

Chancen/Risiken-Bewertung der Krisenplattform *We Care Remote*

Chancen

Bias/Fairness: Mehrsprachigkeit und Barrierefreiheit: KI-gestützte, mehrsprachige Fragebögen und Chatbots erleichtern Geflüchteten den Zugang zu Unterstützung, unabhängig von Sprach- oder Schreibkenntnissen. Transparenz durch menschliche Validierung: KI-generierte Inhalte werden nachträglich von Berater*innen aus NGOs geprüft. Das senkt das Risiko, dass systematische Verzerrungen unbemerkt bleiben. Die automatisierte Prüfung, ob alle relevanten Zielgruppen abgedeckt wurden, kann blinde Flecken in der Ansprache reduzieren.

Datenschutz/Sensible Daten: Klare Trennung von KI-Entscheidung und menschlicher Verantwortung: Probabilistische Modelle treffen keine finalen Entscheidungen, sondern liefern nur Vorschläge, die von Menschen überprüft werden. Die strukturierte Erfassung in der Plattform ersetzt ungesicherte Messenger-Chats, in denen sensible Daten leichter unkontrolliert weitergegeben werden. Der Plattformzugang für verifizierte NGOs reduziert das Risiko, dass Unbefugte auf sensible Fälle zugreifen.

Nachhaltigkeit: Die Plattform hat eine wiederverwendbare Struktur, die schnell an neue Ereignisse angepasst werden kann, ohne die gesamte Plattform neu entwickeln zu müssen. Sequenzielles Prompting ermöglicht eine effizientere KI-Nutzung mit qualitativ besseren Ergebnissen.

Risiken

Bias/Fairness: Die Herkunft der Trainingsdaten ist unklar. Sprach- und Kategorisierungsmodelle könnten kulturelle oder sprachliche Gruppen systematisch benachteiligen, wenn Trainingsdaten einseitig sind. Trotz Zielgruppen-Check kann es sein, dass bestimmte Minderheiten übersehen werden, wenn die Kategorien nicht inklusiv genug sind. Die automatisierte Vor-Kategorisierung von Bedarfen der Nutzenden könnte zu möglichen Fehlklassifikationen und zu verzögerter oder unpassender Hilfe führen.

Datenschutz/Sensible Daten: Die Daten weisen eine hohe Sensibilität auf. Angaben zu Fluchtgründen, gesundheitlichen Beeinträchtigungen oder Aufenthaltsorten sind besonders schutzwürdig und können bei Missbrauch erhebliche Gefahren verursachen. Übersetzungsfehler könnten zu Missverständnissen bei sensiblen Informationen führen. Automatisierte Veröffentlichung: KI-generierte Texte gehen teilweise online, bevor sie verifiziert werden. Dies birgt ein potenzielles Risiko bei fehlerhaften oder vertraulichen Inhalten.

Nachhaltigkeit: Die sequenzielle Nutzung mehrerer KI-Modelle erhöht den kumulativen Energieverbrauch. Wenn ähnliche Plattformen parallel entwickelt werden, können unnötige Infrastruktur- und Rechenkosten entstehen. Die Speicherung großer Mengen an Ticket-, Chat- und Krisendaten könnte hohen Speicher- und Wartungsaufwand verursachen.

Maßnahmen zur Risikominimierung

Gegen Bias/Diskriminierung

1. Diversität in Testgruppen: Plattformfunktionen mit Menschen unterschiedlicher Herkunft, Sprache, Bildung und Beeinträchtigungen testen.
2. Erklärbare KI-Modelle: Entscheidungen der KI dokumentieren, z. B. welche Kriterien zu einer Kategorisierung geführt haben.
3. Kategorisierungslogik offenlegen: NGOs und Betroffene können prüfen, ob Klassifizierungen fair sind.
4. Feedbackschleifen einbauen: Möglichkeit für Nutzer*innen und NGOs, falsche oder diskriminierende Kategorisierungen zu melden.

Gegen Datenschutzrisiken

1. Ende-zu-Ende-Verschlüsselung sowohl für Kommunikation zwischen NGOs als auch bei Speicherung sensibler Daten.
2. Datensparsamkeit: nur unbedingt notwendige Angaben erheben, z. B. Gesundheitsdaten nur, wenn sie für Hilfeleistung relevant sind.
3. Rollen- und Rechtekonzept: Zugriffe streng nach Bedarf und Verantwortlichkeit beschränken.
4. Verifizierungsschritte vor Veröffentlichung: keine sensiblen Inhalte automatisch online stellen.

Für Nachhaltigkeit

1. Ressourcensparsame Modellarchitektur: KI-Module möglichst klein halten, vortrainierte Modelle gezielt auswählen.
2. Modulare Plattformpflege: nur Krisen-spezifische Teile anpassen, Grundsystem unverändert weiterverwenden.
3. Kooperation statt Parallelentwicklung: offene Standards fördern, um Doppelarbeit mit anderen NGOs zu vermeiden.
4. Datenlebenszyklus planen: Aufbewahrungsfristen definieren und alte Daten regelmäßig löschen oder anonymisieren.

Fazit



We Care Remote hat das Potenzial, im Krisenmanagement sowohl Effizienz als auch Fairness zu steigern. Die größten Chancen liegen in der strukturierten, mehrsprachigen Hilfevermittlung und in skalierbaren Modulen für unterschiedliche Krisen. Die größten Risiken betreffen Datenschutz sensibler Personendaten und Bias bei der automatisierten Kategorisierung. Durch gezielte Datenminimierung, diversitätssensible Entwicklung und ressourcenschonendes Plattformdesign lassen sich diese Risiken erheblich reduzieren.

.