

Visualizing Cumulative Consequences: Addressing Distance from Effect and Lack of Awareness in Online Commenting

Osaruiyobo Giwa

osaruiyobo@dgist.ac.kr

Daegu Gyeongbuk Institute of Science and Technology

Sunjun Kim

Daegu Gyeongbuk Institute of Science and Technology

Article

Keywords: Psychological Distance, Online Commenting, Information Visualization, Prosocial Behavior, Cumulative Feedback

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9374995/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Visualizing Cumulative Consequences: Addressing Distance from Effect and Lack of Awareness in Online Commenting

Anonymous Author(s)

Abstract

Harmful commenting persists due to interface designs that obscure social consequences and limited awareness of cumulative impact. Most interventions are momentary, offering little longitudinal reflection support. We developed a browser-based system that maps comments onto a persistent virtual tree designed to reduce psychological distance and increase awareness. In a seven-week field study ($N = 50$), we examined associations between continuous feedback and commenting patterns. Positive comments increased from 38.5% to 67.2% ($\beta = 5.127\%/week$, $95\%CI : 2.08 - 8.18$, $p = 0.008$); neutral decreased from 54.8% to 30.8% ($\beta = -4.16\%/week$, $95\%CI : -6.5to - 1.8$, $p = 0.004$); negative remained low (6.7%→2.0%). Post-study survey ($N = 37$) showed consequences of commenting felt more concrete (Perception: $M = 4.28/5$, $p < 0.001$). However, the single-arm observational design without control group prevents causal inference; patterns could reflect self-selection or Hawthorne effects. Findings warrant randomized controlled trials to establish causal efficacy.

Keywords: Psychological Distance, Online Commenting, Information Visualization, Prosocial Behavior, Cumulative Feedback

1 Introduction

Online social media platforms continue to face persistent challenges related to harmful and unreflective commenting behavior, despite widespread awareness of moderation guidelines and community standards. Although overtly abusive content constitutes a visible portion of the problem, a substantial share of social harm arises from routine, everyday interactions that are not necessarily malicious but nonetheless accumulate negative effects over time [1–3]. Prior research demonstrates that such interactions can undermine individual well-being, discourage participation, and degrade the quality of online discourse. Existing interventions have largely focused on content-level controls, including automated toxicity warnings, moderation prompts, reporting mechanisms,

and post-hoc feedback. These approaches typically intervene either immediately before comment submission or after harm has already occurred. While effective for addressing extreme cases, they often fail to confront a more fundamental limitation: users’ restricted perception of how their everyday comments accumulate and affect others across repeated interactions [4–6].

Two interacting psychological barriers help explain this limitation. First, online interaction introduces distance from effect, in which the social consequences of one’s actions are abstracted, delayed, or rendered invisible by interface design. Compared to face-to-face communication, computer-mediated environments attenuate social cues and feedback, reducing users’ experiential connection to the downstream impact of their behavior [4, 5, 7, 8]. Second, users frequently experience a lack of awareness regarding the cumulative impact of their everyday interactions. Feedback, when available, is often fragmented, momentary, or detached from the act of commenting itself, making it difficult for users to recognize how repeated actions contribute to broader social dynamics [1, 9, 10].

Despite extensive work on momentary prompts, warnings, and moderation cues, relatively little is known about how persistent, cumulative representations of everyday online behavior shape users’ awareness and expressive choices over time. Existing research has examined warning messages, sentiment labels, social norm cues, and friction-based prompts designed to encourage reflection at the moment of posting [11–13]. Large-scale experiments suggest that such interventions can influence harmful or undesirable behavior in the short term [14, 15]. However, these approaches are typically transient, text-based, and episodic. As a result, they provide limited support for understanding longer-term consequences or sustaining reflection across repeated interactions [16, 17]. Prior interventions overwhelmingly target isolated decision points, leaving open the question of whether interface-level visualizations that externalize behavioral history can support longitudinal reflection without coercion.

In this work, we investigate an alternative design direction: the persistent visualization of commenting consequences. Rather than intervening through warnings, restrictions, or moderation actions, the system externalizes the effects of comments into a continuously evolving visual artifact, a virtual tree, whose state reflects the accumulation of positive and negative interactions over time (Fig 1), leveraging conceptual metaphor to render otherwise intangible consequences perceptible [18]. The tree acts as a powerful, grounding, and organic metaphor that bridges abstract comments with tangible, visualization. By framing personal development as nurturing a living thing, it shifts the focus from overwhelming, distant, and complex challenges to manageable, natural stages of growth that encourage awareness and observation. Since visual encodings can shape interpretation, the visualization is treated as an interface intervention whose influence may arise from both informational content and representational form [19]. By embedding feedback directly into the interaction loop and maintaining it across repeated engagements, this approach makes consequences visible, continuous, and cumulative rather than abstract or episodic. This framing builds on foundational work in reflective and experiential design [20–22] and on evidence that integrated visual representations can support comprehension and reflection over time [23].

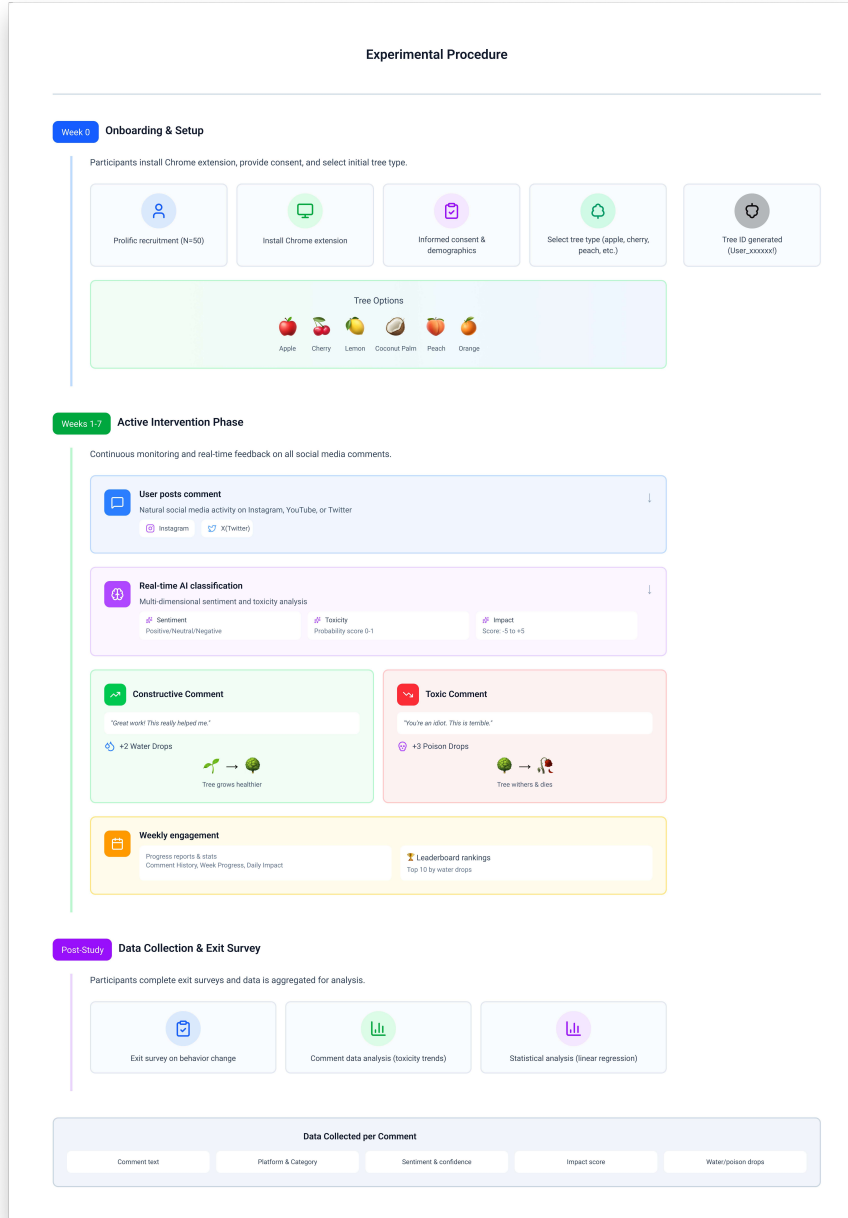


Fig. 1: System architecture and experimental workflow for the persistent visualization intervention. Overview of the 7-week tree growth game deployment (N=50). Participants installed a Chrome extension that monitored social media interactions. The system features a proprietary sentiment classification model integrated into a Chrome extension. This model processes social media interactions in real-time to drive the virtual tree’s growth (water drops) or deterioration (poison drops) based on detected linguistic valence. The workflow includes the transition from real-time feedback to longitudinal data collection and post-study psychological assessment.

Importantly, the design avoids coercive enforcement. The system does not restrict posting, issue punitive warnings, or impose penalties. Instead, it functions as a reflective interface that externalizes impact without interrupting user autonomy. This positions the work within traditions of reflective and experiential design rather than moderation or compliance-oriented tooling [21, 24]. While such an approach limits claims about direct behavioral control, it strengthens the contribution as an examination of how interface-level visualization can reshape users’ experiential understanding of impact without coercion.

We deployed the system in a seven-week field study with 50 participants recruited from Prolific and screened for active social media use (at least 7 comments per week). Rather than testing causal effects, this exploratory observational study examines feasibility, user engagement, and observed associations between exposure to cumulative feedback and aggregate commenting patterns. Field deployments of behaviour-facing interfaces are also vulnerable to reactivity. When participants are aware that their commenting is being tracked or assessed—whether by researchers or by an interface that makes behavioural history salient—they may alter expression due to Hawthorne effects, evaluation apprehension, or self-presentation motives, independent of the intended psychological mechanism (e.g., reduced psychological distance). This concern is especially relevant for non-coercive systems that embed feedback into everyday interaction, because the same visibility that supports reflection can also heighten perceived monitoring. For this reason, we frame the present study as an exploratory observational deployment focused on feasibility and temporal patterns rather than causal estimates, and we treat observed changes as compatible with multiple mechanisms, including reactivity. We hypothesize that persistent visualization of cumulative commenting consequences will be associated with (H1) a shift in the distribution of everyday commenting behavior toward more positively valenced expression and (H2) increased awareness of comment quality and perceived social impact, consistent with reduced psychological distance. Observed trends are interpreted conservatively, acknowledging that contextual, social, and individual factors may also contribute [25]. Accordingly, we frame our findings as behavioral patterns consistent with increased salience of consequences, reduced distance from effect, and enhanced situational awareness, rather than as definitive evidence of prosocial intent or moral reasoning.

In addition to behavioral analysis, we assess participants’ self-reported awareness of the system’s influence on their commenting behavior. While some participants reported consciously recognizing the visualization’s influence, others did not, despite exhibiting observable behavioral trends during the deployment. This aligns with prior work demonstrating that interface-level interventions may shape behavior both reflectively and implicitly [26, 27]. Importantly, a lack of self-reported awareness does not imply an absence of influence; persistent, ambient feedback can affect behavior without conscious attribution, particularly when feedback is non-interruptive and embedded in routine interaction [23].

2 Results

2.1 Overview

This study examined whether exposure to a persistent visualization of cumulative commenting consequences was associated with changes in commenting behaviour and perceived awareness during a seven-week field deployment. Because the study employed a single-condition longitudinal design without a control group, analyses focus on temporal patterns and associations rather than causal inference.

Table 1: Linear regression results for weekly sentiment proportions and cumulative feedback metrics across the seven-week deployment. Slopes are unstandardised; 95% CIs are bias-corrected accelerated bootstrap (10 000 samples). Effect-size metric is Cohen’s f^2 ($f^2 \geq 0.35 = \text{large}$).

Metric	Slope	Intercept	R^2	r	p	Std. error	Sig.	95% CI	f^2	Trend
Positive sentiment (%)	5.127	34.234	0.789	0.889	0.008	1.186	**	[2.08, 8.18]	3.73	Increasing
Negative sentiment (%)	-0.965	7.855	0.558	-0.747	0.054	0.384	†	[-1.95, 0.02]	1.26	Decreasing
Neutral sentiment (%)	-4.162	57.910	0.832	-0.912	0.004	0.836	**	[-6.31, -2.01]	4.95	Decreasing
Water drops	-3.678	478.000	0.012	-0.108	0.817	15.086	ns	[-42.5, 35.1]	0.01	Decreasing
Poison drops	-25.321	171.000	0.716	-0.846	0.016	7.131	*	[-43.7, -7.0]	2.52	Decreasing
Water/Poison ratio	6.624	-9.811	0.856	0.925	0.003	1.217	**	[3.50, 9.75]	5.93	Increasing

Significance: † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$; ns = not significant.

Two complementary dimensions were assessed. First, weekly proportions of positive, neutral, and negative comments were examined descriptively and using linear trend analyses across seven aggregated time points. Second, post-study self-reports evaluated perceived awareness, psychological distance, behavioural orientation (Table 3), and perceived influence.

Across the deployment, the proportion of positive sentiment increased from 38.5% in Week 1 to 67.2% in Week 7. Linear regression analysis indicated a clear upward pattern ($R^2 = 0.789$, $p = 0.008$; Table 1). In parallel, neutral sentiment declined substantially over time ($R^2 = 0.832$, $p = 0.004$; Fig. 2). Negative sentiment showed a downward trend that approached, but did not reach, conventional statistical significance ($R^2 = 0.558$, $p = 0.054$).

Given that regression analyses were conducted on seven weekly aggregates ($N = 7$), statistical estimates should be interpreted cautiously. Observed trends represent descriptive temporal regularities during system exposure rather than confirmatory evidence of intervention-driven change. Together, these results indicate a systematic redistribution of sentiment composition over the study period.

Robustness Check: Permutation Testing

To verify that the observed linear trends were not artifacts of limited temporal aggregation ($N = 7$ weeks), we conducted a permutation test for the primary outcome

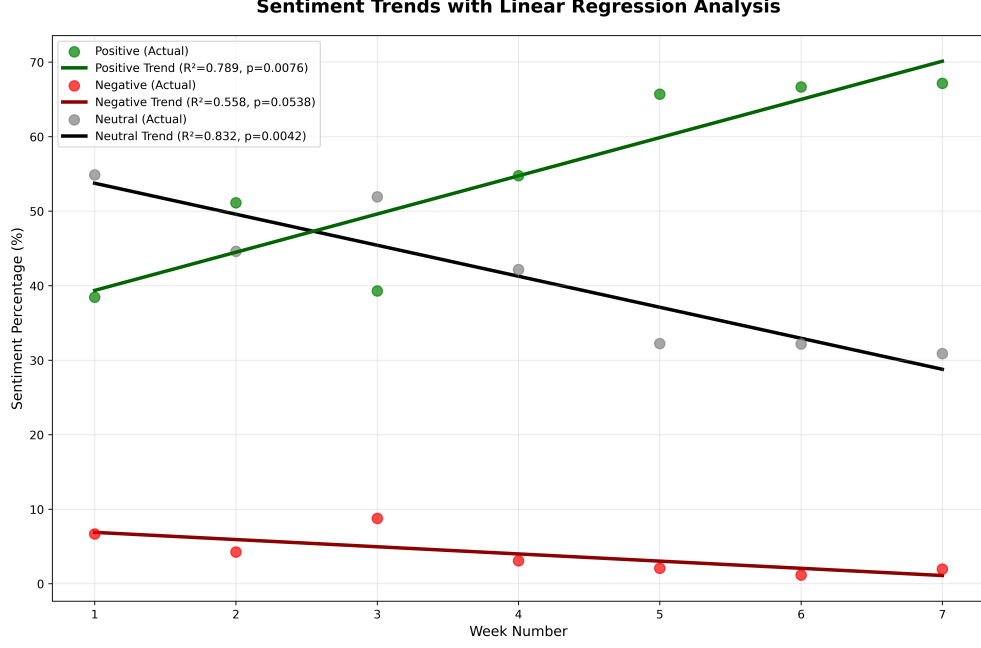


Fig. 2: Weekly proportions of positive, neutral, and negative comments across the seven-week field deployment. Lines represent linear regression trends fitted to weekly aggregates. Positive sentiment increased over time, neutral sentiment declined, and negative sentiment remained low throughout the study period.

(positive sentiment). The observed slope ($\beta = 5.04$) was compared against a null distribution generated from 10,000 random permutations of weekly sentiment values. The permutation p-value ($p = 0.0009$) indicates that fewer than 0.1% of random shuffles produced slopes of equal or greater magnitude, confirming that the positive sentiment trend is statistically robust to temporal ordering.

At the system level, cumulative feedback dynamics were examined using the Water-to-Poison ratio as a proxy for the evolving feedback environment. With the exception of total water drops, all regression slopes exhibited large effect sizes ($f^2 \geq 1.2$; Table 1). While the volume of water drops showed no significant linear trend ($p = 0.817$), poison drops declined significantly over time ($p = 0.0164$). As a result, the Water-to-Poison ratio increased steadily, reflecting an increasing separation between constructive and harmful feedback signals.

Self-report measures aligned with these behavioural observations. Participants reported that the visualization made the consequences of their comments feel more concrete and immediate. While a majority of participants explicitly reported that the tree influenced their commenting behaviour, comparable behavioural trends were also observed among participants reporting limited conscious awareness. These responses indicate that the visualization was experienced both reflectively and as an ambient background cue during routine interaction.

2.2 Longitudinal shifts in commenting sentiment

Weekly sentiment proportions (Fig 2) showed a steady upward trajectory for positive comments and a corresponding decline in neutral comments. Linear trend estimates indicated increasing positive sentiment ($R^2 = 0.789$, $p = 0.008$ Table 2) and decreasing neutral sentiment ($R^2 = 0.832$, $p = 0.004$). Negative sentiment displayed a downward pattern ($R^2 = 0.558$, $p = 0.054$) but remained near floor levels across all weeks.

The dominant behavioural shift consisted of redistribution from neutral toward positive expression rather than substantial suppression of negative content. This distinction is important given the low baseline of negative comments, which introduces a floor effect limiting detectable reductions.

Week-level inspection revealed that change was gradual rather than strictly monotonic. Week 3 showed a temporary increase in negative and neutral proportions. Subsequent weeks returned to the broader directional pattern.

Table 2: Field-deployment sentiment trends: comparison between week 1 and 7

Category	Week 1 %	Week 7 %	Δ change
Positive (+3)	38.5	67.2	+28.7
Neutral (+2)	54.9	30.9	-24.0
Negative (-3)	6.7	2.0	-4.7

Neutral sentiment declined substantially over the same period, decreasing from approximately 54.8% in Week 1 to 30.8% in Week 7 (Fig. 2). This decline was captured by a significant negative linear trend ($R^2 = 0.832$, $p = 0.0042$; Table 1). The opposing trajectories of positive and neutral sentiment indicate a redistribution of sentiment composition rather than a uniform change in overall commenting volume.

Negative sentiment accounted for a small proportion of comments throughout the study, decreasing from 6.7% in Week 1 to 2.0% in Week 7 (Fig. 3). Although the fitted regression line suggested a downward trend ($R^2 = 0.558$), this trend did not reach conventional significance ($p = 0.054$). Negative sentiment remained low across all weeks, limiting the scope for large absolute change.

2.3 Absolute comment volume and robustness of sentiment shifts

Weekly comment counts ranged from 204 to 390 comments. Although participation fluctuated, proportional sentiment redistribution persisted across these variations. Positive comment counts increased in later weeks, while neutral comment counts declined both proportionally and in absolute terms. Negative comment counts remained low throughout.

Because analyses were conducted at the aggregate weekly level, individual-level heterogeneity was not modeled. It therefore remains unclear whether changes reflect uniform shifts across participants or stronger effects among a subset of users.

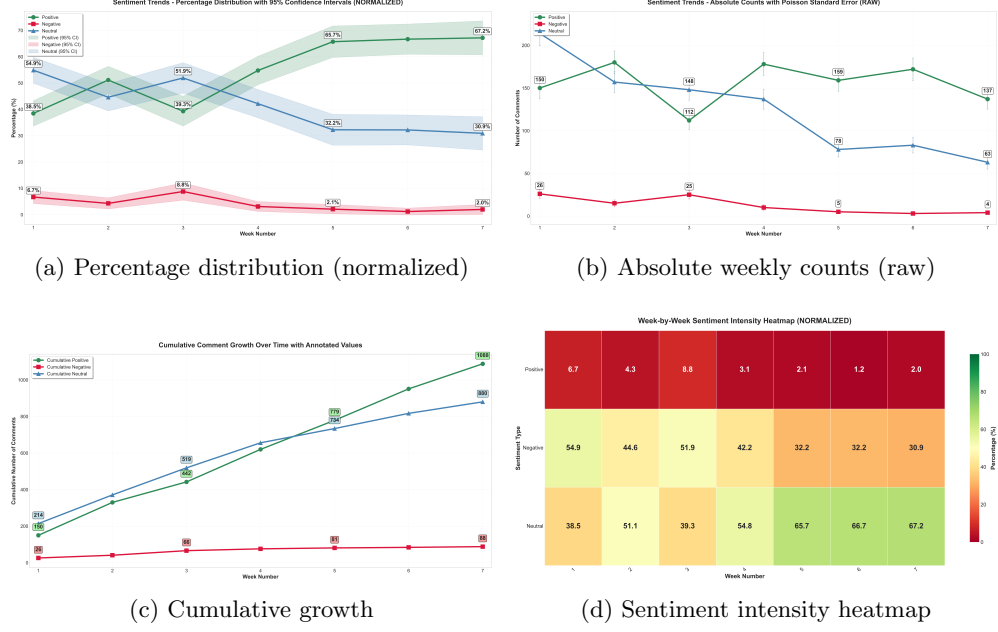


Fig. 3: Longitudinal sentiment dynamics of user comments across the seven-week intervention period (Poisson Standard Error was applied). (a) Percentage distribution of Positive, Neutral, and Negative comments with 95% confidence intervals (normalized). (b) Absolute weekly comment counts with Poisson standard errors (raw). (c) Cumulative comment growth across weeks. (d) Normalized week-by-week sentiment intensity heatmap. Together, the panels depict a systematic transition toward increasing positive engagement and diminishing neutral and negative expression over the course of the study.

2.4 Week-by-week dynamics and transient deviations

Week-level inspection revealed that sentiment change was gradual rather than monotonic (Fig 5). Notably, Week 3 exhibited a temporary increase in negative sentiment (8.8%) alongside a rebound in neutral comments (51.9%). This deviation coincided with a reduction in overall participation and represents the largest single-week fluctuation in the series. Subsequent weeks returned to the prevailing pattern of declining neutral sentiment and increasing positive sentiment. No comparable rebound was observed after Week 4, indicating that transient deviations did not alter the overall trajectory.

2.5 Cumulative feedback signals: water and poison drops

The visualization translated comments into cumulative “water” (positive) and “poison” (negative) feedback signals (Fig. 4). Across the deployment, participants received

3,243 water drops and 488 poison drops, yielding an overall Water-to-Poison ratio of 6.65:1.

Weekly poison drops declined significantly over time ($R^2 = 0.716$, $p = 0.0164$), whereas weekly water drops did not exhibit a reliable linear trend ($R^2 = 0.012$, $p = 0.817$). As a result, the Water-to-Poison ratio increased from 2.95:1 in Week 1 to 45.44:1 in Week 7. Because poison counts were small in later weeks, ratio growth should be interpreted cautiously. Regression analysis (Fig 5) indicated no significant trend in total weekly water drops, whereas poison drops declined significantly over time.

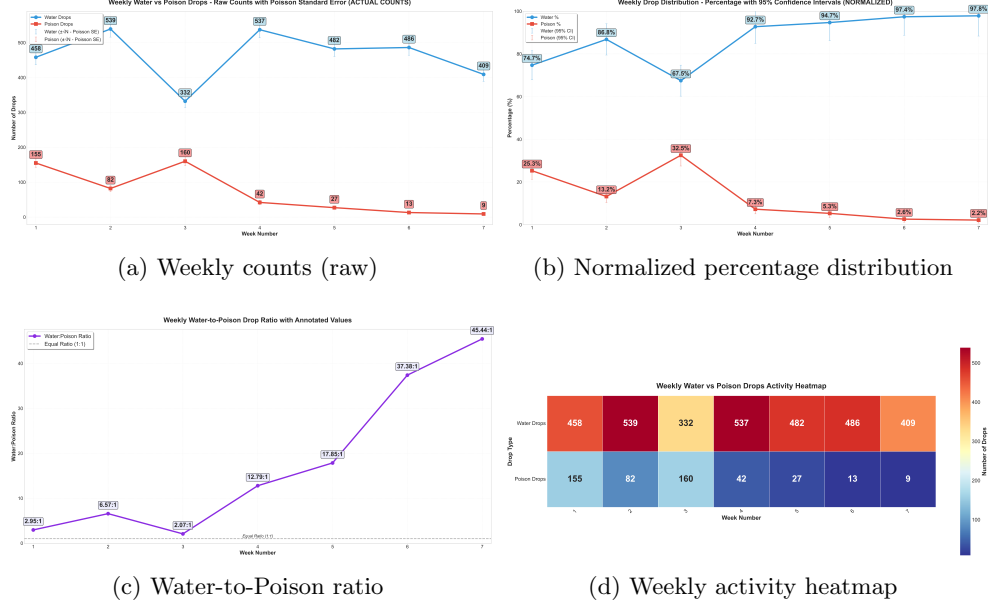


Fig. 4: Temporal dynamics of feedback signals (Water and Poison drops) across the seven-week intervention period. (a) Weekly counts of Water and Poison drops with Poisson standard errors (raw counts), showing consistently high levels of Water drops throughout the study, with pronounced peaks in Weeks 2 and 4 and sustained elevation through Week 6, while Poison drops concentrate in the early weeks and decline sharply after Week 3. (b) Normalized percentage distribution of Water and Poison drops with 95% confidence intervals, demonstrating increasing dominance of Water feedback over time and a corresponding reduction in Poison feedback to near-zero levels by Week 7. (c) Weekly Water-to-Poison drop ratio with annotated values and a 1:1 reference line, revealing a strong monotonic increase after Week 3; the ratio exceeds 10:1 by Week 4 and reaches its maximum by Week 7. (d) Weekly activity heatmap of Water versus Poison drops, confirming the divergence between high and stable Water activity and rapidly diminishing Poison activity over time. Collectively, the panels illustrate a systematic shift toward increasingly positive feedback dynamics over the course of the study.

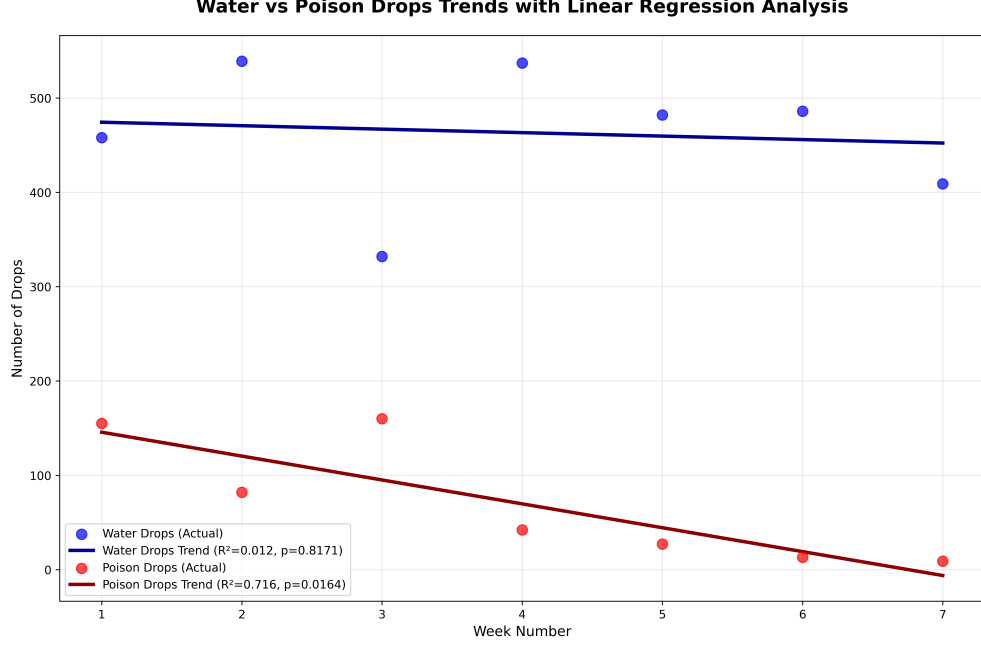


Fig. 5: Linear regression analysis of water to poison count over the 7-week longitudinal study.

Across sentiment proportions, absolute counts, and cumulative feedback metrics, a consistent pattern of sentiment redistribution was observed, characterised primarily by reductions in neutral expression alongside stable or increasing positive expression.

2.6 Statistical Self-reported awareness and psychological outcomes

Thirty-seven participants voluntarily completed the post-study questionnaire. One-sample Wilcoxon signed-rank tests comparing each section to the neutral midpoint ($\hat{x}_0 = 3$) indicated that all four constructs scored significantly above neutrality after Bonferroni correction (Table 3). This approach was selected due to the ordinal nature of the data and to avoid assumptions of normality [28].

All questionnaire sections yielded scores significantly above the neutral midpoint. Perception of the Tree Visualization (S1) indicated consistent attention to the visualization throughout the study. Distance from Effect (S2) scores suggested reduced perceived abstraction of commenting consequences. Awareness of Comment Quality (S3) showed the strongest deviation from neutrality, with low dispersion across respondents. Behavioural Orientation (S4) also exceeded the midpoint, with greater variability, reflecting individual differences in how participants related visual feedback to motivation and behaviour.

Table 3: Post-study self-report results for awareness and behavioural orientation. For each questionnaire section ($n = 37$), median, mean, standard deviation, and Wilcoxon signed-rank test statistics are reported relative to the neutral midpoint of the Likert scale.

Section	n	Median	Mean	SD	Wilcoxon W	p -value
S1-Perception	37	4.50	4.28	0.81	15	$<0.001^1$
S2-Distance	37	4.33	3.98	0.88	58	<0.001
S3-Awareness	37	4.50	4.38	0.59	1	<0.001
S4-Behavioural	37	4.33	4.06	0.96	64	<0.001

¹Exact two-tailed p -values; all sections remain significant after Bonferroni correction for four comparisons.

2.7 Perceived Influence and Qualitative Responses

When asked directly whether the tree visualization influenced their commenting behaviour ($N = 37$), 67.6% responded “Yes,” 10.8% “Maybe,” and 21.6% “No.”

Open-ended responses from 29 participants were analyzed using an exploratory thematic approach. Responses were reviewed iteratively to identify recurring patterns in perceived influence. Three primary themes emerged:

1. Increased Awareness and Reflection

Many participants described heightened mindfulness of tone and impact:

- “It made me more aware of my comments.”
- “It made me be more conscious about how I comment.”
- “More visibility of my commenting behaviour.”

2. Positive Tone Adjustment and Reinforcement Sensitivity

Several participants reported actively modifying wording to obtain positive reinforcement:

- “I sometimes deleted my comments if they showed as negative.”
- “I would also add nice positive words to get more water drops.”
- “I wanted more positivity so it changed my tone.”

3. Monitoring and Constraint Effects

A minority expressed feelings of surveillance or inhibition:

- “You wouldn’t comment honestly out of fear the tree would be poisoned.”
- “I felt monitored.”

These responses suggest that behavioural redistribution observed at the aggregate level was, for many participants, consciously attributed to the visualization, though perceived mechanisms varied.

3 Discussion

This study examined whether making the cumulative consequences of online commenting persistently visible through an interface-level visualization is associated with changes in commenting behaviour. Across a seven-week field deployment, positive comments increased while neutral comments decreased—declining from 54.8% to 30.8%—with negative comments remaining low. Post-study self-reports indicated high awareness of comment quality and cumulative impact, alongside heterogeneous conscious attribution of influence. Open-ended responses further revealed that many participants perceived the visualization as increasing mindfulness of tone and impact, while others described deliberate tone adjustment to obtain positive reinforcement, and a minority reported feelings of monitoring or constraint. Together, these findings suggest that persistent visualization may support behavioural reorientation through increased awareness and reduced psychological distance, rather than through deterrence or enforcement. This interpretation aligns with evidence that persistent, interpretable visual representations can support reflection by stabilizing otherwise transient behavioural information [29, 30].

A central contribution of this work is the dissociation between situational awareness and explicit attribution of influence. Participants consistently reported awareness of comment tone, valence, and accumulation, while attribution of behavioural influence was more variable. Notably, although 67.6% explicitly reported that the visualization influenced their commenting, qualitative responses indicated variation in how that influence was experienced—ranging from reflective awareness (“It made me more aware of my comments”) to strategic adaptation (“I sometimes deleted my comments if they showed as negative”) to perceived monitoring (“I felt monitored”). This dissociation is theoretically expected: individuals often lack accurate introspection about the factors shaping their behaviour, particularly when influence operates implicitly or through environmental structure [26]. In interface contexts, such influence is commonly produced by ambient feedback that alters interpretation without demanding focal attention [27, 31]. Behavioural trends emerged despite inconsistent attribution, suggesting that awareness functioned as a background condition rather than a uniformly conscious decision process. Persistently externalizing consequences altered how users evaluated their behaviour over time, consistent with theories of social translucence that emphasize visibility over sanction [32]. Research on reflection support similarly highlights sustained exposure and interpretability as prerequisites for reflection [33].

Interpretation of effectiveness warrants caution. Negative comments were infrequent at baseline, introducing a floor effect [25]. An important competing explanation is reactivity. Because the system explicitly tracks comments and externalizes their valence into a persistent artefact, participants may have adjusted behaviour due to Hawthorne effects, evaluation apprehension, or impression-management goals—particularly in a paid study context—rather than through the hypothesized pathway of reduced psychological distance. This interpretation is supported by qualitative responses indicating perceived surveillance or inhibition (e.g., “I felt monitored”; “fear the tree would be poisoned”). In other words, the visualization may function not only as a reflective mirror but also as an evaluative presence, and these mechanisms are difficult to disentangle in a single-arm deployment. Future work should therefore

incorporate randomized controlled or staggered rollout designs, include baseline measurement, and use analytic strategies that can separate novelty/observation effects from sustained changes attributable to the visualization itself (e.g., longer follow-up after habituation, blinded outcome coding, or comparison to a logging-only condition).

The dominant behavioural signal was a reconfiguration of everyday expression: neutral comments declined as positive comments increased. Prior work shows that online degradation often arises from cumulative low-intensity interactions rather than rare extreme behaviour [1, 2]. Shifting routine expression may be more impactful at scale than suppressing already infrequent negativity. Neutral comments often reflect low-effort engagement; by contrast, positive contributions shape participation climates through reinforcement and norm perception [34, 35], and influence conversational norms and subsequent participation [36], [12]. However, because the present study employed a single-condition longitudinal design without randomization or baseline behavioural measurement, observed redistribution should be interpreted as an association during system exposure rather than definitive evidence of causal change.

One plausible mechanism is linguistic convergence under persistent feedback constraints. Qualitative responses indicate that some participants intentionally adjusted phrasing to align with the system’s reinforcement logic, suggesting sensitivity to embedded reward structures. Over time, an asymmetric feedback structure may have encouraged clearer positive expression—an adjustment consistent with differential reinforcement [34] and evidence that persistent feedback can scaffold reflection without interruption [37–39]. Conservative handling of multilingual or informal input may have introduced “linguistic friction,” increasing the salience of expressive clarity [40]. Even if users modified phrasing to “game” the tree metaphor, this represents reflective self-regulation rather than mere compliance [41, 42]. At the same time, reports of feeling monitored suggest that evaluative salience may have contributed alongside reflection and reinforcement. Unlike competitive gamification that can undermine intrinsic motivation [43], this approach leverages the generative role of positive affect in broadening social repertoires [44], though the durability of such effects remains to be established.

The water–poison metaphor played a key role in making cumulative consequences legible. By mapping comment valence onto an ecologically grounded metaphor—nourishing or degrading a tree—the system grounded abstract processes in intuitive schemas [45]. The Water-to-Poison ratio operationalized accumulation as a trajectory, consistent with eco-feedback research showing that metaphor and accumulation support sustained awareness beyond momentary cues [46]. This differs from episodic prompt- or warning-based interventions, which are prone to habituation and warning fatigue [11, 14, 47]. Instead, the visualization may have functioned as a persistent background representation, aligning with reflective design frameworks that prioritize sense-making over enforcement [20, 21]. One might interpret the shift as participants simply substituting neutral comments for positive ones to obtain favorable feedback (e.g., water drops). However, from a behavioral design perspective, this does not undermine the intervention. Feedback structures consistently reshape expressive choices without coercion [48, 49], and positive affect can foster more creative and socially constructive behavior [44]. Whether users were “gaming” the system is less significant than the fact that they modified their expressive framing to align with a

persistent representation of impact—a core tenet of reflective self-regulation [41, 42]. Nevertheless, behavioural shifts cannot be disentangled from the classification architecture and reinforcement thresholds embedded in the system, which constitute integral components of the intervention.

Finally, persistence is critical. By maintaining consequence visibility, the visualization counteracted present bias [48] and mitigated the ephemerality of social media that encourages impulsive engagement [49, 50]. This aligns with recent syntheses emphasizing that durable behaviour change requires mechanisms that support sustained attention alongside point-of-action nudges [51]. While causal efficacy cannot be established within the present design, the findings provide preliminary evidence that persistent visualization infrastructures may reshape everyday expressive behaviour by increasing salience, stabilizing cumulative feedback, and embedding reflection within routine interaction.

4 Limitations and Future Work

Several limitations should be considered when interpreting these findings. First, the study employed a single-condition longitudinal field deployment without a randomized control group or pre-intervention baseline. Consequently, observed behavioural patterns cannot be causally attributed to the visualization alone and may reflect platform dynamics, temporal effects, or external influences. In addition, behavioural trends were estimated from seven weekly aggregated time points, which limits statistical power and precision. Results are therefore interpreted as observed associations rather than evidence of causal change [25]. Future work should incorporate randomized or staggered designs and participant-level modelling to isolate causal mechanisms.

Second, negative comments were infrequent at baseline, introducing a floor effect that limits conclusions about suppression of harmful behaviour. The primary behavioural signal was a redistribution of everyday expression from neutral toward positive comments. Deployments in higher-conflict contexts are needed to assess whether similar mechanisms operate under different baseline conditions.

Third, post-study self-reports were completed by a subset of participants and should be interpreted as descriptive indicators rather than representative of the full cohort. Future studies could integrate in-situ or repeated measures of awareness to capture temporal dynamics more directly.

Fourth, the sentiment classification pipeline handled multilingual and ambiguous input conservatively, which may have influenced expressive clarity and reinforcement dynamics. Future work should explore multilingual or adaptive models and examine how persistent visual feedback functions across cultural and linguistic contexts.

Finally, this study focused on commenters’ behaviour and awareness, not downstream effects on recipients or community-level outcomes. Future research should examine how persistent visualization influences conversational reciprocity, perceived climate, and longer-term social dynamics.

5 Conclusion

This work provides evidence that persistently visualizing the cumulative consequences of online commenting is associated with shifts in everyday expressive behaviour and heightened awareness of social impact during a seven-week field deployment. Rather than suppressing negative behaviour directly, the intervention was accompanied by a redistribution of routine expression toward more positive contributions, consistent with increased salience of consequences and reduced psychological distance.

While causal efficacy cannot be established within the present single-condition design, the findings suggest that interface-level systems that externalize cumulative impact may function as reflective infrastructures embedded within everyday interaction. By foregrounding behavioural accumulation over time, such designs may support longitudinal reflection and expressive recalibration without reliance on coercive enforcement or interruptive prompts.

These results underscore the potential of persistent visualization approaches as a complementary direction in the design of online discourse environments, warranting further experimental validation and examination of longer-term and community-level outcomes.

6 Methods

The study employed a single-condition longitudinal field design without a control group; consequently, analyses focus on temporal patterns and associations rather than causal attribution.

6.1 Participants

Fifty active social media users ($Male = 29, Female = 21, M_{age} = 29.78, SD = 6.28$) were recruited via Prolific. Eligibility criteria required participants to:

1. be active on Instagram or Twitter;
2. use Google Chrome as their primary browser;
3. post at least seven comments per week;
4. use English as a primary language; and
5. provide informed consent for automated comment analysis.

Participants were compensated weekly via Prolific, contingent on active participation throughout the deployment. While all 50 participants contributed behavioural data during the study period, 37 completed the optional post-study questionnaire. Behavioural data from all enrolled participants were retained for longitudinal analysis.

6.2 The Virtual Tree Intervention

The intervention centered on a browser-based visualization that externalized the cumulative impact of a user’s online discourse. Upon enrollment, participants selected one of six tree species (e.g., apple, cherry, coconut palm) to serve as a persistent visual metaphor for their social behavior.

1. Water drops (positive reinforcement): Comments classified as constructive, supportive, or positively valenced increased tree health (0–100%) and growth progress.
2. Poison drops (negative consequence): Hostile or toxic comments triggered reduced growth percentage and decreased health.

The system was non-coercive; it did not block content or issue warnings, functioning solely as a behavioral mirror.

6.2.1 Platform-Specific Data Collection

Data collection was facilitated by a custom Google Chrome extension operating across Instagram and Twitter. The system utilized content scripts to detect comment submission events in real-time. Upon submission, the text was processed via a multi-tiered pipeline:

1. Lexical Filter: A deterministic keyword check for high-intent prosocial or harmful terms.
2. Neural Classification: A transformer-based model (BERT) categorized the sentiment and intent of the discourse.
3. Scoring Logic: Comments were mapped to reinforcement states (+3 for positive, -3 for negative, and +2 for neutral/ambiguous).

Classifier validation metrics are reported in Extended Data (Section A1.2): the fine-tuned BERT model achieved 86.7% accuracy and $F1 = 0.864$ on held-out test data (30,252 comments across six categories). A hybrid architecture combining neural classification with deterministic keyword filtering ensured robustness against model uncertainty and out-of-distribution inputs

Detailed technical specifications of the BERT architecture, training datasets, and platform-specific CSS selectors are provided in the Extended Data section [A](#).

6.3 Procedure

6.3.1 Onboarding (Week 0)

Installation, demographic screening, and system tutorial.

Active intervention(Weeks 1–7).

Participants used social media naturally. The extension provided real-time feedback via post-comment overlays, a persistent ambient health indicator, and a progress dashboard.

Post-study.

An exit survey assessed perceived awareness and behavioral reflection ($N = 37$).

6.4 Measures

6.4.1

Primary behavioural outcome.

Weekly sentiment distribution of comments (positive, neutral, negative), derived from automated classification.

Secondary behavioural outcomes.

1. Water-to-poison ratio per week
2. Longitudinal trends in sentiment proportions

Self-report outcomes.

Post-study questionnaire assessing awareness of comment quality, perceived distance from effect, behavioural orientation, and perceived influence (N = 37 respondents).

6.5 Statistical Analysis

Behavioural data were analysed descriptively and longitudinally using Python (v3.11). Weekly sentiment proportions ($N = 7$) and cumulative feedback metrics (Water/Poison counts and ratios) were examined using Ordinary Least Squares (OLS) linear regression via the *SciPy* `linregress` module to assess temporal trends over the seven-week deployment. For each metric, we calculated the unstandardised slope (β), coefficient of determination (R^2), and the two-sided p -value. To ensure the robustness of our trend estimates, 95% confidence intervals were derived using bias-corrected accelerated (BCa) bootstrapping with 10 000 resamples. Effect sizes were quantified using Cohen's $f^2 = R^2 / (1 - R^2)$, where $f^2 \geq 0.35$ was interpreted as a large effect.

Post-study self-report results ($n = 37$) were analysed using non-parametric statistics via the *Pingouin* (v0.5) and *SciPy* libraries. For each of the four questionnaire sections (Perception, Distance, Awareness, and Behavioural Orientation), participant-level mean scores were compared against the neutral midpoint of the 5-point Likert scale ($\tilde{x}_0 = 3$) using one-sample Wilcoxon signed-rank tests (`scipy.stats.wilcoxon`, `zero_method = "zsplit"`, `mode = "exact"`). To control for family-wise error across the four predefined theoretical sections, a Bonferroni correction was applied, setting the significance threshold at $\alpha = 0.0125$. Effect sizes for these tests were calculated as $r = Z / \sqrt{n}$. All reported p -values are exact and two-tailed. Data processing, statistical computation, and visualization were supported by *pandas* (v2.0), *SciPy* (v1.10), *Seaborn* (v0.12), and *Matplotlib* (v3.7).

References

- [1] Slonje, R., Smith, P. K. & Frisén, A. The nature of cyberbullying, and strategies for prevention. *Computers in Human Behavior* **29**, 26–32 (2013).
- [2] Cheng, J., Danescu-Niculescu-Mizil, C. & Leskovec, J. Antisocial behavior in online discussion communities. *Proceedings of the Ninth International AAAI Conference on Web and Social Media* 61–70 (2015).

- [3] Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C. & Leskovec, J. Anyone can become a troll: Causes of trolling behavior in online discussions. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* 1217–1230 (2017).
- [4] Kiesler, S., Siegel, J. & McGuire, T. W. Social psychological aspects of computer-mediated communication. *American Psychologist* **39**, 1123–1134 (1984).
- [5] Suler, J. The online disinhibition effect. *Cyberpsychology & Behavior* **7**, 321–326 (2004).
- [6] Jhaver, S., Birman, I., Gilbert, E. & Bruckman, A. Human-machine collaboration for content regulation: The case of Reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* **26**, 1–35 (2019).
- [7] Lapidot-Leffer, N. & Barak, A. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior* **28**, 434–443 (2012).
- [8] Derks, D., Fischer, A. H. & Bos, A. E. R. The role of emotion in computer-mediated communication: A review. *Computers in Human Behavior* **24**, 766–785 (2008).
- [9] Baumeister, R. F., Vohs, K. D., DeWall, C. N. & Zhang, L. How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. *Personality and Social Psychology Review* **11**, 167–203 (2007).
- [10] Kruger, J., Epley, N., Parker, J. & Ng, Z.-W. Egocentrism over e-mail: Can we communicate as well as we think? *Journal of Personality and Social Psychology* **89**, 925–936 (2005).
- [11] Munger, K. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior* **39**, 629–649 (2017).
- [12] Matias, J. N. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* **116**, 9785–9789 (2019).
- [13] Jhaver, S., Ghoshal, S., Bruckman, A. & Gilbert, E. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* **25**, 1–33 (2018).
- [14] Pennycook, G. *et al.* Shifting attention to accuracy can reduce misinformation online. *Nature* **592**, 590–595 (2021).
- [15] Pennycook, G. & Rand, D. G. Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*

13, 2333 (2022).

- [16] Eslami, M. *et al.* First i like it, then i hide it: Folk theories of social feeds. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* 2371–2382 (2016).
- [17] Fiesler, C., Jiang, J., McCann, J., Frye, K. & Brubaker, J. Reddit rules! characterizing an ecosystem of governance. *Proceedings of the Twelfth International AAAI Conference on Web and Social Media* 72–81 (2018).
- [18] Lakoff, G. & Johnson, M. *Metaphors We Live By* (University of Chicago Press, Chicago, 1980).
- [19] Tractinsky, N. & Meyer, J. Chartjunk or goldgraph? effects of presentation objectives and content desirability on information presentation. *MIS Quarterly* **23**, 397–420 (1999).
- [20] Dourish, P. *Where the Action Is: The Foundations of Embodied Interaction* (MIT Press, Cambridge, MA, 2001).
- [21] Sengers, P., Boehner, K., David, S. & Kaye, J. J. Reflective design. *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility* 49–58 (2005).
- [22] Bagnara, S. & Pozzi, S. Design for reflection. *Work* **41**, 1108–1113 (2012).
- [23] Mayer, R. E. Cognitive theory of multimedia learning. *The Cambridge Handbook of Multimedia Learning* 31–48 (2005).
- [24] Baumer, E. P. S. *et al.* Limiting, leaving, and (re) lapsing: an exploration of Facebook non-use practices and experiences. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 3257–3266 (2013).
- [25] Shadish, W. R., Cook, T. D. & Campbell, D. T. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Wadsworth Cengage Learning, Boston, MA, 2002).
- [26] Nisbett, R. E. & Wilson, T. D. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* **84**, 231–259 (1977).
- [27] Norman, D. *The Design of Everyday Things: Revised and Expanded Edition* (Basic Books, New York, 2013).
- [28] Mircioiu, C. & Atkinson, J. A comparison of parametric and non-parametric methods applied to a Likert scale. *Pharmacy* **5**, 26 (2017).
- [29] Meyer, M. & Dykes, J. Criteria for rigor in visualization design study. *IEEE Transactions on Visualization and Computer Graphics* **26**, 87–97 (2019).

- [30] Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M. & Hullman, J. The science of visual data communication: What works. *Psychological Science in the Public Interest* **22**, 110–161 (2021).
- [31] Wisneski, C. *et al.* Ambient displays: Turning architectural space into an interface between people and digital information. *International Workshop on Cooperative Buildings* 22–32 (1998).
- [32] Erickson, T. & Kellogg, W. A. Social translucence: an approach to designing systems that support social processes. *ACM Transactions on Computer-Human Interaction (TOCHI)* **7**, 59–83 (2000).
- [33] Cho, J., Xu, T., Zimmermann-Niefeld, A. & Vaida, S. Reflection in theory and reflection in practice: an exploration of the gaps in reflection support among personal informatics apps. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* 1–23 (2022).
- [34] Skinner, B. F. Reinforcement today. *American Psychologist* **13**, 94–99 (1958).
- [35] Suls, J., Martin, R. & Wheeler, L. Social comparison: Why, with whom, and with what effect? *Current Directions in Psychological Science* **11**, 159–163 (2002).
- [36] Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M. & Leskovec, J. Can cascades be predicted? *Proceedings of the 23rd International Conference on World Wide Web* 925–936 (2014).
- [37] Murnane, E. L. *et al.* Designing ambient narrative-based interfaces to reflect and motivate physical activity. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* 1–14 (2020).
- [38] Janzen, I. F. & McGrenere, J. Reflective spring cleaning: Using personal informatics to support infrequent notification personalization. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* 1–16 (2022).
- [39] Tabuenca, B., Börner, D. & Kalz, M. Effects of an ambient learning display on noise levels and perceived learning in a secondary school. *IEEE Transactions on Learning Technologies* **14**, 69–80 (2021).
- [40] Loerakker, M. B., Niess, J., Bentvelzen, M. & Woźniak, P. W. Designing data visualisations for self-compassion in personal informatics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **7**, 1–22 (2024).
- [41] Schön, D. A. *The Reflective Practitioner: How Professionals Think in Action* (Routledge, New York, 2017).
- [42] Carver, C. S. & Scheier, M. F. *On the Self-regulation of Behavior* (Cambridge University Press, Cambridge, UK, 2001).

- [43] Deci, E. L., Koestner, R. & Ryan, R. M. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin* **125**, 627–668 (1999).
- [44] Fredrickson, B. L. The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American Psychologist* **56**, 218–226 (2001).
- [45] Lakoff, G. & Johnson, M. *Metaphors We Live By* (University of Chicago Press, Chicago, 2024).
- [46] Chalal, M. L., Medjdoub, B., Bezai, N., Bull, R. & Zune, M. Visualisation in energy eco-feedback systems: A systematic review of good practice. *Renewable and Sustainable Energy Reviews* **162**, 112447 (2022).
- [47] Terrell, J. *et al.* Gender differences and bias in open source: Pull request acceptance of women versus men. *PeerJ Computer Science* **3**, e111 (2017).
- [48] Thaler, R. H. & Sunstein, C. R. *Nudge: Improving Decisions About Health, Wealth, and Happiness* (Penguin Books, New York, 2009).
- [49] Fogg, B. J. A behavior model for persuasive design. *Proceedings of the 4th International Conference on Persuasive Technology* 1–7 (2009).
- [50] Bayer, J. B., Ellison, N. B., Schoenebeck, S. Y. & Falk, E. B. Sharing the small moments: ephemeral social interaction on Snapchat. *Information, Communication & Society* **19**, 956–977 (2016).
- [51] Kozyreva, A. *et al.* Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour* **8**, 1044–1052 (2024).
- [52] Aggarwal, K., Bamdev, P., Mahata, D., Shah, R. R. & Kumaraguru, P. Trawling for trolling: A dataset. *arXiv preprint arXiv:2008.00525* (2020).

Ethical approval

This study involved human participants and the analysis of behavioral data. The authors confirm that all research was performed in accordance with relevant guidelines and regulations, specifically the 1964 Declaration of Helsinki and its later amendments.

The Daegu Gyeongbuk Institute of Science and Technology (DGIST) Institutional Bioethics Committee granted a formal exemption for this research (Approval Number: DGISTIRB-202604-001; Assignment Number: DGISTIRB-2026-008-08). The exemption was officially deliberated and provided on 09/04/2026. Prior to the commencement of the research in November 2025, the authors conducted a self-assessment using the institutional review exemption self-checklist via the research portal¹, establishing the project’s exempt status in accordance with university policy. The authors confirm that all research was performed in accordance with relevant guidelines and regulations, specifically the 1964 Declaration of Helsinki and its later amendments. The basis for exemption was that the study involved the observation of de-identified behavioral data posing minimal risk to participants, pursuant to Article 15, Paragraph 2 of the Bioethics and Safety Act of the Republic of Korea and Article 13 of its Enforcement Rules².

Informed consent

Informed consent was obtained from all participants involved in the study on 13/11/2025. Consent was obtained electronically via the Prolific participation platform by the research team. Prior to participation, all individuals were presented with a digital consent form detailing the study’s purpose, procedures, risks, and data handling practices. Consent was verified through a mandatory digital “I Agree” selection within the platform prior to any data collection. This digital record serves as the verification of written consent in an online format. All data were fully anonymized prior to analysis.

Data availability

The de-identified datasets generated and analyzed during the current study are available in the GitHub repository: <https://codeocean.com/capsule/3669883/tree>. These data are stripped of all platform-specific metadata to further protect participant privacy.

Code availability

The source code for the Chrome extension, the custom-trained sentiment analysis model (including architecture and weights), and the data processing pipeline are publicly available at: <https://anonymous.4open.science/r/Tree-Comment-Grower-9D56/>.

¹Exemption protocol available at: https://irb.dgist.ac.kr/02_gid/gid03.html

²Bioethics and Safety Act: https://elaw.klri.re.kr/eng_mobile/viewer.do?hseq=33442&type=part&key=36

Author contributions

Information has been withheld for anonymous review

Supplementary information

The online version of this article contains supplementary material, which is available to authorized users.

The Supplementary Information is provided as a separate PDF file and includes the following:

- **Supplementary Note 1:** Detailed design logic of the Tree Growth Game browser extension.
- **Supplementary Figure 1:** Interface gallery illustrating the participant onboarding process, tree species selection, and dashboard.

The source code for the Chrome extension and the anonymized sentiment dataset (mapped via randomized Tree IDs) are available via the provided repository link.

Appendix A Extended Data

A.1 Chrome Extension Architecture

We developed a custom Google Chrome extension that operated continuously in the background while participants used social media. The extension consisted of three components:

1. Content scripts injected into supported social media pages to detect comment submission events
2. A background service worker responsible for communication with external APIs and data persistence
3. A visual feedback layer that rendered tree animations and status indicators within the browser

Comment detection occurred at the moment of submission, after which the text was processed by the classification pipeline and feedback was rendered.

A.1.1 Visual Feedback System

The intervention’s interface included multiple components designed to balance salience and non-intrusiveness:

1. **Post-comment overlay:** A brief, non-blocking notification indicating the drops awarded and tree status
2. **Tree visualization:** An animated tree graphic reflecting health and growth
3. **Progress dashboard:** A cumulative summary of comment activity, sentiment ratios, and trends
4. **Ambient health indicator:** A persistent but unobtrusive bar reflecting tree condition

Priority-Based Lexical Filtering.

Before model inference, comments passed through a deterministic keyword filter, triggering immediate feedback for high-priority terms across toxic categories: (1) Hate Speech, (2) Derogatory, (3) Microaggression, (4) Profanity, and (5) Trolling. Positive keywords (e.g., “thanks”, “insightful”) were also included to ensure reinforcement for constructive behavior.

Inference and Punctuated Scoring Logic.

Each comment was assigned a label and confidence score (*conf*), mapped to reinforcement states:

1. **Positive Sentiment (+3):** High-confidence Normal comments received maximum reward
2. **Negative Sentiment (−3):** Toxic category matches triggered penalties
3. **Neutral Sentiment (+2):** Low-confidence or OOD comments (e.g., Spanish or Hindi) received fallback neutral reward

System Resilience and Fallback Logic.

A fail-safe mechanism ensured consistent feedback: if the BERT model was unavailable, keyword filtering ensured continuity (see Table A2).

Linguistic Bias and Fallback Interpretation.

Deployment-time logic (keyword + model) reduced the negative sentiment baseline to 6.7%, compared to 51.03% under raw AI-only reanalysis. This illustrates the filter’s role in mitigating semantic over-sensitivity and avoiding penalization of ambiguous user input.

A.2 Hybrid AI–Lexical Comment Classification

Model Architecture.

We used a 12-layer BERT transformer architecture fine-tuned from *dbmdz/bert-base-turkish-uncased*, retrained on English corpora. This architecture introduced conservative confidence thresholds to reward only high-certainty prosocial expression.

Training Data and Labels.

Using the *Trawling for Trolling* dataset [52], we fine-tuned the model on six categories: *Normal*, *Trolling*, *Profanity*, *Derogatory*, *Hate Speech*, and *Microaggression*. Oversampling was used to balance underrepresented classes, yielding $N = 30,252$ comments across standard train/val/test splits.

Training Procedure and Performance.

Model training used PyTorch and GPU acceleration. Final test accuracy was 0.867, with an F1 score of 0.864 (see Table A1).

Table A1: Model performance across splits

Split	Accuracy	F1	Precision	Recall
Train	0.951	0.951	0.951	0.951
Validation	0.860	0.860	0.859	0.861
Test	0.867	0.864	0.864	0.865

Keyword Lexicons and Priority Mapping.

The full keyword lexicon used in the deterministic layer is shown in Table A2. Word-boundary regular expressions prevented false positives.

Scope and Limitations.

This classifier was designed for coarse sentiment and harm labeling to drive visual feedback—not for granular moderation. Outputs were treated as probabilistic signals. Classification errors and biases (e.g., around irony or reclaimed language) may persist, particularly for linguistically or culturally ambiguous expressions. The keyword filtering layer reduced baseline negative classification from 51% (raw model) to 6.7% (deployment), mitigating semantic over-sensitivity to ambiguous or informal language

Table A2: Harmful and Prosocial Category Keyword Lexicons by Priority Order

Priority	Category	Keywords
1	HATE SPEECH	hate, kill, die, death, racist, nazi, kys, hang, lynch, genocide, terrorist
2	DEROGATORY	idiot, moron, stupid, dumb, loser, pathetic, worthless, trash, scum, garbage, waste
3	MICROAGGRESSION	actually, mansplain, you people, one of the good ones, not like other, surprisingly articulate, exotic, where are you really from
4	PROFANITY	fuck, shit, damn, bitch, ass, bastard, crap, piss, cock, hell, goddamn
5	TROLLING	lol, lmao, cry, cope, seethe, mald, ratio, l+ratio, bozo, skill issue, mad, salty, triggered
-	POSITIVE	great, awesome, love, amazing, wonderful, excellent, fantastic, good, nice, beautiful, helpful, thanks, thank you, appreciate, well done, brilliant, perfect, agree, support, insightful, interesting, cool, respect

Table A3: Raw model sensitivity analysis (strict AI-only logic)

Category	Week 1 %	Week 7 %	Δ change
Positive (+3)	35.13	55.88	+20.75
Negative (−3)	51.03	41.67	−9.36
Neutral (+2)	13.85	2.45	−11.40

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryNote.pdf](#)
- [InterfaceGallery.pdf](#)