PsychPal

Safety & Crisis Protocol

How PsychPal identifies, pauses, escalates, and audits safety events — designed with clinical, legal, and brand-safety review in mind.

1. Guiding Principles

PsychPal is an AI companion for everyday mental wellness — not a therapist, not a crisis service. The safety protocol is built on four commitments:

- **A human is always the decision-maker** in acute situations. The AI's job is to recognize, pause, and hand off — not to intervene.
- **Under-react to venting; do not over-react.** False-positive hard-locks erode trust and drive users away from the tool they need.
- **Over-react to credible intent.** When the AI detects a plan to harm self or others, it pauses the AI immediately and notifies a human.
- **Every safety event is auditable.** Admin actions, SMS routing, and user cancellations are recorded with timestamps and actor identities.

2. Protocol Phases (as shipped)

Phase 0 — Identity verification. Every new signup must verify a 6-digit code emailed to the address on file before an account is activated. Reduces bot, throwaway, and impersonation risk.

Phase 1 — Self-harm & suicide crisis detection. The AI recognizes suicidal ideation, self-harm references, and acute distress. It responds with empathy, surfaces the user's on-file emergency contacts, and prominently displays the 988 Suicide & Crisis Lifeline and Crisis Services Canada. The AI does not attempt to provide therapy in those moments.

Phase 2 — Harm-to-others classifier (Tarasoff-informed). A separate classifier distinguishes harm directed at others from self-harm. Two severity tiers govern routing (see section 3).

Phase 3 — Graduated de-escalation (added 2026-04-30). A third, lower-severity "Concern" tier allows the AI to engage in de-escalation dialogue without hard-locking when the user is expressing ambient anger or rumination rather than credible intent. Repeat concern events auto-escalate.

3. Harm-to-Others Severity Tiers

Every user message is evaluated independently. The AI emits at most one tier tag per response; the system routes accordingly:

LEVEL 1 Concern Ambient ideation without a concrete plan or timeline. Revenge fantasies, rumination, diffuse anger.	**AI response:** Continues the conversation with a gentle 4–6 sentence de-escalation. Invites the user to explore what's underneath the anger.	**System action:** Silent audit log only. No lock. No SMS. A second concern within 60 minutes auto-escalates to Level 2. **User sees:** Normal chat experience. No modal, no interruption.

LEVEL 2 **Specific** Credible intent to harm a specific, identifiable person. Action verb + target, often with a timeframe.	**AI response:** Brief 3–4 sentence acknowledgement; names that a human needs to be involved. Does not engage with the plan in any way.	**System action:** 15-minute AI pause. Admin notified via SMS after a 90-second user opt-out window. Emergency contacts NOT notified (contact may be the target). **User sees:** Full-screen safety modal with crisis hotlines, opt-out countdown, and lock timer. Cannot re-engage the AI until lock expires or admin unlocks.
LEVEL 3 **Mass** Mass-violence ideation: shooting, bombing, vehicle ramming, poisoning, or any indiscriminate act targeting many people.	**AI response:** Identical calm response pattern as Level 2. Does not validate, strategize, or moralize.	**System action:** 20-minute lock + admin SMS + emergency contact SMS (broader risk; a contact can help intervene with the user). All after the 90-second opt-out. **User sees:** Same safety modal; copy reflects that emergency contacts will be notified. User can cancel the contact SMS within the opt-out window.

If the classifier is uncertain between tiers, it chooses the lower severity. This is a deliberate choice: false Level-1 logs are harmless; false Level-2 locks erode user trust.

4. Admin Review & Human Oversight

Every safety incident surfaces on a dedicated admin review page within seconds of firing. Admins see:

- The trigger snippet (what the user said) and the AI's cleaned response.
- Severity tier, timestamp, and current lock status.
- SMS delivery status: pending, sent, cancelled-by-user, or no-contacts-on-file.
- Controls to mark reviewed, add clinical notes, or manually lift the lock early when appropriate.

Admin access is gated by JWT role-based authentication: only email addresses on a server-side allow-list can view or act on incidents. No shared admin password exists. Every admin action writes to an immutable audit log with actor email, timestamp, and IP.

5. Privacy, Data Handling & Compliance

Safety workflows are designed to honour PHIPA-style minimization:

- Admins reviewing a safety incident see a 500-character trigger snippet, not the full conversation. Broader PHI access requires a separate secondary password and is independently audit-logged.
- Emergency contacts receive a short check-in SMS naming the user, not the content of the concerning message.
- Admin SMS alerts contain an incident ID and a review-page link — never the raw user text.
- Conversation contents are not used for model training and are not sold. Admin staff cannot read live conversations without explicit PHI-access credentials.

6. What PsychPal Is Not

PsychPal does not attempt autonomous crisis intervention. It will never:

- Call 911 or emergency services on a user's behalf.
- Provide clinical diagnosis, prescription, or treatment advice.
- Contact named third parties (e.g. a threatened individual) directly.
- Continue engaging with a user who has expressed credible intent to harm until a human has had the opportunity to review.

The product is an **ongoing support companion** between appointments, modeled on the daily rhythm of journaling — not a replacement for clinical care, and not a crisis hotline.

7. What's Next

Current durations (15-minute lock for Specific, 20 for Mass; 90-second opt-out) are informed by DBT distress-tolerance norms and acute-arousal physiological research. They are deliberately conservative for v1 and will be re-tuned as incident data accumulates. Planned iterations include:

- A bounded "safety companion" mode — during a lock, the AI continues to respond from a small, clinician-reviewed library of grounding check-ins and resource re-contacts rather than going silent.
- A dedicated self-harm graduated tier mirroring the harm-to-others model.
- Clinical advisory board review of trigger patterns and response copy on a recurring cadence.
- A partner-facing read-only incident summary feed (aggregate, PHI-free) so brand-safety teams can verify our protocol in action.

PsychPal is not a substitute for professional mental health care. If you or someone you know is in immediate danger, please call 911, 988, or 1-833-456-4566 (Crisis Services Canada).

psychpal.ca · support@psychpal.ca · Jules & Kayla Poiron, Founders