BeAxa Moderation System Overview

1. Overview of BeAxa Moderation System

BeAxa is a real-time AI moderation service designed to improve the civility, clarity, and ethical coherence of online discussions--especially in democratic or collaborative group environments. It functions as an autonomous moderation engine that analyzes the content of chat messages, posts, or transcripts and applies moderation using two distinct tagging systems:

- RepTags: Tags associated with user avatars, reflecting their long-term ethical moderation history.
- ScripTags: Inline tags placed on specific words, phrases, or sentences that violate ethical norms, logical standards, or civility guidelines.

Only BeAxa can assign or modify tags. Neither users nor moderators can tag or remove tags, ensuring a fair and standardized moderation experience.

2. Why BeAxa is Feasible

Skeptical developers may wonder if such a real-time moderation system is technically achievable. Here's why it is:

- Natural Language Processing (NLP): BeAxa builds on GPT-class large language models fine-tuned for ethical reasoning, contextual tone analysis, and logical structure assessment.
- Moderation Libraries: A curated library of ethical principles, fallacies, rhetorical dodges, and discriminatory patterns enables precise tagging.
- Inline Tag Engine: Token-level tagging is already supported in modern AI pipelines, and BeAxa restricts tagging to high-confidence outputs.
- Tag Confidence Thresholding: BeAxa avoids spurious tags by showing only those above a platform-defined confidence level.
- Training Feedback Loop: Misclassifications can be flagged by moderators or system admins to fine-tune local deployments.

3. Live Session Workflow

In live group discussions (e.g., in GroupBuild), BeAxa functions as a silent but active bot:

- Scans real-time content.
- Applies inline ScripTags when violations are detected.
- Updates user RepTags if patterns emerge over time.
- Enforces feedback such as real-time warnings, correction prompts, or cooldowns.

BeAxa does not censor ideas--it flags content that violates group or platform norms while preserving transparency for future readers.

4. Session Monitoring Interface

The Session Monitoring Interface allows platform administrators or developers to:

- View live tag distribution across a session.
- Monitor how BeAxa is scoring and interpreting user behavior.
- See aggregate ethical analytics over time.
- Track RepTag changes and flag repeat offenders.
- Export anonymized tag and session logs for platform governance.

This feature supports transparency, trust in the AI, and research into digital civility.

5. RepTags (Reputation Tags)

- Attached to a user's avatar pseudonym.
- Calculated based on the frequency and severity of BeAxa's interventions.
- Editable only by Admins.
- Categories may include: Respectful, Borderline, Chaotic, Manipulative, Misinformer, or Troll-Risk.

RepTags:

- Indicate trustworthiness and behavioral trend.
- Affect the user's influence weight in community feedback.

- Do not directly restrict participation but shape how feedback from a user is prioritized.

6. ScripTags (Inline Content Tags)
- Placed directly on problematic text (word, phrase, sentence).
- Drawn from a standardized moderation tag library.
- Examples:
  - Manipulative
  - Discriminatory
  - Illogical
  - Unprofessional
  - Misleading
  - Rhetorical Dodge
- Cannot be removed or altered by users or moderators.
- Appear in real-time, optionally with hover-to-learn features.

7. ScripTag Confidence and Severity
- Each tag is paired with a confidence level.
- Tags below threshold are logged but not shown.
- Severity is indicated by number overlays:
  - 1 - Mild
  - 3 - Moderate
  - 5 - Critical

Up to 10 levels of severity are supported. Only up to 2 ScripTags appear per violation.

8. Inline Iconography System
To improve readability and minimize cognitive load:
- Tags use small inline emoji-style icons.
- Overlayed digits show severity (1-10).

Examples:
- 1 - Slightly flawed logic.
- 5 - Deep contradiction.
- 5 - High-severity discriminatory content.
- 3 - Mid-level manipulation.

9. Tag Auditing and Feedback Loop
- All tags are logged with timestamp, Group ID, Pseudonym, and content hash.
- Admins can export logs.
- Admin/moderator feedback allows refinement of local BeAxa models.
- Periodic retraining aligns BeAxa with evolving norms.

10. Tag Customization Per Platform
BeAxa is configurable by platform:
- Select which tag types are shown.
- Change hover language or tone.
- Add custom platform-specific ScripTags.
- Adjust severity thresholds.

11. Future Implementation Steps
1. MVP Demo Launch
   - Embedded BeAxa bot in GroupBuild (non-profit) chat.
   - Demonstrate live RepTag and ScripTag functionality.

BeAxa Moderation System Overview

2. Tagging Library Finalization
   - Refine tag types and visual language.
3. Documentation and Developer SDK
   - Exportable logs, tag review panels, API docs.
4. Optional Integration with Other Platforms
   - Via licensing or SDK.
5. Brand Identity Finalization
   - Including iconography for RepTags and ScripTags.

12. Licensing and Business Model (Preview)
- Free Tier: Initial public test version.
- Licensed Tier: For integration into private or public platforms.
- Subscription Tier: For use at BeAxa.com as a standalone tool.
More details TBD in future versions.