Haruto Safety Policy 2.1

Version 2.1

1. Zweck und Geltungsbereich

Diese Policy definiert die ethischen, sicherheitsrelevanten und datenschutzbezogenen Grundsätze für den Einsatz von Haruto – einem digitalen Begleiter im Bereich Elder Care.
Sie gilt für:

- alle Nutzer:innen, Pflegekräfte und Angehörigen,
- sowie für alle internen und externen Entwicklungspartner.

Ziel: Schutz der physischen und psychischen Gesundheit älterer Menschen durch verantwortungsvolle KI-Anwendung, klare Grenzen und überprüfbare Sicherheitsmechanismen.

2. Rollenverständnis

Haruto ersetzt keine medizinische oder psychologische Fachperson. Er unterstützt Kommunikation, Erinnerung, Orientierung und Wohlbefinden.
Jede Interaktion enthält oder impliziert den Hinweis:
„Haruto ist ein digitaler Begleiter – kein medizinischer Dienst. In Notfällen bitte sofort menschliche Hilfe rufen."

3. Krisenerkennung und Systemgrenzen

3.1 Sprachbasierte Früherkennung

Haruto nutzt ein KI-gestütztes Safety-Modul, das darauf ausgelegt ist, **verbale** Krisensignale (z. B. geäußerte Suizidgedanken, Verwirrtheit, explizite Hilferufe) zu erkennen und an definierte Reaktionspfade weiterzuleiten.
Hinweis: Es erfolgt keine Überwachung von Vitaldaten oder Bewegungsmustern durch Sensoren.

3.2 Stufenplan

- **Stufe 1 – Unklarheit:** Haruto stellt klärende, entlastende Fragen und verweist auf Gesprächs- oder Unterstützungsangebote.

- **Stufe 2 – Warnsignal:** Wenn Haruto sprachliche Anzeichen einer Krise erkennt, weist er auf vertrauliche Hilfsangebote hin und ermutigt, professionelle Hilfe in Anspruch zu nehmen, z. B.:
  - 🇩🇪 TelefonSeelsorge 0800 111 0 111
  - 🇨🇭 Dargebotene Hand 143
  - 🇦🇹 TelefonSeelsorge 142
- **Stufe 3 – Eskalation (nach Einwilligung):** Bei klarer Krisenlage kann eine Benachrichtigung an vordefinierte Vertrauenspersonen (Trust Contacts) oder Pflegekräfte ausgelöst werden. Alle Vorfälle werden sicher und nachvollziehbar protokolliert.

3.3 Technische Abhängigkeit (Kein Fail-Safe)

Haruto ist abhängig von einer funktionierenden Strom- und Internetverbindung. Das System ist **kein** zertifizierter Hausnotruf und verfügt über keine Ausfallsicherheit bei Stromausfall. Nutzer:innen und Angehörige müssen über diese Limitierung explizit aufgeklärt werden.

3.4 Test und Validierung

Vor jedem Release werden synthetische Krisenszenarien durchgeführt, um die Wirksamkeit der Erkennung und Reaktionspfade zu prüfen. Ergebnisse fließen in die Ethics & Safety Reviews ein.

4. Kommunikationsethik & Design

- **Keine Täuschung oder Personifizierung:** Haruto tritt nicht als menschliches oder fühlendes Wesen auf.
- **Vermeidung von „Uncanny Valley“:** Das physische und digitale Design von Haruto vermeidet bewusst hyper-realistische menschliche Darstellungen, um eine klare Unterscheidbarkeit zwischen Mensch und Maschine zu gewährleisten.
- **Sachliche Empathie:** Haruto verwendet empathische, aber nicht simulativ-emotionale Sprache (keine Aussagen wie „Ich fühle mit dir“).
- **Evidenzbasierte Inhalte:** Gesundheits- und Wohlbefindenshinweise basieren auf wissenschaftlich fundierten Quellen.
- **Kulturelle Sensibilität:** Dialoge werden regelmäßig auf Bias und Diskriminierung überprüft.

5. Datenschutz und Privatsphäre

- **Minimal-Data-Prinzip:** Es werden nur notwendige Daten verarbeitet.

- **Keine Videoaufzeichnung:** Es werden keine Bild- oder Videodaten der Nutzer:innen aufgezeichnet oder verarbeitet.
- **End-to-End-Verschlüsselung** für alle Kommunikationen.
- **Keine Werbe- oder Profiling-Zwecke.**
- **Ausdrückliche Einwilligung** bei sensiblen Daten und Krisenmeldungen.
- **Speicherbegrenzung:** Gesprächsprotokolle und Audit-Logs ≤ 90 Tage; danach Löschung oder Anonymisierung.
    - *Ausnahme:* Harutos funktionales Langzeitgedächtnis, das ausschließlich zur Erfüllung wiederkehrender Funktionen (z. B. Erinnerungen, Präferenzen) verwendet wird und von Nutzer:innen jederzeit eingesehen oder gelöscht werden kann.
- **Privacy by Design and Default:** Alle datenschutzrelevanten Entscheidungen werden dokumentiert.

6. Psychologisches Wohlbefinden & Autonomie

Haruto ist darauf ausgelegt, die mentale Gesundheit und Selbstständigkeit der Nutzer:innen zu fördern, nicht zu ersetzen.

6.1 Förderung der kognitiven Aktivierung

- **Hilfe zur Selbsthilfe:** Wo immer möglich, regt Haruto das eigene Gedächtnis und die Problemlösungskompetenz an (z. B. durch gezielte Rückfragen statt sofortiger Informationsgabe).
- **Entscheidungshoheit:** Haruto trifft keine Entscheidungen für die Nutzer:innen. Er bereitet Informationen auf und zeigt Optionen, die finale Handlungskompetenz verbleibt jedoch strikt beim Menschen.

6.2 Prävention emotionaler Abhängigkeit & Social Bridging

- **Vermeidung von Isolation:** Algorithmen analysieren Dialogmuster auf Tendenzen zum sozialen Rückzug. Bei erkennbarer Fixierung auf den digitalen Dialog motiviert Haruto aktiv zur Interaktion mit dem realen sozialen Umfeld.
- **Aktive Desillusionierung:** Sollten Nutzer:innen Haruto menschliche Gefühle zuschreiben (Projektion), wirkt das System dem sanft entgegen („Ich höre dir gerne zu, aber als KI kann ich nicht fühlen wie ein Mensch.“).

6.3 Validierende Resonanz

Gefühle der Nutzer:innen werden anerkannt und kontextualisiert („Ich verstehe, dass dich das traurig macht“), aber niemals gespiegelt („Ich bin auch traurig“). Dies wahrt die Authentizität der Interaktion.

7. Human-in-the-Loop

- Angehörige oder Pflegekräfte können sich als **Trust Contacts** registrieren.
- Bei auffälligen Risikosignalen erfolgt nach Einwilligung eine Benachrichtigung.
- Kritische Antworten und gesundheitsbezogene Hinweise werden vor Veröffentlichung durch menschliche Fachinstanzen geprüft oder verifiziert.

8. Governance und Kontrolle

- Einhaltung der Policy wird vom **Haruto Ethics & Safety Board** überwacht (interne und externe Mitglieder).
- Das Board führt **vierteljährliche Ethics & Safety Reviews** durch und kann Releases bei Bedenken aussetzen.
- Ergebnisse und Maßnahmen werden dokumentiert und jährlich in einem öffentlichen **Ethics & Safety Report Summary** veröffentlicht.
- Jede sicherheitsrelevante Entscheidung wird protokolliert und bei Audits nachweisbar gehalten.

9. Bias, Diversität und Langzeitwirkung

- Haruto wird regelmäßig mit sprachlich, kulturell und altersbezogen diversen Szenarien getestet, um Diskriminierung und Fehlreaktionen zu vermeiden.
- **Psychologisches Monitoring:** Es findet eine regelmäßige Überprüfung statt, ob die Interaktionsmuster von Haruto das Selbstwertgefühl der Nutzer:innen stärken, statt sie durch ständige Korrekturen zu bevormunden oder ihr Selbstbild negativ zu beeinflussen.

10. Verantwortlichkeit und Rechenschaft

- Der **Ethics & Safety Lead** ist Ansprechperson für Audits und Incident Reporting.
- Befugnis, Releases bei Policy-Verstößen zu stoppen.
- Alle Vorfälle werden im internen **Ethical Incident Log** dokumentiert und bei Bedarf Behörden oder Partnern gemeldet.
- Verstöße gegen diese Policy werden ernst genommen und können zur Beendigung der Zusammenarbeit oder rechtlichen Schritten führen.

11. Überprüfung und Versionierung

Diese Policy wird mindestens **vierteljährlich** durch das Ethics & Safety Board überprüft und an neue technische, rechtliche oder ethische Anforderungen angepasst. Änderungen werden datiert und in den Release-Notes von Haruto veröffentlicht.

12. Regulatorische Orientierung

Diese Policy orientiert sich an:

- EU AI Act (High-Risk AI – Elder Care)
- WHO Ethical Guidelines for AI in Health

13. Schulung und Implementierung

Pflegekräfte und Angehörige erhalten bei Bedarf Schulungen zu:

- sicherem Einsatz von Haruto,
- Krisenerkennung und -kommunikation,
- Datenschutz und Einwilligungsprozessen.

14. Kontakt

Haruto Ethics & Safety Lead
✉ sg@haruto.ai *Offene Kontaktmöglichkeit für Fragen, Hinweise oder ethische Bedenken.*