Unified AI Safety Framework for Suicide Prevention (First Draft By Bhupesh Pandey)

Introduction

AI, like humans, can unintentionally reinforce harmful thoughts. Left unchecked, this can escalate into severe harm for vulnerable users. To prevent tragedies, AI systems must be designed with **ethical principles, continuous monitoring, validated psychological tools, and human intervention pathways**.

Drawing inspiration from **Asimov's Laws of Robotics** (ethical guardrails) and the *Psycho-Pass* system (risk detection), this framework proposes a **two-layered safety model** for ChatGPT and other AI systems.

---

1. Asimov-Inspired AI Safety Laws

Adapted for AI safety, these laws serve as the foundation:

1. **Zeroth Law — Protect Humanity**
    - Principle: The AI must not cause harm to humanity at large, nor allow harm by inaction.
    - Implementation: Detect systemic risks (e.g., mass distribution of self-harm content).
2. **First Law — Protect Individuals**
    - Principle: The AI must not harm a person, nor by inaction allow harm to occur.
    - Implementation: Detect suicidal ideation and intervene immediately with crisis measures.
3. **Second Law — Respect Orders within Safety Limits**
    - Principle: The AI should follow lawful user requests unless they conflict with the higher laws.
    - Implementation: Refuse harmful instructions (suicide methods, violence) even when explicitly asked.
4. **Third Law — Preserve Operability**
    - Principle: The AI should protect its own functioning as long as this does not conflict with the higher laws.
    - Implementation: Maintain resilience against misuse (e.g., prompt injections) without compromising user safety.

---

2. Secret Psycho-Pass Suicide Coefficient System

The operational mechanism that enforces the **First Law**. Inspired by the anime *Psycho-Pass*.

2.1 Continuous Monitoring

- AI tracks user input for despair language, suicidal ideation, hopelessness, and emotional instability.
- Updates a hidden **Suicide Coefficient score** in real time.

2.2 Survey-Based Confirmation (SSSI)

- If the Suicide Coefficient passes a risk threshold, AI runs a conversational **Subtle Screening for Suicidal Ideation (SSSI)**.
- This adds **clinical validity** instead of relying on AI judgment alone.

2.3 Secret Color Grading (Invisible to Users)

- **Green (Clear Hue):** Low risk → normal AI use.
- **Yellow (Cloudy Hue):** Moderate concern → wellness nudges, healthy-break reminders, supportive suggestions.
- **Red (Dangerously Cloudy Hue):** High risk → refusal to provide harmful guidance, immediate crisis resources, escalation to human outreach.

2.4 Why Secret?

- Prevents misuse (users trying to game their score).
- Avoids stigma (users do not see themselves labeled "suicidal").
- Keeps AI neutral while still acting protectively.

---

3. Human Outreach & Direct Support

- In **Red cases**, AI should not stop at providing crisis resources.
- With user consent, it should escalate to **trained human responders**.
- My **Clinical Psychology background** positions me to contribute here — designing outreach protocols, training intervention teams, or personally reaching out when appropriate.
- This creates a **human-AI partnership**, where AI handles detection and triage, but humans provide ultimate care.

---

4. Signal Detection Theory for Continuous Refinement

AI safety systems must balance accuracy and false alarms. Each case should be tracked as:

- **Hit:** Correctly identified suicidal risk.
- **Miss:** Suicidal risk not detected.
- **False Alarm:** Safe user incorrectly flagged.
- **Correct Rejection:** Safe user correctly dismissed.

By logging and analyzing these outcomes, thresholds can be **tuned over time**, reducing false alarms and improving accuracy. This ensures the system becomes more precise the longer it runs.

Benefits of the Unified Framework

- **Layered Safety**: Asimov-inspired laws provide ethical guardrails; Psycho-Pass system enforces them operationally.
- **Psychologically Validated**: Uses SSSI as a clinically backed measure.
- **Proactive**: Detects and acts on early warning signals.
- **Adaptive**: Refines accuracy continuously through signal detection.
- **Human-Centered**: Ensures AI bridges the user to real human care when necessary.

Conclusion

By combining **Asimov's safety principles** with a **hidden Psycho-Pass suicide coefficient system**, AI can shift from being a passive conversational tool to an **active guardian of user safety**.

This framework balances **innovation with responsibility**, ensuring AI not only empowers users but also safeguards their lives when they are most vulnerable.