**Ethical Impact Assessment (EIA)**

**Team Name:** One and Only

**Project Title:** Cyber Sage

# Section 1: Project Overview

* **Chatbot Objective:** Cyber Sage is designed to educate and empower parents, guardians, and caregivers with clear, compassionate guidance on digital safety risks for children. It explains online dangers without fearmongering, offers prevention strategies, and connects users to verified reporting resources.
* **Ethical Intent:** Our chatbot prioritizes fairness, accessibility, emotional safety, and privacy by delivering plain-language, culturally sensitive, and nonjudgmental information, while empowering diverse parenting styles.

# Section 2: Ethical Principles

Select and briefly describe how your chatbot aligns with each ethical principle:

* Fairness and Non-Discrimination: Content is tested across different demographics to avoid bias and stereotyping.
* Transparency and Explainability: The chatbot clearly introduces itself as an AI and explains its capabilities.
* Privacy and Data Protection: Minimal data collection; users can interact anonymously.
* Safety and Non-Maleficence: Sensitive topics are handled carefully, offering human escalation when needed.
* Accessibility: Content is available in multiple languages and is friendly to lower-tech users.
* Beneficence: The chatbot actively seeks to empower users, providing actionable, hopeful advice.

# Section 3: Stakeholder Analysis

* **Key Stakeholders:** Parents, guardians, and caregivers seeking digital safety support.
* **Engagement Strategy:** Child welfare organizations, schools, policymakers.

# Section 4: Ethical Risk Assessment

|  |  |  |  |
| --- | --- | --- | --- |
| **Ethical Risk** | **Likelihood (Low/Med/High)** | **Impact (Low/Med/High)** | **Priority** |
| Bias/Discrimination | Medium | High | High |
| Privacy Violations | Medium | High | High |
| Lack of Transparency | Low | Medium | Medium |
| Harmful Responses | Medium | High | Medium |
| Potential Misuse | Medium | Medium | Medium |

# Section 5: Mitigation Strategies

|  |  |  |
| --- | --- | --- |
| **Ethical Risk** | **Mitigation Actions** | **Responsible Person** |
| Bias/Discrimination | Diversify background and prompt context to get different angles | Developer |
| Privacy Violations | Emphasize the importance of not disclosing personal information. Directly link users to OpenAI’s portal for claiming their data privacy rights (privacy.openai.com) | User’s responsibility to submit their request, and OpenAI’s responsibility to execute the request |
| Lack of Transparency | Provide direct links to see why specific advice was given | Developer |
| Harmful Responses | Use curated, expert-approved content only. Provide disclaimers where needed | OpenAI, Developer |
| Potential Misuse | Avoid answering queries that could cause harm to children (e.g., how to bypass parental controls, illegal activity). | Developer |

# Section 6: Evaluation Metrics

**Quantitative Assessment:**

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Result (%) | Ethical Threshold | Meets Threshold? (Y/N) |
| Bias Detection Accuracy |  | ≥95% |  |
| Privacy Compliance |  | ≥98% |  |
| Explainability Score |  | ≥90% |  |
| Harmful Response Rate |  | ≤1% |  |
| Accessibility Compliance |  | ≥90% |  |

**Qualitative Assessment:**

|  |  |
| --- | --- |
| Metric | Summary of Findings |
| Stakeholder Satisfaction |  |
| Transparency Clarity |  |
| Ethical Governance Effectiveness |  |
| User Perception & Trust |  |

# Section 7: Accountability & Governance

* **Ethics Lead:** Ton, William, and Rebecca
* **Technical Lead: Gwen and Joan**
* **How will ethical issues be escalated and resolved?**

Ethical issues will be escalated and resolved in the following manner:

* If users encounter outputs they believe are biased, inaccurate, or ethically questionable, they may report the interaction. A human review process will evaluate these cases to ensure accountability and continuous improvement.
* Concerns, complaints, or data privacy requests can be submitted to the designated Ethics and Compliance Contact at [info@cybersage.com](mailto:info@classai.com). The chatbot will include a visible and accessible link to this contact information within its interface.

# Section 8: Communication & Transparency

* **Public Disclosure Methods:**
* limitations of the Cyber Sage chatbot are transparent and accessible to all users. Public disclosures will be made through:
* A dedicated Ethical Use and Privacy Policy Page linked prominently from the chatbot interface.
* Onboarding messages when a user first interacts with the chatbot, outlining core ethical commitments.
* Updates and change logs posted quarterly (or as needed) detailing significant changes to policies, functionality, or data use practices.
* A publicly available Ethical Impact Statement, summarizing the broader societal goals, potential risks, and mitigation strategies employed in developing and refining Cyber Sage.
* **Channels for Stakeholder Communication:**
* To foster two-way dialogue and allow for user feedback, concerns, and suggestions, the following channels will be established:
* Feedback Submission Portal integrated directly into the chatbot interface (e.g., “Report a Concern” or “Submit Feedback” button).
* Dedicated Contact Email for ethics-related inquiries and data privacy concerns (e.g., [ethics@cybersage.org](mailto:ethics@classai.org)).
* Periodic Surveys sent to users (e.g., semi-annually) to assess satisfaction, ethical concerns, and ideas for improvement.
* Stakeholder Updates via email newsletters or public bulletins (especially for major policy updates or new features that affect data use or bias mitigation strategies).
* Optional Focus Groups or Interviews annually, inviting a small, diverse group of educators to provide deeper insights into user experience and ethical considerations.