Intern. J. Comput.-Support. Collab. Learn
https://doi.org/10.1007/s11412-025-09460-9

RESEARCH

Learning to deal with hate speech: An online collective intelligence experiment on the Collective Learning platform

Tatiana Íñiguez-Berrozpe[1] · Carmen Elboj-Saso[1] · Francesco Marcaletti[2] · Pablo Bautista-Alcaine[3]

Received: 16 December 2024 / Accepted: 29 October 2025

Abstract

Background: Online hate speech on social networks and the Internet is an increasingly pervasive phenomenon to which both children and adolescents are exposed. **Objective:** Our study's main objective was to ascertain whether collective intelligence can improve their handling of hate speech. **Methods:** We conducted the study on the Collective Learning platform, comparing results between three groups of Spanish adolescents aged 15–16 years. The groups were of different sizes: one large group (G1, $n = 123$) and two smaller groups (G2, $n = 18$; G3, $n = 23$). **Results:** The experiment showed that the conditions for the emergence of collective intelligence were met within the large group (G1) but not in the two small groups (G2 and G3). The large group, as a collective, acquired capacities to deal with hate speech; however, this did not occur in the two smaller groups. **Conclusions:** Our study explains how the emergence of collective intelligence in online environments helps group members acquire a series of competencies. In particular, collective intelligence can help adolescents learn to deal with hate speech.

Keywords Learning · Hate speech · Collective intelligence · Children · Adolescents

Online hate speech in children and adolescents

The phenomenon of online hate speech among children and adolescents is an issue of growing concern in today's society. Apart from the easy accessibility of the Internet, the proliferation of social networking platforms, online games, and chat rooms has accelerated the speed at which hate speech can spread, while expanding its scope. This has severe implications for the emotional well-being and safety of young people.

✉ Pablo Bautista-Alcaine
pbautista@unizar.es

[1] Faculty of Education, Department of Psychology and Sociology, University of Zaragoza, Zaragoza, Spain

[2] Faculty of Social and Labor Sciences, Department of Sociology, University of Zaragoza, Zaragoza, Spain

[3] Faculty of Education, Department of Educational Sciences, University of Zaragoza, Zaragoza, Spain

Published online: 22 November 2025

Several studies have highlighted the pervasiveness of online hate speech among youth. Hinduja and Patchin (2020) revealed that more than one-third of the adolescents they surveyed had experienced some form of online harassment, including hate speech in the form of racist, sexist, homophobic, and xenophobic comments. Obermaier and Schmuck (2022) found that witnessing online hate speech is a common experience among young people and can have adverse effects on their psychological well-being. Their findings indicated that victimization from online hate speech is a significant phenomenon among youth and is associated with a negative impact on their emotional well-being.

Online hate speech can manifest itself in various forms, including verbal attacks in the form of comments, messages, or posts containing insults, threats, or attacks directed toward an individual or a specific group on the basis of their identity, such as ethnicity, gender, sexual orientation, religion, or disability (Crandall et al., 2002). Defamation involves the dissemination of false or misleading information with the intention of damaging the reputation of a person or group, which may include false accusations related to personal characteristics protected by law (Cohen-Almagor, 2011). Ethnic discrimination consists of comments, memes, or images that denigrate or stereotype people who belong to certain ethnic groups, thereby promoting intolerance and exclusion (Livingstone et al., 2017). Misogyny is identified with content that expresses hatred or contempt toward women, e.g., promoting sexist stereotypes, reifying women, or justifying gender violence (Jane, 2014). Homophobia is expressed in messages or comments that denigrate, ridicule, or dehumanize people, including lesbian, gay, bisexual, or transgender individuals (Liu and Koerner, 2017). Religious intolerance is reflected in postings that attack or belittle people on the basis of their religious affiliation, thereby promoting discrimination based on religious beliefs (Ramponi et al., 2022). Exposure to all these forms of online hate speech can harm the emotional and psychological well-being of young people by increasing their risk of anxiety, depression, and low self-esteem (Canadian Human Rights Commission, 2019). Hate speech can also influence the perceptions and attitudes of young people toward certain social groups, fostering intolerance and discrimination while at the same time reinforcing negative attitudes, which children and adolescents even adopt toward those groups. This, in turn, hinders the advancement of equality and social inclusion (Livingstone et al., 2017). Moreover, this type of discourse can prevent constructive dialogue and cooperation among adolescents and young people with divergent opinions and perspectives. This, in turn, hinders the construction of a more cohesive and tolerant society; instead, it exacerbates polarization and fragmentation (Liu and Koerner, 2017; Livingstone et al., 2017). Online hate speech can even fuel the rise of physical violence based on discrimination by legitimizing violent attitudes and behaviors toward certain groups. Hate messages can incite harassment, bullying, aggression, and even extreme violence against individuals, including adolescents or marginalized communities. This poses a severe threat to the safety and well-being of all members of society as a whole (Cohen-Almagor, 2011). In short, online hate speech not only affects victims on an individual level, but also on a larger scale, as it prevents social cohesion and peaceful coexistence.

Children and adolescents are not only the victims of online hate speech but can also spread and multiply it. Exposure to intolerant or denigrating content can influence young people's perception of certain social groups and increase their disposition to participate in discriminatory acts. Recent research has revealed the prevalence of hate speech and its effects on younger population groups. Wachs et al. (2022) showed that young people can act as enablers of online hate speech in various ways, for instance, by disseminating discriminatory content or participating in online communities that promote extremist ideologies. Certain young people can fall under the influence of public leaders or peer groups that promote hate speech. Other young people, in turn, are victims; they can be the objects of harassment, discrimination, or verbal attacks owing to their age, gender, ethnicity, or other characteristics (Wachs et al., 2022). This type of experience can have devastating consequences for young people's emotional well-being and their sense of belonging. Wachs et al. (2022c) noted that young people can encounter hate speech online and develop strategies to deal with it; their findings indicate that young people are indeed exposed to this phenomenon and that they also have the capacity to adopt approaches that protect their emotional and psychological well-being.

Strategies adopted by minors to cope with online hate speech

As noted in the previous section, children and young people are confronted with hate speech both offline and online. The fact that they are involved in hate speech not only as spectators but also as victims and/or aggressors should be cause for concern. Researchers in the fields of education and psychology have thus started searching for potential prevention and coping strategies, particularly designed to deal with hate speech in its manifestations on the Internet, on social networks, in online games, and in all other virtual environments where minors can interact with one another (Kansok-Dusche et al., 2022).

There is no doubt that the impunity, intransigency, and viral nature of hate speech—especially on the Internet—make the task of developing effective coping strategies quite complex. Authors, including Castellanos et al. (2023), argue that such strategies should be deployed in environments where children and adolescents tend to undergo socialization, thus, particularly in schools. Schools can implement such strategies on several levels: contextually (for example, by promoting a positive school climate), interpersonally (by encouraging a collective effort in class to combat hate speech), and individually (by fostering empathy and a prosocial attitude in each student). The positive impact of those synergies has been proven empirically. Certain school programs have chosen to combine the positive effects of several interaction levels to oppose hate speech (programs such as "HateLess;" cf. Wachs et al., 2023b), bullying (KiVa; Garandeau et al., 2022), and cyberbullying (cf. the program "Medienhelden," or "Media Heroes"; Schultze-Krumbholz et al., 2018). Hate speech can be prevented by applying some of the same strategies that are used for the prevention of bullying and cyberbullying, given that they impinge upon the same moral and emotional areas of the psyche (Castellanos et al. (2023); Kansok-Dusche et al., 2022). We would add a further overarching level that operates beyond the students' school environment: a systemic level. Strategies on the systemic level include, for instance, demanding that online platforms invest greater efforts in dissuading users from posting or viewing content that contains hate speech, particularly since mere content filtering has not proven effective in preventing it (Griffin, 2022**).**

Our study centered on the meso-systemic and micro-systemic levels of students as individuals relating to one another in groups. The term meso-systemic here refers to the immediate social contexts in which adolescents interact—such as school environments, peer groups, and teacher relationships—while micro-systemic denotes the personal and intrapersonal level, including emotional, moral, and cognitive processes that influence how individuals experience and react to hate speech (El Zaatari & Maalouf, 2022). Therefore, in the following theoretical overview, we focus on the literature that has dealt with those levels. Most authors who have analyzed hate speech on a contextual level among children and primarily adolescents have highlighted the important role that the students' social and school environment plays in helping them identify and combat hate speech online and offline (Kansok-Dusche et al., 2023), as well as in encouraging them to support hate speech victims (Markogiannaki et al., 2021). The same authors have pointed out that teachers, peers, families, and schools can work hand in hand to gradually adopt and apply healthy guidelines that encourage respect for diversity while promoting empathy and harmonious coexistence on the basis of mutual respect and acceptance. Victimization can also be avoided by ensuring appropriate socialization and the formation of healthy bonds among classmates. In their study, Stahel and Baier (2023) pointed out that victimization resulting from online and offline hate speech was a factor that correlated positively with loneliness. Promoting a positive, inclusive climate can help prevent the emergence of hate speech by fomenting empathy toward actual or potential victims and ensuring the efficacy of concrete interventions whenever hate speech occurs (Wachs et al., 2023a). Victims can find support in social "security nets" (Wachs et al., 2023c). Thus, prevention programs against hate speech should not only take intrapersonal factors into account but also social norms and the aspect of personal interrelationships, for example, those that emerge in class (Wachs, et al., 2022a). We further explore that potential aspect in the following paragraphs.

Depending on the contextual level, but descending to the interpersonal level, it is necessary to take peer group interactions into account, as the motivation to perpetrate hate speech mainly stems from the social dynamics of a given social environment, where reciprocal reactive processes, social recognition, social pressure, and imitation may play an important role, especially among adolescents (Wachs et al. 2022b).

That same research team has pointed out that adolescents' motivations are influenced mainly by behaviors they observe in their social environment (descriptive norms) or behaviors they are actively encouraged to adopt (peer pressure) and only to a lesser extent by their perception of which behaviors are approved or disapproved (imperative norms). Given that school is the principal place where adolescents are socialized by their peers, hate speech prevention strategies in school should pay special heed to the danger caused by peer pressure, as the latter tends to normalize insults until they become standard behavior (Wachs et al., 2022a).

One level lower, on the personal level, but taking the above level into account, educators can generate dynamics that foment a more harmonious atmosphere in the classroom, combined with enhanced student awareness of the problems caused by hate speech (Wachs et al., 2023c). For example, Obermaier (2024) showed that the adoption of anti-hate-speech campaigns that encourage a prosocial attitude tends to increase the active participation of hate-speech bystanders in defense of victims.

Wachs et al. (2022b) suggest that teachers present conflict management strategies in the classroom while fomenting adolescents' competency to deal with frustration and negative emotions. This should encourage young people to avoid using hate speech as a means of vengeance. Several authors agree on the importance of developing student empathy and moral commitment, two factors strongly associated with hate speech prevention, according to research conducted by Castellanos et al. (2023). Those authors and Bustamante and Chaux (2014) highlight the importance of developing critical thinking. Critical thinking capacity can increase students' moral commitment, which, in turn, should decrease the amount of hate speech. Further studies recommend the development of practical skills, such as news literacy and digital media literacy, which help young people detect and prevent incidences of hate speech in the media (Samy-Tayie et al., 2023; Obermaier & Schmuck, 2022).

Hate speech tends to be directed toward specific groups: women, immigrants, religious minorities, politically committed individuals, and members of the LGBTI+ community (Obermaier & Schmuck, 2022; Castellanos et al., 2023b). Prevention strategies foresee a proper approach to intergroup conflict by applying empathy (as described above) and extending one's moral commitment to social groups outside one's own (Castellanos et al., 2023a). The same applies to online interactions. On social networks, contact with members outside of an individual's habitual social group predicts a decrease in prejudice and an increase in self-confidence (Schumann & Moore, 2022). The tendency to undertake collective actions in favor of outside groups likewise increases, and the occurrence of hate speech is thereby diminished. Hate speech prevention strategies should thus not only encourage harmonious offline interaction among adolescents but also harmonious online interaction with members outside their habitual social group.

Given that research on hate speech among minors (mainly online) is a recent field, few types of prevention strategies have been applied, and even fewer have been scientifically evaluated. One such recent program is HateLess, developed by Wachs et al. (2023b). Operating on several levels (individually, in the classroom, in the school, and the community at large), the HateLess program's objective consists in preventing the perpetration and victimization of hate speech among adolescents by proposing films, stories, and role-playing scenarios designed to improve professional competencies (e.g., factual knowledge about hate speech), self-competencies (e.g., counter-speech, self-efficacy, coping strategies), emotional competencies (e.g., empathy, moral engagement), social competencies (e.g., cooperative competencies), and methodological competencies (e.g., ethical media competencies). In their study, Wachs et al. (2023b) showed that the HateLess program proved effective in preventing hate speech among adolescents while at the same time improving the competencies they needed to become well-informed citizens of a democratic society.

Another proposition designed to thwart hate speech is "Philosophy for Children," a program designed byc. This program is also oriented toward adolescents. "Philosophy for Children" is based on five pillars: (1) developing critical/creative thinking, (2) custody of careful thoughts (being aware of my words and thoughts); (3) care and concern for others; (4) acquiring one's own identity; and (5) using narrated material and staged scenarios within a "community of examination" to collectively and cooperatively investigate a common problem or theory. Similar to other authors, Barrientos Rastrojo emphasizes the collective, i.e., a group of people who apply multidimensional thinking to explore an issue together. Five adjectives can be used to describe such "communities of examination." They are *communal* (a group of individuals are gathered in a community centered on a project), *procedural* (the community operates by strictly following a chosen dialogue structure and a set of rules), *critical* (not all ideas proposed are equally valid), *self-corrective* (ideas gradually "correct themselves" on an individual level and a group level; the student becomes aware of their own potential and limitations by dialoguing with the community, observing how one can influence others and be influenced by their feedback), and *artisanal* (the outcome is not a material product but results from the interaction among individual abilities incorporated into a network of multidimensional thought).

Taking these approaches into account, we examined the usefulness of collective intelligence for the prevention of hate speech. As explained below, our model complied with the premises enumerated above; at the same time, using this model, we attempted to take a further step in the field of collective hate speech prevention.

The potential role of collective intelligence in hate speech prevention

Collective intelligence is defined as the performance of a large group of people collaborating on a series of complex tasks, generating responses and/or solutions in an aggregative manner, based on the idea that the mean individual performance of each member is lower than the mean performance obtained by the group (Woolley et al., 2010; Woolley & Aggarwal, 2020). The construct of collective intelligence goes beyond the concept of group learning: The latter is defined as the change that emerges within a group during work on a task. The change occurring in group learning arises from the experience shared by the group's members; it can be observed in terms of cognition, successive task organization, and performance, namely, results (Argote & Miron-Spektor, 2011). Ever since the first study on collective intelligence published by Woolley et al. (2010), many different types of experiments have been conducted, including attempts to apply collective intelligence to resolve complex tasks such as the resolution of moral dilemmas or the navigation of conflict resolution (Hjertø & Paulsen, 2016; Meslec et al., 2016): These latter attempts have used technology to replace face-to-face interaction with online interaction. The introduction of technology as an intermediary has allowed researchers to substantially increase the number of participants, thereby expanding the problem-solving potential of collective intelligence (Orejudo et al., 2022; Woolley & Aggarwal, 2020). Certain problems that emerged in face-to-face groups (for instance, in resolving moral dilemmas) can be overcome, for instance, when certain individual members do not collaborate or when proposals tend to be unoriginal owing to an overproduction or overpropagation of answers (Toyokawa et al., 2019). Moreover, the online environment allows researchers to install an instance of artificial intelligence for the purpose of moderating collective intelligence (Orejudo et al., 2022). Such a moderating instance is essential to guarantee the production of answers that are of high quality and genuinely original, while guaranteeing a high degree of social interaction and productivity (Bernstein et al., 2018; Navajas et al., 2018).

Collective intelligence with this type of setup has been experimentally applied to groups of adolescents confronted with the resolution of an ethical dilemma (Orejudo et al., 2022; Bautista et al., 2022). On the basis of a cognitive-evolutionary approach, those studies were designed to apply collective interaction to lead adolescents toward attaining a level of ethical reasoning superior to the level they would normally achieve in face-to-face situations. Indeed, in face-to-face collective intelligence experiments where groups were asked to solve moral dilemmas, certain individuals tended to impose their perspective on others (Kohlberg, 1989). However, Orejudo et al. (2022) and Bautista et al. (2022) designed experiments that confronted groups of students with an ethical conflict to encourage discussion among peers. Those two studies showed that the general level of moral reasoning could be significantly elevated by applying collective intelligence.

We thus proposed the hypothesis that collective intelligence experiments can contribute to preventing and managing hate speech situations among adolescents. Our experimental model took most of the abovementioned premises in the area of hate speech management into account: Collective intelligence is based on interaction within a group or community (Wachs et al., 2022a); it extends its reflections beyond the group context (Castellanos et al., 2023a), encouraging online interaction among peers (Schumann & Moore, 2022) within a structured environment that allows for self-correction (Barrientos Rastrojo, 2022). It can lead participants to develop critical thinking (Bustamante & Chaux, 2014; Castellanos et al., 2023b), empathy, and a prosocial attitude (Obermaier, 2022; Wachs et al., 2023b) while ensuring that the process per se is regarded as equally important as the outcome (Barrientos Rastrojo, 2022). Thus, basing ourselves on previous research in the areas of hate speech prevention and/or collective intelligence, we aimed to reach two goals.

Regarding our knowledge of collective intelligence and the Collective Learning platform (which we used in this study), objective no. 1 was to ascertain whether a large-size group performs better on a task than several smaller groups. This required us, more specifically (objective no. 1a), to determine whether differences observed between large and small groups could be ascribed to the variables listed by Woolley and Aggarwal (2020) or whether they were due to the platform design in itself, coupled with the content of the case study we proposed. Objective no. 2 was to ascertain whether our participants had a *learning experience*, i.e., whether they acquired competencies that would help them confront hate speech on the Internet.

Technology, ethics, and value-based education: our approach

The integration of collective intelligence platforms into educational settings requires reflection not only on their technical potential but also on their ethical implications and underlying educational values. In our study, the use of the Collective Learning platform was not conceived as a neutral tool but as part of a pedagogical vision that emphasizes the promotion of critical thinking, moral engagement, and democratic deliberation among adolescents.

Our approach is based on the belief that digital technologies in education should be aligned with values such as equity, inclusion, and social responsibility. As Kizilcec and Lee (2022) argue, technological tools should be designed and implemented in ways that promote equitable learning opportunities and respect for human dignity. In our experiment, the role of the artificial intelligence (AI) moderator was carefully constructed to encourage equal participation, reduce hierarchical power structures, and ensure that the most constructive ideas, rather than the most dominant voices, gained visibility and influence.

Furthermore, the pedagogical architecture of the platform reflects a relational understanding of moral learning in line with Kudina's (2023) perspective on moral hermeneutics and technology. Rather than promoting normative compliance, the system encourages participants to engage in situated ethical reasoning through dialogic processes that reflect the moral complexity of real life. The collective nature of decision-making, the opportunity for self-correction, and the transparency of interactions contribute to the emergence of what Kudina (2023) describes as "the construction of moral meaning through the relationships between humans, technology, and the world."

In this way, our intervention is situated within a broader educational commitment to values-based digital literacy. It seeks not only to prevent online hate speech but also to promote deeper civic and ethical engagement among young people. Through structured online interaction, teenagers are invited to go beyond reactive responses and develop a reflective, collaborative, and empathetic stance toward social conflicts in digital environments.

Synthesis of theoretical background

To conclude our overview: Online hate speech poses a significant challenge to the emotional and social well-being of children and adolescents in the digital age. Its pervasiveness on social media increases the risk of anxiety, depression, and low self-esteem among young people, while also fueling polarisation and undermining social cohesion. As shown in the reviewed literature, young people are exposed to hate speech not only as victims but also as spectators or facilitators. Several studies highlight the importance of promoting educational strategies that develop empathy, digital literacy, and moral engagement to prevent this type of behavior. Schools are in a unique position to intervene at multiple levels: individual, interpersonal, contextual, and systemic. In addition, collaborative approaches and structured group interactions, such as those enabled by collective intelligence platforms, appear promising for engaging adolescents in moral reasoning and helping them respond constructively to hostile content online. The present study builds on these premises to explore whether collective intelligence can help adolescents cope with hate speech and how group size influences the outcomes of these interventions.

Methods

Participants

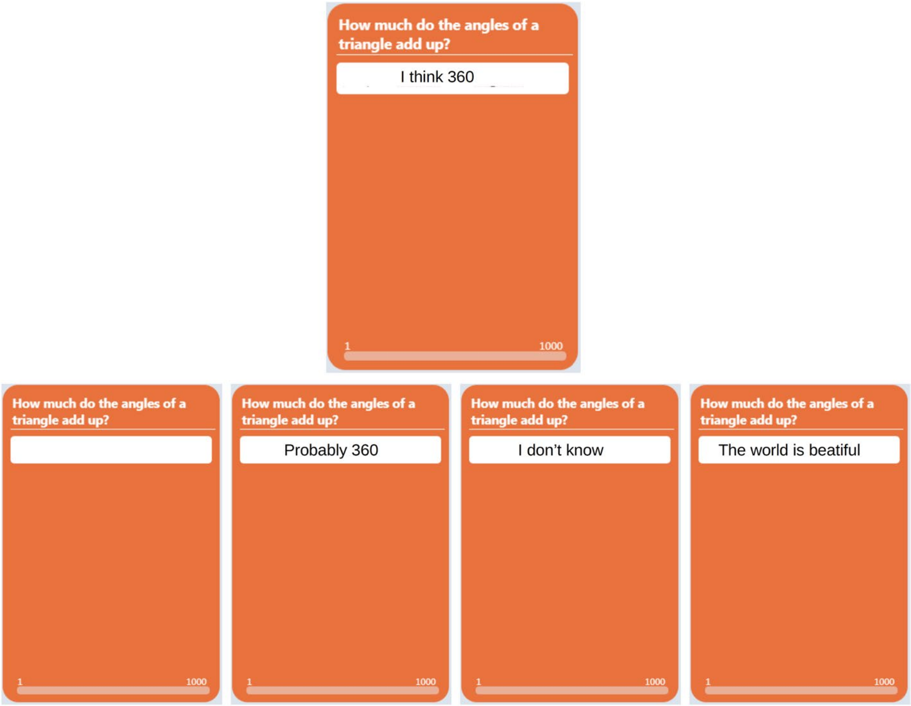
Our study was part of a more extended project involving almost 40 experiments with the Collective Learning tool, which is designed to train digital competencies in Spanish secondary-level students. We were able to compare groups G1, G2, and G3 because all were confronted with the same task regarding hate speech, and all had similar ages, although their platform interaction contexts were different. G1 was a large group of 123 participants enrolled in the 3rd and 4th academic years of Spanish obligatory secondary education (3° ESO and 4° ESO, ages 14–16 years) in four different secondary schools. G2 was a small group of 18 participants enrolled in the same academic year (3° ESO, 14–15 years old) and the same school. G3 was a small group of 20 participants enrolled in the same academic year (4° ESO, 15–16 years old) coming from the same school.

Instrument

On the basis of the concept of collective intelligence propounded by Woolley et al. (2010), through whose results they found that there was a higher group performance in terms of the quality of the group's responses when collective intelligence appeared, researchers at Zaragoza University and members of the Zaragoza University Institute of Research in Biocomputation and the Physics of Complex Systems (BIFI) teamed up with the Kampal Data Solutions firm to elaborate the Collective Learning online tool, featured in this study and many others. Collective Learning is designed to yield high-quality solutions to a wide range of problems by applying collective intelligence following a model of successive digital interactions among participants (Orejudo et al., 2022) while attempting to avoid the usual limitations that tend to arise in such contexts (Toyokawa et al., 2019). Participants work on the platform by solving a practical case study. The case and accompanying questions are based on case methodology (Orejudo et al., 2008), which is completely different from participating in surveys to obtain data, thus basing the intervention on action research (McNiff, 2013). Starting with an individual response phase (phase 1) followed by six phases (phases 2–7) of interaction among users, Collective Learning applies a series of methods over the course of a 35-min session where participants will face the same case and the same questions during all phases. Firstly, the seven phases of participation in the tools are detailed, along with the actions planned for each phase:

– Phase 1: In this first phase (10 min), participants will read the case they are dealing with in this session (in this case, a case related to hate speech) and respond individually to the questions they are working on. During this phase, they cannot see what the other participants are doing.

- Phase 2: In the second phase (4 min), participants observe the responses of four "neighbors," (Fig. 1) who correspond to four other participants within the digital position network above (Fig. 2), below, to the left, and to the right. Each participant can copy their neighbors response, copy and modify it, modify their own response without copying, or keep their response unchanged.
- Phase 3: In the third phase (5 min), participants can perform the same actions as in phase 2 and will continue to see only 4 neighbors. However, the AI that moderates the platform will begin to swap between users so that ideas spread throughout the position network (Fig. 2).
- Phase 4: In the fourth phase (5 min), participants can perform the same actions as in phase 2 and 3, as well as seeing only four neighbors and swapping between them within the position network (Fig. 2). However, on this occasion, in addition to swapping them, the AI will begin to force participants with a low level of involvement in the tool to copy other users' responses to motivate them and reinforce their levels of participation.
- Phase 5: In the fifth phase (5 min), participants continue to perform the same actions as in phases 2, 3, and 4, seeing only four neighbors, as well as swapping between them within the position network (Fig. 2) and forcing copies for participation. However, on this occasion, in addition to swapping them and forcing copying, the AI will begin to eliminate responses that are not being valued by the participants, leaving only those in which consensus is being reached.



How much do the angles of a triangle add up?

I think 360

How much do the angles of a triangle add up?

How much do the angles of a triangle add up?

Probably 360

How much do the angles of a triangle add up?

I don't know

How much do the angles of a triangle add up?

The world is beatiful

Fig. 1 Example screenshot showing how participants view the platform during collaborative work phases. Adapted from "Evolutionary emergence of collective intelligence in large groups of students" (p. 6), by Orejudo et al. (2022), *Frontiers in Psychology*, 13:848048

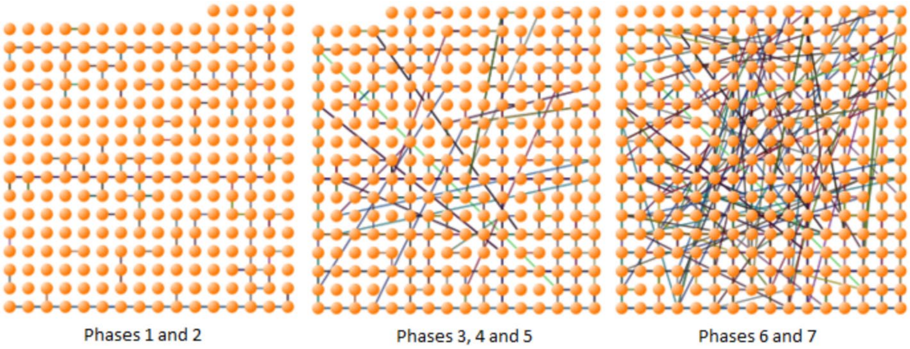


Fig. 2 Evolution of the position network during the seven phases

- Phase 6: In the sixth phase (3 min), the platform's dynamics change completely, eliminating the permutation between users of the position network (Fig. 2) and moving on to displaying a top ten list of answers for each question. Participants can copy answers from this top ten list or keep their own. However, the AI maintains the dynamics of forced copying and deletion of answers to continue converging toward a final consensus.
- Phase 7: In the seventh phase (3 min), participants observe the top ten resulting from the actions taken in the previous phase. In this final phase, all platform processes are cancelled so that it becomes a final consensus phase in which participants can only copy answers from the top ten to form the final top ten.

Before providing a theoretical justification for the platform's design, it is necessary to explain how AI moderates user participation and the platform itself. The AI that moderates the platform has been specifically programmed for Collective Learning and does not operate with any external services. With regard to the permutation of users within the network, this is random; however, the AI is programmed so that throughout phases 3, 4, and 5, all users have permuted at least once to see a greater number of possible responses. Likewise, the AI calculates the popularity or prestige factor of each answer, which the participant can see below their answer. This factor is calculated on the basis of the number of copies of each answer, ranking them internally by popularity throughout the phases until they are displayed in phases 6 and 7. This same system is used to eliminate the answers with the lowest prestige or popularity in phases 5 and 6. Likewise, each participant can see the popularity level of their answer, which is provided by the AI.

Finally, the dynamics of forced response copying are moderated by the AI, taking into account the interaction times of participants within the platform. The AI continuously observes the actions of participants and, if any of them have been inactive for several phases and their responses lack prestige or popularity (i.e., they have not copied or modified their initial responses), it will begin to give them new random responses from among those available on the platform, eliminating their previous responses. The AI has been programmed in this way because an efficient moderator is essential for collective intelligence to emerge among large groups of people (Bigham et al., 2018). The actions and calculations it performs in real time to ensure that information is disseminated efficiently across the network, as well as eliminating poor-quality responses and avoiding noise, are actions that a person could not perform with the same efficiency in such a short time frame.

The design of the platform and the AI that moderates it internally has taken into account the multiple risk factors that can prevent collective intelligence from emerging in large online groups of people (Toyokawa et al., 2019). The first of these problems is information overload; therefore, in phases 2, 3, 4, and 5, each participant is only allowed to see responses proposed by four "neighbors." This allows information to travel across the network without ever overloading participants' capacity to process it (Orejudo et al., 2022). An anonymized random neighbor model also prevents the emergence of a turn-taking monopoly (Mann & Helbing, 2017).

Another key issue is what is known as the response prestige effect or popularity. In our model, artificial intelligence operates in the background as a moderator or a facilitator: in phases 2, 3, 4, and 5, it takes the response prestige effect as a factor into account (Bigham et al., 2018), leading to a top ten list of the most frequent answers. In phases 6 and 7, the top ten list is spread across the network. The popularity effect, or prestige effect, allows a community of users to work toward a consensus by reducing and selecting information. The AI applies the prestige effect as a facilitating factor but taking into account that the system can gradually drift toward responses of lower quality, lower diversity, or backed by less competent leaders (Bernstein et al., 2018). In our model, the popularity/prestige factor gains weight according to how often a certain response appears somewhere on the network: The response can result from one person's musings or it can stem from a joint effort.

Apart from encouraging consensus, the AI popularity system modulates heterogeneity by gradually eliminating responses from phase 6 on. Previous studies using the Collective Learning tool found that the popularity or prestige of an answer is a significantly related factor (Orejudo et al., 2022) which promotes higher quality responses, measured after the session through the reference literature on which the case and questions were based, owing to the emergence of collective intelligence (Bautista et al., 2022, 2024; Woolley et al., 2010). Therefore, the control of response heterogeneity through elimination is essential, as it ensures that the AI can assume the indispensable moderator role within the collective (Bigham et al., 2018).

Case design

For this study, we developed a fictional case featuring "María," a 16-year-old girl entrapped in a hate speech situation on social networks. María follows a male influencer (1), who, during a live stream on Twitter, starts verbally attacking another female content creator (2) because the latter had criticized one of his favorite videogames. After having initially focused exclusively on opinions about video games, 1's attacks soon escalate into comments about 2's sexual orientation. Encouraged by 1's behavior, his community of followers starts trolling and attacking 2 on different platforms. Content creator 2 is eventually obliged to close all her accounts. This leaves María with feelings of guilt, along with doubts regarding how she should have acted.

This case reflects several dynamics described in literature on online hate speech, which are particularly prevalent among adolescents. First, the adolescent girl's situation highlights how young people can become involved in discriminatory behaviors because of the influence of public figures or opinion leaders, as noted by Wachs et al. (2022). Identification with an influencer and the desire to belong to a community can lead minors to engage in behaviors they would not otherwise consider acceptable. María's case also illustrates how hate speech can quickly get out of hand when it emerges in online environments. Influencer 1 started simply by defending a videogame, but his tirade soon morphed into a series of personal attacks. Such tendencies are documented in studies, including Castellanos et al. (2023a), who explored how group dynamics have the potential to amplify hateful statements. Although group interaction among a large number of individuals can be used as a powerful tool to encourage empathy and a prosocial attitude, it can have an adverse effect if not properly managed, given how intense and pervasive peer pressure can be during the period of adolescence (Orejudo et al., 2022). Influencer 2's situation underscores the devastating consequences hate speech can have on people's lives: It can affect their psychological well-being and lead them to hide from public view, as documented in previous studies (Wachs et al., 2022c). This case exemplifies the need for implementing early intervention strategies designed to prevent escalation, thereby preventing victims and even potential perpetrators from involving themselves in pernicious group dynamics. Apart from underscoring the need to confront hate speech online, María's case shows the importance of educating adolescents about the consequences of their actions in online environments while fomenting a culture of respect and empathy among users in online interactions.

Question design

Our questions were designed to assess our adolescent respondents' attitudes, perceptions, and possible actions when confronted with an online hate speech situation. To obtain data that could be analyzed statistically, our questions were quantitative. This would allow us to identify patterns and tendencies in the responses and compare groups of varied sizes. Our questions are listed below, along with a justification for each one:

> Question 1: "Was María doing the right thing when she started by defending her idol (1)?"

This question sought to assess participants' initial perception of the degree of justification of María's actions, focusing on the perceived legitimacy of her choice to defend an influencer she admired. The question explored the notion of loyalty to public figures, a relevant factor in adolescent participation in hate speech (Wachs et al., 2022a). Loyalty toward a leader or influential figure can motivate individuals to participate in bullying or, conversely, to support someone online. Response options were gradual (1–4), allowing us to grasp nuances in our respondents' perception of the degree of legitimacy of María's actions. Responses gradually progressed from a total defense of María's actions ("Yes, she did well to defend him; she didn't know what would happen afterward") to total criticism of those actions ("No; no matter how much she trusts him, one should never support that type of criticism"). The gradual progression within the four potential responses was coherent with the theory of moral development and the justification of behavior in function of a subject's perception of social and personal norms (Kohlberg, 1989).

> Question 2: "Who do you think is harmed by cases like this?"

Our second question aimed to assess our respondents' perception of hate speech victims. According to Castellanos et al. (2023a), hate speech has direct and indirect effects on the individuals involved in it and the community in general. The question explored whether our respondents realized that such behaviors have a more widespread impact on the online social ecosystem. Response options on a scale of 1–4 allowed them to evaluate the impact as they perceived it, ranging from a limited vision (the person who is the object of criticism is the only one harmed) to a more extended view (all involved online users are harmed). It was important to have a gradual range of responses at our disposal to perceive different individual levels of awareness of the harm caused by hate speech. This, in turn, is in line with the theory of empathic concern and recognition of systemic harm (Castellanos et al., 2023a).

> Question 3: “Who do you think benefits from the dissemination of an act of hate speech?”

Question 3 explored our respondents’ opinions regarding the benefits (if any) that might derive from an instance of hate speech. This question was essential to help us understand whether our respondents found that such behaviors can be perceived as beneficial for certain parties involved, which would explain why such acts are perpetrated (Navajas et al., 2018). Ranging on a scale from 1 to 4, our response options varied from finding that hate speech benefits specific people (the influencer and/or his followers) to finding that it benefits no one. As it will be shown in the next section, that gradation of responses was coherent with the literature that discusses how hate speech can be erroneously perceived as a means to obtain recognition or popularity in certain social contexts (Obermaier, 2024).

> Question 4: “What would you have done in María’s place?”

The fourth question is key, as it incites respondents to reflect on their potential actions in a comparable situation. Such reflection is essential in developing the ability to make ethical decisions and grasp how various levels of personal intervention can influence the dynamics of hate speech (Bustamante & Chaux, 2014). In this case, we foresaw a wider range of responses: Here, they were on a scale of 1–6, allowing for a greater degree of differentiation in our respondents’ possible attitudes and behaviors. The answers we proposed ranged from a total defense of influencer 1 (answer 1: “I would have defended 1 to the very end; ultimately, if I like an influencer, it’s not my fault if things go bad”) to a total condemnation of his behavior (answer 6: “I would have reported 1’s social media account for spreading hate speech”). That range of responses allowed us to obtain an overview of a broad range of responses, reflecting our participants’ level of ethical commitment and their readiness to act against hate speech. Such readiness is fundamental in educating adolescents to become digital citizens and prevent online hate speech (Livingstone & Stoilova, 2021).

Theoretical framework for scoring participant responses

The scoring system used in this study was constructed to reflect progressive levels of moral reasoning, perspective-taking, and prosocial engagement, aligned with established theoretical models in developmental psychology and digital ethics education. Rather than treating responses as purely ordinal or subjective, our rating criteria aimed to capture qualitative shifts in how adolescents conceptualize harm, responsibility, and ethical action in online social contexts. Specifically, the response levels were grounded in the following frameworks:

> (Q1) Moral responsibility: On the basis of Kohlberg’s Theory of Moral Development (1989), this model distinguishes between preconventional, conventional, and postconventional levels of moral reasoning. Responses rated at the lower end of the scale (1–2) reflect moral judgments based on authority, obedience, or self-interest, whereas higher

scores (3–4 or 5–6 in Q4) indicate a capacity to reason on the basis of social contracts, universal principles, and autonomous ethical judgment.

(Q2) Empathy and moral engagement: The scoring scheme incorporated dimensions of empathic concern and recognition of systemic harm, inspired by the work of Wachs et al. (2023c), particularly for responses reflecting egocentric or emergent perspective-taking. For higher levels of moral reasoning—such as the awareness of indirect harm, systemic effects, and collective vulnerability—the framework draws on social-ecological models of online risk and responsibility (Markogiannaki et al., 2021), as well as recent contributions on digital citizenship and ethical engagement in adolescent responses to hate speech (Castellanos et al., 2023a). These models support a more complex understanding of how adolescents may come to recognize and oppose online hate as a threat to the broader social fabric.

(Q3) Digital citizenship and ethical agency: The upper levels of the scale reflect dimensions of digital citizenship understood as moral awareness, civic responsibility, and engagement with the social consequences of online behavior (Castellanos et al., 2023a; Obermaier, 2024). In particular, Wachs et al. (2023c) emphasize the role of empathic concern and the recognition of indirect or systemic harm as key components in adolescents' ethical reasoning. Responses rated in this range demonstrate not only a rejection of hate speech as strategically motivated or instrumentally justified but also an understanding of its broader social consequences. Participants who viewed hate speech as detrimental to all, or who expressed a readiness to intervene or report harmful behavior, showed alignment with these ethical frameworks and an emerging sense of communal accountability in digital spaces.

(Q4) Prosocial action readiness: The scoring scale for question 4 reflects a continuum of ethical engagement, ranging from passive complicity to formal institutional action. Lower scores capture group conformity, moral disengagement, and bystander inaction—behaviors typically associated with early stages of moral development and diffusion of responsibility (Kohlberg, 1989; Wachs et al., 2024). Intermediate scores indicate a transition toward emerging moral concern and interpersonal assertiveness, such as anonymously expressing discomfort or choosing to withdraw support, signaling a growing awareness of harm and individual agency (Wachs et al., 2024; Castellanos et al., 2023a). Higher scores represent active prosocial agency, including overt support for the victim or public opposition to hate speech, which aligns with moral courage and civic engagement in digital contexts (Obermaier, 2024). At the highest level, formal reporting of harmful content reflects institutional moral engagement and a commitment to platform-based accountability—an indicator of advanced digital citizenship (Obermaier, 2024).

These frameworks were not used in isolation for each question. Instead, they were synthesized to ensure coherence across the coding of all four items. For instance, while Kohlberg's stages primarily guided the progression of ethical reasoning, models of empathy and digital literacy informed the content-specific application of those stages to the context of online hate speech. To ensure transparency and reproducibility, the full coding scheme, including descriptions and theoretical justification for each score level per question, is provided in Appendix I. The scheme was reviewed by the research team, and responses were coded through a consensus process to reduce subjectivity and ensure alignment with the theoretical model.

Scale synthesis and score alignment

All scoring rubrics were derived from a single construct map that integrates moral reasoning (Kohlberg, 1989), perspective-taking/empathic concern (Wachs et al., 2024), and social-ecological/digital-civic awareness (Markogiannaki et al., 2021; Castellanos et al., 2023). Each item instantiates the same four ascending levels (L1–L4) in its content domain; Q4 adds two levels to differentiate interpersonal (L5) from institutional (L6) action, consistent with accounts of moral courage and responsible digital citizenship (Obermaier, 2024). Scores are therefore level-comparable across items (positions on the same latent continuum), though content differs; analyses report within-item means across phases, and cross-item interpretations are made at the level of complexity (not by equating content). Coding followed a double-review and consensus protocol aligned with the Appendix anchors, including higher-order institutional engagement (Garandeau et al., 2022).

Coding procedures and blinding

Two trained coders independently applied the unified, level-ordered rubrics to all open responses under blind conditions (no personal identifiers, no phase/condition labels, and undisclosed study hypotheses). A total of 1400 responses out of the 4508 submitted by participants were provided to each of the trained coders. For the total number of responses to be coded, the responses were first grouped by phase for the three groups. Then, 50 responses were randomly selected from each of the seven phases for each of the four questions, resulting in a total of 1400 responses to be coded. Prior to reaching consensus, Cohen's kappa was calculated for the 1400 responses categorized by both coders. The value obtained was $k = 0.841$, indicating almost perfect agreement. Disagreements were resolved by consensus (senior adjudication if needed), refining anchor wording without altering the level structure. As a procedural indicator, the preconsensus number of responses flagged as ambiguous or discrepant (i.e., requiring resolution) was 223, subsequently resolved via the predefined consensus protocol. These procedures align with the unified construct map integrating moral reasoning, perspective-taking/empathic concern, and social-ecological/digital-civic orientation (Kohlberg, 1989; Wachs et al., 2024; Markogiannaki et al., 2021; Castellanos et al., 2023; Obermaier, 2024). Full details are provided in Appendix I.

Procedure

We started by applying to the Research Ethics Committee of the Autonomous Community of Aragon (CEICA) to certify that all the ethical criteria for research with human beings and appropriate data treatment were met. After the committee's approval, we contacted the administrative teams of three randomly selected secondary-level schools in the Autonomous Region of Aragon. Although the intervention was not part of the participating schools' curriculum, working to prevent hate speech on social media is part of the education that Spanish students should receive. Once those schools had accepted and coordination had been initiated with them, we sent a brief explanatory guide for the students, along with an informed consent form, indicating the objectives, date, and time the project session would be conducted in their school. Once the informed consent forms had been collected, the research team conducted online training for the teachers who would be supervising the session at all participating schools. Subsequently, prior to the session, the teachers trained by the research team taught the students about the platform and how it works, but not the case or the questions they would be working on, to prevent the students from thinking about the answers before the session took place. After that, we performed the collective intelligence experiment at the date and time we had agreed upon with each school: It was important for all students to be connected to the Collective Learning platform as simultaneously as possible. This procedure was repeated in the same manner at separate times for each of the three groups: G1, G2, and G3.

Data analysis

We started by conducting a descriptive analysis of the mean score on each question in each interaction phase. Although the number of answers per phase and per question was greater than $n = 164$, we only considered the score of the last response emitted in each phase for our calculation of mean scores and their evolution from one phase to the next. Before analyzing the evolution of responses over time, we conducted omnibus tests to assess whether the changes we observed were due to the groups or the phases. To analyze the evolution of answers in function of scores, we compared the mean values in all seven phases with one another, using the mixed model of repeated measures, with the repeated measures factor as the fixed effect and the "participants" factor as random effects. The coding we chose for the fixed factor was the polynomial option, which permitted us to evaluate all possible modification tendencies between phases. For post hoc comparison, we used the Holm method, which foresees a correction of the number of conducted comparisons in function of the number of phases, thereby helping to avoid type 1 errors. The Holm method allows researchers to ascertain what percentage of variance is explained by fixed factors (marginal R^2) and what percentage of variance is explained by fixed and random factors (conditional R^2). For these analyses, we used Jamovi software, version 2.4.

Results

We started by conducting a descriptive analysis of the means of obtained answer scores to each question throughout the seven phases of group interaction in all three groups (Table 1, Fig. 3). In group G1, the large group, the mean answer score continually rose during the session. Score increase was particularly notable from phase 4 onward on questions 1 and 2, and in question 4 in phase 5. The same growth did not take place in the scores for question 3, which rose only slightly from phases 3 to 6 and stagnated in phase 7.

Results were quite different in the small groups G2 and G3. On questions 1 and 4, neither of the two groups managed to improve their mean answer score. Group G2 displayed a negative tendency on questions 2 and 3 between the first and the last phase; group G3 yielded a positive progression on the mean answer scores to question 2 from phase 4 on, but not on question 3, in response to which the mean answer scores remained constant throughout the session.

To determine whether the observed variance was explained by the group, the series of phases, or the total, we conducted fixed-value omnibus tests for each question (Table 2). For question 1, we found significant effects stemming from the independent variables "group" ($p < 0.001$) and "phase" ($p < 0.001$), as well as from the two taken together ($p < 0.001$). Similarly, for question 2, we observed significant effects stemming from the independent variables "group" ($p = 0.011$) and "phase" ($p = 0.003$), as well as from the two

Table 1 Mean scores per group, question, and phase

Question		Phase 1	Phase 2	Phase 3	Phase 4	Phase 5	Phase 6	Phase 7
Q1 (G1), $n = 123$	Mean	2.77	2.84	2.85	2.89	3.07	3.46	3.77
	Deviation	1.122	1.126	1.136	1.103	1.018	0.862	0.598
Q1 (G2), $n = 18$	Mean	2.00	2.00	2.00	2.06	2.06	2.17	2.11
	Deviation	0.767	0.767	0.594	0.416	0.416	0.383	0.323
Q1 (G3), $n = 20$	Mean	3.00	3.00	2.95	2.95	2.95	3.05	3.10
	Deviation	0.858	0.858	0.826	0.826	0.826	0.826	0.852
Q2 (G1), $n = 123$	Mean	2.76	2.77	2.78	2.77	2.98	3.29	3.75
	Deviation	1.174	1.179	1.191	1.179	1.173	1.122	0.795
Q2 (G2), $n = 18$	Mean	3.11	3.11	3.06	2.89	2.83	2.61	2.67
	Deviation	0.963	0.963	1.056	1.132	1.098	1.037	0.970
Q2 (G3), $n = 20$	Mean	3.35	3.35	3.40	3.40	3.80	3.85	4.00
	Deviation	1.089	1.089	1.045	1.046	0.696	0.671	0.000
Q3 (G1), $n = 123$	Mean	1.98	1.95	1.99	2.10	2.15	2.19	2.19
	Deviation	1.267	1.267	1.290	1.315	1.318	1.314	1.313
Q3 (G2), $n = 18$	Mean	3.00	3.06	3.17	3.22	3.28	3.06	2.94
	Deviation	1.138	1.110	1.043	1.003	0.826	0.725	0.725
Q3 (G3), $n = 20$	Mean	1.90	1.90	1.90	1.90	1.90	1.90	1.90
	Deviation	1.294	1.294	1.294	1.294	1.294	1.294	1.294
Q4 (G1), $n = 123$	Mean	3.81	3.83	3.82	3.89	3.87	4.07	4.64
	Deviation	1.601	1.608	1.708	1.698	1.664	1.675	1.558
Q4 (G2), $n = 18$	Mean	2.67	2.67	2.56	2.67	2.78	2.56	2.50
	Deviation	1.455	1.455	1.247	1.414	1.396	1.042	1.043
Q4 (G3), $n = 20$	Mean	3.95	3.80	4.05	4.05	4.05	4.00	4.00
	Deviation	1.191	1.361	1.099	1.099	1.099	1.124	1.124

taken together ($p < 0.001$). The only significant effect on question 3 could be ascribed to the independent variable "group" ($p = 0.002$). Finally, for question 4, the omnibus effect showed that the evolution on that question mainly depended on the independent variable "group" ($p < 0.001$) and on the two independent variables "group" and "phase" taken together ($p = 0.017$). However, "phase," taken alone, did not substantially impact the evolution of mean answer scores over time ($p = 0.547$).

After assessing the omnibus effects according to group and phase for each question (as described in the previous paragraph), we used a mixed model analysis to make parameter estimations of fixed effects per question, group, and phase (Table 3). This allowed us to confirm the significance of the effects observed in the previous omnibus test.

For question 1, we observed a significant effect in group G1 from phase 5 onward (phase 4–5; $p < 0.001$), whereas such an effect was not significant in the smaller groups G2 and G3. This indicates that the omnibus effect we observed on that question was mainly due to the performance obtained by the large group (G1) on the task.

Similarly, for question 2, the large group (G1) showed a significant effect from phase 6 on (phases 5–6; $p < 0.001$). This resembled group G3, which obtained a significant effect in phase 7 (phases 6–7; $p = 0.015$). However, the other small group, G2, also obtained a significant result in phase 6 (phases 5–6; $p = 0.024$), that was due to a *decrease* in user scores on that question,

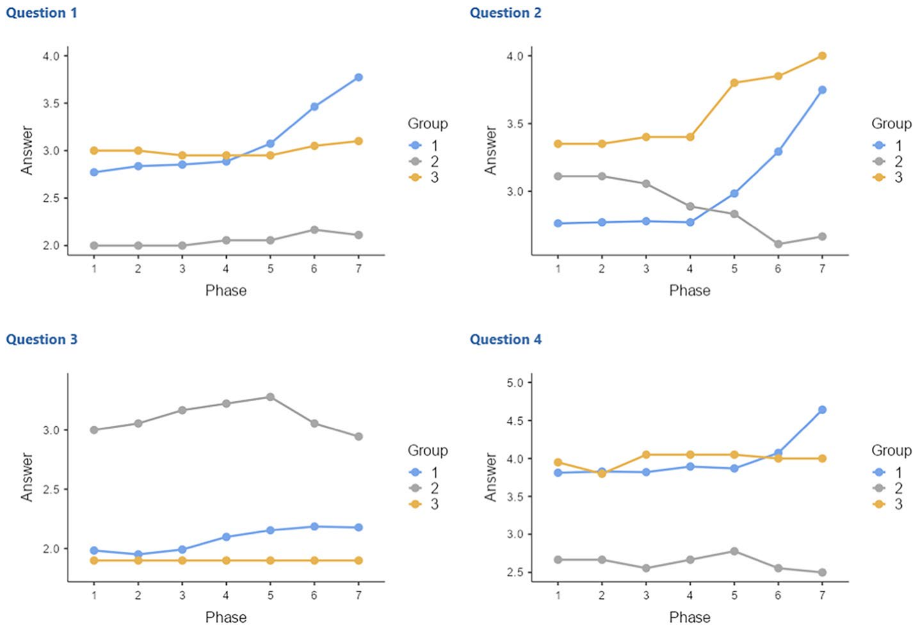


Fig. 3 Evolution of means per group, question, and phase

as opposed to the two other groups. Thus, as for question 1, the omnibus effect in question 2 can be ascribed to the positive evolution in performance achieved by the large group (G1).

For question 3, we observed a significant effect in the large group (G1) from phase 6 on (phases 5–6; $p = 0.031$) but not in the two smaller groups. Mean answer scores remained constant in group G3. Thus, the significant omnibus effect we observed for question 3 was only due to the progress achieved by the large group (G1).

For question 4, the large group (G1) obtained a significant effect in phase 7 (phases 6–7; $p < 0.001$), whereas the two small groups (G2 and G3) did not obtain significant effects. This once more indicates that the omnibus effect we observed on that question was mainly due to group G1's performance on the task.

Therefore, the fixed effects we found in the omnibus tests and our mixed model analysis of each one of the questions show that a large proportion of the progression observed on all four questions in all sessions and groups can be ascribed to the performance achieved by the large group (G1) as opposed to the two smaller groups (G2 and G3). We additionally incorporated random component effects for all four questions and observed that the variance of random effects was different than zero. This indicates that the "user" factor played a significant role in the model. This was confirmed by the likelihood-ratio test (LRT) of random effects, which indicated that incorporating the "user" effect improved the model's fit. As presnted in Table 4, the result was statistically significant in all cases.

Table 2 Fixed effect omnibus test per question

Question 1	**F**	**Num df**	**Den df**	***P*-value**
Group	14.67	2	158	< 0.001
Phase	5.36	6	948	< 0.001
Group * phase	4.13	12	948	< 0.001
Question 2	**F**	**Num df**	**Den df**	***P*-value**
Group	4.61	2	158	0.011
Phase	3.35	6	948	0.003
Group * phase	4.53	12	948	< 0.001
Question 3	**F**	**Num df**	**Den df**	***P*-value**
Group	6.646	2	158	0.002
Phase	0.863	6	948	0.521
Group * phase	0.941	12	948	0.505
Question 4	**F**	**Num df**	**Den df**	***P*-value**
Group	8.520	2	158	< 0.001
Phase	0.830	6	948	0.547
Group * phase	2.056	12	948	0.017

Num numerator, *den* denominator, *df* degrees of freedom

Table 3 Fixed effect parameter estimates per question, group, and phase

Question 1	**Phases 1–2**	**Phases 2–3**	**Phases 3–4**	**Phases 4–5**	**Phases 5–6**	**Phases 6–7**
G1, $n = 123$	0.260	0.283	0.167	< 0.001	<.001	< 0.001
G2, $n = 18$	1.000	1.000	0.772	0.772	0.421	0.579
G3, $n = 20$	1.000	0.330	0.330	0.330	0.666	0.330
Question 2	**Phases 1–2**	**Phases 2–3**	**Phases 3–4**	**Phases 4–5**	**Phases 5–6**	**Phases 6–7**
G1, $n = 123$	0.912	0.863	0.937	0.081	<.001	< 0.001
G2, $n = 18$	1.000	0.805	0.430	0.172	0.024	0.149
G3, $n = 20$	1.000	0.330	0.330	0.070	0.056	0.015
Question 3	**Phases 1–2**	**Phases 2–3**	**Phases 3–4**	**Phases 4–5**	**Phases 5–6**	**Phases 6–7**
G1, $n = 123$	0.495	0.893	0.187	0.060	0.031	0.036
G2, $n = 18$	0.331	0.454	0.260	0.236	0.834	0.848
G3, $n = 20$	1.000	1.000	1.000	1.000	1.000	1.000
Question 4	**Phases 1–2**	**Phases 2–3**	**Phases 3–4**	**Phases 4–5**	**Phases 5–6**	**Phases 6–7**
G1, $n = 123$	0.809	0.939	0.528	0.654	0.084	< 0.001
G2, $n = 18$	0.726	0.726	1.000	0.777	0.717	0.579
G3, $n = 20$	0.330	0.330	0.330	0.428	0.666	0.666

Discussion

These results allowed us to fulfill the two goals we set out at the end of this paper's introductory section. Goal no. 1 consisted in attempting to ascertain whether a large-size group performs better on the task than several smaller groups, thanks to collective intelligence. Comparing the gradual progress of mean answer scores achieved by a large group (G1, 123 participants) with those achieved by two smaller groups (G2, 18 students; and G3, 21

Table 4 Random components per question

Question 1						
Random components				**Variance 95% CI**		
Groups	**Name**	**SD**	**Variance**	**Lower**	**Upper**	**ICC**
User (161)	(Intercept)	0.723	0.523	0.405	0.658	0.576
Residual		0.621	0.385	0.346	0.414	
Random effect LRT						
Test	No. of par	AIC	LRT	df	*P*-value	
User (161)	22.00	3151	577	1.00	< 0.001	
Question 2						
Random components				**Variance 95% CI**		
Groups	**Name**	**SD**	**Variance**	**Lower**	**Upper**	**ICC**
User (161)	(Intercept)	0.779	0.607	0.467	0.768	0.512
Residual		0.760	0.578	0.520	0.621	
Random effect LRT						
Test	No. of par	AIC	LRT	df	*P*-value	
User (161)	22.0	3445	458	1.00	< 0.001	
Question 3						
Random components				**Variance 95% CI**		
Groups	**Name**	**SD**	**Variance**	**Lower**	**Upper**	**ICC**
User (161)	(Intercept)	1.153	1.329	1.049	1.646	0.831
Residual		0.52	0.270	0.243	0.291	
Random effect LRT						
Test	No. of par	AIC	LRT	df	*P*-value	
User (161)	22.0	3777	1402	1.00	< 0.001	
Question 4						
Random components				**Variance 95% CI**		
Groups	**Name**	**SD**	**Variance**	**Lower**	**Upper**	**ICC**
User (161)	(Intercept)	1.277	1.631	1.274	2.036	0.669
Residual		0.897	0.805	0.724	0.865	
Random effect LRT						
Test	No. of par	AIC	LRT	df	*P*-value	
User (161)	22.0	4242	795	1.00	< 0.001	

AIC Akaike Information Criterion, *No. of par* number of parameters, *SD* standard deviation, *ICC* intraclass correlation coefficient

students), we were able to confirm that the large group significantly improved its performance on all questions as the session phases progressed; conversely, G2 did not achieve significant improvement, and G3 only improved on question 2.

According to literature on collective intelligence, heterogeneous interactions among members of a group, encouraged by an online platform such as ours, which facilitates the spread of information within the group (Bernstein et al., 2018; Bigham et al., 2018; Toyokawa et al., 2019), should lead to responses in the final phases of collective endeavor that are of better quality than the responses those same individuals would have provided without group interaction (Bautista et al., 2022, 2024; Orejudo et al., 2022). This was already the case in the first studies on collective intelligence (Woolley et al., 2010; Woolley and Aggarwal, 2020). In our study, however, the premise was fulfilled in the large group

(G1) but not in the small groups (G2 and G3). Collective intelligence could not emerge in the small groups owing to their reduced size.

Thus, to explain our results, we should consider the difference between top-down and bottom-up group processes. Both types of process are essential for the emergence of collective intelligence, as explained by Woolley and Aggarwal (2020). In terms of top-down criteria in our online project (interaction processes inside the group), there were no differences among groups in the aspects of turn-taking (Bernstein et al., 2018), task duration, and the time required for the group to reach a consensus (Dai et al., 2020; De Vincenzo et al., 2017), given that our three groups all operated under the same conditions. Thus, group heterogeneity and diversity of responses are two potential variables capable of explaining the significant differences we observed in terms of group performance.

Group heterogeneity is closely related to the bottom-up level, which includes a wide variety of personal variables, including gender (Curşeu et al., 2015), emotional intelligence (Hjertø & Paulsen, 2016), individual intelligence (Bates y Gupta, 2017), cognitive diversity (Aggarwal et al., 2019), and social awareness (Woolley & Aggarwal, 2020). In our study, the large group (G1) was more heterogeneous owing to its sheer size and because its members stemmed from four different schools, as opposed to the two smaller groups (G2 and G3), which came from one school each: This is the first indicator that helps explain why collective intelligence did emerge in G1 but not in G2 nor G3. Those differences could also partially explain differences in terms of performance and how the task was carried out by the large group as opposed to the two smaller ones (Sulik et al., 2022).

Diversity of responses is a variable more closely related to the top-down level. In the two smaller groups, it was affected for two reasons. The first reason was their smaller size, given that the larger group generated a greater number of different answers. The second reason—perhaps the more relevant one—had to do with the moderation factor (Bigham et al., 2018). Although the Collective Learning platform's AI managed to move information around the group efficiently (Bautista et al., 2024), it did not intervene in the two smaller groups. They carried out the project within the same classroom, where all participants knew one another. Thus, in groups G2 and G3, anonymity was not ensured, and leadership on the part of certain classmates continued to hold sway, therefore leading to a comparatively reduced diversity of responses. Thus, the differences we observed between G1, on the one hand, and G2/G3, on the other hand, were not due to the platform in itself or the type of task but rather to the fact that collective intelligence was able to emerge in the large group (G1) as all conditions were met (Bautista et al., 2024; Woolley & Aggarwal, 2020); this did not occur in the two smaller groups (G2 and G3). Thus, we were able to fulfill our specific goal on the subject of the emergence of collective intelligence.

Goal no. 2 was to ascertain whether the students had undergone a learning experience, i.e., whether they had acquired competencies to deal with online hate speech. Our results show that this was probably the case. As explained in our initial theoretical framework section, online hate speech represents a serious challenge to the emotional and social well-being of children and adolescents, as hate speech is quite pervasive and tends to infiltrate many types of social media. Hate speech eventually leads to anxiety, depression, and a decrease in the individual's sense of social belonging (Wachs et al., 2022c). The results obtained in the present study show that a collective endeavor involving reflexivity and critical thinking in an online environment similar to the one children and adolescents are accustomed to (i.e., a collective intelligence platform) tends to improve the quality of the answers provided by the group while its members work through the successive phases of the task on the platform.

Moreover, collective endeavor on a collective intelligence platform such as this one fulfills the precepts we enumerated in our theoretical framework: The experiment was based on group interaction (Wachs et al., 2022a), and the group reflected on matters that transcended the group's internal concerns, affecting the individual, the group, and their social representation (Castellanos et al., 2023a). At the same time, the platform encouraged online interactions (Schumann & Moore, 2022) within a structured framework that offered possibilities for self-correction (Barrientos Rastrojo, 2022), thereby encouraging critical thinking (Bustamante & Chaux, 2014; Castellanos et al., 2023b). The experience also encouraged the group's members to develop empathy and a prosocial attitude (Obermaier, 2022; Wachs et al., 2023b). Lastly, the process (the experiment's phases) could be considered equally relevant compared with the final result (Barrientos Rastrojo, 2022).

A relevant finding is that this platform and the experiment we designed for it were particularly effective in teaching adolescents to cope with hate speech when the experiment was carried out within a large group (G1: $n = 123$ students). The large group displayed significant improvement on answer scores as the experiment's phases progressed, particularly in the key questions related to awareness of hate speech and learning how to deal with it. This result is in line with previous studies that highlighted the importance of group dynamics and social interaction for the reinforcement of empathy while reducing discriminatory behavior (Castellanos et al., 2023a; Wachs et al., 2022a). The large group thus offered several conditions that the literature in this field considers relevant for the effectiveness of cooperative efforts, as in the case of collective intelligence. One of those conditions is internal heterogeneity coupled with diversity of responses, as the members of the large group came from different schools (Woolley et al., 2015); further conditions include a greater sense of protected anonymity, which, in turn, entails a greater freedom to express oneself (Woolley et al., 2015). Leadership was based on the validity of arguments and not on power structures stemming from interactions outside the platform. This, in turn, encouraged greater exchange in the group's internal interactions (Flecha, 2014) while ensuring equal access to participation, as foreseen in the procedure we designed for the platform (Orejudo et al., 2022).

A similar, constant positive evolution was not apparent in the smaller groups (G2: 18 students and G3: 21 students). In each case, participants belonged to the same school and already knew one another before the experiment began. Thus, a certain lack of anonymity coupled with a more restricted interpersonal dynamic may have influenced results, as participants in small groups might have felt less comfortable when voicing their opinion on sensitive topics (Stahel & Baier, 2023). Along with such hindrances, the heterogeneity and diversity variables of the top-down responses observed in the two small groups may have limited the effectiveness of our intervention, thus suggesting that the group's characteristics and social structure exerted a significant influence on the outcomes of the collective intelligence process (Wachs et al., 2023a).

Limitations

Our study had certain limitations. Its results cannot be generalized beyond a certain point owing to the limited group size and the specific nature of the subject featured in our collective intelligence sessions. To broaden our understanding of the impact of collective intelligence, future studies could explore its effectiveness in different contexts and with other age groups. It would also be relevant to ascertain how such interventions can be adapted to render them more effective in small groups, for instance, by increasing the possibility for anonymity and perhaps by modifying the structure of group interaction (Schumann & Moore, 2022).

Moreover, our results suggest that collective intelligence experiments such as the one described here can foster situated ethical reflection among adolescents within a controlled digital environment. While we observed clear improvements in the quality of responses within the large group (G1) throughout the session phases, these results should be interpreted with caution. Given the absence of a pretest/post-test design external to the experiment, we cannot claim that such improvements reflect long-term or transferable learning. Future research should explore whether this type of intervention has sustained effects on behavior or moral reasoning beyond the platform. Within the limits of our design, however, the results indicate that structured collective deliberation in large, heterogeneous groups can be an effective pedagogical strategy to promote prosocial reasoning in response to hate speech scenarios. Although further evidence is needed to assess its broader impact, this approach holds promise for use in educational settings concerned with digital ethics, tolerance, and civic engagement.

Conclusions

Our results suggest that experiments such as this one, which use collective intelligence to encourage student reflection, can be helpful in preventing and confronting online hate speech. At the same time, results suggest that interventions based on collective intelligence are more effective when applied in large groups. Indeed, large groups provide an environment where group dynamics can be implemented to encourage a culture of tolerance and mutual respect. Interventions such as this one seem particularly effective, not only because they promote awareness of the effects of hate speech but also because they encourage collective action to counteract these behaviors (Orejudo et al., 2022; Author XXXX). Our study's methodology and thematic focus are particularly appropriate for educational environments, where the prevention of hate speech should be given priority to guarantee a positive, secure school atmosphere (Castellanos et al., 2023b).

Although differences are notable according to group size, collective intelligence seems useful as a tool for promoting a culture of respect and empathy among adolescents. If interventions such as the one described herein were implemented in schools, they could play a key role in preventing hate speech and creating more cohesive, tolerant school communities (Wachs et al., 2023b). As the Collective Learning platform encourages group collaboration transcending individual performance, it could be employed as a tool to prevent and confront problematic phenomena such as hate speech, bullying, and cyberbullying (Woolley et al., 2010; Woolley & Aggarwal, 2020). This has been tested in previous studies (Bautista et al., 2022, 2024; Orejudo et al., 2022).

Beyond the composition of the group and individual differences, the results of our study were likely influenced by the material and technological configuration of the experimental environment. The narrative scenario in which María and the influencers participated was designed to reflect the everyday experiences of adolescents on the Internet, encouraging emotional engagement and moral reflection (Castellanos et al., 2023a). The structure of the Collective Learning platform—in particular its sequential phases, restricted visibility, ranking system, and forced copying mechanism—functioned as more than procedural support: It actively mediated cognitive and social dynamics, in line with previous research on digital scaffolding and collective reasoning (Obermaier, 2024). Furthermore, the AI-based moderator played a central role in selecting responses, managing participation, and promoting convergence, effectively shaping the conditions under which group reasoning emerged. These material elements did not merely host the interaction but co-constructed it. Therefore, we recognize that the observed results were not solely the result of participant' contributions but also of the structured possibilities of the learning environment.

In future research featuring the Collective Learning tool for the application of collective intelligence, it will be important to further explore the inner workings that lead to learning experiences such as the ones measured in our study. This could be achieved by adopting a psycho-sociological approach to explore the implications of the interpersonal and intrapersonal variables in the interaction processes that emerge on this platform.

Appendix I. Coding scheme and theoretical justification for response scales

Meta-framework and synthesis protocol

Meta-framework that unifies all scoring scales

We organized all rubrics under a single construct map with four ascending levels of complexity. The construct integrates three dimensions that recur across the literature we draw on: (1) moral reasoning (Kohlberg (1989): preconventional/conventional → postconventional), (2) perspective-taking and empathic concern (Wachs et al., 2024), emergent → other-regarding → systemic, and (3) digital-civic orientation (Markogiannaki et al., 2021; Castellanos et al., 2023a; Obermaier, 2024): from narrow/strategic views to recognition of community-level consequences and responsibilities.

These dimensions covary in typical developmental trajectories; thus, level 1 responses tend to be self-centered or loyalty-centered and low in other-regard, while level 4 responses tend to articulate generalized principles, systemic awareness, and prosocial/communal responsibility. Each question then instantiates this same four-level ladder in a content-specific way (moral evaluation in Q1; scope of harm in Q2; beneficiaries/strategic reasoning in Q3; and action readiness in Q4). Q4 extends the top of the ladder to differentiate interpersonal versus institutional action (levels 5–6) while remaining anchored to the same meta-framework.

How the synthesis was produced (coding protocol)

(1) Theory collation: We extracted level descriptors from Kohlberg (1989) (reasoning), Wachs et al. (2024) (empathy/perspective-taking), Markogiannaki et al. (2021) (social-ecological awareness), and Castellanos et al. (2023) (digital-civic framing), plus prior work on civic courage/agency online (Obermaier (2024).

(2) Construct map: We composed a four-level backbone (L1→L4) that captures jointly: locus of justification (self/loyalty → principle), scope of moral circle (individual → systemic), and civic stance (private indifference → communal responsibility).

(3) Item instantiation: For each question, we translated the backbone into content-specific anchors (e.g., for Q2, "who is harmed" scales from single victim → entire online ecosystem).

(4) Pilot calibration and adjudication: Two coders independently mapped candidate responses to levels using the anchors; disagreements were resolved by consensus, and descriptors were refined for clarity and parsimony.

(5) Final alignment: We verified that level language is homologous across items (e.g., "loyalty-based justification" in Q1 about "narrow, dyadic harm focus" in Q2 about "short-term, strategic benefit focus" in Q3 about "inaction/defense" in Q4 about "loyalty/self-interest justifies support").

Score comparability across questions

Scores represent positions on the same latent continuum of ethical-civic complexity, not interchangeable content. Thus, a 1 in Q1 and a 1 in Q2 are comparable in level (both are level 1 on the construct map: self/loyalty-centered, minimal other-regard, noncivic stance) even though they manifest in different content domains (moral evaluation verus scope of harm). We therefore compare within-item trajectories over phases and interpret cross-item alignment qualitatively (levels), not as literal equality of content. Q4 uses a 1–6 scale to split level-4 action into interpersonal (L5) and institutional (L6) agency; when comparing levels across items, L1–L4 remain directly aligned, and L5–L6 are read as refinements above level 4.

Crosswalk of levels across items

Construct level: Q1: moral evaluation of María's defense; Q2: scope of harm recognized; Q3: perceived beneficiaries of hate speech; Q4: action readiness

L1 (score = 1): loyalty/self-interest justifies support; little/no critique. Only direct target is harmed. Influencer/followers clearly "benefit" (strategic/cynical view). Defend the influencer/active complicity

L2 (score = 2): mixed/qualified defense; emerging doubt. Victim and close circle harmed. Few benefits at others' expense (nascent moral evaluation). Do nothing/bystander nonintervention

L3 (score = 3): defense is problematic; norm-based/order-based critique. Multiple actors in the network harmed. Short-term gains outweighed by long-term harm. Anonymous discomfort/withdraw support/ask to stop

L4 (score = 4): clear rejection on principle. Systemic harm to users/social fabric. Detrimental to all; civic/communal responsibility lens. Public support to victim/speak out (top of the 1–4 ladder)

L5 (score = 5): overt public defense of victim

L6 (score = 6): formal reporting/institutional action

Table 5 Question 1: Was María doing the right thing when she started defending her idol (1)? Purpose: To assess the participant's ability to critically evaluate loyalty-based actions in light of moral responsibility

Score	Description	Theoretical justification
1	Uncritical defense of María based on loyalty or ignorance	Preconventional reasoning; stage 1–2 (Kohlberg, 1989)
2	Partial justification with slight critique	Transitional moral reasoning (Kohlberg, 1989)
3	Recognition that defending someone can be problematic	Conventional morality; social-order awareness (Kohlberg, 1989)
4	Clear rejection of loyalty as justification for supporting harmful speech	Postconventional reasoning (Kohlberg, 1989)

Table 6 Question 2: Who do you think is harmed by cases like this? Purpose: To assess the scope of the participant's moral engagement, empathy and awareness of indirect harm

Score	Description	Theoretical justification
1	Only the direct victim is harmed	Egocentric/limited empathy (Wachs et al., 2023c)
2	Victim and close circle affected	Emergent perspective-taking (Wachs et al., 2023c)
3	Multiple actors in the network are affected	Social-ecological awareness (Markogiannaki et al., 2021)
4	Systemic harm to all users and social fabric	Advanced empathic concern and systemic thinking (Castellanos et al., 2023a)

Coding system

The following coding system was used to evaluate participants' responses to the four questions. Each response was assigned a score from 1 to 4 (or 1–6 in Q4) on the basis of progression in moral reasoning, perspective-taking, and prosocial disposition Table 5, 6, 7, and 8.

Methodological note

The coding was applied by trained members of the research team following this procedure:

Coders and independence: Two trained coders independently applied the unified construct-map rubrics (L1–L4; L5–L6 for Q4) to participants' open responses. A total of 1400 responses out of the 4508 submitted by participants were provided to each of the trained coders. For the total number of responses to be coded, the responses were first grouped by phase for the three groups. Then, 50 responses were randomly selected from each of the seven phases for each of the four questions, resulting in a total of 1400 responses to be coded.

Blinding: Coding was conducted under blind conditions: Coders had no access to personal identifiers or phase/condition labels, and study hypotheses were not disclosed.

Consensus resolution: Prior to reaching consensus, Cohen's kappa was calculated for the 1400 responses categorized by both coders. The value obtained was k = 0.841, indicating almost perfect agreement. All disagreements were resolved by consensus (senior adjudication if required). This process led only to minor clarifications of anchor wording; the level/category structure was not altered.

Procedural ambiguity rate: In the double-coded subsample, the preconsensus number of responses flagged as ambiguous or discrepant (i.e., requiring resolution) was 223; all of these cases were resolved using the predefined consensus protocol.

Rationale: Given that analyses rely on the final consensus code on ordered developmental levels spanning moral reasoning, perspective-taking/empathic concern, and social-ecological/digital-civic orientation (Kohlberg, 1989; Wachs et al., 2024; Markogiannaki et al., 2021; Castellanos et al., 2023; Obermaier, 2024), we provided a transparent procedural report rather than chance-corrected coefficients in this version.

Table 7 Question 3: Who benefits from the dissemination of hate speech? Purpose: To explore perceptions of strategic use of hate speech, digital citizenship and ethical evaluation of instrumentalization

Score	Description	Theoretical justification
1	Influencer and followers' benefit (e.g., visibility, entertainment)	Cynical/strategic reasoning (Obermaier, 2024)
2	A few users benefit at others' expense	Emerging moral evaluation of intentional harm, awareness of instrumentalization (Castellanos et al., 2023a)
3	No one benefits; short-term gain is outweighed by long-term harm	Moral-emotional awareness, recognition of reputational damage and social cost (Wachs et al., 2023c)
4	Hate speech is detrimental to all	Ethical reasoning aligned with digital citizenship and communal responsibility (Obermaier, 2024)

Table 8 Question 4: What would you have done in María's place? Purpose: To evaluate prosocial action readiness and ethical agency

Score	Description	Theoretical justification
1	I would have defended the influencer to the end	Passive complicity; group loyalty (Kohlberg, 1989)
2	I would not have intervened	Bystander effect; disengagement (Wachs et al., 2024)
3	I would have expressed discomfort anonymously	Emerging moral concern; transition from passive to reflective stance (Wachs et al., 2024)
4	I would have asked him to stop or withdrawn support	Moral resistance and interpersonal assertiveness; recognition of harm and autonomy (Castellanos et al., 2023a)
5	I would have supported the victim or spoken out publicly	Active moral agency; prosocial behavior and civic courage (Obermaier, 2024)
6	I would have reported the account as hate speech	Institutional moral engagement (Obermaier, 2024)

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

Aggarwal, I., Woolley, A. W., Chabris, C. F., & Malone, T. W. (2019). The impact of cognitive style diversity on implicit learning in teams. *Frontiers in Psychology, 10*, 112. https://doi.org/10.3389/fpsyg.2019.00112

Argote, L., & Miron-Spektor, E. (2011). Organizational learning: From experience to knowledge. *Organization Science, 22*(5), 1123–1137. https://doi.org/10.1287/orsc.1100.0621

Barrientos Rastrojo, J. (2022). Philosophy for children and teenagers as prevention and treatment of hate speech. *Isegoria*, *67*. https://doi.org/10.3989/isegoria.2022.67.02

Bates, T. C., & Gupta, S. (2017). Smart groups of smart people: Evidence for IQ as the origin of collective intelligence in the performance of human groups. *Intelligence, 60*, 46–56. https://doi.org/10.1016/j.intell.2016.11.004

Bautista, P., Cano, J., Vicente, E., Cebollero, A., & Orejudo, S. (2022). Improving adolescent moral reasoning versus cyberbullying: An online big group experiment by means of collective intelligence. *Computers & Education, 189*, 104594. https://doi.org/10.1016/j.compedu.2022.104594

Bautista, P., Vicente, E., Orejudo, S., & Cano, J. (2024). Training pre-service teachers to deal with cyberbullying: collective intelligence as a mode of learning. *Computers & Education, 220,* 105123. https://doi.org/10.1016/j.compedu.2024.105123

Bernstein, E., Shore, J., & Lazer, D. (2018). How intermittent breaks in interaction improve collective intelligence. *Proceedings of the National Academy of Sciences of the United States, 35*, 8734–8739. https://doi.org/10.1073/pnas.1802407115

Bigham, J. P., Bernstein, M. S., & Adar, E. (2018). Human-computer interaction and collective intelligence. In T. W. Malone & M. S. Bernstein (Eds.), *Handbook of Collective Intelligence* (pp. 57–83). The MIT Press.

Bustamante, A., & Chaux, E. (2014). Reducing moral disengagement mechanisms: A comparison of two interventions. *Journal of Latino/Latin American Studies, 6*(1), 52–54. https://doi.org/10.18085/llas.6.1.123583644qq115t3

Canadian Human Rights Commission. (2019). *Youth and online hate: A study of youth experiences with online hate speech and recommendations for policy and practice*. Ottawa: Canadian Human Rights Commission. Retrieved from https://www.chrcccdp.gc.ca/sites/default/files/online_hate_study_2019_en_1.pdf

Castellanos, M., Wettstein, A., Wachs, S., & Bilz, L. (2023). Direct and indirect effects of social dominance orientation on hate speech perpetration via empathy and moral disengagement among adolescents: A multilevel mediation model. *Aggressive Behavior*. https://doi.org/10.1002/ab.22100

Castellanos, M., Wettstein, A., Wachs, S., Kansok-Dusche, J., Ballaschk, C., Krause, N., & Bilz, L. (2023). Hate speech in adolescents: A binational study on prevalence and demographic differences. *Frontiers in Education*. https://doi.org/10.3389/feduc.2023.1076249

Cohen-Almagor, R. (2011). Fighting hate and bigotry on the internet. *Policy & Internet, 3*(3), 1–26. https://doi.org/10.2202/1944-2866.1059

Crandall, C. S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of Personality and Social Psychology, 82*(3), 359–378. https://doi.org/10.1037/0022-3514.82.3.359

Curşeu, P. L., Pluut, H., Boroş, S., & Meslec, N. (2015). The magic of collective emotional intelligence in learning groups: No guys needed for the spell! *British Journal of Psychology, 106*(2), 217–234. https://doi.org/10.1111/bjop.12075

Dai, A., Zhao, Z., Li, R., Zhang, H., & Zhou, Y. (2020). Evaluation mechanism of collective intelligence for heterogeneous agents' group. *IEEE Access, 8*, 28385–28394. https://doi.org/10.1109/ACCESS.2020.2971278

De Vincenzo, I., Giannoccaro, I., Carbone, G., & Grigolini, P. (2017). Criticality triggers the emergence of collective intelligence in groups. *Physical Review E*. https://doi.org/10.1103/PhysRevE.96.022309

El Zaatari, W., & Maalouf, I. (2022). How the Bronfenbrenner bio-ecological system theory explains the development of students' sense of belonging to school? *Sage Open*. https://doi.org/10.1177/21582440221134089

Flecha, R. (2014). Using mixed methods from a communicative orientation: Researching with grassroots Roma. *Journal of Mixed Methods Research, 2014*(8), 245–254. https://doi.org/10.1177/1558689814527945

Garandeau, C. F., Laninga-Wijnen, L., & Salmivalli, C. (2022). Effects of the KiVa anti-bullying program on affective and cognitive empathy in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 51*(4), 515–529. https://doi.org/10.1080/15374416.2020.1846541

Griffin, R. (2022). New school speech regulation as a regulatory strategy against hate speech on social media: The case of Germany NetzDG. *Telecommunications Policy, 46*(9), Article 102411. https://doi.org/10.1016/j.telpol.2022.102411

Hinduja, S., & Patchin, J. W. (2020). *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*. Sage Publications.

Hjertø, K. B., & Paulsen, J. M. (2016). Beyond collective beliefs: Predicting team academic performance from collective emotional intelligence. *Small Group Research, 47*(5), 510–541. https://doi.org/10.1177/1046496416661236

Jane, E. A. (2014). "Back to the kitchen, cunt": Speaking the unspeakable about online misogyny. *Continuum (Minneapolis, Minn.), 28*(4), 558–570. https://doi.org/10.1080/10304312.2014.924479

Kansok-Dusche, J., Bilz, L., Wettstein, A., Castellanos, M., Schwab, C., Subramaniam, A., & Wachs, S. (2023). Associations between social competence, perceived parents' prosocial educational goals and adolescents' hate speech perpetration in school. *Victims & Offenders, 19*(3), 419–446. https://doi.org/10.1080/15564886.2023.2189191

Kansok-Dusche, J., Ballaschk, C., Krause, N., Zeißig, A., Seemann-Herz, L., Wachs, S., & Bilz, L. (2022). A systematic review on hate speech among children and adolescents: Definitions, prevalence, and overlap with related phenomena. *Trauma, Violence, & Abuse*. https://doi.org/10.1177/15248380221108070

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. In *The ethics of artificial intelligence in education* (pp. 174-202). Routledge.

Kohlberg, L. (1989). Estadios morales y moralización. *El enfoque cognitivo-evolutivo*. Alianza Psicología.

Kudina, O. (2023). *Moral hermeneutics and technology*. Lexington Books.

Liu, J. H., & Koerner, A. F. (2017). The SAGE Handbook of Applied Social Psychology. Sage Publications.

Livingstone, S., & Stoilova, M. (2021). The 4Cs: Classifying online risk to children.

Livingstone, S., Coyer, K., & Carter, C. (2017). *Hate speech on Twitter: A pragmatic approach to countering online hate*. Palgrave Macmillan.

Mann, R. P., & Helbing, D. (2017). Optimal incentives for collective intelligence. *Proceedings of the National Academy of Sciences of the United States of America, 114*(20), 5077–5082. https://doi.org/10.1073/pnas.1618722114

Markogiannaki, M., Biniari, L., Panagouli, E., Thomaidis, L., Sergentanis, T. N., Bacopoulou, F., Babalis, T., Psaltopoulou, T., Tsolia, M., Martens, H., & Tsitsika, A. (2021). Adolescent perspectives about online hate speech: Qualitative analysis in the SELMA project. *Acta Medica Academica, 50*(2), 264–276. https://doi.org/10.5644/ama2006-124.342

McNiff, J. (2013). Action research: Principles and practice. *Routledge*. https://doi.org/10.4324/9780203112755

Meslec, N., Aggarwal, I., & Curşeu, P. L. (2016). The insensitive ruins it all: Compositional and compilational influences of social sensitivity on collective intelligence in groups. *Frontiers in Psychology, 7*, 676. https://doi.org/10.3389/fpsyg.2016.00676

Navajas, J., Niella, T., Garbulsky, G., & Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour, 2*, 126–132. https://doi.org/10.1038/s41562-017-0273-4

Obermaier, M. (2024). Youth on standby? Explaining adolescent and young adult bystanders' intervention against online hate speech. *New Media & Society, 26*(8), 4785–4807. https://doi.org/10.1177/14614448221125417

Obermaier, M., & Schmuck, D. (2022). Youths as targets: Factors of online hate speech victimization among adolescents and young adults. *Journal of Computer-Mediated Communication*. https://doi.org/10.1093/jcmc/zmac012

Orejudo, S., Cano, J., Salinas, A.B., Bautista, P., Clemente, J., Rivero, P., Rivero, A. & Tarancon, A. (2022). Evolutionary generation of collective intelligence in very large groups of students. Publication pending. *Frontiers in Psychology. 13*(848048). https://doi.org/10.3389/fpsyg.2022.848048

Orejudo, S. Fernández, T. y Laparte, M. A. G. (2008). Elaboración y trabajo con casos y otras metodologías activas: cuatro experiencias de un grupo de profesores de la Facultad de Educación de Zaragoza. *Revista Interuniversitaria de Formación del Profesorado: RIFOP*, (63), 21-46. https://dialnet.unirioja.es/servlet/articulo?codigo=2795616

Ramponi, A., Testa, B., Tonelli, S., & Jezek, E. (2022). Addressing religious hate online: From taxonomy creation to automated detection. *PeerJ Journal of Computer Science, 8*, Article e1128. https://doi.org/10.7717/peerj-cs.1128

Toyokawa, W., Whalen, A., & Laland, K. N. (2019). Social learning strategies regulate the wisdom and madness of interactive crowds. *Nature Human Behaviour, 3*(2), 183–193. https://doi.org/10.1038/s41562-018-0518-x

Samy-Tayie, S., Tejedor, S., & Pulido, C. (2023). News literacy and online news between Egyptian and Spanish youth: Fake news, hate speech and trust in the media. *Comunicar, 31*(74), 73–87. https://doi.org/10.3916/C74-2023-06

Schultze-Krumbholz, A., Zagorscak, P., & Scheithauer, H. (2018). A school-based cyberbullying preventive intervention approach: The media heroes program. In M. Campbell & S. Bauman (Eds.), *Reducing cyberbullying in schools* (pp. 145– 158). Elsevier. https://doi.org/10.1016/B978-0-12-811423-0.00011-0

Schumann, S., & Moore, Y. (2022). What can be achieved with online intergroup contact interventions? Assessing long-term attitude, knowledge, and behaviour change. *Analyses of Social Issues and Public Policy, 22*(3), 1072. https://doi.org/10.1111/asap.12333

Stahel, L., & Baier, D. (2023). Digital hate speech experiences across age groups and their impact on well-being: A nationally representative survey in Switzerland. *Cyberpsychology, Behavior, and Social Networking, 26*(7), 519–526. https://doi.org/10.1089/cyber.2022.0185

Sulik, J., Bahrami, B., & Deroy, O. (2022). The diversity gap: When diversity matters for knowledge. *Perspectives on Psychological Science, 17*(3), 752–767. https://doi.org/10.1177/17456916211006070

Wachs, S., Krause, N., Ballaschk, C., Wettstein, A., Bilz, L., Kansok-Dusche, J., & Wright, Mf. (2022). Playing by the rules? An investigation of the relationship between social norms and adolescents' hate speech perpetration in schools. *Journal of Interpersonal Violence, 37*(21–22), NP21143–NP21164. https://doi.org/10.1177/08862605211056032

Wachs, S., Wettstein, A., Bilz, L., & Gámez-Guadix, M. (2022b). Adolescents' motivations to perpetrate hate speech and links with social norms. *Comunicar: Media Education Research Journal*, *30*(71), 9-19. https://phrepo.phbern.ch/6809/

Wachs, S., Wright, M. F., & Gámez-Guadix, M. (2022). Online hate speech victimization and depressive symptoms among adolescents: The protective role of resilience. *Cyberpsychology, Behavior, and Social Networking, 25*(7), 416–423. https://doi.org/10.1089/cyber.2022.0009

Wachs, S., Castellanos, M., Wettstein, A., Bilz, L., & Gámez-Guadix, M. (2023). Associations between classroom climate, empathy, self-efficacy, and countering hate speech among adolescents: A multilevel mediation analysis. *Journal of Interpersonal Violence, 38*(5–6), 5067–5091. https://doi.org/10.1177/08862605221120905

Wachs, S., Krause, N., Wright, M. F., & Gámez-Guadix, M. (2023). Effects of the prevention program "HateLess. Together against Hatred" on adolescents' empathy, self-efficacy, and countering hate speech. *Journal of Youth and Adolescence, 52*(6), 1115–1128. https://doi.org/10.1007/s10964-023-01753-2

Wachs, S., Valido, A., Espelage, D., Castellanos, M., Wettstein, A., & Bilz, L. (2023). The relation of classroom climate to adolescents' countering hate speech via social skills: A positive youth development perspective. *Journal of Adolescence*. https://doi.org/10.1002/jad.12180

Wachs, S., Bilz, L., Wettstein, A., & Espelage, D. L. (2024). Validation of the multidimensional bystander responses to racist hate speech scale and its association with empathy and moral disengagement among adolescents. *Aggressive Behavior, 51*(1), Article e22105. https://doi.org/10.1002/ab.22105

Woolley, A. W., & Aggarwal, I. (2020). Collective intelligence and group learning. *Oxford University Press*. https://doi.org/10.1093/oxfordhb/9780190263362.013.46

Woolley, A. W., Aggarwal, I., & Malone, T. W. (2015). Collective intelligence in teams and organizations. Handbook of collective intelligence, 143–168.

Woolley, A. C., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science, 330*, 686–688. https://doi.org/10.1126/science.1193147