TYPE Hypothesis and Theory
PUBLISHED 09 January 2023
DOI 10.3389/frai.2022.654930

Challenging social media threats using collective well-being-aware recommendation algorithms and an educational virtual companion

Dimitri Ognibene[1,2]*, Rodrigo Wilkens[3], Davide Taibi[4]*, Davinia Hernández-Leo[5], Udo Kruschwitz[6], Gregor Donabauer[6], Emily Theophilou[5], Francesco Lomonaco[1], Sathya Bursic[1], Rene Alejandro Lobo[5], J. Roberto Sánchez-Reina[5], Lidia Scifo[4], Veronica Schwarze[7], Johanna Börsting[7], Ulrich Hoppe[8], Farbod Aprin[8], Nils Malzahn[8] and Sabrina Eimler[7]

[1]Department of Psychology, University of Milano-Bicocca, Milan, Italy, [2]Faculty of Science and Health, School of Computer Science and Electronic Engineering, University of Essex, Colchester, United Kingdom, [3]Cental, Institut Langage et Communication (IL&C), Université catholique de Louvain (UCLouvain), Ottignies-Louvain-la-Neuve, Belgium, [4]Institute for Education Technology, National Research Council of Italy, Palermo, Italy, [5]Department of Information and Communication Technologies, Pompeu Fabra University, Barcelona, Spain, [6]Faculty of Information Science, University of Regensburg, Regensburg, Germany, [7]Institute of Computer Science, Ruhr West University of Applied Science, Bottrop, Germany, [8]Rhein-Ruhr Institut für Angewandte Systeminnovation, Duisburg, Germany

OPEN ACCESS

EDITED BY
Julita Vassileva,
University of Saskatchewan, Canada

REVIEWED BY
Chien-Sing Lee,
Sunway University, Malaysia
Li Liu,
Chongqing University, China
Jorge Luis Bacca Acosta,
Konrad Lorenz University Foundation,
Colombia

*CORRESPONDENCE
Dimitri Ognibene
✉ dimitri.ognibene@unimib.it
Davide Taibi
✉ davide.taibi@itd.cnr.it

SPECIALTY SECTION
This article was submitted to AI for Human Learning and Behavior Change, a section of the journal Frontiers in Artificial Intelligence

RECEIVED 17 January 2021
ACCEPTED 14 December 2022
PUBLISHED 09 January 2023

CITATION
Ognibene D, Wilkens R, Taibi D, Hernández-Leo D, Kruschwitz U, Donabauer G, Theophilou E, Lomonaco F, Bursic S, Lobo RA, Sánchez-Reina JR, Scifo L, Schwarze V, Börsting J, Hoppe U, Aprin F, Malzahn N and Eimler S (2023) Challenging social media threats using collective well-being-aware recommendation algorithms and an educational virtual companion. *Front. Artif. Intell.* 5:654930. doi: 10.3389/frai.2022.654930

Social media have become an integral part of our lives, expanding our interlinking capabilities to new levels. There is plenty to be said about their positive effects. On the other hand, however, some serious negative implications of social media have been repeatedly highlighted in recent years, pointing at various threats to society and its more vulnerable members, such as teenagers, in particular, ranging from much-discussed problems such as digital addiction and polarization to manipulative influences of algorithms and further to more teenager-specific issues (e.g., body stereotyping). The impact of social media—both at an individual and societal level—is characterized by the complex interplay between the users' interactions and the intelligent components of the platform. Thus, users' understanding of social media mechanisms plays a determinant role. We thus propose a theoretical framework based on an adaptive "*Social Media Virtual Companion*" for educating and supporting an entire community, teenage students, to interact in social media environments in order to achieve desirable conditions, defined in terms of a community-specific and participatory designed measure of Collective Well-Being (CWB). This Companion combines automatic processing with expert intervention and guidance. The virtual Companion will be powered by a *Recommender System* (*CWB-RS*) that will optimize a *CWB* metric instead of engagement or platform profit, which currently largely drives recommender systems thereby disregarding any societal collateral effect.

CWB-RS will optimize CWB both in the short term by balancing the level of social media threats the users are exposed to, and in the long term by adopting an *Intelligent Tutor System* role and enabling adaptive and personalized sequencing of playful learning activities. We put an emphasis on *experts* and *educators* in the *educationally managed social media community* of the Companion. They play five key roles: (a) use the Companion in classroom-based educational activities; (b) guide the definition of the CWB; (c) provide a hierarchical structure of learning strategies, objectives and activities that will support and contain the adaptive sequencing algorithms of the CWB-RS based on hierarchical reinforcement learning; (d) act as moderators of direct conflicts between the members of the community; and, finally, (e) monitor and address ethical and educational issues that are beyond the intelligent agent's competence and control. This framework offers a possible approach to understanding how to design social media systems and embedded educational interventions that favor a more healthy and positive society. Preliminary results on the performance of the Companion's components and studies of the educational and psychological underlying principles are presented.

KEYWORDS

collective well-being, recommender systems, social media, virtual companion, social media threats, hierarchical reinforcement learning

1. Introduction

Social media (SM) have become an integral part of our everyday lives. Looking at the field more broadly, the freedom to post whatever someone judges useful has been described as nothing less than a shift in the communication paradigm (Baeza-Yates and Ribeiro-Neto, 1999), or in other words, *the freedom to publish* marks the birth of a new era altogether (Baeza-Yates and Ribeiro-Neto, 2010). There is ample evidence of positive effects of SM that goes beyond just-in-time connectivity with a network of friends and like-minded people, including, but not limited to, improved relationship maintenance (Ellison et al., 2014), increased intimacy (Jiang et al., 2011), reduced loneliness (Khosravi et al., 2016; Ryan et al., 2017), and reduced depression (Grieve et al., 2013). It has become a highly accessible and increasingly popular means of sharing content and immediately re-sharing others' content. Supported by personalizing recommendation algorithms, which suggest content and contacts, SM allows information of any quality to spread at an exponentially faster rate than the traditional "word of mouth" (Murthy, 2012; Webb et al., 2016). However, far from creating a global space for mutual understanding, truthful, and objective information, the large-scale growth of SM has also fostered negative social phenomena, e.g., (cyber)bullying to pick just one (Cowie, 2013; Mladenović et al., 2021), that only existed on a limited scale and slow pace before the digital revolution. These issues are escalated by impulsive, alienating and excessive usage that can be associated with digital addiction (Almourad et al., 2020). These phenomena, enabled by the rapid spread of information on SM can affect the well-being of more vulnerable members of our society, such as teenagers, in particular (Talwar et al., 2014; Ozimek et al., 2017; Gao et al., 2020). Ever since the Cambridge Analytica scandal (Isaak and Hanna, 2018), we have become more sensitive to the negative implications of social media. One might go as far as to suggest that SM may have become so dangerous that we would be in a better place without them, but that is clearly an unrealistic idea.

It can be argued that the impact of online experience, especially in SM, intrinsically depends on the mutual attitudes and interactions between the members of the community (Jones and Mitchell, 2016) and their interplay with the intelligent components of the platforms. This calls for a holistic approach that on one side provides *educational interventions* supporting users in understanding the impact of their actions on the experience of the other members of the community (Jones and Mitchell, 2016; Xu et al., 2019; Taibi et al., 2022) and their role in the *Collective well-being of their social media community (CWB)* (Ahn and Shin, 2013; Roy et al., 2018; Allcott et al., 2020). CWB operationally combines the different aspects of what a community considers its "desirable condition," while also crucially considering individual differences and conflicting interests. Moreover, the lack of users' *"new media literacy"* (Scolari et al., 2018) (i.e., understanding of social media mechanisms) has a strong role in escalating SM threats. For example, a study with middle-school students found that more than 80% of them believed that the "sponsored content" articles shown to them were true stories (Wineburg et al., 2016). On the other side a multifaceted approach needs to provide technological support to reduce the strain cognitive resources of social media users. An important question is how this can be realized considering the complexity of the involved phenomena, the diverse attitudes and interests of the users, the cost of an intervention with a coverage and impact comparable to that of social media.

With this motivation, in this paper, we articulate a framework for educating teenagers in their interaction with SM and synergetically improve and support their experience based on a "*Social Media Virtual Companion*." Inside an external SM platform, it will create an *educationally managed social media community* where playful learning activities and healthy content will be integrated into participants' SM experience. Educational goals and interventions will be designed by experts and educators, e.g., to raise awareness about potential threats and to show alternative healthy interactions. To select the most suitable content and effective interventions based on experts' and educators' designs, the companion will incorporate functions of an Intelligent Tutor System (ITS).

Due to the cognitively burdening and overloading information flow of the current SM platform (see Section 2.4 and Weng et al., 2012; Kramer et al., 2014; Lee et al., 2019; Almourad et al., 2020), the Companion will also have to balance and ignore engagement-driven external platform recommendations to target for a fairer and healthier objective (Rastegarpanah et al., 2019). The community of users of the SM platform is both the producer and consumer of SM content. We affirm that the objective pursued by SM algorithms should be closer to the community's needs than those of the SM platform. As the CWB reflects the global impact of MS on the condition of the individual and the community, we propose that a suitable objective is a measure that formalizes community-specific and participatory designed CWB expanded with student-specific educational objectives. This shifts to a *CWB* metric evaluated directly on the Companion and used as an optimization target by its integrated recommendation engine (*CWB-RS*), which will allow the support and educational management of the local social media community (see Figure 1).

This framework can be seen as a top-down vision that combines education and technology complements, integrates and helps to balance the diverse efforts targeting specific SM issues. We think that social media phenomena' complexity, their intrinsically interlinked nature, and their impact on our society demand the production and discussion of such overarching views in the scientific community. Furthermore, our framework, by proposing educational social media that can be separated or linked with the main social media, would improve the problems that arise from platform-enforced restrictions that hinder experimentation and analysis, especially in extended longitudinal studies. The data collected could be the first step to enabling the definition of adequate regulations and revising SM platform designs to improve their impact on society.

In the next section, a concise overview of the SM threats is presented. In Section 3, we present the educational Companion approach for the increase of digital literacy and the enhancement of CWB. In Section 4, we discuss the CWB metrics. The CWB-RS is presented in Section 5 while, in Section 6, we present a use case exemplifying the interaction of the SM users with the Companion and CWB-RS. In Section 7.1 we present the current advances of this line of research.

2. Social media threats

With the advent of social media, the speed and number of interactions escalated beyond users' ability to monitor and understand their impact. This resulted in challenging threats with a broad range and variability over time, compounded by crucial ethical and practical issues, like preserving freedom of speech and allowing users to be collectively satisfied while dealing with the conflicts generated by their different opinions and contrasting interests. These are magnified by the complex dynamics of information on social media due to the interaction between myriads of users and intelligent artificial systems.

Critical cases are the pervasive diffusion of fake news and biased content and the growing trend of hate practices. Indeed, hate propagators were among the early adopters of the Internet (Schafer, 2002; Gerstenfeld et al., 2003; Chan et al., 2016). Even though SM platforms presenting policies against hate speech and discrimination, these new media have been shown to be powerful tools to reach new audiences and spread racist propaganda and incite violence offline. This gave rise to concern of several human rights associations about the platforms' usage to spread all forms of discrimination[1] (Chris Hale, 2012; Bliuc et al., 2018).

The social media threats can be broadly classified into three categories: 1. content; 2. algorithmic; network, and attacks; and 3. dynamics. However, sharply separating these types of threats is not trivial as they strongly interact and mutually reinforce while often leveraging on several cognitive aspects and limits of the users. In the rest of this section, we briefly discuss SM threats, and we present in Table 2 a list of examples per category. We focus on threats that specifically affect the vulnerable population of teenagers and the related threats, such as bullying (Talwar et al., 2014; Mladenović et al., 2021; Fulantelli et al., 2022), addiction (Tariq et al., 2012; Shensa et al., 2017), body stereotypes, and others (Clarke, 2009; Mcandrew and Jeong, 2012; Ozimek et al., 2017).

[1] Simon Wiesenthal Center: http://www.digitalhate.net, Online Hate and Harassment Report: The American Experience 2020: https://www.adl.org/online-hate-2020.

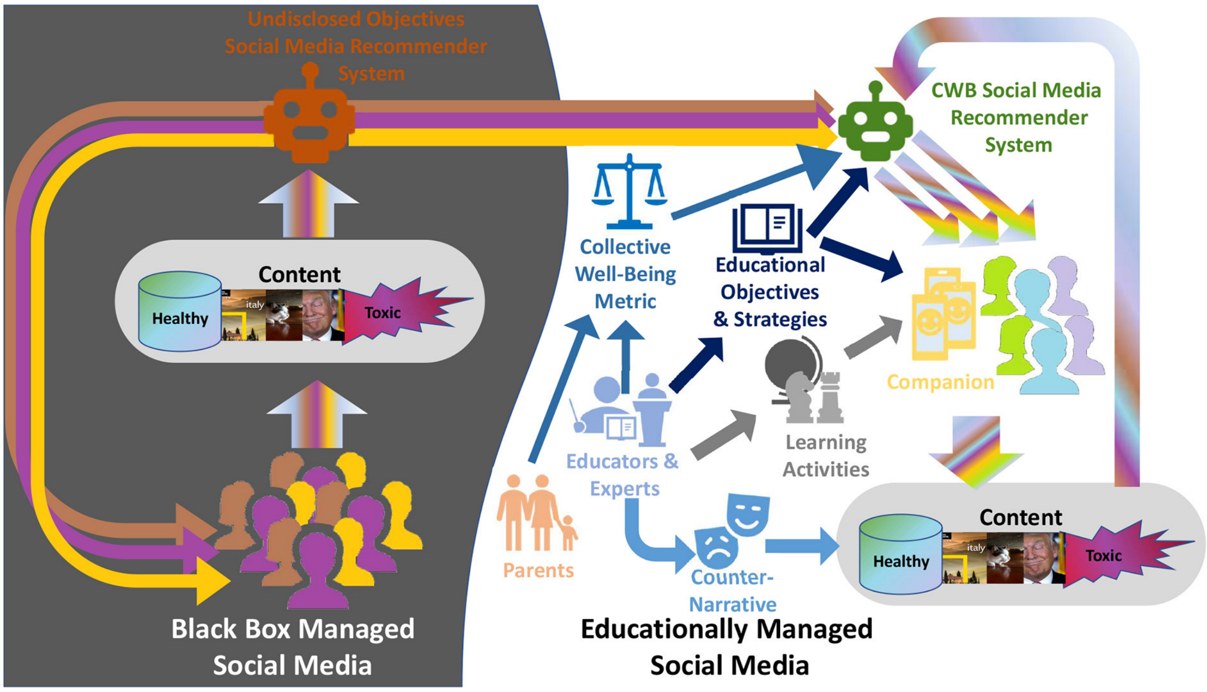
Undisclosed Objectives Social Media Recommender System
Content
Healthy
Toxic
Black Box Managed Social Media
CWB Social Media Recommender System
Collective Well-Being Metric
Educational Objectives & Strategies
Companion
Educators & Experts
Learning Activities
Parents
Counter-Narrative
Content
Healthy
Toxic
Educationally Managed Social Media

FIGURE 1
The virtual *Social Media Companion* enables continue educational and interaction support for a community of students with the involvement of educators. This generates an *Educationally Managed Social Media Community* whose Collective Well-Being is actively improved by the CWB-RS powering the Companion under the guidance of the educational objectives and strategies provided by the educators.

2.1. Content-based social media threats

The content-based threats are common to classical media, but specific issues thrive on the web and social media in particular. Examples of content-based threats include toxic content (Kozyreva et al., 2020), fake news/disinformation (de Cock Buning, 2018), beauty stereotypes (Verrastro et al., 2020), and bullying (Grigg, 2010). Given the importance of these threats, various research is focused on the development of dedicated detection systems as discussed in Section 5.4.

2.2. Algorithmic social media threats

The SM algorithms may create additional threats. For example, the selective exposure of digital media users to news sources (Schmidt et al., 2017), risks creating a permanent distorting state of isolation from different ideas and perspectives, i.e., “filter bubbles” (Nikolov et al., 2015; Geschke et al., 2019), and form closed-group polarized social structures, i.e., “echo chambers” (Del Vicario et al., 2016; Gillani et al., 2018). Another undesired network condition is gerrymandering (Stewart et al., 2019), where users are exposed to unbalanced neighborhood configurations.

2.3. Social media dynamics induced threats

The social media dynamics induced by the extended and fast-paced interaction between their algorithms, common social tendencies, and stakeholders’ interests may also be a source of threats (Anderson and McLaren, 2012; Milano et al., 2021). These factors may escalate the acceptance of toxic beliefs (Neubaum and Krämer, 2017; Stewart et al., 2019), make social media users’ opinions susceptible to phenomena such as the diffusion of hateful content, and induce violent outbreaks of fake news on a large scale (Del Vicario et al., 2016; Webb et al., 2016).

2.4. Social media cognitive and socio-emotional threats

While many studies that analyze the mechanisms of content propagation in social media exist, how to model the effects of users’ emotional and cognitive states or traits on the propagating malicious content is unclear, especially in light of the significant contribution of their cognitive limits (Weng et al., 2012; Allcott and Gentzkow, 2017; Pennycook and Rand, 2018). Important cognitive factors are users’ limited attention and error-prone information processing (Weng et al., 2012) that may be worsened by the emotional features of the messages (Kramer et al., 2014; Brady et al., 2017). Moreover, the lack of non-verbal communication and limited social presence (Gunawardena, 1995; Rourke et al., 1999; Mehari et al., 2014) often exasperates carelessness and misbehaviors, as the users perceive themselves as anonymous (Diener et al., 1980; Postmes and Spears, 1998), do not feel judged or exposed (Whittaker and Kowalski, 2015) and deindividualize themselves and other users (Lowry et al., 2016).

Over time, users' behaviors can deteriorate and show highly impulsive and addictive traits (Kuss and Griffiths, 2011). Indeed, social media usage presents many neurocognitive characteristics (e.g., the presence of impulsivity) typical of more established forms of pharmacological and behavioral addictions (Lee et al., 2019). This recently recognized threat, named Digital Addiction (DA) (Lavenia, 2012; Nakayama and Higuchi, 2015; Almourad et al., 2020), has several harmful consequences, such as unconscious and hasty actions (Ali et al., 2015; Alrobai et al., 2016). Some of them are especially relevant for teenagers affecting their school performance and mood (Aboujaoude et al., 2006). In the last few years, it emerged that recognizing addiction to social media cannot be based only on the "connection time" criterion but also on how people behave (Taymur et al., 2016; Musetti and Corsano, 2018). Like in the other behavioral addictions, a crucial role may be played by the environment structure (Kurth-Nelson and Redish, 2009; Ognibene et al., 2019), more than by biochemical failures of the decision system (Lim et al., 2019). Indeed, many, if not all, aspects of social media environments are under the control of the recommender systems, which may help reduce the condition with specific strategies, such as higher delays for more impulsive users as well as detecting and curbing its triggers, e.g., feelings of Fear of Missing Out (Alutaybi et al., 2019).

2.5. Limited social media literacy

Finally, the lack of digital literacy, common among teenagers (Meyers et al., 2013), can strongly contribute to other threats escalation, for example by favoring the spread of content-based threats and engaging in toxic dynamics (Wineburg et al., 2016). Teenagers also show over-reliance on algorithmic recommendations and a lack of awareness of the unwitting use of toxic content. Thus, reducing their ability to make choices and increasingly deviating toward dangerous behaviors (Walker, 2016; Banker and Khetani, 2019).

This diverse set of phenomena and threats, the latter in particular, motivates our educational approach combining educational methods to rise digital citizenship and new median literacy while supporting the user with a smart companion that can also counter the cognitive burden of interacting with social media.

3. Educational social media companion

Social media have been shown to contribute to our collective well-being enhancing our levels of social connectivity. However, our well-being, and in particular teenagers' one, is vulnerable to social media threats, such as exposure to many types of unwanted or toxic content (Costello et al., 2019; Mladenović et al., 2021). Increasing social media users' digital literacy (Fedorov, 2015) and citizenship (Jones and Mitchell, 2016; Xu et al., 2019) may counter most SM threats that thrive due to users' lack of awareness and over-reliance on algorithmic recommendations (Meyers et al., 2013; Walker, 2016; Banker and Khetani, 2019).

The traditional media literacy approaches were based on the idea that media had adverse effects on children. Therefore, it was necessary to "immunize" young people so they can resist such negative influence. As the media ecosystem evolved, so did media literacy. It soon included a paradigm shift toward education and risk prevention concerning the web, video games, social networks and mobile devices. Recently, new concepts have been developed to name these new forms of literacy, from "digital literacy" or "digital citizenship" to "new media literacy" (Scolari et al., 2018; Xu et al., 2019). With the objective of contrasting social media threats, several countries have introduced educational initiatives to increase the awareness of students with respect to the detection of fake news and misleading information on the web.[2] Still, due to their limited duration and their high costs compared to purely entertaining use of social media, the effects of these programs may be limited.

We propose a framework based on a virtual *Educational Social Media Companion* that enables continued, both in the classroom and outside, educational and interaction support for a community of learners, creating an *Educationally Managed Social Media Community* aimed at improving users' new media literacy and social media experience. Through companion support, the students can safely learn by doing how to deal with social media content, leveraging the positive aspects and counteracting the inherent threats. The relation between those elements is shown in Figure 1.

While previous educational attempts have focused on literacy activities mainly about *external* threats, improving the impact of social media on our society is challenging essentially because the interactions between users determine the quality and consequences of their experience. Rising awareness about the effects of own actions on the community members' experience and the importance of performing healthy interactions to realize

[2] Retrieved from here: https://www.bbc.co.uk/programmes/articles/4fRwvHcfr5hYMMltFqvP6qF/help-your-students-spot-false-news BBC, (UK), https://literacytrust.org.uk/programmes/news-wise/ NewsWise (UK).

a desirable condition notwithstanding the anonymity (Peddinti et al., 2014; Schlesinger et al., 2017) and deindividuation that social media may foster (Diener et al., 1980; Postmes and Spears, 1998; Lowry et al., 2016) is central in the presented educational endeavor.

We propose that the educationally managed communities participate in the description of a shared vision of a "desirable social media community" in terms of an operational Collective Well-Being (CWB) definition specific for their community (see Section 4). This will support the coherent formulation of community regulations, objectives and educational activities that involve several ethical issues entailing the definition of boundaries and trade-offs to own personal behavior online (see Section 4.1.1), such as enabling collective satisfaction and preserving the right to free speech (Webb et al., 2016) while facing the conflicts generated by users' different attitudes, opinions, personal history, and conflicting interests. A formalization of the CWB informs the CWB-RS, the companion recommender system aimed at recommending educational activities and content while balancing the recommendation incoming from the external social media platforms to improve the community's collective well-being, see Section 5.

3.1. An educationally managed social media community

The Companion safeguards teens' interactions on social media and implements *playful adaptive educational strategies* to engage and scaffold them considering personalized *educational needs and objectives*. These strategies comprise *scripted learning designs* (Amarasinghe et al., 2019) that informing by the CWB-RS will articulate the behavior of the Companion presenting teens with the right level of educational scaffolding (Beed et al., 1991) through an adaptive, personalized and contextualized sequence of *learning activities* and supported social media interaction—incorporating behavioral and cognitive interventions (*nudges* and *boosts*) that are grounded in behavioral psychology (Thaler and Sunstein, 2009; Hertwig and Grüne-Yanoff, 2017; Purohit et al., 2020). Game mechanics based on a *counter-narrative* (Davies et al., 2016) approach will support learning activities related to rising awareness: motivation, perspective taking, external thinking, empathy, and responsibility. These narrative scripts pursue collective and individual *engagement* with the Companion, offering motivating challenges and rewards aimed at keeping users' interest even in the presence of non-educational social media platforms (Van Staalduinen and de Freitas, 2011) while maintaining awareness of the digital addiction threat. The autonomous capabilities provided by the CWB-RS to the Companion can be particularly helpful outside of the classroom to avoid the cognitive overload, addiction or over-exposure to toxic content that the recommender system of an external, non-educational, social media platform may select. Moreover, they allow achieving a level of availability comparable with that of non-educational social media while reducing the moderating effort requested from the moderating educators.

3.1.1. Educators and the companion: A human in the loop view

In our framework, the educators not only use the companion for delivering tailored educational activities in the classroom but, together with the experts, participate in the moderation and support of the community as well as in the definition of its CWB and related educational strategies, which drive the Companion by informing the CWB-RS. The educators oversee the CWB-RS behavior playing a key "human in the loop" role (Nunes et al., 2015; Zanzotto, 2019). This alleviates the complexities faced by the CWB-RS, such as noise in the estimation of content toxicity (see Section 5.4), which may also lead to misinterpreting users' needs and possibly exacerbating their condition. While the CWB-RS will have implicit moderating behaviors, e.g., reducing the presentation priority of users' confrontational interactions, the educators will have a central role in arbitrating users' disputes as well as solving the conflicts that may emerge between different components of an "under-construction" CWB measure, such as between emotional health (Roy et al., 2018) of one user and freedom of speech of another.

3.1.2. Adopting behavioral economics to support collective well-being

This educational effort aims to help users of social media make the right decision and teach them the necessary skills to get to that point. Strategies developed in the context of behavioral and cognitive sciences offer a well-founded framework to address this issue. In particular, we consider nudging (Thaler and Sunstein, 2009) and boosting (Hertwig and Grüne-Yanoff, 2017) to be two paradigms that have both been developed to minimize risk and harm—and doing this in a way that makes use of behavioral patterns and is as unintrusive as possible.

Nudging (Thaler and Sunstein, 2009) is a behavioral-public-policy approach aiming to push people toward more beneficial decisions through the "choice architecture" of people's environment (e.g., default settings). In the Companion context, such beneficial decisions could be to explore a broad range of different opinions about a specific topic and check understandable but scientifically correct pieces of information. In this working example, nudges could be implemented through a visual layout of the feed that allows easy exploration of such information (see Figure 2). Other forms of nudging are warning lights and information nutrition labels as they offer the potential to reduce harm and risks in web searches, e.g., Zimmerman et al. (2020).

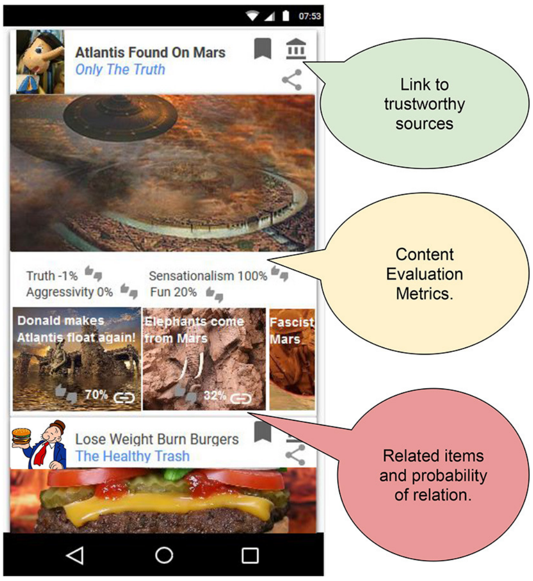

FIGURE 2
Sketch of Companion User Interface. The companion will support the students' interaction with social media by contextualizing the content to increase the students' awareness and allow them to access a more diverse set of perspectives (Bozdag and van den Hoven, 2015) and sources. It also explicitly and visually provides the students with an evaluation of the content's harmfulness (Fuhr et al., 2018). The example shows how a piece of imaginary fake news would be contextualized.

The limitation of nudges is that they do not typically teach any competencies, i.e., when a nudge is removed, the user will behave as before (and not have learned anything). This is where boosts come in as an alternative approach. Boosts focus on interventions as an approach to improve people's competence in making their own choices (Hertwig and Grüne-Yanoff, 2017). In the Companion context, specific educational activities have been designed aimed at teaching people skills that help them make healthy decisions, e.g., select/read/trust articles from authoritative resources rather than those reflecting (possibly extreme) individual opinions (see Section 7.1).

The critical difference between a boosting and nudging approach is that boosting assumes that people are not merely "irrational" and therefore need to be nudged toward better decisions. However, such new competencies can be acquired without too much time and effort and may be hindered by the presence of stress and other sources of reduced cognitive resources. Both approaches nicely fit into the overall approach proposed here. Nudges offer a way to push content to users, making them notice. Boosting is a particularly promising paradigm to strengthen online users' competencies and counteract the challenges of the digital world. It also appears to be a good scenario for addressing misinformation and false information, among others. Both paradigms help us educate online users rather than imposing rules, restrictions, or suggestions on them. They have massive potential as general pathways to minimize and address harm in the modern online world (Kozyreva et al., 2020; Lorenz-Spreen et al., 2020).

3.1.3. Educational activities

The Companion must also provide a satisfying and engaging experience by using *novel hand-defined educational games and activities* based on the interactive counter-narrative concept and educational games. SM's entertainment aspect is preserved during the navigation modulated in taking into account CWB, suggesting activities, content, and contacts for the user but managing the exposure to potential threats and addiction.

The Narrative Scripts help raise users' awareness about SM threats and train the students against them. They are sequences of adaptive learning tasks that provide the right level of educational scaffolding to individuals in developing critical thinking skills, including awareness—perspective taking, motivation, external thinking, empathy, and responsibility) by interacting with narratives, counter-narratives, and peers. These tasks can be different activities, including free-roaming inside the platform, guided roaming following a narrative, quizzes, playing minigames, or participating in group tasks. Different counter-narratives can be triggered depending on students' detected behavior (Lobo et al., 2022).

Counter-narrative are used to challenge biased content and discrimination, highlight toxic aspects of messages and attitudes, challenge their assumptions, uncover limits and fallacies, and dismantle associated conspiracy and pseudo-science theories.

Through a game-oriented setup, the companion bridges the "us" vs. "them" gap that is fostered by hate speech and other expressions of bias (e.g., gendered) and brings forward the positive aspects of an open society and focuses more on "what we are for" and less on "what we are against." The users will be informed and requested to actively and socially contribute to creating and sharing content and material that fosters and supports the idea of an open, unbiased and tolerant society. Thus, the games can also offer the chance to build connections between the users, which, when isolated, are more vulnerable to online toxic content. One approach is to propose periodically specific tests and activities related to each threat, such as Szymanski et al. (2011).

A use case scenario is presented in Section 6 and the outcomes of several pilot studies that lie the basis for the educational activities are presented in Section 7.1.

3.1.4. External and internal SM communities separation allows for educational opportunities

The Companion's location allows it to act as an interface between the educationally managed social media community and the external one. It permits mitigating the effect of external toxic content and offers the opportunity to recreate different interesting experiments about SM phenomena, such as the ones presented in Bail et al. (2018) and Stewart et al. (2019). A controlled environment in which social network dynamics are emulated can be adopted to stimulate students to understand SM mechanisms better, e.g., see Lomonaco et al. (2022a). Nowadays, the interactions intervening in social media are often mediated by automatic algorithms. Most teenagers ignore these dynamics that heavily influence their content and behavior when virtually interacting (Kuss et al., 2013). For example, in a classroom, it may expose sub-groups to recommendations with different biases or allow the students to change the recommender parameters (Bhargava et al., 2019; Lomonaco et al., 2022a).

3.1.5. Companion exposes social media threats

The Companion's autonomous mechanisms will support the students in interacting with the social media content both inside (as a support learning activities) and outside (students' daily social network use) of the classroom. The Companion interface exposes its filtering and recommendation algorithms by allowing direct control of their parameters (Bhargava et al., 2019). It will contextualize the content to increase the students' awareness and allow them to access a more diverse set of perspectives (Bozdag and van den Hoven, 2015) and sources (see Figure 2). It also explicitly and visually will provide the students with an evaluation of the content's harmfulness (Fuhr et al., 2018) (see Section 5.4).

4. Defining a collective well-being metric for social media

Social media is an integral part of our everyday lives that is having both negative and positive effects (Wang et al., 2014; Chen et al., 2017). Hence, as positive aspects rely on the same mechanisms exploited by threats, and because each user's behavior will affect the other members of the community while values can differ between communities, it is desirable and necessary to explicitly and collaboratively define shared community principles corresponding to the desired condition of the community. These community principles will constitute the foundation to define a specific measure of the overall impact of social media in the community at an individual and a societal level, that is, to measure the desirability or Collective Well-Being (CWB) of a certain condition of the social media community (Roy et al., 2018). These community principles, formalized in the CWB measure, together with an understanding of the virtual and physical social dynamics in the community, should drive the definition of users' behavior guidelines and connected educational objectives to reach and maintain the community in the desired condition, or in other words, to achieve a high level of CWB. A quantitative measure of CWB allows for a more accurate evaluation of the impact of different aspects of the interaction on the community while taking into account the complex and fast dynamics of social media. When CWB is estimated directly on the SM platform it could allow directing its autonomous components, e.g., recommenders, to collaborate in achieving the desired community condition. This would be a more democratic and transparent objective than the ones currently pursued by the social media platforms (Gorwa, 2019). In our framework, it is used to direct the algorithms at the interface between the educationally managed community and the external social media.

4.1. Research on collective well-being and social media

The literature presents several definitions and measures of well-being (Topp et al., 2015; Gerson, 2018). Some of them were applied in the context of social media to estimate their effects (Mitchell et al., 2011; Kross et al., 2013; Wang et al., 2014; Chen et al., 2017; Verduyn et al., 2017) but mostly considering the single individual with limited consideration for the overarching social aspects (Helliwell, 2003).

Gross Domestic Product (GDP) has been proposed as an index of the economic well-being of a community.[3] In such contexts, inequality is also an important factor, and it is common practice to use the Gini index to measure it (Osberg, 2017). While the economics view is difficult to connect to a social media context, they share similar key issues: which aspects to measure and, above of all, how to compare and aggregate measures of individuals' well-being to synthesize that of the whole society (Costanza et al., 2014), even if in this work we consider only the local educational community.

Multidisciplinary notions of CWB extend that of individual well-being to measure a group-level property (construct). They include community members' individual well-being incorporating diverse domains, such as physical and mental health, often stressing the presence of positive conditions. They study which properties of the community affect the members and how much each of these properties adds to a comprehensive measure of collective well-being. We already stressed the importance of education and educational objectives to support constructive interactions and achieve desirable community conditions, i.e., a high level of well-being. However, education itself is often already part of well-being frameworks (White, 2007; Michalos, 2017; Spratt, 2017b; Roy et al., 2018). The connection between education and well-being has been analyzed from several perspectives. In our framework, the most relevant

[3] Retrieved from: https://voxeu.org/article/defence-gdp-measure-wellbeing.

TABLE 1 Categories of properties of social media communities relevant for collective well-being and education extracted from the framework presented in Roy et al. (2018).

Categories	Description	References
Vitality	"The vitality domain includes... emotional health, with positive and negative affect, optimism and emotional intelligence."	Hong et al., 2017
Opportunity	the "perceived opportunity to achieve life goals and socioeconomic mobility," "influenced by ... access to education and training"	
Connectedness	"The connectedness domain assesses the level of connection and support among community members... Human relationships and relatedness are fundamental for the achievement of well-being according to many foundational theories of well-being.... Connectedness includes dimensions of social acceptance (i.e., positive attitudes toward people) and social integration (i.e., feeling a sense of belonging to the community)."	Dunn, 1959; Cohen and Wills, 1985; Fredrickson, 2004; Ryff et al., 2004; Lopez and Snyder, 2009; Seligman, 2011; Van Der Maesen and Walker, 2011
Contribution	"The contribution domain incorporates residents' feelings of meaning and purpose attributed to community engagement and belonging (e.g., volunteering, civic engagement, or belonging to a religious or community group). Sense of purpose is a cognitive process that provides personal meaning and defines life goals."	Forgeard et al., 2011; Keyes, 2012; Roy et al., 2018
Inspiration	"The inspiration domain includes community members' perceived access to activities that are intrinsically motivating and stimulating... [such as] life-long learning, goal-striving, creativity, and intrinsic motivation."	Meier and Schäfer, 2018; Roy et al., 2018

one is the one defined as *social and emotional literacy* in Spratt (2017a).

Roy et al. (2018) present a CWB framework divided into different domains and comprising health-care and non-health-care-related community factors where the contribution of the latter ones is supported by evidence of their effects on health. This framework can help to define a checklist for the definition of a community-specific CWB and related measures and indicators. We show in Table 1 the properties that may be relevant for education and social media communities for the following reasons:

- *Opportunity* domain is related to "the perceived opportunity to achieve life goals and socioeconomic mobility" (Diener and Seligman, 2006) as well as the access to education. Social media can be a powerful tool for accessing many opportunities. Feeling in control while using them, instead of just a distraction or worse an addiction, may be an important part of CWB for SM;
- *Connectedness* domain is related to the presence of supportive, high-quality, reciprocal relationships with secure attachments. Includes dimensions of social acceptance and social integration that depend on the behavior of other members of the community (Van Der Maesen and Walker, 2011);
- *Vitality* domain covers many emotional aspects of several individual well-being definitions, such as Fredrickson's one and Seligman's model of flourishing (Fredrickson, 2004; Seligman, 2012). However, spillover effects (Helliwell, 2003) and emotional influence make vitality an important aspect also at a social level;
- The threats presented in Section 2 would impact negatively the affects component of the *Vitality* and *Connectedness* domains;
- The *Contribution* domain relates to community engagement and related feelings of meaning and purpose. Contribution can improve other members' experience but may also have negative effects;
- The *Inspiration* domain relates to creativity and lifelong learning, areas where social media have a huge potential.
- The psychosocial *Community* characteristic that is clearly relevant for social media settings:

> *"A community with a negative psychosocial environment is one that is segregated and has high levels of perceived discrimination and crime, high levels of social isolation and low community engagement, and low levels of trust in government and fellow citizens."* (Mair et al., 2010; Klein, 2013; Engel et al., 2016).

Community is partially overlapping with the Connectedness and Contribution domains but describes aspects that are easier to concretely measure in social media networks.

While these formulations of CWB can inspire a guideline to define social media communities' principles and CWB metrics, they must be extended and formalized to better take into account the specific issues and opportunities of SM and in particular, the threats reported in Table 2. Another important aspect to address is combining contrasting factors or, in other words, formalizing the complex ethical decisions induced by the conflicts and trade-offs that emerge in any social context (Müller, 2020).

TABLE 2 Examples of social media threats distinguished into three categories (content, algorithmic, network, attacks, and dynamics) and examples of cognitive phenomena that may exasperate them.

Content based social media threats	Social media cognitive and socioemotional threats
Toxic content (Kozyreva et al., 2020)	Impulsivity (Lee et al., 2019)
Fake news/disinformation (de Cock Buning, 2018)	Fear of Missing Out (Alutaybi et al., 2019)
Bullying (Grigg, 2010; Mladenović et al., 2021)	Confirmation bias (Knobloch-Westerwick and Kleinman, 2012; Del Vicario et al., 2017)
Hate speech (Zimmerman et al., 2018)	Social reinforcement (Liu et al., 2018)
Stalking (Tartari, 2015)	Backfire effect (Bail et al., 2018)
Discrimination (Stoica et al., 2018)	Attention limit (Weng et al., 2012)
Radicalization (Johnson et al., 2016)	Emotional load (Kramer et al., 2014; Brady et al., 2017)
Smoke (Christakis and Fowler, 2008)	Anonymity (Urena et al., 2019)
Sexism/sexual harassment (Barak, 2005)	Depersonalization (Diener et al., 1980; Postmes and Spears, 1998)
Objectification (Ozimek et al., 2017)	Digital addiction (Kuss and Griffiths, 2011; Brand et al., 2014; Almourad et al., 2020)
Beauty stereotypes (Verrastro et al., 2020)	Lack of digital literacy (Whittaker and Kowalski, 2015; Xu et al., 2019)
Social media dynamics induced threats	**Algorithmic social media threats**
Filter bubbles (Bozdag and van den Hoven, 2015; Nikolov et al., 2015; Geschke et al., 2019)	Content diversity (Adomavicius et al., 2013)
Echo chambers (Gillani et al., 2018)	Misclassification (Stöcker and Preuss, 2020)
Digital wildfire Webb et al. (2016)	Algorithmic bias (Chen et al., 2020)
	Malicious users (Zhou Y. et al., 2017)
	Gerrymandering (Stewart et al., 2019)

4.1.1. Challenges of defining collective well-being for social media

Defining a CWB metric for SM is an ambitious endeavor that requires a combined effort of different disciplines. It would range from political sciences, sociology and psychology over ethical considerations all the way to computer science, machine learning and network theory. Besides CWB aspects for physical societies, the impact of integrated intelligent agents must also be taken into account in the context of social media, as discussed in Sections 2.2, 2.3. A CWB measure for virtual communities has to take into account the conflicts between members as they are frequent and algorithmically augmented. Therefore, the conflict between the right to freedom of expression, user satisfaction, and social impact must be stressed more when defining a social media CWB than with physical societies where these factors have slower and better-understood effects and may have regulations already in place (Webb et al., 2016).

Conflicts between members' interests pose serious ethical concerns that are out of the scope of this paper and have been the focus of recent research in AI and ethics in different domains (Cath et al., 2018; King et al., 2020; Milano et al., 2021). When social media are integrated into an educational framework, the problem may be mitigated by involving educators and experts as moderators. We propose that such an educational setup can also allow initial studies of the implications of a social media platform that aims to improve CWB.

4.2. Participative definition of social media community principles and CWB factors

Social media community principles and corresponding CWB factors must be shared by the members of the community. While research in the field can inform about common social aspects, internationally acknowledged human rights, or social media-specific phenomena, a community would most likely have the freedom to define tailored principles. To achieve this human-centered approaches to the participatory design of technology are being explored by the researchers. These approaches involve the stakeholders in the analysis of relevant factors and the co-design of technological solutions. One of the main challenges is bridging the gap between the community members' knowledge and the complexity of cyber-social systems like social media (DeVito et al., 2018). An example is a qualitative study to explore adolescents' representations of social media based on pictorial metaphors, reported in Sánchez-Reina et al. (2022). The study proposed and analyzed the outcomes of a school project entitled "The Social Media of the future." Discourses and visual representations of a total of 168 drawings about their visions for their ideal Social Media tools were analyzed. The results of the analysis pointed out that the relevant CWB factors shared by the adolescents participating in the study were: care about additive features, transparency in the conflict of interest behind the SM business, also in terms of agency to be able to monitor and control privacy and security facets.

4.3. Toward the automatic estimation of collective well-being in social media communities

Social media are strongly integrated with information systems that can affordably offer a huge amount of data with a high frequency. Transforming this data for the estimation of suitable collective well-being measures through machine learning methodologies would open the way to many research and applicative opportunities, such as autonomous systems that maximize CWB and avoid current issues induced by profit-based objectives.

Current CWB formulations are not easy to estimate directly using data available in real-time on social media, which is necessary to support an autonomous system optimizing CWB. Moreover, such formulations need to be extended to take into account specific social media issues. For example, most of the available formulations of collective well-being focus on positive aspects. Nevertheless, the positive aspects (see Section 4.1) and negative ones (see Section 2) need to be explicitly considered as part of the CWB as they strongly affect social media users and in particular teenagers.

We propose to define a *collective well-being metric for social media* by combining suitable components of classical CWB and SM threat measures. The measures of these components could be measured by periodically proposing specific surveys and activities (Loughnan et al., 2013). However, we propose that additional richer and more transparent measurements can be performed by developing intelligent components that analyze users' behaviors. In this definition, for each user, event, i.e., content or connection related, and aspect defined relevant for the CWB three terms are computed:

- **CS(aspect, user)** Content Shared measures the aspect-specific value of the content shared by the user;
- **CE(aspect, user)** Content Exposure measures the aspect-specific value of the content observed by the user;
- **CC(aspect, user a, user b)** Contact Creation measures the aspect-specific value of new connections based on the participants' CS and CE.

These elements account for the double role of each member of the social media community as both receivers and producers of content. In our educational setup, where only the community of interest is in contact with an external social media community, we distinguish between "endogenous" and "exogenous" aspects. The community can be exposed to threats that are generated outside but a community can also generate such threats inside as part of the interactions in the social medium. In this case, the feeds from external sources may be weighted differently.

While the CS can be seen as a direct expression of the state of the user, it strongly depends on the user's style of interaction. Moreover, only relying on the content shared by users would induce a substantial delay compared to the moment when a user got actually affected by observing a piece of content (CE). Conversely, the user is exposed to a multitude of diverse inputs hindering the interpretation of the overall effect only from the CE, while the user's reactions (CS) may be more indicative of the most impacting events. Indeed, current affective state estimators and toxic/positive content detectors can only provide noisy estimations of the current user state and the content quality. However, the availability of complementary data with higher reliability is limited.

Once each event is scored for each aspect of interest, it must be decided how to aggregate these terms over users, time, and the different aspects to obtain an estimation of the total CWB of the community. Indeed, the definition of an actual metric following this strategy requires making a number of choices. For example, about the scale for the terms of different aspects considered. Regarding aggregation over time, CC, CE, and CS values could be simply averaged. Other approaches could be considered to take into account the frequency of the events or the diversity of opinions presented or give more relevance to extreme events, which may be more accurately detected and evaluated. In particular, the value of being exposed to multiple opinions (time-aggregated CE) may be augmented with a measure of diversity (e.g., entropy) (Garimella et al., 2017; Matakos et al., 2022).

Clearly, the design of the CWB metric presents a number of challenges requiring careful consideration even for small educational communities that our framework targets. In devising their solutions often the naive approach may at best be ineffective, and at worst exacerbate the issues it was intended to solve. For example, the aggregation over the aspects dimension may not seem complex when considering the aspects to be independent. In reality, the impact of the various aspects on the users may be interlinked, for example over exposure to content focused on one aspect (e.g., videogames) may lead to overuse of the platform or tire the user who will lose the opportunity to learn about more important content (e.g., social issues).

The most complex aggregation to design is over users because it has to balance the well-being of different individuals and groups of users taking into account their conflicting interactions along different dimensions. It is important to consider the different features of each user while respecting privacy constraints. For example, vulnerable users are often victims of toxic content but also producers (Bessi, 2016; Bronstein et al., 2019; March and Springer, 2019), which affects the CS value. It is important that they are not isolated (Burrow and Rainone, 2017) and that, at the same time, the toxic content should not be fed to those who could be more affected and instead presented to educators or other community members that have shown constructive reactions to such type of content. This means that the content exposure (CE) should be differently weighted for different community members based on their resilience and that supportive connection creation (CC) should be favored between people with high and lower resilience. Still, it is important that resilient members are not overloaded with toxic content and support responsibilities (Steiger et al., 2021).

Apart from the weighting issue issues, another important format issue is the selection of the actual aggregation function across users. Adopting the naive average a society where a few radicalized users share extremely hateful content may have a higher CWB score than one with a number of users sharing content about action movies with slightly violent scenes. Another reason why a linear combination of components may not be suitable in the definition of a well-being measure is that it will simply induce maximizing the terms with positive weights and minimizing terms with negative ones, without allowing a balance. For example, if interactions between drastically opposite opinions are considered negative because of possible backfire effects and flames (Bail et al., 2018), and interactions between excessively similar opinions are also considered negative because of the echo chambers they may give place, then also interactions between moderately different opinions will have a negative value even when they may lead to a reduced polarization. Other aggregation functions may be chosen but it is still difficult to find general solutions. For example, defining the well-being of society as the well-being of the member with lower well-being (i.e., minimum instead of an average) could lead to focusing all the resources on factors that may not be actually changed.

4.4. Network measures for collective well-being on social media

Network-specific measures (Rayfield et al., 2011) can be an important part of an actionable CWB measure for social media. Several threats and well-being-related phenomena are implicitly defined in terms of network measures. These measures may also be particularly useful as proxies of future critical conditions without having to execute expensive simulations. For example, Moore et al. (2021) show that the increase of a network measure of inclusiveness promotes the efficiency and robustness of a society. Stewart et al. (2019) show that an unbalanced network structure may lead to suboptimal collective decisions. Effects of positive and negative interactions at a network level have been studied in Leskovec et al. (2010). Concepts like social influence and homophily (Aral et al., 2009; Guo et al., 2015) play an important role in the formation of different network conditions, like segregation, that are crucial for CWB. The diversity measures already proposed as part of the CE, CS, and CC elements would also contribute to a higher CWB rating for diversified and integrated communities than polarized and segregated ones. Other measures viable to characterize user roles, such as centrality and closeness, can also be used to aggregate the individual users' threat scores over the network (Drachsler et al., 2009; Manouselis et al., 2011).

5. An educational collective well-being recommender system

Recommendation systems (RSs) are ubiquitous in online activities and are crucial for interacting with the endless sea of information that the Internet and social media present today. In social media platforms, they have introduced the possibility of personalizing suggestions of both content and connections based on the use of user profiles containing also social features (Heimbach et al., 2015; Chen et al., 2018; Eirinaki et al., 2018). Their goal has been to maximize the users' engagement in activities that support the platform itself. However, these self-referential objectives fail to consider repercussions on users and society, such as digital addiction (Almourad et al., 2020), filter bubbles (Bozdag and van den Hoven, 2015), disinformation wildfire (Webb et al., 2016), polarization (Rastegarpanah et al., 2019), fairness (Abdollahpouri and Burke, 2019; Ranjbar Kermany et al., 2021), and other issues discussed in Section 2. To address this, we propose the concept of *Collective Well-Being aware Recommender Systems (CWB-RS)*. The CWB-RS extends social media RS intending to maximize the cumulative long-term *CWB metric* instead of self-referential platform objectives. Compared to previous efforts in dealing with possible negative effects of RSs (Abdollahpouri and Burke, 2019; Rastegarpanah et al., 2019; Ranjbar Kermany et al., 2021), the CWBRS takes into account multiple issues and, to reduce their cumulative impact on society, it adopts longer terms strategies fitting into our educational framework.

Integrating educational objectives aimed at achieving *CWB* in the longer term the CWB-RS will also have functions similar to those of a (collective) *Intelligent Tutoring System* (Greer and Mark, 2016). RSs have been widely used in educational settings

(Manouselis et al., 2011), and they are receiving increasing attention due also to the fast growth of MOOC (Romero and Ventura, 2017) and the availability of big data in education (Seufert et al., 2019). In educational contexts, recommendations are sequential and functional to achieving learning goals (Tarus et al., 2017). Similarly to the social media context, they have also employed social information (Kopeinik et al., 2017; Elghomary and Bouzidi, 2019). However, they are usually acting on the content provided by educators with educational aims, while CWB-RS also has to redirect disparate content flowing from external Social Media toward achieving educational objectives.

As shown in Figure 1, the CWB-RS creates new recommendations presented through the Companion by processing both the content generated *internally* by the members of the *educationally managed social media* community and the content recommended for them by the RSs of the *external* platform. *Content Analyzers and Threat Detectors* (see Figure 3 and Section 5.4) will analyze each piece of content to evaluate the level of threat and other relevant information for the CWB metric, such as the users' opinions and emotions (see Section 4.3). This information will be used to: 1) evaluate the current condition of the users; 2) *augment and contextualize* the content provided to the users; 3) *evaluate* the future effects of different sequences of content re-rankings and recommendations through predictive models of users' conditions; 4) *select* the actions that account for the highest expected, long-term, cumulative CWB metric.

5.1. Educational directions for the CWB-RS

CWB-RS educational objectives are designed by educators and experts (see Section 3). They can be encoded in terms of measures related to specific threats or other well-being variables, such as those extracted by *content analyzers and threat detectors* allowing to easily combine educational and regular CWB objectives (Van Seijen et al., 2017). Different approaches have been proposed to effectively combine and scale multiple terms in objective functions (Harutyunyan et al., 2015; Marom and Rosman, 2018). These objectives express how much each student: (a) is conscious of his role in other users' well-being, (b) improves his behavior, and (c) is having a healthy experience. For example, an objective would be "curb obsessive selfies posting" (Ridgway and Clayton, 2016), which would act on the content shared (CS) for the aspect "selfies." Another example could be breaking the filter bubbles focused on racist content and helping users hold an unbiased mindset (reduce both CE and CS on the aspect "racism"). In this case, the connected recommendation strategy will be to provide content with opposite but not confrontational perspectives (Bozdag and van den Hoven, 2015; Garimella et al., 2017; Matakos et al., 2022). This strategy can be combined with educational games proposing specifically themed challenges, such as finding pictures of achievements performed by people of different ethnicities, suggesting changing the recommender filter parameters directly, or just reducing the racist content presented and substituting it with low harm feeds. The CWB-RS can also recommend content to asses the current student's condition (Zhou et al., 2010; Kunaver and Požrl, 2017) to inform successive personalized interaction.

Educators and experts will also define interaction strategies specific to each objective (Griffith et al., 2013). Sketches of *high-level CWB-RS educational strategies* will be hand defined by the educators and experts to choose between the different educational objectives for each student in an effective and contextualized manner. *Lower-level educational strategies* for the CWB-RS comprise hand-defined *learning activities* and *minigames* as well as modulation of the recommendations, for example, showing diverse content as tests to explore students' preferences.

Engagement is an important factor for both social media platforms (Wu et al., 2017; Zheng et al., 2018) and educational activities (Sawyer et al., 2017). The CWB-RS must prevent students from "dropping out" (Eagle and Barnes, 2014; Yukselturk et al., 2014) and moving to non-educational social media. In a complementary manner to the game-oriented motivational mechanisms of the Companion (Van Staalduinen and de Freitas, 2011), the CWB-RS must therefore preserve a healthy level of engagement (Arroyo et al., 2007; Chaouachi and Frasson, 2012; Mostafavi and Barnes, 2017; Zou et al., 2019) while avoiding excessive exposure to toxic content as well as any form of addictive use (Lavenia, 2012; Nakayama and Higuchi, 2015; Almourad et al., 2020).

5.2. Challenges in social media RS and CWB-RS

The realization of effective social media recommendation systems, as reviewed in Chen et al. (2018) and Eirinaki et al. (2018), presents several challenges that in recent years have brought drastic changes to the field. In particular, some of the biggest challenges are the highly diverse information they process (e.g., content, trust, connections), the complex dynamics of the interactions, the fast pace of growth of the social graph, and the enormous amount of multimedia and textual elements to process (Covington et al., 2016; Eksombatchai et al., 2018). In the case of the CWB-RS, the size of the internal social network is limited (i.e., the number of students) and a big part of the data will come preselected by the external RS, thus forming an implicit two stages approach (Borisyuk et al., 2016; Covington et al., 2016; Ma et al., 2020) with only the second stage in charge of the CWB-RS. However, the creation of a CWB-RS presents several other theoretical,

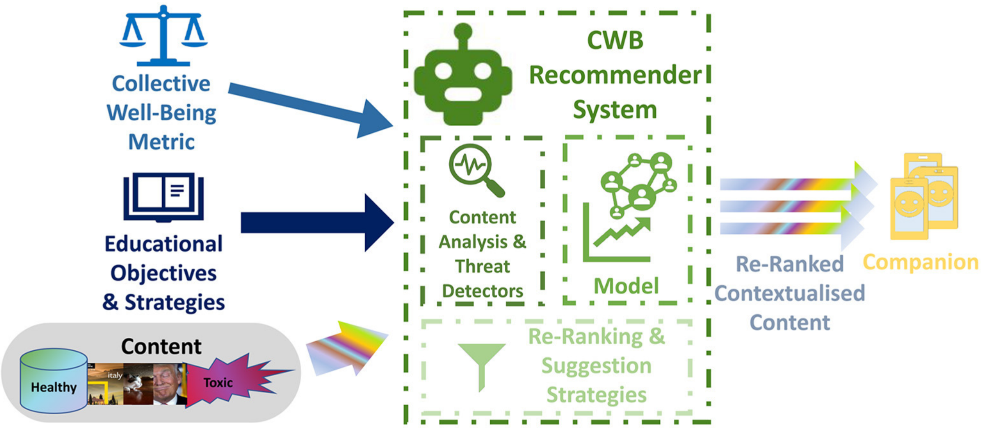
Collective Well-Being Metric
Educational Objectives & Strategies
Content
Healthy
Toxic
CWB Recommender System
Content Analysis & Threat Detectors
Model
Re-Ranking & Suggestion Strategies
Re-Ranked Contextualised Content
Companion

FIGURE 3
Role of the CWB-RS in the Companion. CWB-RS will process the *content generated by the users* of the *educationally managed social media* and the *content externally recommended* for them by the RSs of the external social media platform to create new recommendations aimed at maximizing the cumulative long-term *collective well-being metric*. *Content Analyzers and Threat Detectors* will analyze and evaluate the level of threat for each piece of content and other relevant information as the users' emotional state. This information will be used to: 1) *augment* the information provided to the users by the companion interface; 2) *evaluate* through *predictive models of users' opinions and reactions* the future effects of different sequences of re-ranking and recommending actions; 3) *select* the re-ranking and recommending actions that resulted in the highest expected cumulative improvement in terms of learning objectives, CWB metrics, agreement with selected educational strategies and user engagement.

technical and ethical challenges that are mostly not faced by classical RS.

5.2.1. Diverse internal and external content

A first demand for the CWB-RS is to combine content defined by the members of the educationally managed social media with recommendations from the external social media. While this controlled separation from the external platforms offers the opportunity for novel educational experiences, the heterogeneous nature of signals and structures poses the question of how to combine them. This is all conceptually similar to some of the major challenges and opportunities of enterprise and intranet search compared to general web search (Hawking, 2010; Kruschwitz and Hull, 2017).

5.2.2. Social information

In classical social media RSs, the use of social information is relatively straightforward. For example, connections between users can be interpreted as a cue of similarity between their interests. For a CWB-RS, sharing content based on social connections may spread toxic content, however, it can be useful if one of the connected users has exemplary behavior. Moreover, social network structures affect not only information propagation but also decision and behavior (Stewart et al., 2019). Thus in CWB-RS, some properties of the structure of the social connection graph of the internal community may be part of the objective (e.g., CC in Section 4.3). Still, the recommendation and creation of connections between diverse groups may sometimes lead to toxic behaviors, e.g., backfiring (Bail et al., 2018).

5.2.3. Lack of direct reference information for the CWB-RS

Classical RSs maximize the users' satisfaction and engagement, usually estimated through accessible proxy measures, such as time of usage or likes. These allow the definition of reference information or teaching signals to improve the RSs behavior based on the similarity between items or between users' previous selections (Wu et al., 2017; Eirinaki et al., 2018). These signals do not inform about the level of CWB or achievement of user-specific educational objectives. The CWB-RS needs both to estimate less accessible quantities, such as knowledge acquired or behavioral improvement, and to recommend content taking into account the users' learning trajectories, comprising their current state and assigned objectives. Still, these measures do not easily translate into future recommendations. For example, if a recommendation led a student to achieve an educational goal, this does not imply that it would be useful to suggest similar content to the same student again, as it will not provide him with new educational information. It may still indicate that it is useful to suggest similar content to other students who have to achieve the same goal.

5.2.4. Temporal aspects and sequence of recommendations

Classical RSs regard recommending as a static process mainly focusing on "the immediate feedback and do not consider long term reward" (Liu et al., 2018; Zhao et al., 2019a). Instead, to achieve lasting CWB and the related educational processes, it is necessary to account for the effects of sequences of recommendations. For example, sequencing of lectures, tests, and feedback, is common in most educational strategies. In addition, a classical RS does not consider the interdependence between users' preferences and the RS recommendations, which is crucial to model and counter the filter bubble and echo chamber phenomena. Another reason for the CWB-RS to consider a temporal dimension is to enable the use of an accurate dynamic model of the students and the natural variation of their preferences (Zeng et al., 2016). This allows, for example, to prepare the conditions and select the best time for exposure to content aimed at improving students' empathy as well as avoiding wrong conditions, such as those with a high level of user stress, when such content would be ignored or even lead to backfire (Bail et al., 2018).

5.3. CWB-RS adaptation and personalization through Reinforcement Learning

The Reinforcement Learning (RL) paradigm adoption to drive the adaptation and personalization of the CWB-RS behavior (Zhao et al., 2019a; Zou et al., 2019) is a natural solution to the sequential control, lack of supervised teaching signal, and the other technical issues described above. RL-based recommender systems are recently gaining attention in the community (Shani et al., 2005; Liu et al., 2018; Zheng et al., 2018) because of their flexibility, and the growth of the deep reinforcement learning field (Mnih et al., 2015; Zheng et al., 2018). As suggested in Zhao et al. (2019a), RL-based RSs allow solving not only the problem of frequent updates of the user profile, typical of RS in social media, and offer also a precise formulation of the initialization problem in terms of exploitation-exploration (Iglesias et al., 2009; Hron et al., 2020).

From a machine learning perspective, *CWB-RS* educational objectives, learning strategies and activities, can be respectively seen as manually defined rewards, sub-goals, and sub-policies in a Hierarchical Reinforcement Learning (HRL) framework (Zhou et al., 2019, 2020) which improves its adaptation performance by breaking down the high-level decisions (e.g., the educational objective a student must achieve) and the step-by-step decisions (e.g., which activity or content to show at the moment). This reduces the computational costs and amount of data necessary to derive the educational policy and objectives directly from the long-term optimization of the CWB metric (Barto and Mahadevan, 2003).

Both classical RL (Iglesias et al., 2009; Dorça et al., 2013; Zhou G. et al., 2017) and HRL have been used in ITS (Zhou et al., 2019, 2020) and RS. To our knowledge, this is the first time they are combined. While the field of RL-based ITS is still young and presents several limits (Zawacki-Richter et al., 2019), it could address the complex problem of supporting students dealing with the diverse and enormous environment of social media. Still, the additional flexibility of RL-based RS comes at the cost of higher complexity, particularly in terms of training and evaluation setup (Henderson et al., 2018), as well as deploying in real-world applications (Dulac-Arnold et al., 2019; Rotman et al., 2020).

5.3.1. Difficulty of creating CWB-RS datasets

Reinforcement Learning systems developed to act in real-world conditions are usually pretrained offline on available datasets. Much of the solution quality depends on the similarity between the dataset and the application setting (Rotman et al., 2020). The creation of real-world reinforcement learning datasets most often requires *ad-hoc* solutions.

The collection of CWB-RS datasets must take into account the users' profiles, which may be gathered using a self-reported survey, as in Khwaja et al. (2019), as well as users' neighborhood information, behaviors (e.g., posts) and observations (e.g., recommendations). Mining this information, however, needs to comply with privacy and company policies. Additional challenges are presented by the necessity to cover the various reactions that students may have under exposition to combinations of disparate social media (Zhao et al., 2019a). Social media show a complex interplay between the individual, social, and technological levels of filtering (Gillani et al., 2018; Geschke et al., 2019), with substantial effects on users' behaviors. Therefore, one of the strongest challenges is washing out the effects of the RS adopted during the data collection, which functioning is usually unknown, enabling the use of the dataset to train a CWB-RS that could propose diverse recommendations and induce different selections.

Crowdsourcing (Boudreau and Lakhani, 2013) can be used for large-scale evaluations or for creating datasets under limited periods (Kittur et al., 2008). However, special care needs to be taken to ensure the reliability of crowd data (Buhrmester et al., 2018) as the seriousness with which volunteers take their interactions with the system can be limited. These complexities demand to devise an effective strategy to build a real-world dataset that considers including the micro-, meso-, and macro-structure, different sources, and modalities.

Model-Based RL For the specific setting of the educationally managed social media community, the task is simplified considering the reduced content variety compared to the external community. Also, while a CWB-RS must be aware of the condition and behavior of the entire community, this may be factored in terms of the dynamic models of its members. Using different combinations of the same members' models, it could be possible to create different community models that allow a broader set of training conditions for the CWB-RS in simulation. They will also enable online simulations for estimating the results of a sequence of recommendations (see Figure 3 and Zhao et al., 2019b; Schrittwieser et al., 2020). The literature on interaction models for social media is extensive. Szabo and Huberman (2010) were one of the first to show the importance of cognitive and content factors. The models proposed in Guo et al. (2015); He et al. (2015) reason simultaneously on the patterns of propagation and the topics. Most of these models do not account for user adaptation, which is crucial in this context. However, the solution could be to adopt generative models of adaptive user behaviors, such as Das and Lavoie (2014), Lindström et al. (2019), and Ognibene et al. (2019). While these studies and many more led to improved forecasting systems, there is a consensus that there are intrinsic problems that limit the predictive power with both sufficient accuracy and anticipation, see for example Cheng et al. (2014). A significant improvement of baseline algorithms requires very detailed information about the community (Watts, 2011). However, the CWB-RS has access to rich information about the educationally managed network. This, together with its limited, size will improve the efficacy of the predictive models.

5.3.2. Risks in the exploration phase of RS based on RL

Reinforcement learning can provide online adaptation to conditions that detach from the training set used for offline pretraining. However, this comes with exploration costs that in real environments can pose prohibitive risks (Rotman et al., 2020). Even if the CWB-RS is not facing critical safety tasks like those of self-driving systems, repeated sub-optimal recommendations may just reinforce the threats the Companion is trying to address. To alleviate these issues adaptive novelty detection methods (Rotman et al., 2020) will be implemented in the CWB-RS to recognize situations far from the agent experience and hand over the control to educators or a safe controller. Moreover, the HRL paradigm has been adopted for the CWB-RS to constrain and minimize exploration risks and costs (Nachum et al., 2018; Steccanella et al., 2020) while providing direct control and interpretability to the educators (Shu et al., 2017; Lyu et al., 2019). Ultimately, under the direction of learning objectives and strategies, the set of problems that the CWB-RS will have to solve would be limited to balancing reranking requests from different active strategies and prioritizing one objective over the few others defined in the current high-level learning strategy.

5.3.3. Noisy rewards and action results

An additional constraint comes from the difficulty of characterizing the toxicity of the social media content (see Section 5.4) on which the RS must act. This results both in erroneous recommendations (e.g., content that was mistakenly supposed to be toxic undergoes reduced propagation speed) and stochastic rewards (toxic content is evaluated by error as healthy and a positive reward is provided to the CWB-RS from the CE and CS estimation). While the RL method accounts for noisy actions' results, they still affect the performance of the system, both in terms of execution and learning time. Regarding noisy rewards, literature has only recently started to provide solutions (Huang and Zhu, 2019; Wang et al., 2020). Still, it must be noted that in our setting, getting a positive reward for something that was considered positive should not crucially impair the acquired RS policy as the system allowed the propagation of something that it evaluated healthy (or toxic) and accordingly evaluated its reception by other SM users. Thus, in this case, the two errors may cancel each other out and take advantage of improvements in the detectors. Moreover, when applying RL for ITS, an additional strategy that can be leveraged to counter these issues is to use more reliable tests that would allow for evaluating the state of the users and provide more reliable rewards. Due to social media complexities, the effects of detectors' failures on the performance of CWB-RS can be heavy, with backfiring as the worse-case scenario. Extensive tests would be necessary both in simulation (e.g., Geschke et al., 2019 and real-life as well as comparisons with classical recommender systems for social media that are not sensitive to content toxicity.

5.4. Threat detectors and content analyzers

Social media threat detectors and content analyzers have multiple roles in the platform already described in Section 5. Given the importance of social media threats, as described in Section 2, researchers have been studying how to automatically identify them (some examples can be seen in Table 3). Several shared tasks have been proposed and each year they become more challenging. Moreover, new evaluation criteria, such as multilingual detection at Task 5 in Semeval 2019 (Basile et al., 2019), different domains at HaSpeeDe in Evalita 2020 (Hoffmann and Kruschwitz, 2020; Sanguinetti et al., 2020), detections at the spam level at Task 5 in Semeval-2021 (Pavlopoulos et al., 2021), and generalization to social media platforms other than those used in training at EXIST in IberLEF 2021 (Rodríguez-Sánchez et al., 2021), have been included in the datasets.

Those detectors are usually defined as a classification task commonly solved using deep learning. Different features are used as parameters for the models. For example, in fake news

TABLE 3 Short list of works on social media threat detection and content analysis exemplifying the variety of approaches and works.

Type of detector	References
Stance detection	Augenstein et al., 2016; Zarrella and Marsh, 2016
Controversy identification	Hessel and Lee, 2019; Zhong et al., 2020
Fact-checking	Dale, 2017; Long, 2017; Wang, 2017; Jobanputra, 2019; Liu and Lapata, 2019; Nie et al., 2019; Atanasova et al., 2020
Hate speech	Cer et al., 2018; Basile et al., 2019; Indurthi et al., 2019; Nikolov and Radivchev, 2019
Violence recognition	Perronnin et al., 2010; Nievas et al., 2011; Bilinski and Bremond, 2016; Zhou et al., 2018
Gender bias	Prost et al., 2019
Offensive content	Hosseini et al., 2017; Zampieri et al., 2019

identification, Hessel and Lee (2019) explored the combination of different models and features, including hand-designed features, word embeddings, ratings, number of comments and structural aspects of discussion trees. In addition, another key element of the detectors is the datasets. For some threats (e.g., hate speech and fake news), few standard datasets target social media, but that is not the case for all threats. For violent content detection, for example, there is not a standard dataset focused on SM to the best of our knowledge. In order to overcome these limitations, works such as Bilinski and Bremond (2016) and Zhou et al. (2018) use a proxy dataset, such as Hockey Violence Dataset (Nievas et al., 2011).

Regarding the content analysis to extract users' affective state, beliefs and opinions, similar approaches are viable. Affective Computing aims to recognize, infer and interpret human emotions (Poria et al., 2017), distinguishing between sentiment analysis, polarity of content (e.g., Liu et al., 2017; Guo et al., 2018; Gupta et al., 2018), and recognition of the emotions present in a piece of information (e.g., Baziotis et al., 2018; Ahmad et al., 2020). In comparison, Opinion Extraction aims at discovering users' interests and their corresponding opinions (Wang et al., 2019). In general, the systems extract the entity or the target, the aspect of the entity, the opinion holder, the time when the opinion was expressed, and the opinion (Liu, 2012). Similarly, the positive aspects of social media interaction, crucial for estimating the CWB, could be extracted. Still, they have attracted less attention, but see Wang et al. (2014) and Chen et al. (2017).

Despite the success achieved by these efforts, the robustness of these systems is still limited. For instance, seldom they can generalize to new datasets and resist attacks (for example, word injection) (Hosseini et al., 2017; Gröndahl et al., 2018). An example of that is the case that occurred in the OffensEval shared task (Zampieri et al., 2019), where different hate speech classification models were compared in different subtasks. The best system in Subtask B (i.e., Han et al., 2019) ranked the 76th position in Subtask A that is a general and simple case of Subtask B.[4] This example stresses how small changes in these tasks may drastically impact system performance informing on the challenge of applying these approaches in the dynamic contexts of social media. Some recent models can generalize the task while maintaining similar results in different platforms and languages under certain conditions (Wilkens and Ognibene, 2021b).

6. Use case

The following scenario is an example of how the Companion enables the personalization of educational interventions to help develop users' resilience against social media threats. The focus of this use case scenario is on the algorithmic threat of filter bubbles and how it can affect the users' perspective of healthiness. The content threat is associated with body image concerns (Marengo et al., 2018).

Alex is a 15-year-old high school student who spends a fair amount of his free time on his phone on a daily basis.

Without the Companion: Alex scrolls through his social network newsfeed and encounters a photo of an influencer that promotes masculinity. As summer is approaching, he decides to check the influencer's profile for possible tips to help him tone his body. Alex spends the next hour watching videos in the influencer's profile and starts following similar profiles. The social media platform algorithms learn that Alex is interested in posts related to masculinity, and he can spend hours interacting with this type of content. Thus, to maximize engagement, the platform starts displaying more content related to masculinity. Occasionally, the platform presents an advertisement in the form of a post to indulge Alex to buy a related product. Alex now finds his newsfeed to be filled up with fitness influencers and fitness products. Day by day, he likes and follows more fitness influencers, slowly leading his newsfeed to be full of fitness influencers that promote a specific body type. Through time, Alex's opinion regarding beauty standards starts to shift. He starts to believe that the male body needs to be muscular to be considered attractive and healthy. When looking in the mirror, he now feels that his body is far away from being considered attractive, and he will never be able to reach the beauty standards that have been set. He starts feeling unhappy with his body and seeks comfort through his social media platform. He comes across an influencer that promotes a product for rapid muscle growth and decides to look further into his profile. There he encounters photos that show a drastic change in the influencer's physical appearance claimed to be the result of the product. Alex

[4] We highlight that, despite this extreme case, systems tended to maintain a similar performance across the different subtasks.

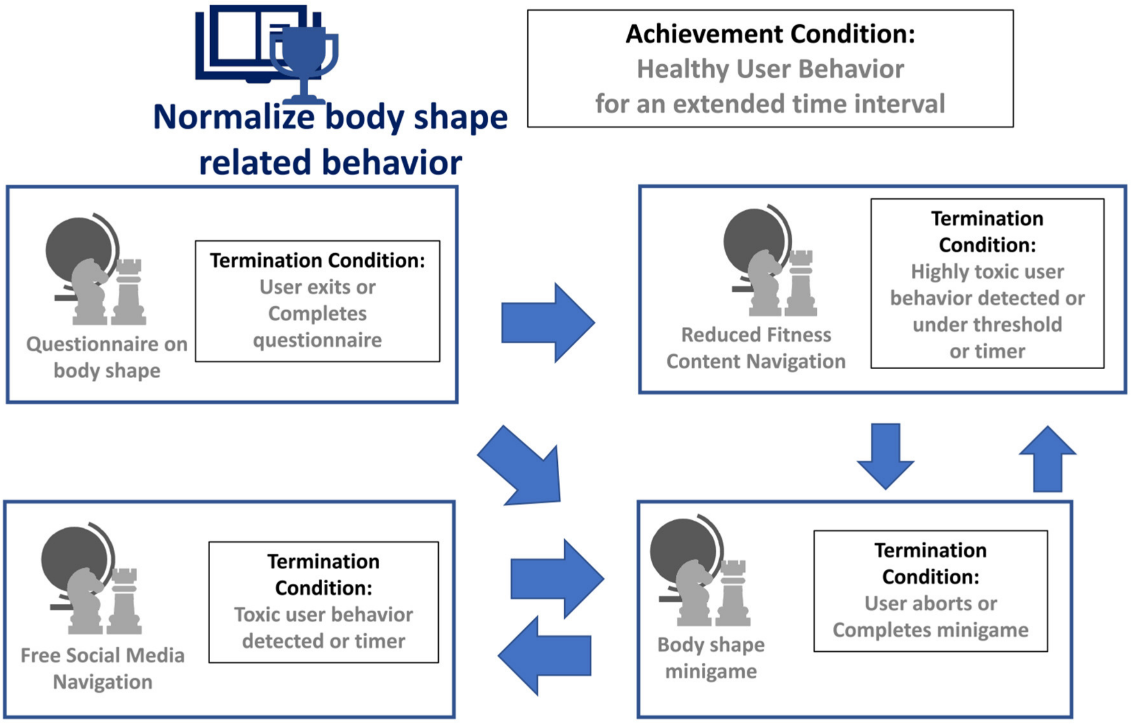
Normalize body shape related behavior
Achievement Condition: Healthy User Behavior for an extended time interval
Questionnaire on body shape
Termination Condition: User exits or Completes questionnaire
Reduced Fitness Content Navigation
Termination Condition: Highly toxic user behavior detected or under threshold or timer
Free Social Media Navigation
Termination Condition: Toxic user behavior detected or timer
Body shape minigame
Termination Condition: User aborts or Completes minigame

FIGURE 4
Educational strategy example. A visual example of how the policy to improve body shape-related behavior is accomplished within the platform. An initial questionnaire is completed by the user to determine if their behavior is classified as healthy or toxic. In the scenario that the questionnaire results come back as healthy, the user is placed into a free social media navigation state. This state will be terminated when the system detects that the user's behavior is no longer classified as healthy. This classification is done by analyzing the profiles the user has been following based on their category and further analyzing them with image classifiers. In this case, the system detects that the user's behavior has shifted from healthy to toxic a learning activity is initiated. The user is then placed into a state where the system alters the content they receive in their newsfeed.

decides that this product is the solution to his problem and buys it.

With the Companion: Alex scrolls through his social network newsfeed and encounters a photo of an influencer that promotes masculinity. As summer is approaching, he decides to check the influencer's profile for possible tips to help him tone his body. Alex spends the next hour watching videos in the influencer's profile and starts following similar profiles. The Companion runs in the background and detects that the majority of profiles Alex has started to follow fall under the category of fitness. Image classifiers further identify that those profiles promote a specific body type. Then, the Companion triggers a narrative script and notifies Alex that a new game (the script) is available (Figure 4). Alex accesses the game and initiates the narrative script. The narrative script mechanisms assign him to an influencer that supports the opposite perspective (counter-narrative) than the one he triggered. He is instructed to navigate through the profile and self-reflect on how this profile makes him feel. Alex is asked to participate in an online collaborative game showing the impact of social media influence and filter bubbles on our decision capabilities (e.g., see Lomonaco et al., 2022b and Section 7.1). He is then shown a brief video of how SM algorithms work and how they can place a user into filter bubbles. In the next screen, Alex enters a mini-game where he is instructed to manipulate a filter bubble by following and unfollowing profiles and by liking and unliking posts. During the game, Alex can see how the newsfeed of the user changes according to his behavior. Alex starts to understand how social media works and how algorithms can learn from our behavior. Once the game is over, the narrative script ends, and Alex receives a badge for completing it. The educational component registers Alex's signs of progress and marks the learning objective of filter bubbles as complete (Figure 5). Alex returns to his social media profile and receives a notification from the Companion that the content of his newsfeed has been altered by the CWB-RS component to reduce the harmful content that he has been receiving. He has the option to revert this setting, but he decides to continue with it. The CWB-RS component filters Alex's news feed with images unrelated to muscular fitness. Eventually, this alters Alex's content needs and influences him to start following profiles that are not solely related to muscular fitness, which leads to minimizing his exposure to influencers promoting a

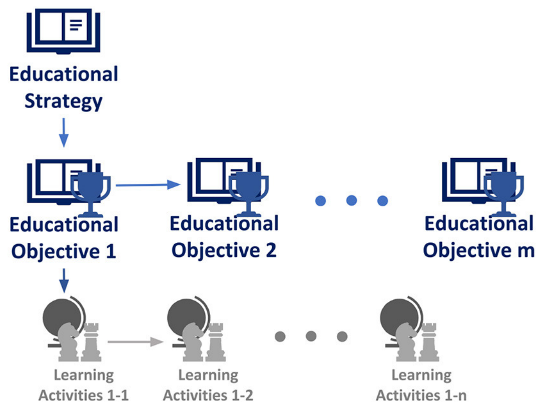

FIGURE 5
A visualization of the hierarchical structure of the educational strategy. Each educational strategy (narrative script) has a set of educational objectives that can be reached by a sequence of adaptive learning activities. The learning activities can be in the form of free-roaming, guided roaming, quizzes, minigames, or participating in group tasks. They are triggered based on the user's behavior within the platform.

perfect body. To confirm that Alex is staying on the right track, a few days later, the Companion operates a further inspection to analyze the content being followed. The Companion verifies that Alex's online behavior has improved after the completion of the mini-game and it does not trigger any further mini-games for him. Alex receives a notification informing him that the CWB-RS component has stopped altering his newsfeed. His newsfeed content has now become more balanced. Alex has become less obsessed with the idea of having a muscular body.

7. Preliminary experimental results

The realization of the COURAGE companion is progressing through the study of different educational strategies and the development and testing of educational tools and computational components.

7.1. Educational and psychological studies

Educational and psychological studies are the starting point to define objectives, methodologies, and tools that will be integrated into the Companion and guide the participatory design of the *educationally managed social media community*.

Data was collected through online studies to calculate correlations between toxic content tagging (e.g., disagreement measure) and personality traits (e.g., cognitive empathy or authoritarianism). A link between learners' judgments and personality traits could only be weakly found for authoritarianism (Aprin et al., 2022a). We also studied users' intentions to share emotional images. The study was conducted in Italy and Germany, with university students surveyed online in Germany. The evaluation of nearly 200 students is not yet completed. It is expected that results will provide insights into the relationships between socio-emotional competencies, moral values, and the willingness to share images with diverse social groups.

Similarly, a pilot study aimed to investigate the relationships between emotional intelligence and social media threats was conducted involving 110 adolescents of a secondary school in Italy during an extracurricular school activity (Scifo et al., 2022). In particular, two research studies have been conducted within this pilot. The first study had the purpose of investigating the relationships between emotional intelligence and adolescents' ability to detect fake news on social media. The second study included a training path aimed at stimulating emotional intelligence and promoting a conscious use of social media. Moreover, the training path has also contributed to raising adolescents' awareness of bullying and cyberbullying. The analysis of the results is ongoing. These studies will drive the development of new educational components for the companion and help to define the companion's personalized educational strategies.

We tested a game-based educational experience (De Gloria et al., 2014) to increase students' awareness of social media algorithmic threats, focusing on filter bubbles and echo chambers inspired by the "wisdom of crowds" (Lorenz et al., 2011; Becker et al., 2017). It was tested with both University and High school students providing encouraging results (Lomonaco et al., 2022b). While more data is being collected a specific component is being designed to reproduce the experience inside the companion.

Furthermore, we developed a scenario to inform students about racist content on social media. Here, users are informed about the background of racism by the virtual companion in a closed social media environment. By means of the results of an experimental study in which the virtual learning companion either transmits information on racism (experimental group) or not (control group), we will analyze the effects on users' knowledge and awareness regarding racism.

Also, we are constantly working on a scenario for empathy training which shall sensitize young users regarding the negative effects of cyberbullying. For this training, for instance, a video showing an example and providing a definition of empathy will be shown to students in an experimental study in Germany and Spain. We hypothesize that students who completed the empathy training will be more sensitive to cyberbullying and are less likely to intend to bully in the future.

Finally, in Taibi et al. (2021) we present a platform specifically designed to support the development of competences related to Information and Data Literacy. This platform extends the open-source alternative to Instagram called Pixelfed, with functionalities designed to support students in increasing their awareness of social media mechanisms based upon artificial intelligence algorithms. A pilot with secondary school students has been conducted to experiment educational activities based on the proposed platform.

7.2. Educational components

Several components and applications, that will later be integrated in the Companion, are being developed and tested.

A number of mini-games to increase social media awareness, covering topics such as the digital footprint, social media addiction, misinformation and body image dissatisfaction have been designed and tested. For instance, the serious mini-game "SwipeIt" for sensitizing students to toxic content (e.g., cyberbullying), was endowed with additional features like a multi-language interface.

One of the scenarios using the Companion aims at raising learners' awareness of fake content in an Instagram-like social media environment (Aprin et al., 2022a). The VLC guides the learners through various examples in a chatbot-like dialogue. Additionally, learners are provided with access to other instances of the embedded images that are found through Google Reverse Image Search. The idea is that seeing the image in other contexts provides clues for judging the credibility of the presented content. The Companion in this scenario has been implemented as a Chrome browser plugin, which allows for running the scenario in a familiar web environment. Initial tests with a heterogeneous group of users indicated that the environment is perceived as supportive and usable for the classification task. Subsequently, the scenario has been tested in a secondary school classroom setting with 30 students. Preliminary findings suggest that the Companion was effective in supporting the decision about the veracity of the images shown.

The Narrative Scripts for empowering digital and self-protection skills of users through the use of computer-supported collaborative learning activities and the help of a virtual companion were presented at the Sixteenth European Conference on Technology Enhanced Learning in Italy (Hernández-Leo et al., 2021). Over 200 school workshops were conducted involving over 1.000 adolescents in private and public schools in Barcelona. Simultaneously, a light version of school workshops and the study was replicated at the University of Campo Grande (Brazil). These workshops contributed to the testing and the fine-tuning of the educational tools developed by UPF. Data collection included information derived from the implementation of Narrative Scripts, PyramidApp and EthicsApp, based on the studies Collaborative Learning for digital Environment (CSCL), Sequencing in Learning, and the evaluation of Narrative Scripts to raise teens' social media awareness.

Most workshops have been media education interventions with Narrative Scripts. The final result consisted of a social media simulated environment supported in Pixelfed. The pilots consisted of a six-module intervention with teaching and learning activities supported by Narrative Scripts and other gamifying elements. The interventions were diversified to integrate interactive features supported by AI elements, image decorations and "smart narratives" (allocation of roles/counternarratives; decorations in shared content). The first results of the data collected evaluate how the adaptive educational intervention embedded in the Narrative Scripts facilitates a suitable approach to educating adolescents about body image and stereotyping in social media. In particular, the analysis examines and compares approaches to identify the dominant body image stereotype in students' social media. Results showed that the use of xAPI (tracking user behavior in Pixelfed) combined with self-reported answers can provide a satisfactory detection of adolescents' educational needs, so as to enable automatic distribution of suitable counter-narratives (out of a collection) to students in the scripts (Lobo et al., 2022).

We also developed a visual interface that augments tweets with machine learning-based detectors of different forms of toxic content. To help the interpretation of this information created by state-of-the-art components, a web page was built showing correct and erroneous results produced by the detectors on different types of content, that will soon be tested in educational activities in high schools.

7.3. Computational components

The computational backbone of the Companion, which comprises diverse components such as content popularity predictors, user models and recommenders, is being developed, tested and outlined in Aprin et al. (2022b). Particular effort has been devoted to the development of content-based threat detectors because of their multiple roles: a) triggering specific educational activity, b) evaluating community well-being, and c) supporting recommendation and re-ranking of content.

Models to detect fake news and irony were presented at LREC 2022 (Hartl and Kruschwitz, 2022; Turban and Kruschwitz, 2022). The fake news detection system has established a new-state-of-the-art benchmark performance on the commonly used FakeNewsNet dataset. To improve performance and find the best trade-off with computational cost, the detectors were continuously updated and different architectural patterns (e.g., graph neural networks) were explored. The results were presented at several competitions about fake news detection as well as topics around hate speech (Wilkens and Ognibene, 2021a,b; Lomonaco et al., 2022a), organized within the scope of well-established annual events such as CLEF 2021, CLEF 2022 and GermEval 2021. Although submissions were very competitive, the contributions by UR resulted in winning the German cross-lingual fake news detection challenge at CLEF 2022 "CheckThat!" (Tran and Kruschwitz, 2022) and being runner-up in the fact-claiming comment identification at GermEval 2021 (Tran and Kruschwitz, 2021).

Finally, we experimented with different models of social network connectivity and user behavior. Several computational experiments showed that recommender systems have a substantial impact on the user experience on social media. For example, we simulated the impact of different recommender systems on combinations of users' satisfaction and content diversity exposure as proxies of potential components of the CWB metrics. Satisfaction is assumed as a proxy for the sustainability of the social media platform. Content diversity exposure could play an important role in countering the effects of filter bubbles (Bozdag and van den Hoven, 2015; Nikolov et al., 2015), echo chambers (Wolfowicz, 2015; Bessi, 2016; Gillani et al., 2018), and ultimately society polarization (Cinus et al., 2022). In the results shown in Figure 6. We compare three different new connection recommenders: maximize opinion diversity, random, overlapping third order neighborhood. Users were modeled by extending the model proposed in Geschke et al. (2019) with a backfiring component (Bail et al., 2018), i.e., users exposed to content presenting opinions distant from theirs changed their minds in the opposite direction. The recommender that maximizes the diversity of opinion between the pairs of users to connect showed a slower start but achieved higher exposition to more diverse content and a similar level of satisfaction to the other two RSs. In the near future, we aim at integrating a full CWB-RS with educational objectives in the simulation.

8. Discussion and conclusion

This contribution is motivated by the desire to improve the impact of social media on our society. They have indeed several positive effects (Wang et al., 2014; Chen et al., 2017): they extend our capacity to be connected with our contacts, create new useful social connections, and scale up and accelerate social interactions. Moreover, they supported various forms of activism (Gretzel, 2017; Murphy et al., 2017) and even enabled whistle-blowing in oppressive regimes (Joseph, 2012) as well as protests organization (Gladwell, 2011; Shirky, 2011). However, what can be defined as an explosion of SM has also brought several new negative social phenomena, such as digital addiction (Kuss and Griffiths, 2011; Young, 2017) and exacerbated existing ones, e.g., misinformation (wildfires) (Webb et al., 2016), which existed only on a limited scale and slow pace before.

Teenagers are a group that is particularly affected by numerous social media threats (Clarke, 2009; Ozimek et al., 2017). We propose an educational and support platform, a Companion, focused on rising teenagers' "new media literacy" (Scolari et al., 2018), "digital citizenship" (Jones and Mitchell, 2016; Xu et al., 2019), and awareness of social media threats. The Companion will allow the smooth passage from everyday life use of social media to an educational experience by interfacing with the students to support and guide their interaction with the social media environment both inside and outside the classroom. Several components of the Companion have been developed and successfully tested, as briefly described in Section 7.2.

In social media communities, as in any society, the safety and well-being of its members are determined by their own mutual interactions (Jones and Mitchell, 2016). Therefore, an important endeavor is to increase users' awareness of the consequences of their actions and acceptance of necessary boundaries, especially in such deindividuating environments (Lowry et al., 2016). The presence of a trade-off between users' rights and duties or freedom VS safety introduces ethical issues (EUC, 2019; Ienca and Vayena, 2020) (e.g., defining what is considered hate speech) that require the formulation of a comprehensive and shared view of the values of the social media community. This led to the introduction of the concept of Collective Well-Being (CWB) for Social Media communities, the shared view of the desirability of the conditions of the specific community, which would drive the definition of the educational objectives and the desired behaviors of the community members. To define the desired social media community as well as the corresponding CWB objectives, the explicit community regulations, and the educational objectives necessary to support them, we argued for a collaborative participatory design approach involving experts, educators, and community members, i.e., parents and teenagers (Sánchez-Reina et al., 2022).

In Section 4.3 a methodology is proposed to measure from online behavior the CWB of social media communities. Defining an operational measure of CWB could help deal with the cognitive and algorithmic threats that characterize social media and may hinder the effectiveness of purely educational efforts. A CWB measure could help transfer the community interests and values, as well as the educational objectives, to the recommendation algorithms that drive the users' experience by selecting and ordering feeds and connections. In the Companion this will be realized by the Collective Well-Being Recommender System (CWB-RS), which sequences educational activities and balances the content presented to the students in order to maximize the CWB (see Section 5).

From a technical point of view, the problems are multiple. Starting from the formulation of the CWB measure, the number of aspects to balance and the likely non-linear interactions between the single and the community sub-groups will require an iterative design approach. Moreover, while the state of the art for the components that detect

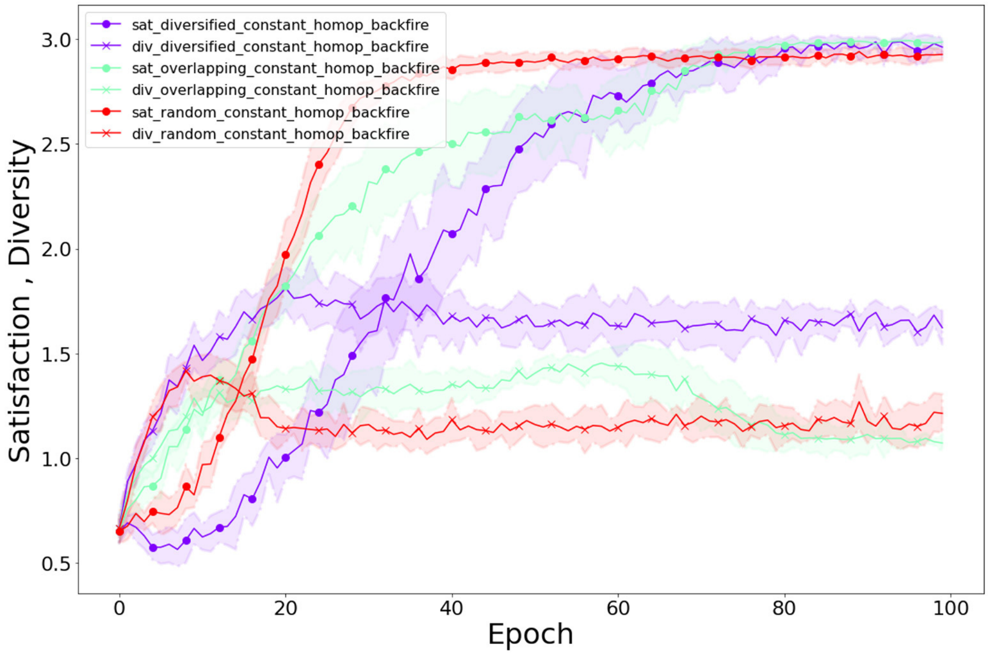

FIGURE 6
Simulated impact of different recommender systems on users' satisfaction and content diversity exposure. Satisfaction is assumed as a proxy for the sustainability of the social media platform. Content diversity exposure could play an important role in countering the effects of filter bubbles (Bozdag and van den Hoven, 2015; Nikolov et al., 2015), echo chambers (Wolfowicz, 2015; Bessi, 2016; Gillani et al., 2018), and ultimately society polarization. Mean and standard deviation over 10 runs with three different new connections *Recommenders*: maximize opinion diversity, random, overlapping neighborhood. For each strategy (colors) Satisfaction ("o") and diversity ("x") are pictured. *Overlapping*: recommend users with the highest number of common friends. *Diversified*: recommend users with the highest opinion difference. *Random*: baseline, recommend random users. *Satisfaction*: the mean distance for each user between his opinion and the ones in his feed. *Diversity*: entropy of binned opinions that populate users' feed in each time step. Highlighted areas represent standard deviation across different runs. Each social network is initialized with 100 users (nodes) and connections (edges) are created with an adaption of preferential attachments (Albert and Barabási, 2002). Differently from Albert and Barabási (2002) the nodes' probability of being connected with an incoming node is not proportionally related to nodes' degree but is related with their opinion distance. Users were modeled by extending the model proposed in Geschke et al. (2019) with a backfiring component (Bail et al., 2018), i.e., users exposed to opinions that were distant from theirs moved in the opposite direction. The recommender that maximizes diversity between the pairs of users to connect showed a slower start but achieved higher exposition to more diverse content and a similar level of satisfaction to the other two RSs.

the relevant quantities is constantly improving (e.g., sharing of hate speech in the community, see Table 3 and Sections 7.3, 5.4), the process is still noisy. The development of active evaluation methodologies, possibly involving educators as humans-in-the-loop, is a possible way forward. The CWB-RS must face additional complexities to evaluate the longer-term impact of its recommendations for the achievement of educational objectives and the future CWB of the community as well as balancing the level of engagement necessary for the educational and social functions (e.g., finding out that a friend needs online support) while avoiding digital addiction. We discussed in Section 5 that these issue may require combining an intelligent tutor system with recommender systems built using the hierarchical reinforcement learning framework.

Our contribution in this paper, in particular our experimental studies, are specifically designed for relatively small communities that can tailor the approach to their own specific needs. One may be tempted to think about scaling up the whole approach and integrating the educator in the loop, the CWB and CWBRS on the global social media platforms. This would prohibitively escalate the moderation costs that are already very demanding (Steiger et al., 2021) and would have to take also the educational aspect into account, which requires wider expertise and user-specific policies. From an ethical point of view, the undertaking would be enormous. While privacy, censorship, freedom of speech, misinformation campaigns and hate speech are strongly involved ethical problems, they are by now very common in the discussion about social media (Webb et al., 2016), especially after Twitter permanently banned Donald Trump (Courty, 2021). However, the formulation of a CWB for social media requires not only formulating a metric that balances many different demands but it justifying the worldwide and cross-cultural adoption of a value set that supports such a metric applied to the social and dynamic version of the *WWW*.

Currently, the international community is undertaking a substantial effort in understanding and regulate the ethical implication of AI systems (EUC, 2019; IEEE, 2019; OCED, 2019; UNESCO, 2020). Unmistakably there are mixed ethical and technical issues that go beyond those currently faced (EUC, 2019; IEEE, 2019; OCED, 2019; UNESCO, 2020).[5] For example, trying to optimize the CWB may induce a further increase in social media complexity. This may reduce even more our control over social dynamics (Floridi, 2014) and backfire with even more threatening, addictive, and unhealthy dystopian situations.

While it is crucial that the international community continues its effort and targets social media (Gorwa, 2019), we highlighted that aiming to improve the CWB of SM local communities implies first and foremost aiming to educate local communities themselves, as the CWB depends on users' attitude, interactions and relationships (Jones and Mitchell, 2016). Education is the best way we know to improve human behavior. Indeed, if the methodology is successfully applied on a sufficient scale improving the members' new media literacy and digital citizenship, it may improve the general impact of social media on our society. Focusing on more controlled communities, e.g., schools, with a very limited scale for the social media domain reduces the ethical burden on the design side as well as the technical demands for accuracy and reliability through the integration of a mediator role for educators and parents, through a "Human in the Loop" paradigm. This approach also allows focusing on the critical educational aspect. The creation of educationally managed social media communities allows supported learning experiences and a full range of new experiments (Amarasinghe et al., 2021; Fulantelli et al., 2021; Hernández-Leo et al., 2021; Malzahn et al., 2021).

Differently from previous other interventions with a similar aim, this paradigm enabled by the educational virtual Companion for social media has indeed the potential to provide an educational experience on a scale comparable to that of the social media platforms. Indeed, it will be challenging to define an educational path that covers most of the numerous points of interest in digital citizenship (Jones and Mitchell, 2016; Xu et al., 2019). Also, the integration of this educational experience in student life is challenging, especially regarding the experience outside the classroom, where the non-educational global platforms will compete for student time and attention. Still, we believe that the combined technological and educational strategy implemented by the Companion has a good chances to be effective in containing many of the current social media threats.

Finally, this approach is the perfect means to bootstrap and test the concept of CWB-RS systems, verify their feasibility, stability and robustness, and create suitable datasets. The data collected from this initiative may not only be useful for replicating and extending this type of educational approach, but it could also be a first step to provide evidence that social media's impact on society can be improved by taking the community needs more into account in their design. A characteristic feature of social media is that they are crucially the result of a community activity, which both consumes and produces their content. It is daunting that platforms' objectives are so detached from those of the community. Therefore, we hope that our results can support the process of introducing new evidence-based regulations both for the platforms and their algorithms, beginning with requesting the platforms to release their data for scientific research and enable large-scale studies, which have been curbed after the limits they recently set following Cambridge Analytica and other scandals (Hemsley, 2019).

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

DO conceived of the presented ideas and concepts and contributed to the design of the educational platform architecture. RW and GD supported, revised the design, and developed the content machine learning aspects. UK proposed, described the integration of behavioral economics methodologies for education, and supervised the information retrieval aspect of the contribution. FL and SB helped with the aspects of network dynamics and dataset collection. DH-L, JS-R, ET, and RL contributed with the proposal of playful educational methodology, the cooperative social media design, and narrative scripts. UH, NM, FA, VS, JB, and SE overviewed the conceptual development. DT contributed to the design of the educational platform architecture, DT and LS contributed with expertise on digital addiction and digital literacy intervention design. All authors contributed to the as article and approved the submitted version.

Funding

This work was developed in the framework of the project COURAGE—A social media companion safeguarding and educating students (Nos. 95563, 95564, 95565, 95566, 95567, and

[5] See https://standards.ieee.org/standard/7010-2020.html IEEE 7010-2020 - IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being.

9B145) and funded by the Volkswagen Foundation in the topic Artificial Intelligence and the Society of the Future.

Acknowledgments

DH-L (Serra Húnter) acknowledges the support by ICREA under the ICREA Academia programme.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdollahpouri, H., and Burke, R. (2019). Multi-stakeholder recommendation and its connection to multi-sided fairness. *CoRR*, abs/1907.13158. doi: 10.48550/arXiv.1907.13158

Aboujaoude, E., Koran, L. M., Gamel, N., Large, M. D., and Serpe, R. T. (2006). Potential markers for problematic internet use: a telephone survey of 2,513 adults. *CNS Spectr*. 11, 750–755. doi: 10.1017/S1092852900014875

Adomavicius, G., Bockstedt, J. C., Curley, S. P., and Zhang, J. (2013). Do recommender systems manipulate consumer preferences? a study of anchoring effects. *Inf. Syst. Res*. 24, 956–975. doi: 10.1287/isre.2013.0497

Ahmad, Z., Jindal, R., Ekbal, A., and Bhattachharyya, P. (2020). Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding. *Expert. Syst. Appl*. 139, 112851. doi: 10.1016/j.eswa.2019.112851

Ahn, D., and Shin, D.-H. (2013). Is the social use of media for seeking connectedness or for avoiding social isolation? mechanisms underlying media use and subjective well-being. *Comput. Hum. Behav*. 29, 2453–2462. doi: 10.1016/j.chb.2012.12.022

Albert, R., and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys*. 74, 47. doi: 10.1103/RevModPhys.74.47

Ali, R., Jiang, N., Phalp, K., Muir, S., and McAlaney, J. (2015). "The emerging requirement for digital addiction labels," in *International Working Conference on Requirements Engineering: Foundation for Software Quality* (Essen: Springer), 198–213.

Allcott, H., Braghieri, L., Eichmeyer, S., and Gentzkow, M. (2020). The welfare effects of social media. *Am. Econ. Rev*. 110, 629–676. doi: 10.1257/aer.20190658

Allcott, H., and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *J. Econ. Perspect*. 31, 211–236. doi: 10.1257/jep.31.2.211

Almourad, B. M., McAlaney, J., Skinner, T., Pleva, M., and Ali, R. (2020). Defining digital addiction: Key features from the literature. *Psihologija* 53, 17–17. doi: 10.2298/PSI191029017A

Alrobai, A., McAlaney, J., Phalp, K., and Ali, R. (2016). "Online peer groups as a persuasive tool to combat digital addiction," in *International Conference on Persuasive Technology* (Salzburg: Springer), 288–300.

Alutaybi, A., McAlaney, J., Arden-Close, E., Stefanidis, A., Phalp, K., and Ali, R. (2019). "Fear of missing out (fomo) as really lived: five classifications and one ecology," in *2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC)* (Beijing), 1–6.

Amarasinghe, I., Hernández-Leo, D., and Jonsson, A. (2019). Data-informed design parameters for adaptive collaborative scripting in across-spaces learning situations. *User Model. User-Adap*. 29, 869–892. doi: 10.1007/s11257-019-09233-8

Amarasinghe, I., Hernández-Leo, D., Theophilou, E., Sanchez-Reina, R., and Lobo, R. (2021). "Learning gains in pyramid computer-supported collaboration scripts: factors and implications for design," in *Proceedings of CollabTech Conference*.

Anderson, S. P., and McLaren, J. (2012). Media mergers and media bias with rational consumers. *J. Eur. Econ. Assoc*. 10, 831–859. doi: 10.1111/j.1542-4774.2012.01069.x

Aprin, F., Chounta, I.-A., and Hoppe, H. U. (2022a). "'See the image in different contexts': using reverse image search to support the identification of fake news in instagram-like social media," in *International Conference on Intelligent Tutoring Systems* (Cham: Springer LNCS), 264–275.

Aprin, F., Malzahn, N., Lomonaco, F., Donabauer, G., Ognibene, D., Kruschwitz, U., et al. (2022b). "The courage virtual learning companion: learning design and technical architecture," in *Proceedings of the 4th International Conference on Higher Education Learning Methodologies and Technologies Online (HELMeTO2022)-Book of Abstracts* (Palermo), 90–92.

Aral, S., Muchnik, L., and Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl. Acad. Sci. U.S.A* 106, 21544–21549. doi: 10.1073/pnas.0908800106

Arroyo, I., Ferguson, K., Johns, J., Dragon, T., Meheranian, H., Fisher, D., et al. (2007). "Repairing disengagement with non-invasive interventions," in *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work, Vol. 2007* (Los Angeles, CA), 195–202.

Atanasova, P., Simonsen, J. G., Lioma, C., and Augenstein, I. (2020). "Generating fact checking explanations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online: Association for Computational Linguistics), 7352–7364.

Augenstein, I., Rocktäschel, T., Vlachos, A., and Bontcheva, K. (2016). "Stance detection with bidirectional conditional encoding," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Austin, TX: Association for Computational Linguistics), 876–885.

Baeza-Yates, R., and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Boston, MA: Addison Wesley.

Baeza-Yates, R., and Ribeiro-Neto, B. (Eds.). (2010). *Modern Information Retrieval, 2nd Edn*. Boston, MA: Addison-Wesley.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., et al. (2018). Exposure to opposing views on social media can increase political polarization. *Proc. Natl. Acad. Sci. U.S.A*. 115, 9216–9221. doi: 10.1073/pnas.1804840115

Banker, S., and Khetani, S. (2019). Algorithm overdependence: how the use of algorithmic recommendation systems can increase risks to consumer well-being. *J. Public Policy Market*. 38, 500–515. doi: 10.1177/0743915619858057

Barak, A. (2005). Sexual harassment on the internet. *Soc. Sci. Comput. Rev*. 23, 77–92. doi: 10.1177/0894439304271540

Barto, A. G., and Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dyn. Syst*. 13, 41–77. doi: 10.1023/A:1022140919877

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., et al. (2019). "Semeval-2019 task 5: multilingual detection of hate speech against immigrants and women in Twitter," in *Proceedings of the 13th International Workshop on Semantic Evaluation* (Minneapolis, MN), 54–63.

Baziotis, C., Athanasiou, N., Chronopoulou, A., Kolovou, A., Paraskevopoulos, G., Ellinas, N., et al. (2018). Ntua-slp at semeval-2018 task 1: predicting affective content in tweets with deep attentive rnns and transfer learning. *arXiv preprint* arXiv:1804.06658. doi: 10.18653/v1/S18-1037

Becker, J., Brackbill, D., and Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proc. Natl. Acad. Sci. U.S.A.* 114, E5070-E5076. doi: 10.1073/pnas.1615978114

Beed, P. L., Hawkins, E. M., and Roller, C. M. (1991). Moving learners toward independence: the power of scaffolded instruction. *Read. Teach.* 44, 648–655.

Bessi, A. (2016). Personality traits and echo chambers on facebook. *Comput. Hum. Behav*. 65, 319–324. doi: 10.1016/j.chb.2016.08.016

Bhargava, R., Chung, A., Gaikwad, N. S., Hope, A., Jen, D., Rubinovitz, J., et al. (2019). "Gobo: asystem for exploring user control of invisible algorithms in social media," in *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing* (Austin, TX), 151–155.

Bilinski, P., and Bremond, F. (2016). "Human violence recognition and detection in surveillance videos," in *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (Colorado Springs, CO: IEEE), 30–36.

Bliuc, A.-M., Faulkner, N., Jakubowicz, A., and McGarty, C. (2018). Online networks of racial hate: a systematic review of 10 years of research on cyber-racism. *Comput. Hum. Behav*. 87, 75–86. doi: 10.1016/j.chb.2018.05.026

Borisyuk, F., Kenthapadi, K., Stein, D., and Zhao, B. (2016). "Casmos: a framework for learning candidate selection models over structured queries and documents," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA), 441–450.

Boudreau, K. J., and Lakhani, K. R. (2013). Using the crowd as an innovation partner. *HBR* 91, 60–69.

Bozdag, E., and van den Hoven, J. (2015). Breaking the filter bubble: democracy and design. *Ethics Inf. Technol.* 17, 249–265. doi: 10.1007/s10676-015-9380-y

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., and Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proc. Natl. Acad. Sci. U.S.A.* 114, 7313–7318. doi: 10.1073/pnas.1618923114

Brand, M., Laier, C., and Young, K. S. (2014). Internet addiction: coping styles, expectancies, and treatment implications. *Front. Psychol* 5, 1256. doi: 10.3389/fpsyg.2014.01256

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., and Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *J. Appl. Res. Mem. Cogn.* 8, 108–117. doi: 10.1037/h0101832

Buhrmester, M. D., Talaifar, S., and Gosling, S. D. (2018). An evaluation of amazon's mechanical turk, its rapid rise, and its effective use. *Perspect. Psychol. Sci.* 13, 149–154. doi: 10.1177/1745691617706516

Burrow, A. L., and Rainone, N. (2017). How many likes did i get?: purpose moderates links between positive social media feedback and self-esteem. *J. Exp. Soc. Psychol.* 69, 232–236. doi: 10.1016/j.jesp.2016.09.005

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., and Floridi, L. (2018). Artificial intelligence and the 'good society': the us, eu, and uk approach. *Sci. Eng. Ethics* 24, 505–528. doi: 10.1007/s11948-017-9901-7

Cer, D., Yang, Y., Kong, S.-Y., Hua, N., Limtiaco, N., John, R. S., et al. (2018). Universal sentence encoder. *arXiv preprint* arXiv:1803.11175. doi: 10.18653/v1/D18-2029

Chan, J., Ghose, A., and Seamans, R. (2016). The internet and racial hate crime: offline spillovers from online access. *MIS Q.* 40, 381–403. doi: 10.25300/MISQ/2016/40.2.05

Chaouachi, M., and Frasson, C. (2012). "Mental workload, engagement and emotions: an exploratory study for intelligent tutoring systems," in *International Conference on Intelligent Tutoring Systems* (Chania: Springer), 65–71.

Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., and He, X. (2020). Bias and debias in recommender system: a survey and future directions. *arXiv preprint* arXiv:2010.03240. doi: 10.48550/arXiv.2010.03240

Chen, L., Gong, T., Kosinski, M., Stillwell, D., and Davidson, R. L. (2017). Building a profile of subjective well-being for social media users. *PLoS ONE* 12, e0187278. doi: 10.1371/journal.pone.0187278

Chen, R., Hua, Q., Chang, Y.-S., Wang, B., Zhang, L., and Kong, X. (2018). A survey of collaborative filtering-based recommender systems: from traditional methods to hybrid methods based on social networks. *IEEE Access* 6, 64301–64320. doi: 10.1109/ACCESS.2018.2877208

Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., and Leskovec, J. (2014). "Can cascades be predicted?" in *WWW '14: Proceedings of the 23rd International Conference on World Wide Web* (Seoul), 925–936.

Chris Hale, W. (2012). Extremism on the world wide web: a research review. *Crim. Justice Stud*. 25, 343–356. doi: 10.1080/1478601X.2012.704723

Christakis, N. A., and Fowler, J. H. (2008). The collective dynamics of smoking in a large social network. *N. Engl. J. Med*. 358, 2249–2258. doi: 10.1056/NEJMsa0706154

Cinus, F., Minici, M., Monti, C., and Bonchi, F. (2022). "The effect of people recommenders on echo chambers and polarization," in *Proceedings of the International AAAI Conference on Web and Social Media, Vol. 16*, eds C. Budak, M. Cha. and D. Quercia (Atlanta, GA: AAAI), 90–101. doi: 10.1609/icwsm.v16i1.19275

Clarke, B. (2009). Early adolescents' use of social networking sites to maintain friendship and explore identity: implications for policy. *Policy Internet* 1, 55–89. doi: 10.2202/1944-2866.1018

Cohen, S., and Wills, T. A. (1985). Stress, social support, and the buffering hypothesis. *Psychol. Bull.* 98, 310. doi: 10.1037/0033-2909.98.2.310

Costanza, R., Kubiszewski, I., Giovannini, E., Lovins, H., McGlade, J., Pickett, K. E., et al. (2014). Development: time to leave gdp behind. *Nat. News* 505, 283. doi: 10.1038/505283a

Costello, M., Hawdon, J., Bernatzky, C., and Mendes, K. (2019). Social group identity and perceptions of online hate*. *Sociol. Inq.* 89, 427–452. doi: 10.1111/soin.12274

Courty, A. (2021). *Despite Being Permanently Banned, Trump's Prolific Twitter Record Lives On*. Available online at: https://theconversation.com/despite-being-permanently-bannedtrumps-prolific-twitter-record-lives-on-152969

Covington, P., Adams, J., and Sargin, E. (2016). "Deep neural networks for youtube recommendations," in *Proceedings of the 10th ACM Conference on Recommender Systems* (Boston, MA), 191–198.

Cowie, H. (2013). Cyberbullying and its impact on young people's emotional health and well-being. *Psychiatrist* 37, 167–170. doi: 10.1192/pb.bp.112.040840

Dale, R. (2017). Nlp in a post-truth world. *Nat. Lang. Eng.* 23, 319–324. doi: 10.1017/S1351324917000018

Das, S., and Lavoie, A. (2014). "The effects of feedback on human behavior in social media: an inverse reinforcement learning model," in *AAMAS '14: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems* (Paris), 653–660.

Davies, G., Neudecker, C., Ouellet, M., Bouchard, M., and Ducol, B. (2016). Toward a framework understanding of online programs for countering violent extremism. *J. Deradicalizat*. 6, 51–86.

de Cock Buning, M. (2018). *A Multi-Dimensional Approach to Disinformation: Report of the Independent High Level Group on Fake News and Online Disinformation*. Publications Office of the European Union.

De Gloria, A., Bellotti, F., and Berta, R. (2014). Serious games for education and training. *Int. J. Serious Games* 1, 11. doi: 10.17083/ijsg.v1i1.11

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., et al. (2016). The spreading of misinformation online. *Proc. Natl. Acad. Sci. U.S.A.* 113, 554–559. doi: 10.1073/pnas.1517441113

Del Vicario, M., Scala, A., Caldarelli, G., Stanley, H. E., and Quattrociocchi, W. (2017). Modeling confirmation bias and polarization. *Sci. Rep*. 7, 40391. doi: 10.1038/srep40391

DeVito, M. A., Birnholtz, J., Hancock, J. T., French, M., and Liu, S. (2018). "How people form folk theories of social media feeds and what it means for how we study self-presentation," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal, QC), 1–12.

Diener, E., Lusk, R., DeFour, D., and Flax, R. (1980). Deindividuation: effects of group size, density, number of observers, and group member similarity on self-consciousness and disinhibited behavior. *J. Pers. Soc. Psychol.* 39, 449. doi: 10.1037/0022-3514.39.3.449

Diener, E., and Seligman, M. E. (2006). Measure for measure: the case for a national well-being index. *Sci. Spirit* 17, 36–38. doi: 10.3200/SSPT.17.2.36-37

Dorça, F. A., Lima, L. V., Fernandes, M. A., and Lopes, C. R. (2013). Comparing strategies for modeling students learning styles through reinforcement learning in adaptive and intelligent educational systems: an experimental analysis. *Expert Syst. Appl*. 40, 2092–2101. doi: 10.1016/j.eswa.2012.10.014

Drachsler, H., Hummel, H., and Koper, R. (2009). Identifying the goal, user model and conditions of recommender systems for formal and informal learning. *J. Digit. Infm*. 10, 4–24.

Dulac-Arnold, G., Mankowitz, D., and Hester, T. (2019). Challenges of real-world reinforcement learning. *arXiv preprint* arXiv:1904.12901. doi: 10.48550/arXiv.1904.12901

Dunn, H. L. (1959). High-level wellness for man and society. *Am. J. Public Health Nations Health* 49, 786–792. doi: 10.2105/AJPH.49.6.786

Eagle, M., and Barnes, T. (2014). "Modeling student dropout in tutoring systems," in *International Conference on Intelligent Tutoring Systems* (Honolulu, HI: Springer), 676–678.

Eirinaki, M., Gao, J., Varlamis, I., and Tserpes, K. (2018). Recommender systems for large-scale social networks: a review of challenges and solutions. *Future Generat. Comput. Syst.* 78, 413–418. doi: 10.1016/j.future.2017.09.015

Eksombatchai, C., Jindal, P., Liu, J. Z., Liu, Y., Sharma, R., Sugnet, C., et al. (2018). "Pixie: a system for recommending 3+ billion items to 200+ million users in real-time," in *Proceedings of the 2018 World Wide Web Conference* (Lyon), 1775–1784.

Elghomary, K., and Bouzidi, D. (2019). "Dynamic peer recommendation system based on trust model for sustainable social tutoring in moocs," in *2019 1st International Conference on Smart Systems and Data Science (ICSSD)* (Rabat: IEEE), 1–9.

Ellison, N. B., Vitak, J., Gray, R., and Lampe, C. (2014). Cultivating social resources on social network sites: Facebook relationship maintenance behaviors and their role in social capital processes. *J. Comput. Mediated Commun.* 19, 855–870. doi: 10.1111/jcc4.12078

Engel, L., Chudyk, A., Ashe, M., McKay, H., Whitehurst, D., and Bryan, S. (2016). Older adults' quality of life-exploring the role of the built environment and social cohesion in community-dwelling seniors on low income. *Soc. Sci. Med.* 164, 1–11. doi: 10.1016/j.socscimed.2016.07.008

EUC (2019). *Ethics guidelines for trustworthy ai*. Technical report, European Commission.

Fedorov, A. (2015). *Media Literacy Education*. ICO: Information for all.

Floridi, L. (2014). *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford: Oxford University Press UK.

Forgeard, M. J., Jayawickreme, E., Kern, M. L., and Seligman, M. E. (2011). Doing the right thing: measuring wellbeing for public policy. *Int. J. Wellbeing* 1, 15. doi: 10.5502/ijw.v1i1.15

Fredrickson, B. L. (2004). The broaden-and-build theory of positive emotions. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 359, 1367–1377. doi: 10.1098/rstb.2004.1512

Fuhr, N., Giachanou, A., Grefenstette, G., Gurevych, I., Hanselowski, A., Jarvelin, K., et al. (2018). "An information nutritional label for online documents," in *ACM SIGIR Forum, Vol. 51* (New York, NY: ACM), 46–66.

Fulantelli, G., Scifo, L., and Taibi, D. (2021). "Training school activities to promote a conscious use of social media and human development according to the ecological systems theory," in *Proceedings of the 13th International Conference on Computer Supported Education*.

Fulantelli, G., Taibi, D., Scifo, L., Schwarze, V., and Eimler, S. C. (2022). Cyberbullying and cyberhate as two interlinked instances of cyber-aggression in adolescence: a systematic review. *Front. Psychol.* 13, 909299. doi: 10.3389/fpsyg.2022.909299

Gao, J., Zheng, P., Jia, Y., Chen, H., Mao, Y., Chen, S., et al. (2020). Mental health problems and social media exposure during COVID-19 outbreak. *PLoS ONE* 15, e0231924. doi: 10.1371/journal.pone.0231924

Garimella, K., Gionis, A., Parotsidis, N., and Tatti, N. (2017). Balancing information exposure in social networks. *arXiv preprint* arXiv:1709.01491. doi: 10.48550/arXiv.1709.01491

Gerson, J. (2018). *Social media use and subjective well-being: an investigation of individual differences in personality, social comparison and Facebook behaviour* (Ph.D. thesis). City, University of London.

Gerstenfeld, P. B., Grant, D. R., and Chiang, C.-P. (2003). Hate online: a content analysis of extremist internet sites. *Anal. Soc. Issues Public Policy* 3, 29–44. doi: 10.1111/j.1530-2415.2003.00013.x

Geschke, D., Lorenz, J., and Holtz, P. (2019). The triple-filter bubble: using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *Br. J. Soc. Psychol.* 58, 129–149. doi: 10.1111/bjso.12286

Gillani, N., Yuan, A., Saveski, M., Vosoughi, S., and Roy, D. (2018). "Me, my echo chamber, and i: introspection on social media polarization," in *Proceedings of the 2018 World Wide Web Conference* (Lyon), 823–831.

Gladwell, M. (2011). From innovation to revolution-do social media made protests possible: an absence of evidence. *Foreign Aff.* 90, 153.

Gorwa, R. (2019). What is platform governance? *Inf. Commun. Soc.* 22, 854–871. doi: 10.1080/1369118X.2019.1573914

Greer, J., and Mark, M. (2016). Evaluation methods for intelligent tutoring systems revisited. *Int. J. Artif. Intell. Educ.* 26, 387–392. doi: 10.1007/s40593-015-0043-2

Gretzel, U. (2017). Social media activism in tourism. *J. Hospit. Tourism* 15, 1–14. doi: 10.4324/9781315659657-38

Grieve, R., Indian, M., Witteveen, K., Tolan, G. A., and Marrington, J. (2013). Face-to-face or facebook: can social connectedness be derived online? *Comput. Hum. Behav.* 29, 604–609. doi: 10.1016/j.chb.2012.11.017

Griffith, S., Subramanian, K., Scholz, J., Isbell, C. L., and Thomaz, A. L. (2013). Policy shaping: integrating human feedback with reinforcement learning. *Adv. Neural Inf. Process Syst.* 26, 2625–2633.

Grigg, D. W. (2010). Cyber-aggression: definition and concept of cyberbullying. *J. Psychol. Counsell. Sch.* 20, 143–156. doi: 10.1375/ajgc.20.2.143

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., and Asokan, N. (2018). "All you need is: evading hate speech detection," in *PWAIS-ACM'18* (New York, NY: ACM), 2–12.

Gunawardena, C. N. (1995). Social presence theory and implications for interaction and collaborative learning in computer conferences. *Int. J. Educ. Telecommun.* 1, 147–166.

Guo, F., Blundell, C., Wallach, H., and Heller, K. (2015). "The bayesian echo chamber: modeling social influence via linguistic accommodation," in *AI&S* (San Diego, CA), 315–323.

Guo, X., Zhu, B., Polanía, L. F., Boncelet, C., and Barner, K. E. (2018). "Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO), 635–639.

Gupta, A., Agrawal, D., Chauhan, H., Dolz, J., and Pedersoli, M. (2018). "An attention model for group-level emotion recognition," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO), 611–615.

Han, J., Wu, S., and Liu, X. (2019). "jhan014 at semeval-2019 task 6: identifying and categorizing offensive language in social media," in *Proceedings of the 13th International Workshop on Semantic Evaluation* (Minneapolis, MN), 652–656.

Hartl, P., and Kruschwitz, U. (2022). "Applying automatic text summarization for fake news detection," in *Proceedings of the Language Resources and Evaluation Conference* (Marseille: European Language Resources Association), 2702–2713.

Harutyunyan, A., Devlin, S., Vrancx, P., and Nowé, A. (2015). "Expressing arbitrary reward functions as potential-based advice," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 29* (Austin, TX).

Hawking, D. (2010). "Enterprise search," in *Modern Information Retrieval, 2nd Edn*, eds R. Baeza-Yates and B. Ribeiro-Neto (Harlow: Addison-Wesley), 645–686.

He, X., Rekatsinas, T., Foulds, J., Getoor, L., and Liu, Y. (2015). "Hawkestopic: a joint model for network inference and topic modeling from text-based cascades," in *ICML* (Lille), 871–880.

Heimbach, I., Gottschlich, J., and Hinz, O. (2015). The value of user's facebook profile data for product recommendation generation. *Electronic Markets* 25, 125–138. doi: 10.1007/s12525-015-0187-9

Helliwell, J. F. (2003). How's life? combining individual and national variables to explain subjective well-being. *Econ. Model.* 20, 331–360. doi: 10.1016/S0264-9993(02)00057-3

Hemsley, J. (2019). *Social Media Giants Are Restricting Research Vital to Journalism*. Columbia Journalism Review. Available online at: https://www.cjr.org/tow_center/facebook-twitter-api-restrictions.php

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. (2018). "Deep reinforcement learning that matters," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32* (New Orleans, LA).

Hernández-Leo, D., Theophilou, E., Lobo, R., Sánchez-Reina, R., and Ognibene, D. (2021). "Narrative scripts embedded in social media towards empowering digital and self-protection skills," in *Proceedings of the European Conference on Technology-Enhanced Learning* (Bozen-Bolzano: Springer), 394–398.

Hertwig, R., and Grüne-Yanoff, T. (2017). Nudging and boosting: steering or empowering good decisions. *Perspect. Psychol. Sci.* 12, 973–986. doi: 10.1177/1745691617702496

Hessel, J., and Lee, L. (2019). "Something's brewing! early prediction of controversy-causing posts from discussion features," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1* (Minneapolis, MN).

Hoffmann, J., and Kruschwitz, U. (2020). "Ur nlp@ haspeede 2 at evalita 2020: Towards robust hate speech detection with contextual embeddings," in *EVALITA*.

Hong, R., He, C., Ge, Y., Wang, M., and Wu, X. (2017). User vitality ranking and prediction in social networking services: a dynamic network perspective. *IEEE Trans. Knowl. Data Eng.* 29, 1343–1356. doi: 10.1109/TKDE.2017.2672749

Hosseini, H., Kannan, S., Zhang, B., and Poovendran, R. (2017). Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint* arXiv:1702.08138. doi: 10.48550/arXiv.1702.08138

Hron, J., Krauth, K., Jordan, M. I., and Kilbertus, N. (2020). Exploration in two-stage recommender systems. *arXiv preprint* arXiv:2009.08956. doi: 10.48550/arXiv.2009.08956

Huang, Y., and Zhu, Q. (2019). "Deceptive reinforcement learning under adversarial manipulations on cost signals," in *International Conference on Decision and Game Theory for Security* (Stockholm: Springer), 217–237.

IEEE (2019). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems*. Technical report, IEEE.

Ienca, M., and Vayena, E. (2020). *Ai ethics guidelines: European and global perspectives*. Technical report, Ad hoc committee on artificial intelligence (CAHAI).

Iglesias, A., Martínez, P., Aler, R., and Fernández, F. (2009). Reinforcement learning of pedagogical policies in adaptive and intelligent educational systems. *Knowl. Based Syst*. 22, 266–270. doi: 10.1016/j.knosys.2009.01.007

Indurthi, V., Syed, B., Shrivastava, M., Chakravartula, N., Gupta, M., and Varma, V. (2019). "Fermi at semeval-2019 task 5: using sentence embeddings to identify hate speech against immigrants and women in Twitter," in *Proceedings of the 13th International Workshop on Semantic Evaluation* (Minneapolis, MN), 70–74.

Isaak, J., and Hanna, M. J. (2018). User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer* 51, 56–59. doi: 10.1109/MC.2018.3191268

Jiang, L. C., Bazarova, N. N., and Hancock, J. T. (2011). The disclosure-intimacy link in computer-mediated communication: an attributional extension of the hyperpersonal model. *Hum. Commun. Res*. 37, 58–77. doi: 10.1111/j.1468-2958.2010.01393.x

Jobanputra, M. (2019). "Unsupervised question answering for fact-checking," in *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), Vol. 2019* (Hong Kong Association for Computational Linguistics), 52–56.

Johnson, N. F., Zheng, M., Vorobyeva, Y., Gabriel, A., Qi, H., Velásquez, N., et al. (2016). New online ecology of adversarial aggregates: isis and beyond. *Science* 352, 1459–1463. doi: 10.1126/science.aaf0675

Jones, L. M., and Mitchell, K. J. (2016). Defining and measuring youth digital citizenship. *New Media Soc*. 18, 2063–2079. doi: 10.1177/1461444815577797

Joseph, S. (2012). Social media, political change, and human rights. *Boston Coll. Int. Compar. Law Rev*. 35, 145. doi: 10.2139/ssrn.1856880

Keyes, C. L. (2012). *Mental Well-Being: International Contributions to the Study of Positive Mental Health*. Dordrecht: Springer Science & Business Media.

Khosravi, P., Rezvani, A., and Wiewiora, A. (2016). The impact of technology on older adults' social isolation. *Comput. Hum. Behav*. 63, 594–603. doi: 10.1016/j.chb.2016.05.092

Khwaja, M., Ferrer, M., Iglesias, J., Faisal, A., and Matic, A. (2019). "Aligning daily activities with personality: towards a recommender system for improving wellbeing," in *RecSys '19: Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen), 368–372.

King, T. C., Aggarwal, N., Taddeo, M., and Floridi, L. (2020). Artificial intelligence crime: an interdisciplinary analysis of foreseeable threats and solutions. *Sci. Eng. Ethics* 26, 89–120. doi: 10.1007/s11948-018-00081-0

Kittur, A., Chi, E. H., and Suh, B. (2008). "Crowdsourcing user studies with mechanical turk," in *CHI '08: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence), 453–456.

Klein, C. (2013). Social capital or social cohesion: what matters for subjective well-being? *Soc. Indic. Res*. 110, 891–911. doi: 10.1007/s11205-011-9963-x

Knobloch-Westerwick, S., and Kleinman, S. B. (2012). Preelection selective exposure: confirmation bias versus informational utility. *Commun. Res*. 39, 170–193. doi: 10.1177/0093650211400597

Kopeinik, S., Lex, E., Seitlinger, P., Albert, D., and Ley, T. (2017). "Supporting collaborative learning with tag recommendations: a real-world study in an inquiry-based classroom project," in *Proceedings of the Seventh International Learning Analytics and Knowledge Conference* (Vancouver, BC), 409–418.

Kozyreva, A., Lewandowsky, S., and Hertwig, R. (2020). Citizens versus the internet: confronting digital challenges with cognitive tools. *Psychol. Sci. Public Interest* 21, 103–156. doi: 10.1177/1529100620946707

Kramer, A. D., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl. Acad. Sci. U.S.A*. 111, 8788–8790. doi: 10.1073/pnas.1320040111

Kross, E., Verduyn, P., Demiralp, E., Park, J., Lee, D. S., Lin, N., et al. (2013). Facebook use predicts declines in subjective well-being in young adults. *PLoS ONE* 8, e69841. doi: 10.1371/journal.pone.0069841

Kruschwitz, U., and Hull, C. (2017). Searching the Enterprise. *Foundat. Trends Inf. Retrieval* 11, 1–142. doi: 10.1561/9781680833058

Kunaver, M., and Požrl, T. (2017). Diversity in recommender systems-a survey. *Knowl. Based Syst*. 123, 154–162. doi: 10.1016/j.knosys.2017.02.009

Kurth-Nelson, Z., and Redish, A. D. (2009). Temporal-difference reinforcement learning with distributed representations. *PLoS ONE* 4, e7362. doi: 10.1371/annotation/4a24a185-3eff-454f-9061-af0bf22c83eb

Kuss, D., Rooij, A. V., Shorter, G., Griffiths, M., and de Mheen, D. V. (2013). Internet addiction in adolescents: prevalence and risk factors. *Comput. Hum. Behav*. 29, 1987–1996. doi: 10.1016/j.chb.2013.04.002

Kuss, D. J., and Griffiths, M. D. (2011). Online social networking and addiction–a review of the psychological literature. *Int. J. Environ. Res. Public Health* 8, 3528–3552. doi: 10.3390/ijerph8093528

Lavenia, G. (2012). *Internet e le sue dipendenze*. Dal coinvolgimento alla psicopatologia. Franco Angeli.

Lee, R. S., Hoppenbrouwers, S., and Franken, I. (2019). A systematic meta-review of impulsivity and compulsivity in addictive behaviors. *Neuropsychol. Rev*. 29, 14–26. doi: 10.1007/s11065-019-09402-x

Leskovec, J., Huttenlocher, D., and Kleinberg, J. (2010). "Signed networks in social media," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, GA), 1361–1370.

Lim, T. V., Cardinal, R. N., Savulich, G., Jones, P. S., Moustafa, A. A., Robbins, T., et al. (2019). Impairments in reinforcement learning do not explain enhanced habit formation in cocaine use disorder. *Psychopharmacology* 236, 2359–2371. doi: 10.1007/s00213-019-05330-z

Lindström, B., Bellander, M., Chang, A., Tobler, P. N., and Amodio, D. M. (2019). A computational reinforcement learning account of social media engagement. *Nat. Commun*. 12, 1311. doi: 10.31234/osf.io/78mh5

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures Hum. Lang. Technol*. 5, 1–167. doi: 10.1007/978-3-031-02145-9

Liu, F., Tang, R., Li, X., Ye, Y., Chen, H., Guo, H., et al. (2018). Deep reinforcement learning based recommendation with explicit user-item interactions modeling. *arXiv [Preprint]*. arXiv: 1810.12027. Available online at: https://arxiv.org/pdf/1810.12027.pdf

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., and Song, L. (2017). "Sphereface: deep hypersphere embedding for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 212–220.

Liu, Y., and Lapata, M. (2019). Text summarization with pretrained encoders. *arXiv preprint* arXiv:1908.08345. doi: 10.18653/v1/D19-1387

Lobo, R., Theophilou, E., Sánchez-Reina, R., and Hernández-Leo, D. (2022). "Evaluating an adaptive intervention in collaboration scripts deconstructing body image narratives in a social media educational platform," in 27th *International Conference, CollabTech* (Santiago: Springer).

Lomonaco, F., Donabauer, G., and Siino, M. (2022a). *Courage at checkthat! 2022: harmful tweet detection using graph neural networks and electra*. Working Notes of CLEF, 1.

Lomonaco, F., Ognibene, D., Trianni, V., and Taibi, D. (2022b). "A game-based educational experience to increase awareness about the threats of social media filter bubbles and echo chambers inspired by "wisdom of the crowd": preliminary results," in *4th International Conference on Higher Education Learning Methodologies and Technologies Online* (Palermo).

Long, Y. (2017). *Fake News Detection Through Multi-Perspective Speaker Profiles*. Taipei: Association for Computational Linguistics.

Lopez, S. J., and Snyder, C. R. (2009). *The Oxford Handbook of Positive Psychology*. Oxford: Oxford University Press.

Lorenz, J., Rauhut, H., Schweitzer, F., and Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proc. Natl. Acad. Sci. U.S.A*. 108, 9020–9025. doi: 10.1073/pnas.1008636108

Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C. R., and Hertwig, R. (2020). How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nat. Hum. Behav*. 4, 1102–1109. doi: 10.1038/s41562-020-0889-7

Loughnan, S., Pina, A., Vasquez, E. A., and Puvia, E. (2013). Sexual objectification increases rape victim blame and decreases perceived suffering. *Psychol. Women Q*. 37, 455–461. doi: 10.1177/0361684313485718

Lowry, P. B., Zhang, J., Wang, C., and Siponen, M. (2016). Why do adults engage in cyberbullying on social media? an integration of online disinhibition and deindividuation effects with the social structure and social learning model. *Inf. Syst. Res*. 27, 962–986. doi: 10.1287/isre.2016.0671

Lyu, D., Yang, F., Liu, B., and Gustafson, S. (2019). Sdrl: interpretable and data-efficient deep reinforcement learning leveraging symbolic planning.

Proc. AAAI Conf. Artif. Intell. 33, 2970–2977. doi: 10.1609/aaai.v33i01.33012970

Ma, J., Zhao, Z., Yi, X., Yang, J., Chen, M., Tang, J., et al. (2020). "Off-policy learning in two-stage recommender systems," in *Proceedings of The Web Conference 2020* (Taipei), 463–473.

Mair, C., Roux, A. V. D., and Morenoff, J. D. (2010). Neighborhood stressors and social support as predictors of depressive symptoms in the chicago community adult health study. *Health Place* 16, 811–819. doi: 10.1016/j.healthplace.2010.04.006

Malzahn, N., Aprin, F., Hoppe, H. U., Eimler, S. C., and Moder, S. (2021). "Measuring disagreement in learning groups as a basis for identifying and discussing controversial judgements," in *Proceedings of the 14th International Conference on Computer-Supported Collaborative Learning-CSCL 2021*.

Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H., and Koper, R. (2011). "Recommender systems in technology enhanced learning," in *Recommender Systems Handbook* (New York, NY: Springer), 387–415.

March, E., and Springer, J. (2019). Belief in conspiracy theories: the predictive role of schizotypy, machiavellianism, and primary psychopathy. *PLoS ONE* 14, e0225964. doi: 10.1371/journal.pone.0225964

Marengo, D., Longobardi, C., Fabris, M., and Settanni, M. (2018). Highly-visual social media and internalizing symptoms in adolescence: the mediating role of body image concerns. *Comput. Hum. Behav.* 82, 63–69. doi: 10.1016/j.chb.2018.01.003

Marom, O., and Rosman, B. (2018). Belief reward shaping in reinforcement learning. *Proc. AAAI Conf. Artif. Intell.* 32, 11741. doi: 10.1609/aaai.v32i1.11741

Matakos, A., Aslay, C., Galbrun, E., and Gionis, A. (2022). Maximizing the diversity of exposure in a social network. *IEEE Trans. Knowl. Data Eng.* 34, 4357–4370. doi: 10.1109/TKDE.2020.3038711

Mcandrew, F. T., and Jeong, H. S. (2012). Who does what on facebook? age, sex, and relationship status as predictors of facebook use. *Comput. Hum. Behav.* 28, 2359–2365. doi: 10.1016/j.chb.2012.07.007

Mehari, K., Farrell, A., and Le, A.-T. (2014). Cyberbullying among adolescents: measures in search of a construct. *Psychol. Violence* 4, 399–415. doi: 10.1037/a0037521

Meier, A., and Schäfer, S. (2018). The positive side of social comparison on social network sites: How envy can drive inspiration on instagram. *Cyberpsychol. Behav. Soc. Network.* 21, 411–417. doi: 10.1089/cyber.2017.0708

Meyers, E. M., Erickson, I., and Small, R. V. (2013). Digital literacy and informal learning environments: an introduction. *Learn. Media Technol.* 38, 355–367. doi: 10.1080/17439884.2013.783597

Michalos, A. C. (2017). "Education, happiness and wellbeing," in *Connecting the Quality of Life Theory to Health, Well-Being and Education* (Cham: Springer), 277–299.

Milano, S., Taddeo, M., and Floridi, L. (2021). Ethical aspects of multi-stakeholder recommendation systems. *Inf. Soc.* 37, 35–45. doi: 10.1080/01972243.2020.1832636

Mitchell, M., Lebow, J., Uribe, R., Grathouse, H., and Shoger, W. (2011). Internet use, happiness, social support and introversion: a more fine grained analysis of person variables and internet activity. *Comput. Hum. Behav.* 27, 1857–1861. doi: 10.1016/j.chb.2011.04.008

Mladenović, M., Ošmjanski, V., and Stanković, S. V. (2021). Cyber-aggression, cyberbullying, and cyber-grooming: a survey and research challenges. *ACM Comput. Surveys* 54, 1–42. doi: 10.1145/3424246

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature* 518, 529–533. doi: 10.1038/nature14236

Moore, J. M., Small, M., and Yan, G. (2021). Inclusivity enhances robustness and efficiency of social networks. *Physica A* 563, 125490. doi: 10.1016/j.physa.2020.125490

Mostafavi, B., and Barnes, T. (2017). Evolution of an intelligent deductive logic tutor using data-driven elements. *Int. J. Artif. Intell. Educ.* 27, 5–36. doi: 10.1007/s40593-016-0112-1

Müller, V. C. (2020). "Ethics of artificial intelligence and robotics," in *Stanford Encyclopedia of Philosophy* (Stanford, CA: Stanford University).

Murphy, J., Hofacker, C., Gretzel, U., et al. (2017). Dawning of the age of robots in hospitality and tourism: challenges for teaching and research. *Eur. J. Tourism Res.* 15, 104–111. doi: 10.54055/ejtr.v15i.265

Murthy, D. (2012). Towards a sociological understanding of social media: theorizing twitter. *Sociology* 46, 1059–1073. doi: 10.1177/0038038511422553

Musetti, A., and Corsano, P. (2018). The internet is not a tool: reappraising the model for internet-addiction disorder based on the constraints and opportunities of the digital environment. *Front. Psychol.* 9, 558. doi: 10.3389/fpsyg.2018.00558

Nachum, O., Gu, S. S., Lee, H., and Levine, S. (2018). "Data-efficient hierarchical reinforcement learning," in *Advances in Neural Information Processing Systems, Vol. 31*, eds S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Montréal, QC: Curran Associates, Inc.), 3303–3313.

Nakayama, H., and Higuchi, S. (2015). Internet addiction. Nihon rinsho. *Jpn J. Clin. Med.* 73, 1559–1566.

Neubaum, G., and Krämer, N. C. (2017). Opinion climates in social media: blending mass and interpersonal communication. *Hum. Commun. Res.* 43, 464–476. doi: 10.1111/hcre.12118

Nie, Y., Chen, H., and Bansal, M. (2019). Combining fact extraction and verification with neural semantic matching networks. *Proc. AAAI Conf. Artif. Intell.* 33, 6859–6866. doi: 10.1609/aaai.v33i01.33016859

Nievas, E. B., Suarez, O. D., García, G. B., and Sukthankar, R. (2011). "Violence detection in video using computer vision techniques," in *International Conference on Computer Analysis of Images and Patterns* (Seville: Springer), 332–339.

Nikolov, A., and Radivchev, V. (2019). "Nikolov-radivchev at semeval-2019 task 6: offensive tweet classification with bert and ensembles," in *Proceedings of the 13th International Workshop on Semantic Evaluation* (Minneapolis, MN), 691–695.

Nikolov, D., Oliveira, D. F., Flammini, A., and Menczer, F. (2015). Measuring online social bubbles. *PeerJ Comput. Sci.* 1, e38. doi: 10.7717/peerj-cs.38

Nunes, D. S., Zhang, P., and Silva, J. S. (2015). A survey on human-in-the-loop applications towards an internet of all. *IEEE Commun. Surveys Tutorials* 17, 944–965. doi: 10.1109/COMST.2015.2398816

OCED (2019). *Recommendation of the council on artificial intelligence.* Technical report, OECD.

Ognibene, D., Fiore, V. G., and Gu, X. (2019). Addiction beyond pharmacological effects: the role of environment complexity and bounded rationality. *Neural Netw.* 116, 269–278. doi: 10.1016/j.neunet.2019.04.022

Osberg, L. (2017). On the limitations of some current usages of the gini index. *Rev. Income Wealth* 63, 574–584. doi: 10.1111/roiw.12256

Ozimek, P., Baer, F., and Förster, J. (2017). Materialists on facebook: the self-regulatory role of social comparisons and the objectification of facebook friends. *Heliyon* 3, e00449. doi: 10.1016/j.heliyon.2017.e00449

Pavlopoulos, J., Laugier, L., Sorensen, J., and Androutsopoulos, I. (2021). "Semeval-2021 task 5: toxic spans detection," in *Proceedings of SemEval* (Austin, TX: Association for Computational Linguistics).

Peddinti, S. T., Ross, K. W., and Cappos, J. (2014). ""On the internet, nobody knows you're a dog" a twitter case study of anonymity in social networks," in *Proceedings of the Second ACM Conference on Online Social Networks* (Dublin), 83–94.

Pennycook, G., and Rand, D. G. (2018). Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *J. Pers.* 88, 185–200. doi: 10.1111/jopy.12476

Perronnin, F., Sánchez, J., and Mensink, T. (2010). "Improving the fisher kernel for large-scale image classification," in *European Conference on Computer Vision* (Heraklion: Springer), 143–156.

Poria, S., Cambria, E., Bajpai, R., and Hussain, A. (2017). A review of affective computing: from unimodal analysis to multimodal fusion. *Inf. Fusion* 37, 98–125. doi: 10.1016/j.inffus.2017.02.003

Postmes, T., and Spears, R. (1998). Deindividuation and antinormative behavior: a meta-analysis. *Psychol. Bull.* 123, 238. doi: 10.1037/0033-2909.123.3.238

Prost, F., Thain, N., and Bolukbasi, T. (2019). "Debiasing embeddings for fairer text classification," in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (Florence: Association for Computational Linguistics).

Purohit, A. K., Barclay, L., and Holzer, A. (2020). "Designing for digital detox: making social media less addictive with digital nudges," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI), 1–9.

Ranjbar Kermany, N., Zhao, W., Yang, J., Wu, J., and Pizzato, L. (2021). A fairness-aware multi-stakeholder recommender system. *World Wide Web* 24, 1995–2018. doi: 10.1007/s11280-021-00946-8

Rastegarpanah, B., Gummadi, K. P., and Crovella, M. (2019). "Fighting fire with fire: using antidote data to improve polarization and fairness of recommender systems," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne, VIC), 231–239.

Rayfield, B., Fortin, M.-J., and Fall, A. (2011). Connectivity for conservation: a framework to classify network measures. *Ecology* 92, 847–858. doi: 10.1890/09-2190.1

Ridgway, J. L., and Clayton, R. B. (2016). Instagram unfiltered: exploring associations of body image satisfaction, instagram# selfie posting, and negative romantic relationship outcomes. *Cyberpsychol. Behav. Soc. Netw*. 19, 2–7. doi: 10.1089/cyber.2015.0433

Rodríguez-Sánchez, F., de Albornoz, J., Plaza, L., Gonzalo, J., R., P., et al. (2021). Overview of exist 2021: sexism identification in social networks. *Proc. IberLEF* 2021, 95–207. doi: 10.26342/2021-67-17

Romero, C., and Ventura, S. (2017). Educational data science in massive open online courses. *Wiley Interdisc. Rev*. 7, e1187. doi: 10.1002/widm.1187

Rotman, N. H., Schapira, M., and Tamar, A. (2020). "Online safety assurance for learning-augmented systems," in *Proceedings of the 19th ACM Workshop on Hot Topics in Networks, HotNets '20* (Association for Computing Machinery: New York, NY), 88–95.

Rourke, L., Anderson, T., Garrison, D. R., and Archer, W. (1999). Assessing social presence in asynchronous text-based computer conferencing. *J. Distance Educ*. 14, 50–71.

Roy, B., Riley, C., Sears, L., and Rula, E. Y. (2018). Collective well-being to improve population health outcomes: an actionable conceptual model and review of the literature. *Am. J. Health Promot*. 32, 1800–1813. doi: 10.1177/0890117118791993

Ryan, T., Allen, K. A., Gray, D. L., and McInerney, D. M. (2017). How social are social media? a review of online social behaviour and connectedness. *J. Relationships Res*. 8, 13. doi: 10.1017/jrr.2017.13

Ryff, C. D., Singer, B. H., and Dienberg Love, G. (2004). Positive health: connecting well-being with biology. *Philos. Trans. R. Soc. Lond. B Biol. Sci*. 359, 1383–1394. doi: 10.1098/rstb.2004.1521

Sánchez-Reina, J., Hernández-Leo, D., Theophilou, E., and Lobo-Quintero, R. (2022). "But i don't wanna share my data'. Analyzing teen's concerns about the use of social media," in *IAMCR 2022 Conference. Communication Research in the Era of Neo-Globalisation: Reorientations, Challenges and Changing Contexts* (Beijing).

Sanguinetti, M., Comandini, G., Di Nuovo, E., Frenda, S., Stranisci, M., Bosco, C., et al. (2020). "Haspeede 2@ evalita2020: overview of the evalita 2020 hate speech detection task," in *EVALITA*.

Sawyer, R., Rowe, J., and Lester, J. (2017). "Balancing learning and engagement in game-based learning environments with multi-objective reinforcement learning," in *International Conference on Artificial Intelligence in Education* (Wuhan: Springer), 323–334.

Schafer, J. (2002). Spinning the Web of hate: Web-based hate propagation by extremist organizations. *J. Crim. Just. Popul. Cult*. 9, 69–88.

Schlesinger, A., Chandrasekharan, E., Masden, C. A., Bruckman, A. S., Edwards, W. K., and Grinter, R. E. (2017). "Situated anonymity: impacts of anonymity, ephemerality, and hyper-locality on social media," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, CO), 6912–6924.

Schmidt, A. L., Zollo, F., Del Vicario, M., Bessi, A., Scala, A., Caldarelli, G., et al. (2017). Anatomy of news consumption on facebook. *Proc. Natl. Acad. Sci. U.S.A*. 114, 3035–3039. doi: 10.1073/pnas.1617052114

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., et al. (2020). Mastering atari, go, chess and shogi by planning with a learned model. *Nature* 588, 604–609. doi: 10.1038/s41586-020-03051-4

Scifo, L., Fulantelli, G., and Taibi, D. (2022). "Adolescents and social media education: the role of emotional intelligence," in *Proceedings of VIII ICEI 2022-International Congress on Emotional Intelligence* (Palermo).

Scolari, C. A., Masanet, M.-J., Guerrero-Pico, M., and Establés, M.-J. (2018). Transmedia literacy in the new media ecology: Teens' transmedia skills and informal learning strategies. *El Profesional de la Informacion* 27, 801–812. doi: 10.3145/epi.2018.jul.09

Seligman, M. E. (2011). Flourish: a visionary new understanding of happiness and well-being. *Policy* 27, 60–61.

Seligman, M. E. (2012). *Flourish: A Visionary New Understanding of Happiness and Well-Being*. New York, NY: Simon and Schuster.

Seufert, S., Meier, C., Soellner, M., and Rietsche, R. (2019). A pedagogical perspective on big data and learning analytics: a conceptual model for digital learning support. *Technol. Knowl. Learn*. 24, 599–619. doi: 10.1007/s10758-019-09399-5

Shani, G., Heckerman, D., and Brafman, R. I. (2005). An mdp-based recommender system. *J. Mach. Learn. Res*. 6, 1265–1295.

Shensa, A., Escobar-Viera, C. G., Sidani, J. E., Bowman, N. D., Marshal, M. P., and Primack, B. A. (2017). Problematic social media use and depressive symptoms among us young adults: a nationally-representative study. *Soc. Sci. Med*. 182, 150–157. doi: 10.1016/j.socscimed.2017.03.061

Shirky, C. (2011). The political power of social media: Technology, the public sphere, and political change. *Foreign Affairs* 90, 28–41.

Shu, T., Xiong, C., and Socher, R. (2017). Hierarchical and interpretable skill acquisition in multi-task reinforcement learning. *arXiv preprint* arXiv:1712.07294. doi: 10.48550/arXiv.1712.07294

Spratt, J. (2017a). "Conceptualising wellbeing," in *Wellbeing, Equity and Education* (Cham: Springer), 35–56.

Spratt, J. (2017b). *Wellbeing, Equity and Education. A Critical Analysis of Policy Discourses of Wellbeing in Schools*. Cham: Springer.

Steccanella, L., Totaro, S., Allonsius, D., and Jonsson, A. (2020). Hierarchical reinforcement learning for efficient exploration and transfer. *arXiv [Preprint]*. arXiv: 2011.06335. Available online at: https://arxiv.org/pdf/2011.06335.pdf

Steiger, M., Bharucha, T. J., Venkatagiri, S., Riedl, M. J., and Lease, M. (2021). "The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama), 1–14.

Stewart, A. J., Mosleh, M., Diakonova, M., Arechar, A. A., Rand, D. G., and Plotkin, J. B. (2019). Information gerrymandering and undemocratic decisions. *Nature* 573, 117–121. doi: 10.1038/s41586-019-1507-6

Stöcker, C., and Preuss, M. (2020). "Riding the wave of misclassification: how we end up with extreme youtube content," in *International Conference on Human-Computer Interaction* (Copenhagen: Springer), 359–375.

Stoica, A.-A., Riederer, C., and Chaintreau, A. (2018). "Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity," in *International World Wide Web Conferences Steering Committee WWW '18, WWW '18* (Geneva: CHE), 923–932.

Szabo, G., and Huberman, B. A. (2010). Predicting the popularity of online content. *Commun. ACM* 53, 80–88. doi: 10.1145/1787234.1787254

Szymanski, D. M., Moffitt, L. B., and Carr, E. R. (2011). Sexual objectification of women: advances to theory and research. *Couns Psychol*. 39, 6–38. doi: 10.1177/0011000010378402

Taibi, D., Börsting, J., Hoppe, H. U., Ognibene, D., Hernández-Leo, D., and Eimler, S. (2022). "Designing educational interventions to increase students' social media awareness - experience from the courage project," in *Proceedings of the 4th International Conference on Higher Education Learning Methodologies and Technologies Online (HELMeTO2022)-Book of Abstracts* (Palermo), 81–83.

Taibi, D., Fulantelli, G., Monteleone, V., Schicchi, D., and Scifo, L. (2021). "An innovative platform to promote social media literacy in school contexts," in *ECEL 2021 20th European Conference on e-Learning* (Berlin: Academic Conferences International Limited), 460.

Talwar, V., Gomez-Garibello, C., and Shariff, S. (2014). Adolescents' moral evaluations and ratings of cyberbullying: The effect of veracity and intentionality behind the event. *Comput. Hum. Behav*. 36, 122–128. doi: 10.1016/j.chb.2014.03.046

Tariq, W., Mehboob, M., Khan, M. A., and Ullah, F. (2012). The impact of social media and social networks on education and students of pakistan. *Int. J. Comput. Sci. Issues* 9, 407.

Tartari, E. (2015). Benefits and risks of children and adolescents using social media. *Eur. Sci. J*. 11.

Tarus, J. K., Niu, Z., and Yousif, A. (2017). A hybrid knowledge-based recommender system for e-learning based on ontology and sequential pattern mining. *Future Generat. Comput. Syst*. 72, 37–48. doi: 10.1016/j.future.2017.02.049

Taymur, I., Budak, E., Demirci, H., Akdağ, H. A., Güngör, B. B., and Özdel, K. (2016). A study of the relationship between internet addiction, psychopathology and dysfunctional beliefs. *Comput. Hum. Behav*. 61, 532–536. doi: 10.1016/j.chb.2016.03.043

Thaler, R. H., and Sunstein, C. R. (2009). *Nudge: Improving Decisions About Health, Wealth, and Happiness. Penguin*. Yale University Press, 304.

Topp, C. W., Østergaard, S. D., Søndergaard, S., and Bech, P. (2015). The who-5 well-being index: a systematic review of the literature. *Psychother Psychosom*. 84, 167–176. doi: 10.1159/000376585

Tran, H. N., and Kruschwitz, U. (2021). "UR-IW-HNT at GermEval 2021: an ensembling strategy with multiple BERT models," in *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments* (Duesseldorf: Association for Computational Linguistics), 83–87.

Tran, H. N., and Kruschwitz, U. (2022). "UR-IW-HNT at checkthat!-2022: cross-lingual text summarization for fake news detection," in *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of CEUR Workshop Proceedings*, eds G. Faggioli, N. Ferro, A. Hanbury, and M. Potthast (Bologna: CEUR-WS.org), 740–748.

Turban, C., and Kruschwitz, U. (2022). "Tackling irony detection using ensemble classifiers and data augmentation," in *Proceedings of the Language Resources and Evaluation Conference* (Marseille: European Language Resources Association), 6976–6984.

UNESCO (2020). *First draft of the recommendation on the ethics of artificial intelligence*. Technical report, UNESCO.

Urena, R., Kou, G., Dong, Y., Chiclana, F., and Herrera-Viedma, E. (2019). A review on trust propagation and opinion dynamics in social networks and group decision making frameworks. *Inf. Sci*. 478, 461–475. doi: 10.1016/j.ins.2018.11.037

Van Der Maesen, L. J., and Walker, A. (2011). *Social Quality: From Theory to Indicators*. Palgrave Macmillan London. doi: 10.1007/978-0-230-36109-6

Van Seijen, H., Fatemi, M., Romoff, J., Laroche, R., Barnes, T., and Tsang, J. (2017). "Hybrid reward architecture for reinforcement learning," in *Advances in Neural Information Processing Systems* (Long Beach, CA), 5392–5402.

Van Staalduinen, J.-P., and de Freitas, S. (2011). A game-based learning framework: Linking game design and learning. *Learn. Play* 53, 29.

Verduyn, P., Ybarra, O., Résibois, M., Jonides, J., and Kross, E. (2017). Do social network sites enhance or undermine subjective well-being? a critical review. *Soc. Issues Policy Rev*. 11, 274–302. doi: 10.1111/sipr.12033

Verrastro, V., Liga, F., Cuzzocrea, F., Gugliandolo, M. C., et al. (2020). Fear the instagram: beauty stereotypes, body image and instagram use in a sample of male and female adolescents. *Qwerty Open Interdisc. J. Technol. Cult. Educ*. 15, 31–49. doi: 10.30557/QW000021

Walker, K. L. (2016). Surrendering information through the looking glass: transparency, trust, and protection. *J. Public Policy Market*. 35, 144–158. doi: 10.1509/jppm.15.020

Wang, J., Liu, Y., and Li, B. (2020). Reinforcement learning with perturbed rewards. *Proc. AAAI Conf. Artif. Intell*. 34, 6202–6209. doi: 10.1609/aaai.v34i04.6086

Wang, J.-L., Jackson, L. A., Gaskin, J., and Wang, H.-Z. (2014). The effects of social networking site (sns) use on college students' friendship and well-being. *Comput. Hum. Behav*. 37, 229–236. doi: 10.1016/j.chb.2014.04.051

Wang, R., Zhou, D., Jiang, M., Si, J., and Yang, Y. (2019). A survey on opinion mining: from stance to product aspect. *IEEE Access* 7, 41101–41124. doi: 10.1109/ACCESS.2019.2906754

Wang, W. Y. (2017). ""liar, liar pants on fire": a new benchmark dataset for fake news detection," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Vancouver, BC: Association for Computational Linguistics), 422–426.

Watts, D. J. (2011). *Everything Is Obvious:* Once you Know the Answer*. New York, NY: Crown Business.

Webb, H., Burnap, P., Procter, R., Rana, O., Stahl, B. C., Williams, M., et al. (2016). Digital wildfires: propagation, verification, regulation, and responsible innovation. *ACM Trans. Inf. Syst*. 34, 15. doi: 10.1145/2893478

Weng, L., Flammini, A., Vespignani, A., and Menczer, F. (2012). Competition among memes in a world with limited attention. *Sci. Rep*. 2, 335. doi: 10.1038/srep00335

White, J. (2007). Wellbeing and education: Issues of culture and authority. *J. Philos. Educ*. 41, 17–28. doi: 10.1111/j.1467-9752.2007.00540.x

Whittaker, E., and Kowalski, R. M. (2015). Cyberbullying via social media. *J. Sch. Violence* 14, 11–29. doi: 10.1080/15388220.2014.949377

Wilkens, R., and Ognibene, D. (2021a). "Bicourage: ngram and syntax gcns for hate speech detection," in *Forum for Information Retrieval Evaluation (Working Notes) (FIRE)* (CEUR-WS.org).

Wilkens, R., and Ognibene, D. (2021b). "Mb-courage@ exist: gcn classification for sexism identification in social networks," in *IberLEF@ SEPLN*, 420–430.

Wineburg, S., McGrew, S., Breakstone, J., and Ortega, T. (2016). Evaluating information: the cornerstone of civic online reasoning. *SDR* 8, 2018.

Wolfowicz, M. (2015). *A Social Learning Theory Examination of the Complicity of Personalization Algorithms in the Creation of Echo Chambers of Online Radicalization to Violent Extremism*. Macquarie University Publisher, 1–14.

Wu, Q., Wang, H., Hong, L., and Shi, Y. (2017). "Returning is believing: optimizing long-term user engagement in recommender systems," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Singapore), 1927–1936.

Xu, S., Yang, H. H., MacLeod, J., and Zhu, S. (2019). Social media competence and digital citizenship among college students. *Convergence* 25, 735–752. doi: 10.1177/1354856517751390

Young, K. S. (2017). The evolution of internet addiction. *Addict. Behav*. 64, 229–230. doi: 10.1016/j.addbeh.2015.05.016

Yukselturk, E., Ozekes, S., and Türel, Y. K. (2014). Predicting dropout student: an application of data mining methods in an online education program. *Eur. J. Open Distance e-learning* 17, 118–133. doi: 10.2478/eurodl-2014-0008

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). "Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)," in *PIWSE '19* (Minneapolis, MN), 75–86.

Zanzotto, F. M. (2019). Human-in-the-loop artificial intelligence. *J. Artif. Intell. Res*. 64, 243–252. doi: 10.1613/jair.1.11345

Zarrella, G., and Marsh, A. (2016). "Mitre at semeval-2016 task 6: transfer learning for stance detection," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (San Diego, CA), 458–463.

Zawacki-Richter, O., Marín, V. I., Bond, M., and Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education-where are the educators? *Int. J. Educ. Technol. Higher Educ*. 16, 39. doi: 10.1186/s41239-019-0171-0

Zeng, C., Wang, Q., Mokhtari, S., and Li, T. (2016). "Online context-aware recommendation with time varying multi-armed bandit," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA), 2025–2034.

Zhao, X., Xia, L., Tang, J., and Yin, D. (2019a). "Deep reinforcement learning for search, recommendation, and online advertising: a survey," in *ACM SIGWEB Newsletter* (Spring), 1–15.

Zhao, X., Xia, L., Yin, D., and Tang, J. (2019b). Model-based reinforcement learning for whole-chain recommendations. *arXiv preprint* arXiv:1902.03987. doi: 10.48550/arXiv.1902.03987

Zheng, G., Zhang, F., Zheng, Z., Xiang, Y., Yuan, N. J., Xie, X., et al. (2018). "DRN: a deep reinforcement learning framework for news recommendation," in *Proceedings of the 2018 World Wide Web Conference* (Lyon), 167–176.

Zhong, L., Cao, J., Sheng, Q., Guo, J., and Wang, Z. (2020). "Integrating semantic and structural information with graph convolutional network for controversy detection," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online: Association for Computational Linguistics), 515–526.

Zhou, G., Azizsoltani, H., Ausin, M. S., Barnes, T., and Chi, M. (2019). "Hierarchical reinforcement learning for pedagogical policy induction," in *International Conference on Artificial Intelligence in Education* (Yokohama: Springer), 544–556.

Zhou, G., Wang, J., Lynch, C. F., and Chi, M. (2017). "Towards closing the loop: bridging machine-induced pedagogical policies to learning theories," in *International Educational Data Mining Society* (Wuhan).

Zhou, G., Yang, X., Azizsoltani, H., Barnes, T., and Chi, M. (2020). "Improving student-system interaction through data-driven explanations of hierarchical reinforcement learning induced pedagogical policies," in *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa), 284–292.

Zhou, P., Ding, Q., Luo, H., and Hou, X. (2018). Violence detection in surveillance video using low-level features. *PLoS ONE* 13, e0203668. doi: 10.1371/journal.pone.0203668

Zhou, T., Kuscsik, Z., Liu, J.-G., Medo, M., Wakeling, J. R., and Zhang, Y.-C. (2010). Solving the apparent diversity-accuracy dilemma of recommender systems. *Proc. Natl. Acad. Sci. U.S.A*. 107, 4511–4515. doi: 10.1073/pnas.1000488107

Zhou, Y., Kim, D. W., Zhang, J., Liu, L., Jin, H., Jin, H., et al. (2017). Proguard: detecting malicious accounts in social-network-based online promotions. *IEEE Access* 5, 1990–1999. doi: 10.1109/ACCESS.2017.2654272

Zimmerman, S., Kruschwitz, U., and Fox, C. (2018). "Improving hate speech detection with deep learning ensembles," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (Miyazaki).

Zimmerman, S., Thorpe, A., Chamberlain, J., and Kruschwitz, U. (2020). "Towards search strategies for better privacy and information," in *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, CHIIR '20* (Vancouver, BC: Association for Computing Machinery), 124–134.

Zou, L., Xia, L., Ding, Z., Song, J., Liu, W., and Yin, D. (2019). "Reinforcement learning to optimize long-term user engagement in recommender systems," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '19* (New York, NY: Association for Computing Machinery), 2810–2818.