How does the app filter and select what can be posted and what can not?

bhapi has a public posting area, where 'memos' (short for memories)can be shared. 'Memories' is a public place where content can be shared. The app also has a private space where users can message people directly. In the public place it is moderated to a 'G' level as set out by the MPAA. If users post something that is not 'G' rated it goes into the users private area as a draft and can be shared via your private messages. If users feel there is an error in their memo or another memo they can reach our moderators who will review and give feedback on the request. A global independent panel is planned to be set up to provide governance and escalation for the moderators.

Could you please specify what content bhapi filters?

All memos (public space) are moderated to MPAA standard of 'G' rating.

Including but not exclusively :

- Cyberbullying - tends to be digital bullying from people who are anonymous but know you.
- Trolling - tends to be harassment from people who are anonymous but who don't know you.
- Hate speech or discrimination of any kind.
- Disinformation, clearly false or misleading information.
- Sexual harassment - the unwanted sexual behaviour where a reasonable person would have anticipated that the person harassed would feel offended, humiliated or intimidated. It has nothing to do with consensual behaviour.
- Catfishing & grooming - deceptive activity where a person creates a fictional persona or fake identity on a social media or gaming network usually targeting a specific victim. Groom is where targeting is specifically of a minor with pedophilic intent.
- Sextortion - a form of blackmail that involves threatening to share digital content usually images or video unless they comply with certain demands normally sexual or violent in nature.
- Doxing - The intentonal exposing of individuals identity , private information and personal data without consent to cause damage or embarrassment.
- Deep fakes - A form of artificial doxing where perperetrators create compromising images and videos of individuals to cause damage or embarrassment.
- Illegal content - This can include images and videos of child sexual abuse, content that advocates terrorist acts, content that promotes, incites or instructs in crime or violence, footage of real violence, cruelty, criminal activity & animal cruelty.
- Pornographic or sexual content
- Violent content - Content glorifying or trivialising violence.
- Self Harm and suicidal content - Content that encourages self harm and suicidal behaviour.

- Drug related content - Content relating to purchase, consumption, manufacturer and distribution of illegal or prescription drugs.

What does bhapi classify as '' Toxic Content''?

Including but not exclusively the following:

- Cyberbullying - tends to be digital bullying from people who are anonymous but know you.
- Trolling - tends to be harassment from people who are anonymous but who don't know you.
- Hate speech, sexism, racism and discrimination of any kind.
- Disinformation, clearly false or misleading information.
- Sexual harassment - the unwanted sexual behaviour where a reasonable person would have anticipated that the person harassed would feel offended, humiliated or intimidated. It has nothing to do with consensual behaviour.
- Catfishing & grooming - deceptive activity where a person creates a fictional persona or fake identity on a social media or gaming network usually targeting a specific victim. Groom is where targeting is specifically of a minor with pedophilic intent.
- Sextortion - a form of blackmail that involves threatening to share digital content usually images or video unless they comply with certain demands normally sexual or violent in nature.
- Doxing - The intentonal exposing of individuals identity , private information and personal data without consent to cause damage or embarrassment.
- Deep fakes - A form of artificial doxing where perperetrators create compromising images and videos of individuals to cause damage or embarrassment.
- Illegal content - This can include images and videos of child sexual abuse, content that advocates terrorist acts, content that promotes, incites or instructs in crime or violence, footage of real violence, cruelty, criminal activity & animal cruelty.
- Pornographic or sexualised content
- Violent content - Content glorifying or trivialising violence.
- Self Harm and suicidal content - Content that encourages self harm and suicidal behaviour
- Drug related content - Content relating to purchase, consumption, manufacturer and distribution of illegal or prescription drugs.

What is the tech bhapi uses to filter content in real time? Would this tech support millions of users?

Memories are not real time but near time this allows for AI process and human moderator intervention. We use several different AI models that are layered that are supported by moderators who monitor our models and adjust them for feedback from our users so bhapi AI algo are continually improving with feedback. We currently use the Perspective API and are in

the process of integrating our own customised versions of Detoxify and Moderator. Our technical approach is fully cloud native and scalable by design.

Why do you affirm this is a g-rated social media channel? How can you ensure it will be safe for kids?

- Clear policy implementation with AI models and moderators.
- Near time not real time public posting.
- Full COPPA compliance (in progress, currently support 13+ years only, once COPPA compliant we will be available for under 13 year olds.)