Safe Online Spaces Using Meronymity & Trust-Based Human Moderation

Nouran Soliman - nouran@mit.edu, Farnaz Jahanbakhsh - farnazj@mit.edu & David Karger - karger@mit.edu

Motivation & Objective

Enhancing Online Spaces for Safety and Free Expression: Our research is dedicated to pioneering new design paradigms and computing systems that bolster safety and trust online, aiming to balance the essentials of free speech with the need for secure environments. Our focused objectives include:

- **Meronymity in Digital Interactions:** We are redefining digital communication by enabling selective identity disclosure through meronymity, offering credibility and protection while fostering open engagement. Proven effective in academic settings, this approach promises wider benefits by reducing social anxieties and facilitating easier exchanges.

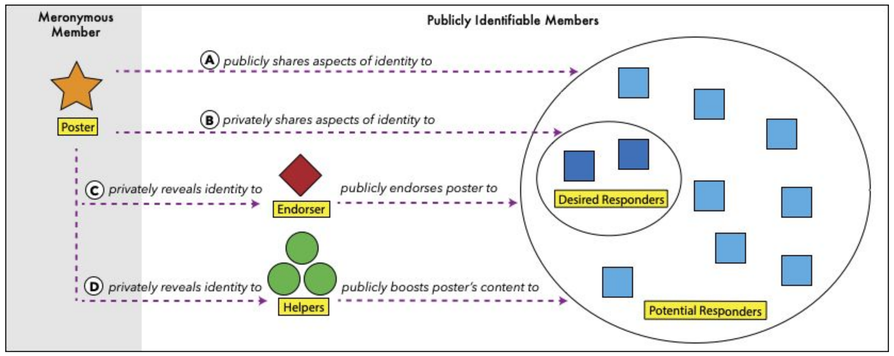
- **Extending Meronymity Across Social Media:** We seek to apply meronymous communication broadly, particularly on social media, to support nuanced dialogues and aid minority and vulnerable groups. A crucial aspect of this expansion is authenticating identity elements through human verification to maintain trust.

- **User-Centric Trust-Based Moderation:** At the core of our approach is a trust-based moderation system, empowering users to customize their online spaces to their comfort levels, thereby making content moderation more relevant and individualized.

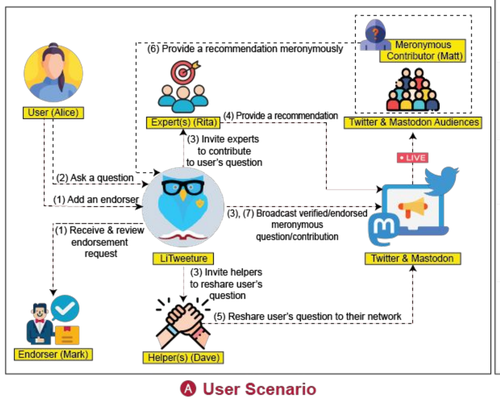
- **AI and LLMs for Enhanced Interactions:** The exploration of leveraging artificial intelligence and large language models raises vital questions about their integration into content moderation, with the goal of achieving personalized moderation that respects user nuances.

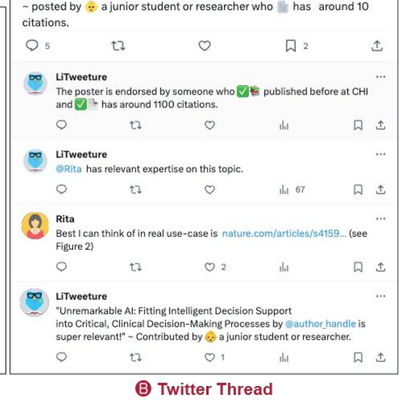
Through these endeavors, we aim to create online platforms that mirror the complexity of human interaction, where the principles of safety, trust, and open dialogue intertwine, ensuring every participant feels safe, heard, and understood.

Meronymity



Meronymity Model & Design. This diagram delineates the roles and interactions among five key stakeholders. The **'Poster'** initiates communication for assistance, choosing varying anonymity levels when sharing identity: **(A)** In full or partial using aspects of identity (meronym) publicly to Potential Responders, **(B)** In full or partial using personalized aspects of identity privately to Desired Responders for targeted communication, **(C)** In full privately to an Endorser for public endorsement, leveraging the Endorser's identity to enhance credibility, **(D)** In full privately to Helpers who amplify the content to Potential Responders.

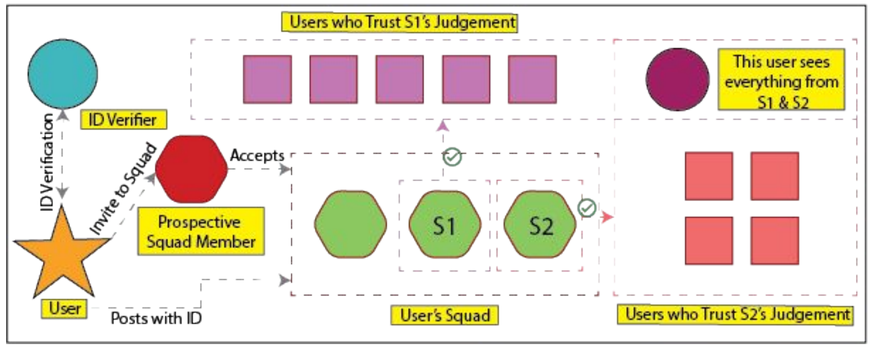




A User Scenario B Twitter Thread

Example User Scenario on Asking For Paper Recommendations Using LiTweeture. Part A illustrates the steps of the user scenario. Alice requests Mark as her Endorser who then approves the request **(A-1)**. Alice composes a question with a meronym, enlists Expert Rita with a personalized meronym, and enlists Helper Dave **(A-2)**. Alice's question is posted with a meronym on Twitter & Mastodon on LiTweeture's accounts **(A-3)**, Rita is privately messaged with a meronym **(A-3)**, and Dave is privately messaged **(A-3)**. Rita then responds publicly on Twitter **(A-4)**. Dave then reshares Alice's question to his Twitter network **(A-5)**. Matt, a potential responder, encounters Alice's question and provides a recommendation meronymously **(A-6)**. After Alice's moderation, his meronymous contribution is posted publicly to Twitter & Mastodon **(A-7)**. Part B illustrates the resulting Twitter thread from this user scenario.

Trust-Based Human Moderation



Initial Design of Trust-Based Human Moderation. The **User** verifies their identity signals through a **human ID Verifier**. The User then builds their **Squad**, a number of Users who see all the posts made by the User with their real Identity regardless of level of anonymity selected by the User. The Squad members can mark the User's post as appropriate or just not interact with it. If marked appropriate by a Squad member, the post becomes visible to other users who chose to **Trust** the judgement of this Squad Member. This repeats between users to propagate content across the network based on **trust in alignment of judgement**.

Using AI To Facilitate Personalized Content Moderation

Leveraging AI, particularly large language models **(LLMs)**, in personalized moderation presents a promising avenue for enhancing online interactions, aiming to utilize AI's potential while minimizing unintended consequences. It emphasizes the importance of **aligning AI with user-defined safety and responsibility standards**, and calls for robust oversight mechanisms. The notion extends to the broader implications of generative AI in social media, focusing on how it can support creators and provide personalized user experiences, without directly involving user-generated content (UGC). Key considerations include ensuring AI's alignment with human values, maintaining transparency and accountability, addressing potential biases, and understanding the societal impacts of these technologies. Some interesting questions around this are:

- How can oversight frameworks be structured to guarantee the responsible deployment of LLMs in personalized moderation?
- What methods can allow users to specify their criteria for 'safe' and 'responsible' AI-driven moderation?
- How will the integration of LLMs in moderation processes impact online community dynamics and user engagement?