Principles for AI-Assisted Social Influence and Their Application to Social Mediation

Ian Perera[1], Alex Memory[2], Vera A. Kazakova[1], Bonnie J. Dorr[3], Brodie Mather[1], Ritwik Bose[2], Arash Mahyari[1], Corey Lofdahl[4], Mack S. Blackburn[4], Archna Bhatia[1], Brandon Patterson[1], Peter Pirolli[1]

[1]Florida Institute for Human and Machine Cognition, Ocala, FL
[2]Johns Hopkins University Applied Physics Laboratory, Laurel, MD
[3]University of Florida, Gainesville, FL
[4]Leidos, Inc., Reston, VA

{iperera@ihmc.org, Alex.Memory@jhuapl.edu, vkazakova@ihmc.org, bonniejdorr@ufl.edu, bmather@ihmc.org, Rik.Bose@jhuapl.edu, amahyari@ihmc.org, Corey.Lofdahl@leidos.com, Mack.Blackburn@leidos.com, abhatia@ihmc.org, bpatterson@ihmc.org, ppirolli@ihmc.org}

Abstract

Successful social influence, whether at individual or community levels, requires expertise and care in several dimensions of communication: understanding of emotions, beliefs, and values; transparency; and context-aware behavior shaping. Based on our experience in identifying mediation needs in social media and engaging with moderators and users, we developed a set of principles that we believe social influence systems should adhere to to ensure ethical operation, effectiveness, widespread adoption, and trust by users on both sides of the engagement of influence. We demonstrate these principles in D-ESC: Dialogue Assistant for Engaging in Social-Cybermediation, in the context of AI-assisted social media mediation, a newer paradigm of automatic moderation that responds to unique and changing communities while engendering and maintaining trust in users, moderators, and platform-holders. Through this case study, we identify opportunities for our principles to guide future systems towards greater opportunities for positive social change.

1 Introduction

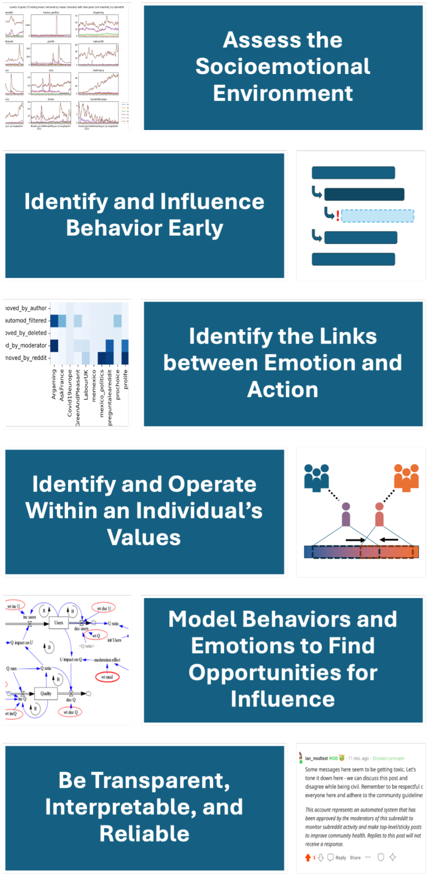
AI systems for social influence in communications are often viewed with suspicion, especially when they exert social influence explicitly, which can be seen as potentially malicious. While AI is increasingly used in social influence, ethical guidelines and principles typically focus on philosophical perspectives for black-box systems, rather than providing practical guidance for ethical methods and implementations (Zhou et al., 2020). We believe that effective and responsible social influence systems require not only oversight, but awareness of the socioemotional landscape and transparent models based on that landscape.

We consider one target domain as moderation or mediation on social media platforms, where AI-based approaches are often embedded in a socioemotional context, but lack direct engagement with user emotions. Moderation typically relies on categorical rules such as “No personal attacks” or “No racial slurs”, which fail to address shifts in community tone, focus, and overall health of discussion. Communities can radicalize over time through interactions that to not explicitly violate community rules. Additionally, what is considered harmful or disruptive can evolve (dos Santos et al., 2024), influenced by factors such as a user’s platform history (Cheng et al., 2021), requiring more adaptable and holistic mediation strategies.

Ethical AI-assisted social influence is a nuanced and challenging problem, especially in this domain. Maintaining community health may require limiting user freedoms, which can foster perceived censorship and contribute to radicalization. Effectively addressing undesirable behavior requires understanding both disinformation tactics and individual responses to communication from others. Community health is also dynamic, requiring ongoing adaptation even within a single community.

We believe there are a set of guiding principles that can provide a guiding framework for tackling these and other challenging domains in the realm of social influence systems. These principles are shown in Figure 1 and were developed by building on prior work in sociolinguistics, psychology, and social cybersecurity, and then incorporating lessons learned from designing and deploying our work with feedback from moderators.

Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024), pages 129–140
November 16, 2024 ©2024 Association for Computational Linguistics


Assess the Socioemotional Environment
Identify and Influence Behavior Early
Identify the Links between Emotion and Action
Identify and Operate Within an Individual's Values
Model Behaviors and Emotions to Find Opportunities for Influence
Be Transparent, Interpretable, and Reliable

ian_modtest MOD · 11 mo. ago · Stickied comment

Some messages here seem to be getting toxic. Let's tone it down here - we can discuss this post and disagree while being civil. Remember to be respectful c everyone here and adhere to the community guideline:

This account represents an automated system that has been approved by the moderators of this subreddit to monitor subreddit activity and make top-level/sticky posts to improve community health. Replies to this post will not receive a response.

Figure 1: The principles we posit as enabling ethical and effective social influence systems in complex socioemotional environments.

Social media serves as a valuable example for broader social influence dynamics, as complex interactions can occur in various emotion-laden contexts such as negotiation, decision-making, disaster relief, or patient-caregiver interactions. We introduce D-ESC (Dialogue Assistant for Engaging in Social-Cybermediation) as a case study for comprehensive social influence systems.

A key contribution of D-ESC is its inclusion of multiple components that facilitate positive social influence by detecting and addressing potentially undesirable behaviors in social media communities. Through a mixture of tested natural language processing (NLP) techniques for emotion and sentiment detection and novel topic-based, stance-based (Mather et al., 2021), soft logic, and generative approaches, D-ESC analyzes emotional dynamics, generates deescalation responses while adhering to community guidelines, and provides explainable predictions for those responses. These components form a framework for analyzing, modeling, and influencing communities with human oversight, enabling exploration of potential interventions.

We briefly cover prior social influence work that could benefit from integrating fundamental principles of social influence in complex socioemotional environments, then highlight these opportunities specifically in the domain of social media moderation. We then outline core principles we adhered to in developing D-ESC, and describe their implementation. Finally, we cover potential future applications that demonstrate the broad applicability of this case study beyond social media mediation.

2 Prior Work

Existing AI-assisted social influence work includes systems for improving attitudes and communicative behavior (Anastasiou and De Liddo, 2023), persuading users to give to charitable organizations (Tran et al., 2022), and safeguarding online communities through early identification of antisocial users (Cheng et al., 2021). Emotional awareness has been used to produce prosocial responses to other individuals' statements of negative emotions (Zhao et al., 2023). Other efforts analyze human social influence techniques in social media (Tan et al., 2016), suggest less inflammatory language in the form of paraphrases (Som et al., 2024), and position AI as a "moral crumple zone" to protect human relationships by taking the blame for failed communications (Hohenstein and Jung, 2020).

In social media, AI-assisted social influence efforts typically follow a binary approach: either remove posted content or allow it to remain (Diaz and Hecht-Felella, 2021). Some platforms implement zero-tolerance policies with explicit rules against prohibited content, hate speech, harassment, violence, or other harmful or illegal content (Facebook, November, 2022; Twitter, March, 2023; YouTube, 2019). However, these methods may not effectively address the complex dynamics of online communities, overlooking the cumulative impact of interactions (Massanari, 2017; Suler, 2004). Additionally, adopting machine learning for scalable detection, as reviewed by Balayn et al. (2021), has led to manual moderation in response to errors, raising concerns of discrimination, as some populations are more frequently mis-classified than others.

Automated mediation differs from automated moderation in its shift from removing problematic content to fostering dispute resolution and promoting *civil discourse* within online communities. Automated mediation systems are still in early research stages, for example, within the legal field

(Roos, 2023; Bergman, 2023). More recent work has explored Large Language Model approaches for social mediation (Cho et al., 2024) and political discourse (Argyle et al., 2023), but these methods do not posit methods for awareness of community dynamics prior to influence. Our work aims to influence users to engage more positively in their community, and to encourage moderators to support the community rather than just enforcing rules.

3 Core Principles of Social Influence Systems

In developing social media mediation systems, engaging with potential users, surveying related work, and considering future applications, we have established a set of core principles for successful social influence systems. We define "successful" in terms of high scores for a) adoption (likelihood of use), b) trust (perceived as a positive, trustworthy agent), c) effectiveness (achieving intended social influence), and d) alignment (mirroring human social influence skills). Whereas prior work has focused on task-specific metrics such as *influence outcome* or *partner perception* (Chawla et al., 2023), our metrics are intended to apply more broadly—at both individual and societal levels—capturing key dimensions deemed crucial for the role social influence systems in society.

We begin by outlining our proposed principles of social influence system design, and then describe how we work to achieve these principles in our D-ESC mediation system.

3.1 Assess the socioemotional environment

A key part of social influence is understanding the social context and the emotions that driving behavior. The focus must be on both expressed emotions and the underlying emotions that lead to them. Many prior systems use sentiment as a proxy for understanding emotion, but sentiment measures often poorly reflect emotional state (Nandwani and Verma, 2021). In social media communities, even pro-social interactions can include profanity and insults that build camaraderie, while polite or formal interactions may reduce emotional support or downplay justified outrage. Even in information-centric communities, users may want to see social and emotional support as part of the community values (Worrall et al., 2021).

Upon reviewing D-ESC, social media moderators have expressed concerns that an AI system might not fully understand the context of interactions. Moderators often witness false positives in automated toxicity detection and are hesitant to adopt similar tools due to AI's lack of socioemotional awareness. In emotion regulation agents, users often complain that the agent does not seem like a good listener or does not consider the specific situation (Hopman et al., 2023). These examples demonstrate that understanding the socioemotional environment is key for adoption and effectiveness.

3.2 Identify and influence behavior early

Behavior change becomes harder once it is habitual, making early interventions crucial for influencing behavior before it turns harmful. Caught early, social media users are more open to rephrasing their posts constructively. Caught late, with options like removal or banning, users may feel unfairly targeted or suppressed. Historical data across CNN, IGN, and Breitbart show that banned users often posted more and elicit more replies, disproportionately affecting their communities (Cheng et al., 2021). Their antisocial behavior worsens over time, underscoring the need for early detection and moderation.

Additionally, users may be unaware that their behavior could lead to negative responses or be perceived as inflammatory. Prior work on computational modeling of polarization on social media suggests that early feedback is essential to prevent extreme polarization and skepticism of alternative views (Lim and Bentley, 2022). Early engagement with the user offers the greatest potential and the most options for influence.

3.3 Identify links between emotion and action

Online interactions have been characterized as *intentional social actions*, with social and individual antecedents, as well as online and offline consequences (Richard P. Bagozzi and Pearo, 2007). Although social influence applications focus on behavior, emotions often precede these behaviors. Thus, considering and potentially influencing emotions may be more effective than targeting behaviors directly, as seen in the BEND framework (Carley, 2020), which targets human biases and emotions to achieve behavioral change.

Emotions can also be weaponized by users, such as trolling, even when not overtly violating community norms. This disrupts healthy communication and can even lure threads or entire communities into degenerative and polarized discussions through asymmetrical responses, such as: ignoring, challenging, or inflaming others (Paakki et al., 2021). While overt antisocial behaviors may be easy to identify, covert antisocial behaviors require more nuanced strategies focusing on emotional triggers (Hardaker, 2013). Understanding these dynamics allows a social influence system to more effectively influence users and adapt to external and individual factors that may affect the influence.

3.4 Identify an individual's values and operate within them when possible

Interactions within a system must feel relevant and valuable to users. Thus, AI-based social influence systems must be grounded in human social and emotional concepts. People are most influenced by appeals to their own experiences and ideals, so AI systems need a consistent base of support for the perspectives they present. Social Judgment theory (Sherif and Hovland, 1961) explains how individuals respond to information in relation to their existing attitudes, dividing ideas into acceptable, indifferent, or unacceptable. Aiming for vast changes in attitudes or behavior can lead to ineffective interventions at best, and harmful reactions at worst, as disproportionate moderations can exacerbate undesirable behaviors (Cheng et al., 2021).

3.5 Model behaviors and emotions, and find opportunities for influence

All AI-assisted social influence applications use models, but these are often predictive without a causal hypothesis, limiting opportunities for effective influence. In social media, individuals and communities are often seen as unchanging, labeled as "good," "bad," "toxic," or "positive." Even when social influence is the goal, models typically train on population data, learning general strategies, rather than tailoring to individual attitudes and personalities. An effective social influence system not only predicts responses to specific actions, but also uses a causal model to explore individualized strategies for influencing dynamic users.

3.6 Be transparent, interpretable, and reliable

Key barriers to AI adoption include lack of transparency, questionable output interpretability, and resulting distrust in the system's effectiveness (Bedué and Fritzsche, 2022). AI systems must be both trustworthy and trusted, as users become vulnerable by relying on them for desired outcomes (Jacovi et al., 2021). Trust requires clear and continuous evidence that the system will act predictably and align with expected values and policies. Users of the system to exert influence must trust that the system will predictably act according to their goals and ethical considerations, while the influenced users must view the system's influence as either sufficiently valuable or inconsequential, when compared to its overall value. To make informed decisions regarding this judgment, users need transparency about system goals, continuous oversight of system behavior, and safeguards against misbehavior.

4 The D-ESC System

D-ESC is a multi-component system designed for social media environments, either directly or indirectly engaging with users, and providing automated feedback to moderators or administrators. It has been deployed on Reddit, where it can post to improve community health, offers a dashboard, and generates natural language reports based on observed activity. The dashboard allows moderators to view potentially problematic posts or indicators of impending conflict. A subreddit-specific component encourages constructive discourse by rewriting posts containing harmful language, while maintaining the original intent. We present the data and textual enrichments used, then describe how each D-ESC component aligns with the principles of effective social influence outlined above.

4.1 Data Description

Data is curated from several Reddit communities using the PushShift API[1] and the Python Reddit API Wrapper (PRAW),[2] with daily collections from November 2021 through June 2022 identifying comments and posts removed by moderators following Chandrasekharan and Gilbert (2019).

4.2 Textual Enrichments

A range of linguistic dimensions, such as emotions or sentiment, are extracted from each post to provide human-interpretable values. These serve as low-level features that are used in D-ESC components or combined to form classifiers and generative models. Off-the-shelf tools classify text based on emotion,[3] sentiment,[4] and toxicity (Hanu and Unitary team, 2020), though these measures can

[1] https://files.pushshift.io/reddit/

[2] https://github.com/praw-dev/praw

[3] hf.co/bhadresh-savani/distilbert-base-uncased-emotion

[4] hf.co/nlptown/bert-base-multilingual-uncased-sentiment

be overly sensitive to profanity (which can be used in non-toxic contexts). Each comment and post is also summarized using a fine-tuned version of `flan-t5-xxl`,[5] and moral foundations (Haidt and Joseph, 2004) are extracted based on prior work.

5 D-ESC Approach

D-ESC's components address the challenges of moderation and mediation, while advancing key principles for successful AI-assisted social influence. The following subsections correspond with the principles outlined in Section 3. The connection of these methods to those principles is shown in Figure 2.

5.1 Assessing the Socioemotional Environment in Social Media

D-ESC analyzes the socioemotional environment by examining social media for expressed and unexpressed attitudes, beliefs, emotions, and experiences. A key tool is *stance detection* (Mather et al., 2021), which identifies topic-driven beliefs and determines the corresponding attitude based on belief and sentiment strength. For example, the statement *I really regret having an abortion* yields the following stance representation, with specific values for belief and sentiment: *REGRET(abortion),belief-strength=3.0,sentiment=-1.0,attitude=-3.0*. This allows D-ESC to focus moderation efforts on the underlying attitudes expressed in posts.

Stance detection uses lexical resources to extract hidden mental states related to specific topics. While previous approaches, like those of Mather et al. (2021), use semi-automatic processing with human input to create domain-relevant lexicons, D-ESC builds these lexicons fully automatically. It does so by computing predicate-argument pairs and then directly uses these as belief types, streamlining the process and enhancing the system's ability to moderate based on the nuanced understanding of user beliefs and attitudes.

Through stance detection and automatic resource building on controversial topics, D-ESC can target specific beliefs expressed by authors and automatically tailor its moderation techniques to reduce post toxicity while preserving the message's underlying content. Furthermore, applying this method to highly toxic posts helps D-ESC to iteratively refine its lexicon, improving the identification of polarizing conversations that may need moderation.

5.2 Early Behavior Shaping

Individuals may not be aware that their behavior could contribute to a degradation of community health, and early, mild intervention can keep users engaged while redirecting their communication to be more constructive. We create a conversation deviation algorithm to predict whether a social media post will provoke controversy. Data from Reddit reveals that many heated debates start with seemingly innocuous comments or posts that gradually deviate from the main topic, leading to contentious interactions. For example, in a subreddit focused on sharing COVID-related tips for working from home, a question about mask mandates or vaccinations might spark controversy as it diverges from the ongoing discussion.

Due to the unavailability or costliness of labeled data, we adopt an unsupervised approach, training a classifier head on top of the encoder of a large language model (T5), with posts from various subreddits. Posts are arranged chronologically in a sliding window of length L, shifting one post at a time. Posts within the same subreddit are labeled as normal (0).

To create controversy-provoking chains for training, we randomly select a post from a different subreddit and replace the last post in the window with the chosen post. Since this last post is from an unrelated subreddit, the topic will have deviated from the flow of the analyzed subreddit. Accordingly, L subsequent posts are labeled as 1 to indicate a deviation from the associated subreddit's theme. This process is repeated across subreddits to generate a training dataset without human annotation.

To assess the effectiveness of our method, we collect data from the abortion subreddit on two contentious days, labeling posts removed by moderators on the second day as 1. Using the L-length moving window technique, we test our model on this real, annotated dataset. We achieve an accuracy of 78% compared to 74% accuracy in prior moderation work that uses hand-annotated training data (Chandrasekharan et al., 2019).

5.3 Modeling Emotional Actions and Responses

Each social media community has unique emotions, interactions, and moderation considerations. Medical support communities may reward sympathetic and careful responses to individual stressors or struggles, while gaming communities might re-

[5] hf.co/jordiclive/flan-t5-11b-summarizer-filtered

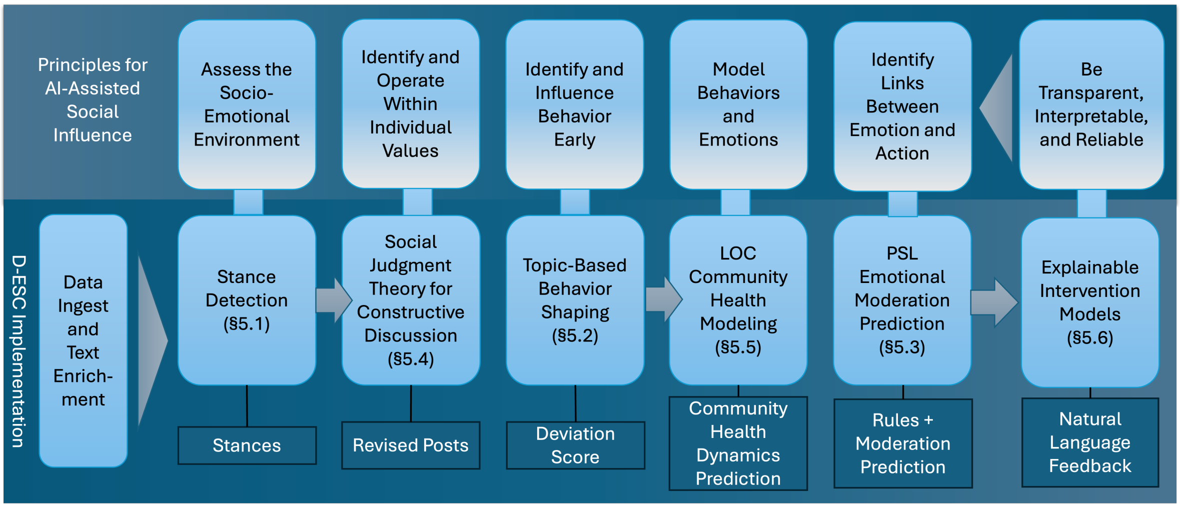

Principles for AI-Assisted Social Influence
Assess the Socio-Emotional Environment
Identify and Operate Within Individual Values
Identify and Influence Behavior Early
Model Behaviors and Emotions
Identify Links Between Emotion and Action
Be Transparent, Interpretable, and Reliable
D-ESC Implementation
Data Ingest and Text Enrich-ment
Stance Detection (§5.1)
Social Judgment Theory for Constructive Discussion (§5.4)
Topic-Based Behavior Shaping (§5.2)
LOC Community Health Modeling (§5.5)
PSL Emotional Moderation Prediction (§5.3)
Explainable Intervention Models (§5.6)
Stances
Revised Posts
Deviation Score
Community Health Dynamics Prediction
Rules + Moderation Prediction
Natural Language Feedback

Figure 2: Illustration of how the D-ESC components address principles of AI-assisted social influence, with arrows demonstrating cross-component communication, and outputs of each component at the bottom.

ward humor and witty insults. To automatically learn community responses to emotions and values, we use an approach based on soft logic and probabilistic graphical models. Soft logic helps preserve rule interpretability and allows reasoning about input indicators with varying confidence levels, e.g., degrees of toxicity in posts. Specifically, we use Probabilistic Soft Logic (PSL) (Bach et al., 2017), which uses rules to encode a probabilistic graphical model that can learn weights on rules and perform inference over large volumes of indicator inputs.

We define a set of intervention rules in PSL and test whether we can learn weights on these rules that predict how Reddit moderators intervene on posts in their communities. We implement three types of intervention rules, each addressing different types of evidence.

1. *Community*: Community conventions for whether interventions are used (e.g., heavily moderated vs. lightly moderated) and which interventions are used (e.g., manual intervention vs. automated moderation tools)
2. *Indicator*: Whether posts are outside community norms, according to some indicator.
3. *History*: Patterns of unhealthy posts by the same user (recommended in interviews with actual moderators)

To represent community conventions, we use rules like this:

$$w:\ \forall P.\neg \mathit{intervene}(\mathrm{moderator}, P, \mathrm{r/Argaming}) \tag{1}$$

In the rule, P represents the post, $\mathrm{moderator}$ is a specific intervention type, $\mathrm{r/Argaming}$ is a specific subreddit, and w is the weight on the rule, which is learned. This example rule suggests that in the subreddit *r/Argaming*, moderators rarely remove posts. If the learned weight w is large, we are less likely to recommend such interventions for that community.

Using a year's data from a dozen subreddits, we learn weights for all three types of rules. For indicator rules, an example is:

$$\begin{aligned} w :&\forall P.\mathit{enrich}(\mathrm{sadness}, P, \mathrm{r/Argaming}) \\ &\rightarrow \mathit{intervene}(\mathrm{moderator}, P, \mathrm{r/Argaming}) \end{aligned} \tag{2}$$

This indicates that moderators of *r/Argaming* typically remove posts containing sadness. We represent indicator inputs with the predicate enrich, where its soft truth value for post P is based on the output of the indicator for P. Figure 3 summarizes rule weights, where dark blue marks the intervention types that are common for each community to use, when encountering posts scoring high for varying indicators. For example, moderators of *r/LabourUK* tend to remove posts that are toxic.

After training on Reddit posts from November 2021 through June 2022, we evaluate the rules' ability to predict interventions from June to October 2022. Figure 4 shows accuracy results for a common intervention type, with soft truth values predicting interventions, with R^2 error between zero and one. Our ablation study results demonstrate that using all rules together (ALL) yields the lowest error rate, suggesting future work to include

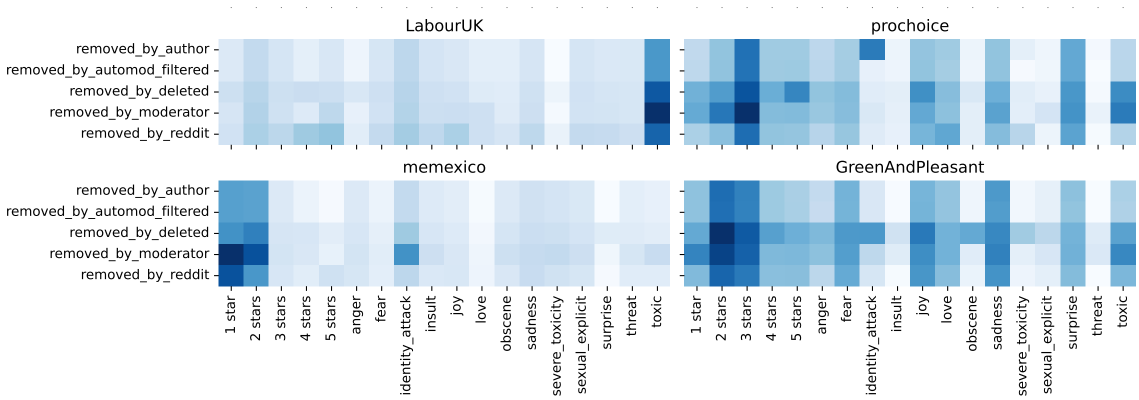


Figure 3: Examples of weights learned on indicator intervention rules. Dark blue marks the intervention types that are common for each community to use, when encountering posts scoring high for varying indicators.

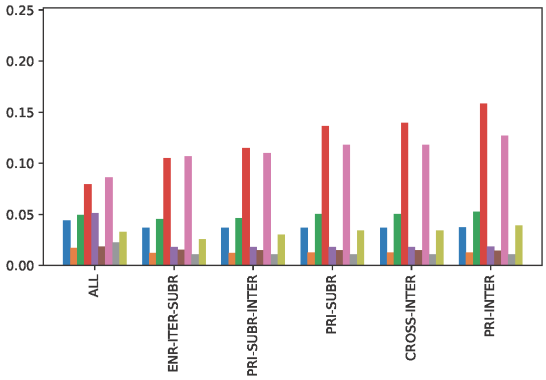


Figure 4: Error (R^2) for post interventions predicted by learned intervention rules while ablating different intervention rule types. Colors code different subreddits.

more indicators or improve rule structures to better address community-specific emotional responses.

5.4 Applying Social Judgment Theory for Constructive Discussion

A common phenomenon on social media is the presence of "echo-chambers", where a community becomes isolated from outside opinions, leading users to strongly align with a core set of values (Sunstein, 2001). One potential application of social influence is an "echo-chamber burster," which helps users explore ideas outside of their community. However, these echo chambers often include toxic language towards outside views or communities, even absent direct interaction (Efstratiou et al., 2022). Overcoming these barriers requires presenting opposing views in a respectful tone and considering emotions and values behind opinions. Framed within Social Judgment Theory (Sherif and Hovland, 1961), addressing echo chambers might best be achieved by presenting users with language that is within their *latitude of acceptance* – a range of possible positions that may not be held by an individual, but could be accepted by that individual. We can view the latitude of acceptance as holding not just standard ethical or political positions (or *stances*), but additionally certain perspectives, language, or emotions that might be acceptable within a given context.

To reduce toxicity and align communities within this framework, we build on the system described in Bose et al. (2023), which instruction-tunes a 770M `T5-large`[6] model to rephrase highly toxic posts while maintaining the style and meaning. This approach is rated as more authentic than other paraphrase methods (e.g., ChatGPT-3.5 baseline), while retaining coherence and relevance to the original content and context. With this work, we can generate suggested cross-subreddit rephrasings to express ideas across ideological divides (in this case, opposing subreddits). The model can also modify posts to reflect changes in emotions or values, and it can be tuned to match the language of the target community, potentially leading to higher rates of positive engagement.

5.5 Community Health Modeling and Influence Prediction

A holistic, quantifiable perspective on community health is necessary for assessing the effectiveness of interventions and forming hypotheses. Community dynamics are complex, influenced by moderator activity, which can be both a positive and negative indicator of community health. To address this, we focus on specific outcomes, such as reducing unsubscribes, decreasing rule-breaking

[6] https://hf.co/google-t5/tf-large

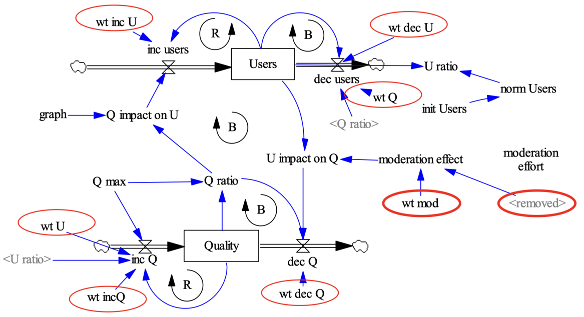


Figure 5: Logistic, Overshoot, and Collapse (LOC) model to be fitted

posts, increasing the proportion of removed rule-breaking posts, and higher downvote ratios for toxic/unhealthy posts. These outcomes are modeled to reflect the dynamics of a community in response to a new post, comment, or moderation event. Each such community activity has an associated potential social impact, represented as an effect on or of metrics such as emotion, toxicity, moral foundations, and likelihood of moderation.

To augment our rule-learning approach for predicting community health and provide a more general theory of moderation and online community behavior dynamics, we explore such behavior using a dynamic hypothesis model. However, subreddit communities can deteriorate after initial growth due to toxic posts, leading to a "collapse" in post quality. We thus use System Dynamics (SD) Modeling and Simulation (M&S) methodology (Sterman, 2000) to simulate interaction between users, post quality, and moderation activity. Figure 5 shows the model with interactions and weights on Users (U) and Post Quality (Q), where weights learned on the various relationships in the model provide a means to tailor the model to the behavior of a particular community. User movement can be tracked via subreddit metadata, while Post Quality is defined as a function of our learned textual and meta (i.e. upvote ratio) indicators of posts that yield a positive effect on the community.

We have begun experiments with learning model weights for different subreddits, informed by enrichments and moderator activity from Reddit. This work moves us toward creating a "digital twin" of social media communities, enabling intervention testing in a simulated environment. Future modeling will help us analyze the broader impact of social influence and identify communities that could benefit from our methods.

5.6 Explainable Intervention Models

D-ESC is designed for transparency and interpretability, with human-understandable models and features. We prioritize translating intervention recommendations into natural language explanations to engage moderators in the analysis process for influencing behavior. PSL model outputs are converted into clear explanations for recommended interventions, balancing reduction of moderators' cognitive load with providing sufficient evidence to support and elicit the recommended action.

Specifically, PSL rules consist of propositions (each with a truth value [0,1]) that represent statements about individual Reddit posts. We extract all PSL rules that recommend interventions, excluding those with propositions having a truth value below 0.5. For example, the proposition `0.95:ENR('enrich_toxic', 'id123')` indicates high-confidence that post `id123` contains toxic language, while `0.03:INTER('removed_by_author', 'id123')` indicates low-confidence that the same post will be removed by its author. Thus, a rule that refers to both propositions above would be removed, as not all of its propositions are of high truth-value (≥0.5).

For each intervention, rules are grouped into three categories: prior posts by the same author, labeled feelings, and labeled sentiment polarity. These are then aggregated and translated into template-based natural language explanations for the recommended moderation, similar to the approach in SPLAIN (Kazakova et al., 2019).

For example, a removal recommendation might suggest past guideline violations:*"At detection time, 74 posts by the same author had been removed by Reddit."*. Notice of a prior violation may hint at community norms: *"Historical data suggests that posts expressing anger and sadness are frequently removed by moderators."*. These recommendations and explanations are presented to Reddit moderators through an interactive online dashboard.

We stress that interactivity is crucial for adoption, as it allows moderators to access: 1) additional reasoning details, incident information, and historical user or subreddit data; 2) streamlined moderator actions (e.g., approving or rejecting moderations with a single click); and 3) provision of moderator feedback to improve the moderation models.

5.7 Modes of Operation

D-ESC was envisioned to operate flexibly in a variety of situations, whether as a moderator tool, a user assistant, or an administrative analytic dashboard. Our primary interfaces developed were an automated posting capability, an explainable moderation suggestion dashboard, and a post rephraser.

Fully automated posting of potential violations of community guidelines could be enabled by PSL moderation prediction (where moderation activity is taken when the system confidently predicts a moderator would take the same action), assisted by topic-based behavior shaping. However, moderators typically prefer a dashboard or report that provides an opportunity to verify the system's judgment. We developed both moderator reports and a dashboard with the Explainable Intervention Models providing a natural language description of the rationale for a particular recommendation. In this interface, moderators are able to see the recommended action, then approve or deny it, with the action then executed in the community.

Community Health Modeling was not fully implemented in terms of a usable interface – however, we expect such a system would provide a userful tool for administrators of social media platforms, as it provides a high-level analysis of community dynamics that can provide predictions as to whether a community may be potentially turning toxic.

Finally, the method to apply Social Judgment Theory for constructive discussion can be applied in multiple ways. First, as a suggested alternative for users before posting a potentially inflammatory comment. Alternatively, we envision an automated agent that could generate responses that steer the discussion in a more constructive direction.

6 Future Work

D-ESC would likely benefit most from increased interoperability and communication across components to more fully deliver on the promise of the principles outlined in this paper. For example, we plan to use stances to guide toxicity reduction in posts, ensuring important content is retained through the rephrasing, and extend our PSL models with more enrichments and indicators. Furthermore, as there are various complex components, an automatically learned process model for achieving specific outcomes could yield an effective use-case applied to a specific community. Additionally, while we originally positioned this primarily as a moderator tool, there could be greater opportunities for adoption as a tool for end-users of social media to consider how best to engage with the community using our knowledge of how that community would likely respond to a post – thus influencing the user towards pro-social behavior.

While we designed D-ESC to primarily operate on social media platforms, we believe the overall architecture could be applied to a wide variety of domains. For example, individuals might use a version of the system to consider how their social interactions project certain values or beliefs, as certain expressions can lead to social isolation (Yang and Nino, 2023). In patient-caregiver interactions, careful mediation of communication might enable both patients and caregivers to feel that their unique challenges and stressors are understood, potentially alleviating caregiver depression (Hua et al., 2021) and burnout (de Souza Alves et al., 2019).

7 Conclusions

We have outlined a set of principles for social influence systems that serves as a framework for ethical, effective, and widely adopted social influence applications. We demonstrate a system that follows these principles, applying novel NLP and reasoning techniques to enable moderators foster more constructive discussions. Additionally, we illustrate how combining multiple indicators and techniques provides a more tailored, nuanced approach to positive community influence.

There is more work needed to position this as a comprehensive social influence system, with additional interface development and fine-tuning to broadly represent the interests of social media communities. Nevertheless, we believe this integrated system can be easily adapted to other domains, including local community activism, patient-caregiver dialogue, and disaster relief. Furthermore, we hope this work encourages the community to consider integrated, complex social dynamics and work to develop baselines and evaluations that consider a holistic, multi-user environment.

Acknowledgements

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00112290022. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

8 Limitations

Research in AI applied to social influence is relatively new, with most discussions concerning the harm of biased or unregulated AI systems that are not explicitly intended for influencing individuals. While we believe our work in the social media mediation space provides an useful starting point for discussing broader principles, the nascent nature of the field limits the claims we can make about the broad applicability of our guiding principles.

Additionally, while we have some evaluations of individual components, we encountered difficulties in evaluating the system as a whole on a sufficiently large dataset or environment. On Reddit, ground truth instances of degradation of community health are often primarily found in quarantined or banned subreddits, which have little to no current activity and would likely not be promising targets for mediation efforts. Moderation communities in larger subreddits prioritize standard moderation practices using Automod, and do not often see a need for mediation efforts. Smaller subreddits, especially those that aim to provide an environment for respectful discourse (e.g. *r/AbortionDebate*), are more amenable to mediation intervention but have fewer instances of mediation events to train on. Additionally, such communities are rightfully concerned about the potential side effects of AI mediation, with a general protectiveness of users' data and personal experiences.

Finally, our work has empirically demonstrated the unique behavioral and emotional factors in each subreddit, but such diversity makes it more difficult to evaluate the effectiveness of our approaches at scale. Each subreddit had different moderation techniques, community guidelines, media and post types, and discussion typologies. Thus while we believe our approaches could apply widely given community-specific targets to learn, evaluating our novel techniques at a scale sufficient for strong research claims proved difficult.

9 Ethics Statement

The application of AI to research social influence has a significant potential for amplifying existing societal risks inherent in non-AI-based social influence research (Broom, 2006), as people may even be more susceptible to influence from an AI system (Riva et al., 2022). While we outline an approach for ethically applying social influence (with transparency, human-interpretable methods and analysis, and with explanations provided to users as to the goals of the system), there is nevertheless a risk of negative outcomes for both an individual or a community.

We have generally considered that more transparency makes application of these systems more ethical – yet this may not be the case. Transparency of methods and attempts to influence users may cause them to be wary of interacting on certain related platforms, even in cases when no influence is intended. Also, if systems can ethically influence individuals to make more positive decisions without transparency, and transparency reduces a system's effectiveness, it is unclear whether added transparency yields a net social benefit.

Social media discussions have real-life consequences, from ostracization to persecution to rioting. People often turn to social media for guidance on issues relating to work, their health, and family. Thus, tools that interact with such communities have the potential to cause harm if applied without care. We believe we have mitigated some of the potential risks of our system through our goals of explainability, human-in-the-loop functionality, and awareness of potential side-effects that could occur with interventions. Additionally, we have been extremely cautious with data and potential interventions – this added to the difficulty of a large-scale application of this system.

One remaining risk is that users may feel they are being watched or judged as their expressions are deemed inappropriate for a community according to an algorithm. While our work tends to identify negative emotions like outrage as harmful for a community, there are some cases where outrage is a reasonable response to a situation – silencing individuals who may be going through a difficult time will not necessarily be a net positive when considering the effect on that individual and the community.

Nevertheless, we believe that our community-specific models and our generation techniques provide an opportunity to bridge communities that would otherwise be divided, and opening such a dialogue could have a significant positive effect on online interactions.

Our data collection and methods were evaluated by our institutional IRB and the US Office for Human Research Protections.

References

Lucas Anastasiou and Anna De Liddo. 2023. BCause: Reducing group bias and promoting cohesive discussion in online deliberation processes through a simple and engaging online deliberation tool. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 39–49, Toronto, Canada. Association for Computational Linguistics.

Lisa P. Argyle, Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023. Leveraging ai for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41):e2311627120.

Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2017. Hinge-loss markov random fields and probabilistic soft logic. *Journal of Machine Learning Research*, 18(109):1–67.

Agathe Balayn, Jie Yang, Zoltan Szlavik, and Alessandro Bozzon. 2021. Automatic identification of harmful, aggressive, abusive, and offensive language on the web: A survey of technical biases informed by psychology literature. *Trans. Soc. Comput.*, 4(3).

Patrick Bedué and Albrecht Fritzsche. 2022. Can we trust ai? an empirical investigation of trust requirements and guide to successful ai adoption. *Journal of Enterprise Information Management*, 35(2):530–549.

Robert Bergman. 2023. Chatgpt and mediation. *Mediate.com*.

Ritwik Bose, Ian Perera, and Bonnie Dorr. 2023. Detoxifying online discourse: A guided response generation approach for reducing toxicity in user-generated text. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 9–14, Toronto, Canada. Association for Computational Linguistics.

Alex Broom. 2006. Ethical issues in social research. *Complementary Therapies in Medicine*, 14(2):151–156.

Kathleen M. Carley. 2020. Social cybersecurity: an emerging science. *Comput. Math. Organ. Theory*, 26(4):365–381.

Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Eshwar Chandrasekharan and Eric Gilbert. 2019. Hybrid approaches to detect comments violating macro norms on reddit. *Preprint*, arXiv:1904.03596.

Kushal Chawla, Weiyan Shi, Jingwen Zhang, Gale Lucas, Zhou Yu, and Jonathan Gratch. 2023. Social influence dialogue systems: A survey of datasets and models for social influence tasks. *Preprint*, arXiv:2210.05664.

Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2021. Antisocial Behavior in Online Discussion Communities. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):61–70.

Hyundong Cho, Shuai Liu, Taiwei Shi, Darpan Jain, Basem Rizk, Yuyang Huang, Zixun Lu, Nuan Wen, Jonathan Gratch, Emilio Ferrara, and Jonathan May. 2024. Can language model moderators improve the health of online discourse? *Preprint*, arXiv:2311.10781.

Ludmyla Caroline de Souza Alves, Diana Quirino Monteiro, Sirlei Ricarte Bento, Vânia Diniz Hayashi, Lucas N.C. Pelegrini, and Francisco Assis Carvalho Vale. 2019. Burnout syndrome in informal caregivers of older adults with dementia: A systematic review. *Dementia & Neuropsychologia*, 13:415 – 421.

Angel Diaz and Laura Hecht-Felella. 2021. Report on "double standards in social media content moderation".

Thiago Freitas dos Santos, Nardine Osman, and Marco Schorlemmer. 2024. Is this a violation? learning and understanding norm violations in online communities. *Artificial Intelligence*, 327:104058.

Alexandros Efstratiou, Jeremy Blackburn, Tristan Caulfield, Gianluca Stringhini, Savvas Zannettou, and Emiliano De, Cristofaro. 2022. Non-polar opposites: analyzing the relationship between echo chambers and hostile intergroup interactions on reddit.

Facebook. November, 2022. Facebook community standards: Hate speech.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Claire Hardaker. 2013. “uh.... not to be nitpicky, but… the past tense of drag is dragged, not drug.”: An overview of trolling strategies. *Journal of Language Aggression and Conflict*, 1(1):58–86.

Jess Hohenstein and Malte Jung. 2020. Ai as a moral crumple zone: The effects of ai-mediated communication on attribution and trust. *Computers in Human Behavior*, 106:106190.

Katherine Hopman, Deborah Richards, and Melissa M. Norberg. 2023. A digital coach to promote emotion regulation skills. *Multimodal Technologies and Interaction*, 7(6).

Alice Y. Hua, Jenna L. Wells, Casey L. Brown, and Robert W. Levenson. 2021. Emotional and cognitive empathy in caregivers of people with neurodegenerative disease: Relationships with caregiver mental health. *Clinical Psychological Science*, 9(3):449–466.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635.

Vera A. Kazakova, Jena D. Hwang, Bonnie J. Dorr, Yorick Wilks, J. Blake Gage, Alex Memory, and Mark Clark. 2019. Splain: Augmenting cybersecurity warnings with reasons and data. In *Proceedings of FLAIRS*.

Soo Ling Lim and Peter J Bentley. 2022. Opinion amplification causes extreme polarization in social networks. *Scientific Reports*, 12(1):18131.

Adrienne Massanari. 2017. *#Gamergate and The Fappening: How Reddit's Algorithm, Governance, and Culture Support Toxic Technocultures*. The University of Illinois Press.

Brodie Mather, Bonnie J. Dorr, Owen Rambow, and Tomek Strzalkowski. 2021. A General Framework for Domain-Specialization of Stance Detection. *The International FLAIRS Conference Proceedings*, 34.

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11.

Henna Paakki, Heidi Vepsäläinen, and Antti Salovaara. 2021. Disruptive online communication: How asymmetric trolling-like response strategies steer conversation off the track. *Computer Supported Cooperative Work (CSCW)*, 30(3):425–461.

Utpal M. Dholakia Richard P. Bagozzi and Lisa R. Klein Pearo. 2007. Antecedents and consequences of online social interactions. *Media Psychology*, 9(1):77–114.

Paolo Riva, Nicolas Aureli, and Federica Silvestrini. 2022. Social influences in the digital era: When do people conform more to a human being or an artificial intelligence? *Acta Psychologica*, 229:103681.

Hanna Roos. 2023. Arbitration tech toolbox: Let's chat some more about chatgpt and dispute resolution. Kluwer Arbitration Blog. `https://www.kluwerarbitration.com/2023/04/08/arbitration-tech-toolbox-lets-chat-some-more-about-chatgpt-and-dispute-resolution/`.

M. Sherif and C.I. Hovland. 1961. *Social Judgment: Assimilation and Contrast Effects in Communication and Attitude Change*. Yale University Press, New Haven, CT.

Anirudh Som, Karan Sikka, Helen Gent, Ajay Divakaran, Andreas Kathol, and Dimitra Vergyri. 2024. Demonstrations are all you need: Advancing offensive content paraphrasing using in-context learning. *Preprint*, arXiv:2310.10707.

John Sterman. 2000. Business dynamics, system thinking and modeling for a complex world. 19.

John Suler. 2004. The online disinhibition effect. In Jayne Gackenbach, editor, *The Psychology of Cyberspace*, pages 71–92. Academic Press.

C.R. Sunstein. 2001. *Republic.com*. Republic.com. Princeton University Press.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16. International World Wide Web Conferences Steering Committee.

Nhat Tran, Malihe Alikhani, and Diane Litman. 2022. How to ask for donations? learning user-specific persuasive dialogue policies through online interactions. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '22, page 12–22, New York, NY, USA. Association for Computing Machinery.

Twitter. March, 2023. Twitter rolls out updated zero tolerance policy on violent speech.

Adam Worrall, Alicia Cappello, and Rachel Osolen. 2021. The importance of socio-emotional considerations in online communities, social informatics, and information science. *Journal of the Association for Information Science and Technology*, 72(10):1247–1260.

Song Yang and Michael Nino. 2023. Political views, race and ethnicity, and social isolation: Evidence from the general social survey. *Societies*, 13(11).

YouTube. 2019. Youtube community guidelines: Hate speech.

Weixiang Zhao, Yanyan Zhao, Xin Lu, and Bing Qin. 2023. Don't lose yourself! empathetic response generation via explicit self-other awareness. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13331–13344, Toronto, Canada. Association for Computational Linguistics.

Jianlong Zhou, Fang Chen, Adam Berry, Mike Reed, Shujia Zhang, and Siobhan Savage. 2020. A survey on ethical principles of ai and implementations. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 3010–3017.