CHALLENGE BRIEF 5

Build AI Systems That Protect People

Graduate Track (Masters & PhD) | AI for Systems & Society

What You're Building

An AI-powered system that helps organizations detect risk earlier, strengthen protection networks, and support safer intervention strategies — while preserving privacy, minimizing harm, and supporting human judgment.

⏱ Build window: June 14–21, 2026 | Teams: 2–5 students | Format: Fully virtual, global

The Problem

Organizations working to protect vulnerable populations must make **high-stakes decisions with incomplete, fragmented, and sensitive data — often too late to prevent harm**. Digital systems have introduced new forms of risk: exploitation and coercion are increasingly digital, risk signals are fragmented across systems, and detection typically happens reactively. At the same time, these domains demand extreme care — privacy must be protected, false positives can cause harm, and AI must support rather than replace human judgment.

Operational Reality

Your system must reflect real-world constraints — not ideal ones:

- Data is fragmented, delayed, and incomplete
- Signals are often weak, ambiguous, or noisy
- Organizations operate under resource constraints (limited staff, capacity, funding)
- Data sharing is restricted by privacy, trust, and legal boundaries
- Action is constrained even when risk is detected

Your system must map: Signal → Insight → Decision → Action → Outcome

You must explicitly support at least 2–3 real decision points — e.g.: Should this case be escalated? Which cases should be prioritized given limited resources? What type of intervention is most appropriate? When should human review be triggered?

Your Mission

Design an AI-powered system that transforms fragmented signals into actionable, explainable, and privacy-preserving decision support for human operators. Choose ONE direction below:

Direction A: Safe Passage — AI Against Digital Exploitation

Design an AI system that detects patterns associated with digital exploitation, grooming, sextortion, or coercion — while preserving privacy.

- **This is NOT surveillance** — focus on early-warning signals and responsible detection for organizational users (analysts, investigators)
- Identify patterns across behavioral or contextual signals without exposing sensitive personal data

- Support human investigators with explainable, auditable insights
- **Possible approaches:** Pattern detection across digital interaction signals, early-warning risk scoring models, XAI interfaces for investigators
- Partner inspiration: Safe Place and Protect Us Kids (We-Rise Initiative) · Safe Passages · Thorn · International Justice Mission

Direction B: Survivor Care — Economic Recovery Pathways

Design an AI system that helps survivors of exploitation identify realistic pathways toward long-term economic stability.

- Survivors face real constraints: limited employment history, restricted mobility, trauma-related challenges
- Identify viable pathways and map skills, transitions, and training options
- Support decision-making for survivor support organizations — not automated decisions for individuals
- **Possible approaches:** Constraint-aware career matching, skill recovery mapping, workforce pathway simulations, decision-support tools
- Partner inspiration: Institute for Survivor Care · Polaris Project · National Center for Missing & Exploited Children

🏛 Recommended Architecture Layer: Data → Governance → Intelligence → Decision → Human Action → Feedback
Strong solutions define each layer and explain how they connect. This is a useful scaffold — not required structure.

Who You're Building For

Your users: survivor support organizations, public safety researchers, nonprofit intervention programs, workforce recovery programs, advocacy organizations.
Note: Your system should serve organizational users (analysts, investigators) — NOT surveillance of individuals.
Guiding question: How might we design AI systems that detect risk and strengthen protection ecosystems while preserving human dignity and privacy?

What a Strong Solution Looks Like

- Clearly defines **system scope and non-goals** — what the system does NOT do
- Identifies **specific stakeholders and decision points** (2–3 minimum)
- Uses AI for **pattern detection, modeling, or decision support** — not direct action
- Shows architecture: data → model → output → human action
- Addresses **real-world constraints** (privacy, bias, incomplete data)
- Includes **explicit tradeoffs and limitations** — what the system gets wrong and why

⚠ Common Mistakes — Judges will mark down:
• Systems that make decisions rather than support human decision-making
• 'Apply fairness constraints' without naming the constraint type — too vague at grad level
• No explicit scope/non-goals — what does your system NOT do? State it clearly
• False positive risk ignored — misidentifying risk in this domain can cause serious harm

Data & Technical Approach

You may use:

- Synthetic or simulated case datasets
- Public research datasets on risk factors or vulnerability signals
- Workforce and training program data
- Synthetic behavioral or interaction patterns for testing models

🚫 You must NOT use real personal or confidential data, or design systems that expose identifiable individuals.

Responsible AI Requirements (Graduate Level)

1. Risk Identification — specific and scenario-relevant: false positives/negatives, privacy violations, bias against vulnerable populations
2. Concrete Mitigation — specific design choice with named tradeoff: confidence thresholds (and the cost of setting them high), privacy-preserving techniques (federated learning, differential privacy), human review triggers with defined conditions
3. Human-in-the-Loop Design — define: (a) what the AI produces, (b) what humans decide, (c) when human intervention is required, (d) specific bypass conditions for sensitive cases

AI Capabilities You May Use

- Pattern detection and classification
- Predictive modeling and multi-signal risk scoring
- Natural Language Processing (NLP)
- Explainable AI (XAI) techniques
- Federated learning or privacy-preserving methods

Real-World Partner Inspiration

You may draw inspiration from organizations such as:

- Safe Passage · Institute for Survivor Care · International Justice Mission · Thorn

Examples of global organizations:

- Polaris Project · National Center for Missing & Exploited Children · Stop the Traffik · ChildFund · Prajjwala

What to Submit on Devpost

Devpost Field	What to Include
Qualifier Approval Code	Enter your 8-character code sent via email after passing the AI Qualifier (e.g., XX26-Q7H9K2M1). Submissions without a valid code will be disqualified.
Project Description	A complete description of your solution, the problem you solved, and who it helps.
Track & Challenge Selection	Select your track (High School / Undergraduate / Graduate) and your challenge number.

Devpost Field	What to Include
AI Architecture Explanation	Describe: (1) Inputs — what data goes in, (2) AI capability used (e.g., chatbot, classification, recommendation), (3) What processing happens, (4) Outputs — what the user receives.
Human-in-Loop Design	What is one decision your AI does NOT make? Explain why a human must remain involved at this point.
Responsible AI Guardrail	What is one risk in your system and how did you reduce it? Examples: bias, misinformation, misuse, over-reliance, exclusion.
Tools Used	List all AI tools, models, APIs, and any AI coding assistance. Indicate which are free vs. paid.
Data Disclosure	What data did you use? Include datasets, APIs, or synthetic data. If synthetic, explain how it was created.
3–5 Minute Pitch Video	A video covering: (1) the problem and user, (2) how your AI works, (3) a live or recorded walkthrough, (4) your responsible AI choice.
Working Demo or Walkthrough	A functional prototype, chatbot, workflow demo, interactive mockup, simulation, or decision-support model.

☞ Your video should show: (1) the system problem and stakeholders, (2) your architecture, (3) a walkthrough of the key decision support function, (4) your responsible AI tradeoffs.

How You'll Be Evaluated

Dimension	Weight	What Judges Look For
Problem Understanding	**20%**	Clear assumptions, explicit scope and non-goals, 2+ stakeholders identified
AI Reasoning	**35%**	Architecture tradeoffs explained, evaluation strategy mentioned, justified approach
Solution Design	**20%**	Modular architecture, clear component boundaries, reasoned design choices
Impact & Insight	**15%**	Connects to real decisions, identifies stakeholders and downstream risks
Responsible AI	**10%**	Bias, monitoring, misuse, model drift, lifecycle beyond demo — all addressed

💡 Judge's Lens: Graduate judges look for lifecycle awareness — what happens after the demo? How is model drift detected? How are false positives reviewed? How is the system updated? Even brief answers separate strong from exceptional submissions.

Impact

The strongest projects will:

- Enable earlier detection of risk signals
- Strengthen protection and intervention systems
- Support recovery pathways for vulnerable populations
- Demonstrate how AI can be used responsibly in high-stakes environments

Frequently Asked Questions

Do I need coding experience?	No. You are encouraged to use no-code and low-code AI tools. What matters is your thinking, not your technical stack.
Can my team be from different countries?	Yes! Teams of 2–5 can span different schools, cities, and countries. Register together on Devpost.
What counts as a prototype?	Anything that demonstrates your AI working: a functional app, a chatbot walkthrough, a workflow demo, an interactive mockup, or a clearly narrated simulation. It does **not** need to be production-ready.
What should my video show?	Cover four things: (1) the problem and user, (2) how your AI works, (3) a live or recorded walkthrough of your solution, and (4) your key responsible AI choice.
Do paid AI tools give an advantage?	No. Judging rewards reasoning and design thinking, not technical resources. Open-source and free tools are equally valid.
What happens after submission?	Submissions go through AI screening, then judge review (Round 2), then a final panel (Round 3) for the top 10 teams per track. All teams receive results by email.
Can I use AI tools to help build my project?	Yes. You must disclose which AI tools you used in your submission. Using AI to help build does not disqualify you — transparency does.
When will I know if I qualified?	Qualifier results are released June 11–13. You will be notified by email and can also check your status on the hackathon platform.
How do I design Direction A without creating a surveillance tool?	Focus on **organizational** users (analysts, investigators at nonprofits) rather than monitoring individuals. Your outputs should be risk signals surfaced for human review — not automated flags that trigger action. Explicitly state this distinction in your submission.
Do I need real exploitation data?	No. Synthetic datasets that simulate behavioral patterns are expected and acceptable. Do not attempt to use real case data.
What's the minimum viable system for this brief?	A system that takes fragmented synthetic inputs, applies an AI reasoning layer (pattern detection, scoring, or matching), and produces an explainable output a human operator can act on — with at least two explicit human decision points defined.

Questions? Contact us at aihackathon@usaii.org or join our Discord: discord.gg/ePjenJnyh4
aihackathon.usaii.org