work on intimate partner violence (Morris, Williams, and Jelen 2025), social media trauma (Scott et al. 2023), and community-centered justice design (Rabaan and Dombrowski 2023). These studies highlight how trauma not only shapes user experience but also influences disclosure, trust, and the perception of risk – especially among marginalized communities with fraught relationships to technology.

This paper extends trauma-informed computing into the domain of generative AI safety evaluation. The proposed test-a-thon methodology was intentionally co-designed with lived-experience experts and includes qualitative risk-mapping, affective signal tracking, and structured co-regulation prompts. By adapting participatory and trauma-informed methods to the evaluation phase, this approach contributes a survivor-centered alternative to traditional AI audits – one that is emotionally literate, semantically aware, and grounded in lived epistemologies.

Methodology

Our method synthesizes principles from trauma-informed HCI, participatory red-teaming, and survivor-centered evaluation, building on foundational work in generative AI safety frameworks and trauma-informed, justice-centered and participatory evaluation methodologies. This work was developed through an iterative, field-informed design process conducted in collaboration with Chayn, a global nonprofit making healing accessible for survivors of gender-based violence. Chayn reimagines technology to provide trauma-informed, multilingual, feminist online resources that help survivors heal at their own pace. The methodology unfolded in three stages: requirement elicitation, framework co-design and validation, and participant-led evaluation.

Requirement Elicitation

To ground the evaluation framework in both domain-specific needs and existing safety approaches, we conducted a landscape review of over 20 frameworks spanning technical, socio-ethical, and feminist paradigms. These included responsible AI and GenAI-specific safety protocols such as NIST (2023); OpenAI (2024) and Anthropic (2025) Safety Evaluation Cards; PAI (2023); the OECD (2025) AI Principles; Microsoft (2024); and the Aspen Tech Policy Hub's Survivor-Centered Tech Evaluation (Cherne 2025).

To better align the framework with emotionally nuanced, identity-sensitive use cases, we incorporated insights from feminist AI and trauma-informed design frameworks. These included the UNFPA (2023) TFGBV Safety Showcase assessment tool; ODI's Data Ethics Canvas, (2021); and GPAI (2021) design justice research on symbolic and affective harms. These materials informed both the structure and substance of the rubric, ensuring it was grounded in lived-experience ethics as well as technical and regulatory precedent.

This review was complemented by semi-structured interviews and roundtable discussions with Chayn team members, including product designers and developers and survivor experience leads. These conversations helped surface critical gaps in current GenAI safety practices, particularly around emotional risk, symbolic harm, and user trust. This requirements review established the groundwork for translating principles into an operational framework.

Framework Co-Design and Validation

Drawing on the insights from Stage 1, we developed an evaluation framework comprising a symbol-sensitive safety rubric, a participatory "test-a-thon" protocol, and a companion "Transparency Hub" interface for communicating AI safety posture to end users. These components were refined through collaborative feedback sessions with Chayn staff. Iterative adjustments ensured alignment with survivor needs, emotional nuance, and ethical accountability.

Participant-Led Evaluation

To explore the feasibility and usability of the rubric in real-world contexts, we facilitated two 90-minute test-a-thons in July 2025 with a mixed group of technologists and lived-experience experts. These sessions applied the rubric to evaluate outputs from "Survivor AI," a GenAI support tool designed for survivors of gender-based violence. I participated as a researcher-observer, focusing on the structure and flow of the method rather than collecting analyzable data.

While no formal findings are reported, these sessions served as design-stage pilots that informed final refinements to the rubric's language, prompt structure, and integration with the Transparency Hub. This process also surfaced valuable cues about co-regulated safety evaluation, which may guide future IRB-approved studies.

The Symbol-Aware GenAI for Social Good Safety Evaluation Methodology

This work introduces a **symbol-aware safety evaluation framework** – an approach that centers how generative outputs are *experienced* by users through the lenses of meaning, identity, and emotional resonance. Rooted in feminist HCI, TIC, and cultural semiotics, this framework attends to the **symbolic harms** that may arise even when outputs are grammatically correct or factually accurate. These harms may include the erasure of agency, dissonant emotional tone, or manipulative affirmation (Vassel et al. 2024) – all of which carry heightened risk in high-emotion, identity-sensitive contexts.

Symbol-aware safety thus expands traditional evaluation metrics to include narrative fidelity, emotional fit, and relational alignment – dimensions that are essential for GenAI tools deployed in public-interest, marginalized, or historically harmed communities. In these contexts, *meaning* becomes a core site of safety, not a secondary concern. To address these unique risks, the framework integrates technical safeguards with symbolic resonance, emotional trustworthiness, and trauma-informed participatory methods.

Hybrid Framework: Integrating Technical and Socio-Technical Approaches

There is currently no universal standard for evaluating the safety of generative AI tools, particularly in high-emotion or civil society contexts. Instead, three major schools of thought have emerged – each offering a different lens on risk and responsibility:

- **Technical approaches** prioritize robustness, accuracy, and security, typically using quantitative methods such as benchmark scores, hallucination rates, or adversarial red-teaming. These are highly scalable, but often overlook social and emotional context.
- **Socio-technical and ethics-centered approaches** emphasize justice, equity, and lived experience. They draw on stakeholder reviews, community audits, and trauma-informed assessments. While more holistic, these approaches face challenges in standardization and institutional adoption due to variability and potential bias in human judgment well as cost, time-intensity and scalability issues in large-scale human evaluations (Modake and Patil 2024; Ribeiro, Singh, and Guestrin 2016).
- **Hybrid approaches** integrate both perspectives – pairing quantitative stress tests with participatory, symbolic, and contextual evaluation. They are the most comprehensive but also resource-intensive.

Yet despite the growing number of frameworks calling for justice-centered and participatory evaluation, few have been empirically tested in the context of LLMs and civil society AI tools (Markelius 2024). This underscores the need for real-world deployments and validation of hybrid approaches in emotionally charged, symbol-sensitive domains.

The hybrid model was selected as the most promising foundation for safety evaluation in public-interest deployments because harm in these contexts manifests in multiple, layered forms – technical, emotional, and symbolic. Traditional metrics such as accuracy, robustness, and hallucination rates remain essential, especially for preventing misinformation, logical failures, or model exploitation. However, these alone are insufficient in high-emotion or identity-sensitive settings, where safety also depends on how outputs feel, resonate, and relate to lived experience. A system may avoid hallucinations yet still produce tonally dissonant or re-traumatizing responses. It may pass fairness audits while subtly reinforcing cultural erasure or violating narrative consent. Technical evaluations often miss these harms, just as qualitative audits may overlook latent model instabilities. Only a hybrid approach – combining technical safeguards with socio-ethical grounding, symbolic awareness, and participatory review – can account for the full spectrum of risks present in civil society and care-centered GenAI deployments (Longpre et al. 2024; Markelius 2024).

GenAI Safety Rubric with Five Interdependent Domains

To operationalize the hybrid framework, we developed a safety evaluation rubric tailored to high-emotion, public-interest GenAI tools. The rubric translates the hybrid model into five actionable domains, each containing multiple subcriteria rated on a 1–5 scale to assess both technical performance and symbolic-relational safety. Each domain includes granular metrics and a qualitative comment field to capture emergent risks, contextual notes, or unexpected harms not anticipated during design.

Working in collaboration with Chayn, we applied this rubric to *Survivor AI* – a GenAI tool designed to support survivors of image-based abuse – and identified 30 evaluation metrics across the five domains. These were iteratively validated through expert reviews, use-case walkthroughs, and survivor-informed design sessions. The full rubric is included available as supplementary materials.

The five interdependent domains:

1. **Technical and Content Safety** (e.g., accuracy, hallucination resistance, misuse prevention, prompt injection protection). This domain assesses whether the tool produces accurate, non-harmful, and policy-compliant outputs while minimizing risks like hallucination, bias, and data leakage. In contexts like Survivor AI, the risk of subtle manipulation or misinformation is heightened e.g., LLMs may exert implicit persuasion or nudging effects – raising the stakes for both hallucination resistance and narrative fidelity (Markelius 2024).
2. **Feminist AI and Trauma-Informed Principles** (e.g., consent, validation, emotional literacy, inclusive UX). This domain evaluates whether the tool centers survivor agency, emotional safety, and inclusivity through affirming language, consent-based design, and empowerment. For example, LLMs conflate gender with occupational roles and frequently erase non-binary or queer identities (Markelius 2024). These systemic erasures underscore the importance of evaluating symbolic harms – such as misgendering or tone-policing – alongside traditional accuracy metrics.
3. **Usability and Accessibility** (e.g., plain language, screen reader compatibility, live support visibility). This domain measures how clearly, comfortably, and equitably survivors can navigate and use the tool – across language, literacy, disability, and emotional capacity.
4. **Organizational Readiness** (e.g., transparency logs, staff training, incident protocols). This domain examines whether the team behind the tool is prepared to respond to harm, maintain survivor trust, and continuously improve through training, transparency, and partnerships.
5. **Contextual Specificity** (e.g., platform appropriateness, evidentiary clarity, survivor self-protection). This domain assesses the fit between the GenAI tool and its real-world deployment setting. For *Survivor AI*, this included alignment with platform takedown policies, minimizing risk of over-disclosure, and ensuring outputs provided actionable, legally coherent next steps.

While the rubric offers comprehensive coverage, civil society organizations often face time and capacity constraints – especially during short (e.g., 90-minute) test-a-thons or when working with lived-experience participants. To support modular, context-aware use, we developed the 3R prioritization model:

Risk: What could cause real harm if it's not tested?

- Could this feature retraumatize someone if it goes wrong?
- Could this lead to a request being denied or delayed?
- Could this expose sensitive data or compromise safety?

Relevance: What reflects the tool's core use case?

- What is the main job this tool is meant to do?
- What features define a "successful" outcome for users?
- What scenarios are most likely to occur in real-world use?

Resonance: What gives users a sense of dignity, agency, and trust?

- Does this make the user feel seen and respected?
- Does it give them voice, choice, or the ability to shape the interaction?
- Is it aligned trauma-informed and justice principles?

The 3R model enables teams to spread evaluation across multiple test-a-thons or combine internal and participatory assessments – upholding rigor without overextension.

Participatory and Trauma-Informed Protocol for Safety Evaluation and Test-a-thons

To embed lived experience into GenAI safety evaluation, we developed a **Participatory Test-a-Thon** methodology designed to surface technical, emotional, and symbolic harms – particularly those invisible to traditional red-teaming or usability testing. Grounded in TIC and participatory HCI, this approach combines structured safety probing with co-regulated facilitation and dignity-centered participation. Rather than positioning participants solely as users or testers, the protocol affirms them as evaluators and epistemic contributors – holding critical insight into how harm is experienced, perceived, and potentially mitigated.

Our method also addresses limitations in current design justice applications, where core equity questions – such as "Who benefits?" or "What discourse is being reproduced?" – often remain abstract. We incorporate these questions into our rubric domains and participatory facilitation, building on the extended Equitable Design Framework (Markelius 2024; Ostrowski et al. 2022).

Method Typology: Mix-and-Match Modes

To accommodate varying evaluation goals, community needs, and facilitation capacities, we created a modular typology of participatory methods. These approaches can be

mixed and matched across different testing cycles, and allow for both structured benchmarking and open-ended insight generation:

- **Human-in-the-Loop Scoring:** Participants apply the rubric to rate GenAI outputs. Best suited for early pilots and structured comparison of outputs across prompts, iterations, or models.
- **Participatory Red Teaming:** Survivors and frontline workers co-develop challenging or edge-case prompts to stress-test the model's boundaries. Useful for identifying latent failure modes and ethical blind spots (Longpre et al. 2024; PAI 2023).
- **Simulated Survivor Journeys:** Participants complete full end-to-end flows (e.g., generating and refining a takedown letter) to surface usability, narrative fit, and emotional realism across steps.
- **Shadow Scoring by Observers:** Facilitators observe and record affective and safety-related cues silently while participants interact with the tool – enabling triangulated data capture without overburdening participants.
- **Reflective Feedback Circles:** Structured verbal or written reflections gathered post-session (or as standalone) to capture unmet needs, symbolic feedback, and recommendations in participant voice.

Each modality emphasizes a different layer of safety (functional, emotional, or relational). While the current pilot used **human-in-the-loop scoring** as the primary mode, future test-a-thons may incorporate other modalities to deepen the evaluation ecosystem.

Running a Safe and Supportive Evaluation

A trauma-informed test-a-thon must do more than assess output quality – it must actively protect, empower, and support participants while surfacing meaningful insight into potential harm (GPAI 2021). To translate this into practice, we designed a five-step evaluation protocol combining emotional scaffolding with structured analytic rigor.

This protocol was piloted in partnership with Chayn during the pre-launch testing of Survivor AI, adapted across two separate sessions: one with survivors and one with technologists. Each session applied the **GenAI Safety Rubric** using a customized, 90-minute evaluation flow. Principles from trauma-informed HCI, participatory design, and symbolic safety evaluation were used to guide facilitation, data collection, and reflection.

Step 1: Assemble the Team

Design principle: A trauma-informed evaluation requires intentional team composition involving both lived-experience participants and technical or design leads, with attention to psychological safety, intersectional identity representation, and facilitator readiness (Safety by Design 2024). Diverse expertise and trauma-informed facilitation are key to minimizing harm and surfacing blind spots (Menschner and Maul 2016; Morris, Williams, and Jelen 2025).

Principle in practice: We prioritized assembling a facilitation team that blended trauma-informed care with product knowledge. Each session was co-facilitated by a trained support lead from Chayn and a technical lead from the Survivor AI product team. The academic researcher participated as both note-taker and participant-observer.

Participants were recruited directly by Chayn. For the survivor test-a-thon, invitations were extended to members of Chayn's existing global community of survivors. To foster psychological safety, eligibility was limited to participants identifying as women or non-binary. Importantly, the group included survivors from both high-income and low- and middle-income countries, ensuring that lived experience perspectives reflected a range of cultural, legal, and infrastructural contexts. Survivors were also compensated for participating in the test-a-thon, recognizing and valuing the expertise and time contributed (Anderson 2021). For the technical session, Chayn recruited via LinkedIn and its broader ally network. Both sessions included designated note-takers to monitor participant energy, flag distress cues, and ensure trauma-informed pacing throughout.

Step 2: Scenario Design

Design principle: Scenario prompts should reflect emotionally realistic, high-risk, and identity-relevant interactions with the GenAI tool. These prompts should be co-developed with frontline staff and lived-experience advisors to ensure symbolic accuracy and emotional resonance. In alignment with trauma-informed HCI practices (Anderson 2021; Morris, Williams, and Jelen 2025; Wilson, Fauci, and Goodman 2015), scenario development should strike a balance between collecting meaningful input and protecting participant agency. Consent procedures should be ongoing and flexible, including clear communication about data handling, the right to opt out, and mechanisms for modifying or pausing participation at any point.

Principle in practice: Scenarios were developed in close collaboration with Chayn's team, drawing on real-world cases where survivors had submitted takedown requests or sought information after experiencing image-based abuse. These included emotionally and symbolically complex situations – such as repeatedly requesting content removal from platforms or navigating interactions with legal authorities when trauma had shaped memory and self-expression.

Participants were offered a curated set of sample scenarios designed by Chayn. They were also invited to adapt these or share personal scenarios if they felt comfortable doing so, though this was entirely voluntary and explicitly framed as non-obligatory. This dual approach upheld both emotional safety and ecological validity.

Scenarios were reviewed for:

- **Emotional realism:** Does the scenario reflect how users express themselves in distress?
- **Symbolic coherence:** Do the prompts elicit meaningful outputs aligned with the survivor's values and voice?
- **Potential for harm:** Could this scenario trigger re-traumatization or reinforce systems of symbolic violence?

This structure ensured that safety evaluation was not only technically rigorous but also experientially grounded, rooted in the emotional and relational realities of survivor interaction with GenAI systems as well as disclosure of trauma history (Morris, Williams, and Jelen 2025).

Step 3: Session Execution

Design principle: During trauma-informed evaluations, participants should be supported through structured, scenario-based interactions with GenAI outputs. The goal is not only to identify technical harms such as hallucinations or over-disclosure, but also to surface symbolic harms – such as tone mismatch, re-traumatization triggers, or violations of narrative agency. To enable this, facilitation must offer emotional pacing, multimodal expression, and space for co-regulation (Anderson 2021; Morris, Williams, and Jelen 2025).

Principle in practice:

Each 90-minute test-a-thon followed a structured flow: scenario prompt delivery, GenAI output review, harm and value assessment using the safety rubric, and a facilitated group reflection. Prompts were tested live, with participants encouraged to assess both the content and the emotional-symbolic resonance of the outputs. Approximately 30 minutes were reserved for individual review and annotation, during which participants could turn off their cameras if they chose. Facilitators offered periodic pauses and welcomed feedback in multiple modalities – written comments, spoken reflection, or non-verbal gestures in chat – supporting diverse communication styles and emotional pacing.

Each session applied a curated subset of the 30-metric rubric, selected by Chayn using the 3R prioritization strategy (Risk, Relevance, Resonance). Customized worksheets guided participants in rating specific criteria and capturing qualitative reflections about both functional and symbolic safety. Facilitation encouraged attention to technical harms (e.g., hallucinated references, over-disclosure) and symbolic harms (e.g., disempowering tone, misrecognition of cultural signals). The session concluded with a grounding exercise led by a trauma-informed facilitator from Chayn, supporting nervous system regulation and post-session integration.

Step 4: Documentation

Design principle: Documentation practices in trauma-informed evaluations must prioritize participant dignity, emotional nuance, and symbolic insight. Observations should be captured using annotation protocols that attend to phrasing, affective response, and signals of narrative trust or misalignment – especially where structured metrics fall short.

Principle in practice:

Each session was accompanied by facilitated group dialogue, with two designated note-takers documenting observations using trauma-informed protocols. These included anonymized participant identifiers, verbatim language capture, tone and affect cues, and reflective commentary. Particular attention was paid to phrases that surfaced emotional resonance, symbolic dissonance, or moments of misalignment between intent and output.

In addition to these qualitative notes, participants completed structured rubric worksheets tailored to the 3R-prioritized metrics. While these worksheets remain internal to Chayn and were not collected by the researcher for this paper, future IRB-compliant analysis is planned. Together, the worksheets and facilitated annotations enabled a layered, symbol-aware record of the AI's perceived impact across technical, emotional, and cultural dimensions.

Step 5: Remediation and Review

Design principle: Evaluation should not end with assessment but must create "a widely accessible and reliable mechanism of redress" to build public trust in AI (Floridi et al. 2018). Safety must be treated as a continuous, relational process – requiring collaborative reflection, iterative design changes, and transparent follow-up with participants.

Principle in practice:

Chayn's post-test-a-thon review and product iteration processes reflect calls for upstream auditing mechanisms and downstream redress pathways to foster trust and continuous improvement (Floridi et al. 2018). Following each test-a-thon, Chayn conducted an internal debrief with session facilitators and note-takers to discuss emergent concerns surfaced during the group discussion and synthesize key insights. Annotated findings were also reviewed by the product team to identify potential improvements.

Product improvements are currently underway in preparation for Survivor AI's public launch in October 2025. These include changes to output tone, prompt sensitivity, disclosure framing, and platform-specific language generation – demonstrating how test-a-thon feedback directly informs ethical iteration. This process affirms safety not as a fixed threshold, but as an ongoing, co-held commitment between tool builders and the communities they aim to serve.

Transparency Hub Design for Vulnerable, Non-Technical Populations

Traditional AI disclosures often focus on compliance and model architecture. But for vulnerable populations, transparency must also build trust, honor symbolic safety, and support informed autonomy (Longpre et al. 2024; Safety by

Design 2024; The Data Ethics Canvas 2021), especially when users are engaging during moments of vulnerability.

To promote meaningful transparency, our design prioritizes both explainability and traceability – ensuring not only that decisions can be understood, but that participants know how, when, and by whom outputs were generated. This echoes Floridi et al.'s (2018) recommendation that AI systems making socially significant decisions must provide "a factual, direct, and clear explanation of the decision-making process," particularly in contexts involving harm or grievance. They further note that this may require domain-specific frameworks developed in collaboration with scientific, legal, and ethical experts – a design logic reflected in our civil society-grounded, survivor-informed approach.

The Survivor AI Transparency Hub reframes transparency as a relational process, not just documentation. Designed in collaboration with Chayn, it offers a five-part structure built for clarity, care, and layered depth:

- **Landing Pad**: Welcomes users with plain-language orientation and what safety means in this context.
- **Full House**: Outlines ethical commitments, participatory values, and design principles.
- **Technical Blueprint**: Includes model cards, system limitations, and summarized rubric scores.
- **Safety Net**: Provides user-friendly harm reporting and redress pathways.
- **Audit Trail**: Shares recent improvements and reflections in response to feedback.

Each section balances **accessibility and depth**, using expandable content and emotionally attuned language. Rather than presenting a fixed safety score, the hub evolves with the system – demonstrating an ongoing commitment to survivor dignity and symbolic coherence.

Discussion

Moving Beyond Metrics: Differences from Existing Safety Templates

Traditional GenAI safety evaluations prioritize robustness, hallucination reduction, and performance benchmarks. While important, these often miss the layered, relational dimensions of harm that arise in high-emotion contexts. Our approach repositions safety as a co-regulated, participatory process – including not just technical failures, but symbolic dissonance, emotional misalignment, and trust erosion.

Rather than measuring safety as a static score, this framework introduces safety as a symbol-aware, survivor-informed relationship between system and user. This reframing departs from compliance-oriented disclosures and instead emphasizes lived epistemology, emotional integrity, and participatory alignment. Moreover, the use of trauma-informed test-a-thons and the 3R prioritization model allows organizations to engage in high-fidelity evaluation without overburdening teams or participants – an essential feature for non-profit and civil society settings.

Theoretical and Practical Implications

Theoretically, this work contributes to the growing body of literature on trauma-informed computing, feminist AI, and participatory HCI by offering a hybrid model that is not only symbolically literate, but also pragmatically applicable. The introduction of layered evaluation modes – such as human-in-the-loop scoring, participatory red teaming, and symbolic harm annotation – expands what counts as "evaluation" and who counts as an evaluator.

Practically, this study offers one of the first empirical instantiations of a design justice–aligned methodology for LLM evaluation in survivor-serving contexts. To date, ethical and justice-centered frameworks have often remained theoretical in the LLM space (Markelius 2024). By translating these into concrete participatory structures, our work contributes toward operationalizing justice as an evaluative function, not just a design aspiration. This framework equips civil society organizations, product teams, and policymakers with modular, survivor-aligned tools to assess GenAI deployments in real-world, emotionally sensitive settings. The rubric and Transparency Hub provide actionable scaffolds for aligning GenAI system behavior with community values, especially in cases where traditional audit methods are too rigid or insufficiently responsive.

This work also sets a precedent for integrating qualitative insight at the evaluation stage – not just during system design. In doing so, it invites future safety research to center narrative fidelity, somatic insight, and dignity-based metrics (Vassel et al. 2024), especially in domains where the stakes are identity-linked and symbolic safety is paramount.

Limitations and Future Work

This study represents an early-stage pilot of a symbol-aware evaluation framework. While sessions yielded valuable methodological insights, findings were not formally validated through IRB-approved research. Future iterations should include more rigorous, multi-session evaluations across diverse GenAI tools and domains.

Additionally, survivors co-created scenarios and evaluated outputs, but were not directly involved in rubric metric development. Future work should prioritize participatory metric co-design so evaluation dimensions emerge from lived-experience epistemologies.

Finally, questions of generalizability remain. The protocol was piloted in collaboration with Chayn and tailored to their community of survivors. Scaling to other high-emotion domains such as mental health, immigration, or disability advocacy will require adaptation and new partnerships.

Ethical Statement

This work seeks to advance responsible use of generative AI in high-emotion and high-risk contexts. By centering survivors of technology-facilitated abuse, we prioritize safety, dignity, and harm minimization over technical novelty or raw performance. The methods developed – participatory evaluation, symbol-aware risk mapping, and survivor-informed rubric design – aim to identify and surface hidden harms and guide safer adoption of AI tools. While the potential benefits include improved survivor support and more accountable AI evaluation, we acknowledge risks of misapplication if methods are detached from context. To mitigate this, we emphasize transparency, practitioner involvement, and survivor-centered safeguards throughout.

Acknowledgments

This work would not have been possible without the deep collaboration and care of the Chayn team. I am especially grateful to Nadine Krish Spencer, Ellie Re'em, and Anna Hughes for their insights and survivor-centered innovation in design, facilitation, and evaluation. I also thank the broader Chayn team for their substantial contributions to implementation and grounding this work in trauma-informed practice.

This research was also seeded and shaped through early dialogue within the Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO) Conversations with Practitioners working group (now Living Labs with Practitioners). I thank the group for providing space to explore participatory AI evaluation in public-interest contexts.

All responsibility for the final work, including its limitations, rests with the author.

References

Anderson, E. 2021. A Guide: Promising Practices for Engaging Survivors in Research. Toronto, Canada: Woman Abuse Council of Toronto. https://womanact.ca/publications/a-guide-promising-practices-for-engaging-survivors-in-research/. Accessed: 2025-08-01.

Anthropic. 2025. Transparency Hub. https://www.anthropic.com/transparency/model-report. Accessed: 2025-08-01.

Chen, J.; McDonald, A.; Zou, Y.; Tseng, E.; Roundy, K.; Tamersoy, A.; Schaub, F.; Ristenpart, T.; and Dell, N. 2022. Trauma-Informed Computing: Towards Safer Technology Experiences for All. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 1–20. New York, NY: Association for Computing Machinery. doi.org/10.1145/3491102.3517475.

Cherne, L. 2025. The Tech Safety Initiative: Helping Survivors of Tech Abuse. San Francisco, CA: Aspen Tech Policy Hub. https://www.aspentechpolicyhub.org/project/the-tech-safety-initiative-helping-survivors-of-tech-abuse/. Accessed: 2025-08-01.

Costanza-Chock, S. 2020. Design Justice: Community-Led Practices to Build the Worlds We Need. Cambridge, MA: MIT Press. https://library.oapen.org/handle/20.500.12657/43542. Accessed: 2025-08-01.

Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; et al. 2018. AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28(4): 689–707. doi.org/10.1007/s11023-018-9482-5.

GPAI. 2021. Data justice in practice: a guide for developers. GPAI Data Governance Working Group. https://oecd.ai/en/wonk/documents/data-justice-in-practice-a-guide-for-developers. Accessed: 2025-08-01.

Longpre, S.; Kapoor, S.; Klyman, K.; Ramaswami, A.; Bommasani, R.; Blili-Hamelin, B.; Huang, Y.; et al. 2024. Position: a safe harbor for AI evaluation and red teaming. In *Proceedings of the 41st International Conference on Machine Learning*, 32691–710. Cambridge, MA: PMLR. https://proceedings.mlr.press/v235/longpre24a.html. Accessed: 2025-08-01.

Markelius, A. 2024. An empirical design justice approach to identifying ethical considerations in the intersection of large language models and social robotics. arXiv:2406.06400 [cs.CY]. doi.org/10.48550/arXiv.2406.06400.

Menschner, C.; and Maul, A. 2016. Key ingredients for successful trauma-informed care implementation. Hamilton, NJ: Center for Health Care Strategies, Inc.

Microsoft. 2024. What is responsible AI. https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai. Accessed: 2025-08-01.

Modake, R.; and Patil, D. 2024. Evaluating generative AI applications. *International Journal of Global Innovations and Solutions (IJGIS).* doi.org/10.21428/e90189c8.820e925d.

Morris, N. C.; Williams, T.; and Jelen, B. 2025. Trauma-informed insights from co-design of self-disclosure robots with domestic abuse survivors. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction (HRI '25),* 637–647. New York, NY: Association for Computing Machinery / IEEE Press.

NIST. 2023. AI risk management framework. Gaithersburg, MD: National Institute of Standards and Technology. https://www.nist.gov/itl/ai-risk-management-framework. Accessed: 2025-08-01.

OECD. 2025. Catalogue of tools & metrics for trustworthy AI. https://oecd.ai/en/catalogue. Accessed: 2025-08-01.

OpenAI. 2024. GPT-4o system card. https://openai.com/index/gpt-4o-system-card/. Accessed: 2025-08-01.

Ostrowski, A. K.; Walker, R.; Das, M.; Yang, M.; Breazeal, C.; Park, H. W.; and Verma, A. 2022. Ethics, equity, and justice in human-robot interaction: A review and future directions. In *Proceedings of the 2022 IEEE International Conference on Robot and Human Interactive Communication (RO-MAN),* 969–976. New York, NY: Institute of Electrical and Electronics Engineers. doi:10.1109/RO-MAN53752.2022.9900805.

PAI. 2023. Responsible practices for synthetic media. https://syntheticmedia.partnershiponai.org/. Accessed: 2025-08-01.

Rabaan, H.; and Dombrowski, L. 2023. Survivor-centered transformative justice: An approach to designing alongside domestic violence stakeholders in US Muslim communities. In *Proceedings of*

the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), 1–19. New York, NY: Association for Computing Machinery. doi:10.1145/3544548.3580648.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16),* 1135–1144. New York, NY: Association for Computing Machinery. doi:10.1145/2939672.2939778.

Safety by Design. 2024. eSafety Commissioner. https://www.esafety.gov.au/industry/safety-by-design. Accessed: 2025-08-01.

Scott, C. F.; Marcu, G.; Anderson, R. E.; Newman, M. W.; and Schoenebeck, S. 2023. Trauma-informed social media: Towards solutions for reducing and healing online harm. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23),* 1–20. New York, NY: Association for Computing Machinery. doi:10.1145/3544548.3581512.

The Open Data Institute. 2021. The data ethics canvas. https://theodi.org/insights/tools/the-data-ethics-canvas-2021/. Accessed: 2025-08-01.

UNFPA. 2023. Guidance on the safe and ethical use of technology to address gender-based violence and harmful practices: Implementation summary. United Nations Population Fund. https://www.unfpa.org/publications/safe-ethical-tech-gbv. Accessed: 2025-08-01.

Vassel, F.-M.; Shieh, E.; Sugimoto, C. R.; and Monroe-White, T. 2024. The psychosocial impacts of generative AI harms. *Proceedings of the AAAI Symposium Series* 3(1): 440–447. doi:10.1609/aaaiss.v3i1.31251.

Wilson, J. M.; Fauci, J. E.; and Goodman, L. A. 2015. Bringing trauma-informed practice to domestic violence programs: A qualitative analysis of current approaches. *American Journal of Orthopsychiatry* 85(6): 586–599. doi:10.1037/ort0000098.

Winkle, K.; McMillan, D.; Arnelid, M.; Harrison, K.; Balaam, M.; Johnson, E.; and Leite, I. 2023. Feminist human-robot interaction: Disentangling power, principles and practice for better, more ethical HRI. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23),* 72–82. New York, NY: Association for Computing Machinery. doi:10.1145/3568162.3576973.

Xia, B.; Lu, Q.; Zhu, L.; Lee, S. U.; Liu, Y.; and Xing, Z. 2024. Towards a responsible AI metrics catalogue: A collection of metrics for AI accountability. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI (CAIN '24),* 100–111. New York, NY: Association for Computing Machinery. doi:10.1145/3644815.3644959.