Participant Takeaway (1 Page)

Who Defines Harm? Survivor and Cultural Expertise in the Age of AI

Ruth Reymundo Mandel — SafetyNexus AI | Safe & Together Institute

1) Core Problem (What's happening right now)

AI tools in child welfare, domestic violence, and justice systems are increasingly trained on survivor behaviour, institutional SOPs, and historical system data (case notes, removals, and past interventions). This can automate old harms when system logic becomes the model's logic.

2) Predictable Failure Modes (What the model learns)

- Protective action → "noncompliance"
- Help-seeking → "instability"
- Staying / returning → "high risk"

Result: Child danger is missed → removal becomes the default "safety" action.

3) Key Insight (For tech, design, ethics, and policy)

AI doesn't "find risk." It learns system priorities. What gets documented, labeled, and measured becomes what the system acts on. In high-stakes family decision-making, **accuracy** ≠ **safety** if the model optimizes institutional outcomes instead of real-world protection and stability.

4) The SafetyNexus Approach (The solution)

Credible Expert QA is a survivor- and culture-led framework for ethical AI development in govtech and social care. It functions as **continuous domain validation** by paid lived + cultural experts with systems literacy, integrated upstream into design, labeling/ontology, evaluation, and monitoring.

5) How Credible Expert QA is Different from "Normal" QA

Typical Human-Centered Design QA	Credible Expert QA (SafetyNexus)
Tests usability + comprehension	Detects system blind spots + failure modes early
Optimizes workflow + efficiency	Re-defines labels + signals (harm vs protection)
Feedback after prototypes	Produces testable harm scenarios (what to check + what to change)
Often advisory	Can stop / redirect harmful features before launch
Success measured by institutional KPIs	Success includes end-user impact + targeted community outcomes

6) Why Impact Feedback Matters

Institutions often measure success using policy-defined outputs (process compliance, throughput, system actions) rather than **end-user satisfaction** and **targeted community impact**. Credible Expert QA creates a continuous feedback loop that surfaces real-world harms and benefits early—and keeps tools aligned to safety, stability, and self-determination over time.

7) Five Questions to Take Back to Your Team

- What are we optimizing for: safety and stability, or system involvement and compliance?
- What are our labels really measuring: harm, or survivor constraint?
- Where are perpetrator patterns represented—or are they missing entirely?
- What cultural needs might the system misread or ignore?
- What will this tool cause workers to do: more removal and surveillance, or more protection and stability?

Closing Question

Who defines harm in your AI system—and who should?