From Comments to Cognition: How online comments on social media influence personal opinion

Doctoral thesis

to obtain a doctorate (PhD)

from the Faculty of Medicine

of the University of Bonn

Federica Nisini

from Napoli, Italy

2025

Written with authorization of

the Faculty of Medicine of the University of Bonn

First reviewer: PD. Dr. Johannes Schultz

Second reviewer: Prof. Dr. rer. Nat. Silke Lux

Day of oral examination: 09.01.2025

Institut für experimentelle Epileptologie und Kognitionsforschung

Director: Prof. Dr. Heinz Beck

Center for Economics and Neuroscience (CENs)

Director: PD Dr. Johannes Schultz and Prof. Dr. Bernd Weber

Table of Contents

List of Abbreviations

AG	Angular gyrus
dACC	Dorsal anterior cingulate
DIMI	Digital maturity inventory
dlPFC	Dorsolateral prefrontal cortex
fMRI	Functional magnetic resonance imaging
FWE	Family wise error
GFM	Gradient field map
GLM	General linear model
IFG	Inferior frontal gyrus
MFG	Middle frontal gyrus
mPFC	Medial prefrontal cortex
MTG	Middle temporal gyrus
OFC	Orbitofrontal cortex
PC	Precuneus
pMFC	Posterior medial frontal cortex
ROI	Regions of interest
SFG	Superior frontal gyrus
STG	Superior temporal gyrus
SVC	Small volume correction
ToM	Theory of mind
TPJ	Temporoparietal Junction
vmPFC	Ventromedial prefrontal cortex

1. General Introduction

The human brain is a remarkably plastic organ, allowing us to continuously learn from signals in our environment and adjust our behavior accordingly. In particular, humans utilize information from the behavior of others to navigate their surrounding and make informed decisions. By observing and interacting with others, individuals can quickly adapt to novel environments, acquire new skills, and at the same time trying to avoid possible costs associated with trial-and-error learning (Molleman et al., 2019). Effectively using social information is crucial to increase knowledge and adapt to social environments, often by integrating diverse information from various sources (Molleman et al., 2019). However, the complexity of social information is amplified in atypical settings such as online social media, where constant and exaggerated information can significantly impact people's lives (Tamir & Ward, 2015). The steady increase in the use of mobile devices led to a surge in information exchange over the Internet (Livingstone et al., 2018). However, online interactions fundamentally differ from face-to-face conversations (Tamir & Ward, 2015), and the effects of online networks on the processing of social information remains largely unknown. In terms of opinion formation, the shift towards digital platforms has resulted in more news being consumed via online media rather than traditional newspaper (Steinfeld et al., 2016). Therefore, the influence of other users on one's perception of a news article is leading researchers to reexamine the dynamics and impact of online social information. This becomes particularly concerning when social information is used to undermine established scientific knowledge, as observed in current debates on vaccines or climate change (Williams & Hsieh, 2021).

This doctoral thesis aims to contribute to the growing research on social influence on social media. In this chapter, we provide an overview on the background and issues that motivated this research. Specifically, we will explore the rise of the Internet and social media, the transition from traditional newspapers to online media, and the impact of user-generated comments on news perception and opinion formation.

1.1 The Rise of Internet and Social Media Sites

Humans are inherently social creatures. Since the origins of our species, we constantly seek connections and community. For this reason, the advent of Internet represented an

historic turning point that radically transformed how we interact and communicate with each other and with the world around (Jordan, 2013). Indeed, if we look around us, we are surrounded by technological devices: computers, smartphones, smart televisions; they all are mediums to access the online social world (Tamir & Ward, 2015). These artificial social environments are not only used to facilitate communication but can even replace face-to-face interactions (Tamir & Ward, 2015). For instance, the COVID-19 pandemic highlighted the Internet's crucial role in sustaining the global economy and supporting people's wellbeing (Kozyreva et al., 2020).

Worldwide, it is estimated that around 5.35 billion people (66.2% of the global population) are active internet users (DataReportal, & Meltwater, & We Are Social, 2024). Among them, a growing number of children use the Internet on a daily basis and grow up as "digital natives", becoming increasingly immersed in digital and mobile technologies (Livingstone et al, 2018). Especially mobile devices as smartphones and tablets enable people to be online everywhere and every time (Basole, 2004) and account for almost 59% of the total internet traffic (Statistica, 2023). The Social Web, represented by social media sites like Facebook, is only 20 years old but already has passed the 5 billion active users, making social media the most popular online activity in the world (DataReportal, & Meltwater, & We Are Social, 2024).

As Tamir and Ward (2015) explained, new media provide seemingly constant social cues and put our social brain into "overdrive", by offering potentially never-ending social interactions without the same constrains of the physical world. In fact, statistics show that everyday people are spending more and more time online, and that they utilise social media mainly to stay connected with family and friends, but also to fill spare time and to read online news (DataReportal, & Meltwater, & We Are Social, 2024). In regard to this last activity, another critical turning point that inevitably shaped the public sphere is the advent of online newspapers on social media as main source of daily news consumption for the general population (Guo et al., 2021). One of the most important features characterizing online news is the ability for the audience to comment the online content they are reading. On social media, for instance, the comment section is immediately below the post which fosters interactivity and discussion among users (Springer et al., 2015).

Therefore, people are now able to both read the news and openly debate with other ordinary users, making reading news a collective activity (Lee & Jang, 2010).

1.2 The Shift from Traditional Media to Digital Platforms

As Steinfeld and colleagues (2016) explained, the transition from traditional newspaper to social media revolutionised the journalistic industry, which was originally characterised by a unilateral relationship where the journalist was the authority and the audience had little possibility to interact and express their opinion. After the advent of the first online newspapers, online interactivity and debate grew exponentially, becoming a fundamental part of public deliberation and engaging citizens into a broader range of opinions (Steinfeld et al., 2016). Unlike in the past, readers are now confronted with both the news and the reactions of other readers to this news at the same time (Lee & Jang, 2010). As these platforms became a vast source of news diffusors, more and more people turned to social media (like Facebook, Twitter and Instagram) to access information, consume news and form their opinions (Guo et al., 2021).

However, a specific format seems to be very popular on social media sites: the “snack news”. Snack news are social media posts that provide a news headline, a picture, and a short preview of the news article (Schäfer, Sülflow, & Müller, 2017). They are fast to read and need a low level of cognitive engagement (Schäfer, Sülflow, & Müller, 2017). As Schäfer (2020) pointed out, snack news is mainly used to get a general overview of a topic without gaining in-depth knowledge of it. Bakshy and colleagues (2015) found that, when it comes to political and world affairs news, just 7% of 10.1 million Facebook users clicked on the links in their news feed to view the full articles, meaning that the great majority of users simply read the news preview, or snack news. A drawback of a heavy “snack news diet” is that users feel more informed about an issue after encountering this format, without a true gain in their actual knowledge (Schäfer, 2020). In this way, people consistently exposed to snack news on social media could develop an illusion of knowledge, feeling more informed than they are and thus holding and expressing stronger convictions on these topics (Park, 2001).

An even darker criticality of online information spread on social media is that often these platforms lack transparency behind their recommendation algorithms (Barbu, 2016). For

instance, research has found evidence of the contribution of Facebook and YouTube on the rise and unification of far right-wing parties in US and Germany (Kaiser & Rauchfleisch, 2018; Rauchfleisch & Kaiser, 2017). This was due to the fact that these platforms automatically recommended more and more polarising and conspiracist material to its users. But why algorithmic recommendation systems prioritise controversial content? Psychologists suggest that sensational content is often emotionally charged and trigger strong reactions like outrage and indignation, thereby increasing engagement metrics, such as shares and likes (Vosoughy et al., 2018). The lack of transparency regarding how social media operate was also particularly evident in the Facebook scandal involving the use of “dark ads” during the 2016 US presidential election and the UK Brexit referendum. Dark ads are advertisements only visible to the predefined targeted audience, designed to exploit psychological characteristics and vulnerabilities of their target audience, in order to influence their attitudes, decision-making and ultimately voting behaviours (Trott et al., 2021). The scandal led to stricter regulations and increased public awareness about privacy rights, however, the illegal acquisition and exploitation of personal data undermined the trust over these platforms and alarmed on possible future manipulations of democratic processes (Saunders, 2020).

The influence on user perception is further exacerbated by the virality of social media, which facilitates the spread of sensational and controversial content at the expenses of more moderate and balanced perspectives, since the former evokes stronger emotions from users (Berger & Milkman, 2012; Vosoughi et al., 2018). This also means that propaganda and misinformation can spread faster than real and professionally curated content. Indeed, research found that false news on social media spread significantly faster, farther and deeper than real news (Vosoughy et al., 2018). Fake news usually employs impressive clickbait headlines, exaggerated stories with emotional language, and dramatic images to draw attention and increase monetarization, or to manipulate ideological beliefs (Baptista & Gradim, 2020).

However, invisible targeted ads and false news are not the only factors to influence users’ attitudes and beliefs on social media. Social engagement metrics, such as likes, comments and shares, have received a lot of attention over the last decade because they appeal to people’s social brains which seek social connection and understanding of what

other people think about an issue (Tamir & Ward, 2015). As evidenced by research, the presence of these social engagement cues can influence the probability for a news to be read and shared (Dvir-Gvirsman, 2019; Segesten et al., 2020). For instance, the presence of social engagement cue increased the attention and selection rates of news, especially when these engagement cues signalled a high level of endorsement (Dvir-Gvirsman, 2019). However, not all social cues are perceived equally by users. Qualitative cues, such as user-generated comments, are more capable to influence opinions and attitudes compared to quantitative cues, such as likes and shares (Dvir-Gvirsman, 2019; Segesten et al., 2020).

1.3 User-generated Comments

The power of user-generated comments created a paradigm shift in the way in which contemporary news are produced and disseminated online. User comments are the most popular form of audience engagement in the contemporary news landscape (Weber, 2014). They are placed directly below the news article posted on social media sites, and thus have the potential to reach as many other users as the journalistic articles (Springer et al., 2015). Unlike the past, where editors were gatekeepers between the writer and the readers and could filter out irrelevant or flaming content, social media offer a platform where everyone can express their opinions and unfiltered reactions through comments (Waddell, 2019). Although user-generated comments were initially seen as a form of easy, active citizen engagement beneficial for deliberative democratic processes, many comments produced on social media are often shallow and sometimes disrespectful (Ksiazek & Springer, 2018). This can have potential negative repercussions on the material they are attached to. For instance, uncivil and poor-quality user comments under a professional news seem to have an adverse effect on the perceived formal quality of the article (Prochazka, Weber, & Schweiger, 2016). Anderson and colleagues (2014) found that uncivil user-generated comments contribute to polarization on the risk perception of unfamiliar topics, such as nanotechnology. The authors pointed out that even the perception of issues with scientific consensus, such as climate change, might be shaped and polarised not only by the mere information provided by the articles but by uncivil online comments as well (Anderson et al., 2014).

The effect of online comments becomes even more unsettling in light of recent studies that found that users often read the comment sections before reading the news article (Jones et al., 2019) and that some of them spend more time reading comments than the article itself (Stroud, Duyn & Peacock, 2016). Scholars offered various suggestions of why user-generated comments might be so engaging. According to Lee and Tandoc (2017), online comments are so engaging and effective because they are a hybrid between mass and interpersonal communication: they are messages from individuals who express their personal thoughts and feelings but are visible and can reach a mass audience. Indeed, laypeople without proper expertise or reputation can now express their opinions on social media and potentially reach as many users as respectable journals like CNN or The New York Times (Allcott & Gentzkow, 2017). According to Lee and Jang (2010), who cited the Exemplification Theory (Zillmann & Brosius, 2000), user comments are seen as "exemplars": anecdotal and subjective opinion cues that make an abstract issue more concrete, vivid and easy to comprehend. According to this theory, exemplars can shape attitudes and behaviors because they are perceived as representing the public sentiment, which in turn can lead to social conformity (Lee & Jang, 2010; Zillmann & Brosius, 2000). Indeed, the exposure to these comments, misinterpreted as the crowd opinion, produces an illusion of representativeness - and because people tend to see the majority's beliefs as more accurate (via the "bandwagon effect"; Sundar, 2008) - they tend to endorse and adjust their judgments to reflect the crown sentiment (Axsom et al., 1987; Waddell, 2019). However, Neubaum and Krämer (2017), who applied the Spiral of Silence Theory (Noelle-Neumann, 1974) to online media, explained how the comment sections can actually represent a rather distorted picture of the public sentiment. Specifically, they explain that individuals who have the same view as the majority tend to openly voice their perspectives, instead individuals with incongruent viewpoints tend to withhold their opinions because they fear repercussions and public shame. Over time, this process perpetuates into a "spiral" effect, where the minority's view gets more and more marginalized and the majority becomes even more predominant, creating a distorted picture of the public sentiment (Neubaum & Krämer, 2017; Noelle-Neumann, 1974).

This process is problematic for multiple reasons. Firstly, online news commenters do not represent the public climate, but rather a limited, non-representative and possibly biased sample of the general population (Lee, Jang, & Chung, 2021). Indeed, recent studies have

shown that specific socio-demographic characteristics can be observed in people that comment news online (although with geographically localized contexts that should not be generalized to the global population). For instance, US commenters tend to exhibit lower levels of education compared to passive readers (Stroud et al., 2016). In Germany, online commenters tend to be part of an older age cohort compared to readers (Springer et al., 2015) and hold more conservative ideologies, often supporting far right-wing parties such as AfD (Köcher, 2016). Secondly, because these individuals are more active and more prominent in the comment sections, people holding moderate divergent perspectives, or unsure people, might be dissuaded to expose themselves and engage in a discussion, risking a polarization of the online public discourse (Neubaum & Krämer, 2017). Several scholars underlined how news on social media might increase polarizing views, making people more intolerant to discordant opinions, less willing to interact with individuals holding opposing views, and increasing friction for societal decision-making (see review from Kubin & Sikorski, 2021; Sude et al., 2019; Waddell, 2019). Therefore, it is of critical importance to study the new challenges that social media and user-generated comments might create for the public sphere and overall for society.

1.4 Digital Maturity

To counteract these influences, research efforts are increasingly focused on mitigating the effects of mobile devices and the Internet, while trying to identify protective factors against their pervasiveness. Recently, new metrics have been developed to evaluate and quantify individuals' maturity in using digital technologies. For example, Laaber and colleagues (2023) addressed the new challenges of digital environments by conceptualizing and operationalizing a new construct of digital maturity. They defined digital maturity as a comprehensive concept encompassing the attitudes and capabilities that individuals need in order to thrive in online environments and potentially shield themselves against digital threats. The digital maturity index consists of ten sub-dimensions which aim to address different characteristics of the construct, including digital literacy, digital risk awareness, and autonomy within digital contexts. Laaber and colleagues (2023) developed and validated the Digital Maturity Inventory (DIMI) as part of the project DIGYMATEX (https://digymatex.eu/), which is funded by the European Union and also financed this doctoral dissertation. This inventory is specifically designed to measure levels of digital

maturity in younger generations, who are more vulnerable to online risks and influences (Ahmed et al., 2020). Within the scope of this thesis, the DIMI is a useful tool for determining a person's potential susceptibility to online social influences because it specifically examines how actively and self-determined individuals use technology and online information.

1.5 Aim of the Thesis

The primary objective of this doctoral thesis is to employ a multimethodological approach to investigate the impact of user-generated comments on how personal opinions regarding news posted on social media platforms are formed. Additionally, we were interested in determining whether digital maturity might serve as a buffer against the social influence of online comments. Specifically, we developed and implemented a novel social influence paradigm designed to address existing gaps in the literature and contributing to scientific findings about opinion change on social media. The validation of this new behavioral paradigm involved comprehensive testing across diverse populations, located both in US and Germany. After refining and testing the behavioral paradigm, our research extended to neuroimaging methods, such as functional magnetic resonance imaging (fMRI), adapting the behavioral paradigm to investigate the neural correlates involved in the processing of user-generated comments posted on social media. In summary, this doctoral thesis advances the literature on the influence that online comments exert on social media regarding online news, by creating a novel social influence paradigm and leveraging on a multimethodological approach.

2. Creation and Validation of a New Social Influence Paradigm

2.1 Introduction

2.1.1 Influence of Comments on News Perception

Over the years, social media became a virtual agora where citizens participate in the public discourse, exchange ideas and foster a sense of collective engagement in real time and without geographical boundaries (Papacharissi, 2002). The influence that online comments exert on individuals' personal opinion has been documented in several studies. The earliest investigations (for example Anderson et al., 2014; Lee & Jang, 2010) focused on the impact of comments written below blog articles and online news websites in shaping redears' perception. On these specific platforms, individuals are able to first read the full article and then read the comments of other users. These studies demonstrated a clear influence of comments in shaping people's perception and attitude. However, social media differ in that since they allow readers to view other users' comments even before opening the articles, potentially molding and influencing even more individuals' opinions and perceptions of the upcoming content they will read.

For instance, Winter and collogues (2015) found that readers had more negative attitudes towards news posted on Facebook after reading opposing comments to the post, but their attitudes did not change following supporting comments. This effect was stronger after relevant argumentative comments rather than subjective ones. A limitation of this study was to only test one news, which makes more complex to assess whether the influence effect could be generalized to other news topics. Recent research on scientific publications posted on Reddit found that reading negative comments before the article could reduce people's interest in reading the study and influenced how they agreed with the study's methodology and findings (Williams & Hsieh, 2021). The authors addressed how effective attempts to discredit scientific discoveries may be in persuading readers to mistrust scientific findings based on the negative, low-quality comments left below the post. This phenomenon has broader implications, which could elucidate the emerging and proliferation of the anti-vaccination movement during the COVID-19 pandemic. Indeed, during the COVID-19 pandemic, negative and skeptic messages towards vaccines were a large proportion of the content spread on social media (Cascini et al., 2022). These

messages created a fertile ground for the consolidation of anti-vaccination echo chambers (Van Raemdonck, 2019), the polarization of users' attitudes (Schmidt et al., 2018), and had a crucial role in shaping people's perception towards government measures, therefore having a tangible impact on society (Cascini et al., 2022; Van Raemdonck, 2019).

Interestingly, while negative comments might exacerbate skepticism towards legitimate scientific findings, this same effect could reveal to be beneficial in debunking fake news posted on social media. Indeed, user-generated comments are so influential that they seem to be even more effective at flagging fake-news to other users than official disclaimers attached to the post made by the platforms themselves (Colliander, 2019). Research by Colliander (2019) revealed that these critical comments worsened people's attitude to the news and made them significantly less likely to positively comment or share the post. It is worth noting that this study, as others before, only employed comment sections featuring unanimous comments towards the post, which could magnify the effect. However, real-world comment sections typically have a combination of both critical and supportive comments, which emphasizes the importance to include mixed-comment conditions to obtain a comprehensive understanding of their impact. Wijenayake and colleagues (2020) tried to address this limitation by introducing an extra condition where there was no majority within the comment section of a news posted on Facebook; instead, an equal distribution between supportive and critical comments was created. They found a tendency among individuals to conform their personal view to the majority's opinion (compared to the no-majority condition), especially when this majority expressed criticism towards the post. Additionally, the researchers found an upward trend between the size of the majority and the likelihood to conform to their opinion. Moreover, when individuals were more uncertain regarding their initial opinion towards the post, they were more likely influenced by the majority's opinion (Wijenayake et al., 2020).

Some scholars pointed out that the overwhelming and intricate nature of social media make people rely on the first available piece of information, such as other people's comments, in order to save cognitive resources, thereby making themselves more susceptible to persuasion (Sude et al., 2019). For instance, the Limited Cognitive Model pustulates that individuals have limited cognitive capacities when it comes to encode, interpret and evaluate media content, therefore they don't always process media

exhaustively, but rely on heuristics or mental shortcuts to conserve cognitive resources (Lang, 2000). According to this theory, individuals may vary their processing considering factors such as: prior topic knowledge, cognitive abilities, interest in the topic, and situational factors like task demands and distractions. This becomes particularly evident in the context of social media, where the enormous amount of diverse information completely saturates our capacity to process information (Sülflow et al., 2019). Lee and Young (2010) found that individuals with more need for cognition appear to be less susceptible to the influence of anonymous commenters compared to those less prone to effortful thinking. However, the social influence effect found in this study was due to only presenting participants with negative comments and researchers did not test whether participants had strong or weak pre-existing opinions towards the topic of the article, thus making it difficult to quantify a systemic shift in opinion before and after exposure to user comments.

Some of these limitations are present among studies within this domain. At times, these investigations adopt between-subject designs, focusing on a single topic per condition, and often lack systematic measures to test pre-existing attitudes and opinions towards both the issue and the specific news item. Another potential limitation is that the stimuli employed in some studies, while attempting to increase ecological validity by mimicking as closely as possible the overwhelming nature of social media environments, may include non-pertinent information, potentially introducing confounding variables that complicate the isolation of the effects of user-generated comments. These constrains require the necessity for more refined experimental designs aimed at isolating and precisely quantifying the effect of interest. In our pursuit to address these constrains, our objective was to craft a novel task capable of systematically assessing opinions both before and after exposure to user comments. Our task was designed to quantitatively measure not only general pre-existing attitudes toward the different topics used but also the initial opinions regarding each stimulus employed before participants would be potentially influenced by the experimental manipulation, such as the different valence of the comments presented below the news posts. Therefore, by computing a quantitative index of opinion updating, we would not only be able to evaluate the degree to which individuals are influenced by online comments but also to explore correlations with diverse variables,

allowing us to discern the conditions that increase susceptibility to be influenced by online comments.

2.1.2 Aim of the Project

This project aimed to investigate the impact of user-generated comments written below news headlines posted on social media on individuals' opinions regarding controversial contemporary topics. Our main objective was to develop and validate a novel behavioral task capable to quantify changes in opinion following exposure to other people's comments. To achieve this, we aimed to systematically measure opinions before and after exposure to other people's opinions (written in form of comments to the news posts). Our goal was to replicate previous findings on opinion change following exposure to user-generated comments and bridge the gaps in literature mentioned in the previous section. Several hypotheses were formulated. First, we hypothesized that participants would adjust their personal opinions in the direction of the sentiment of the comments, which could be supportive, critical or mixed. Second, we hypothesized that greater shifts in personal opinions would follow incongruent comments to participants' personal opinions, compared to congruent comments. Third, we hypothesized that the strength of participants' pre-existing attitudes towards the topics would modulate the magnitude of opinion updating, with stronger attitude towards the topics leading to smaller opinion changes. Fourth, we hypothesized that participants' confidence in their opinion would diminish following incongruent comments to own's opinion compared to congruent comments. Lastly, building on the conceptualization of digital maturity by Laaber and colleagues (2023), we expected that high levels of digital maturity would act as protective factors against the social influence of online comments.

2.2 Creation of a New Social Influence Paradigm

With this novel social influence task, we aimed at quantifying the degree of social influence participants would experience when reading comments written by others about news headlines posted on social media. In the following sections, we describe in details the rationale behind the stimuli creation and task design.

2.2.1 Creation of the Stimuli

In order to address how online comments can influence opinion formation when reading news on social media, we had to first determine what social media platform would be the most suitable for our experiment. At the time of the task development in 2021, Facebook, YouTube and Instagram have been the most popular social media platforms with respectively 71%, 74% and 38% Americans using them (Shearer & Grieco, 2019). However, when we looked at social media as a pathway to access news, Facebook was the leading platform with 52% of Americans reading news on it, compared to 28% on YouTube and 17% on Twitter (Pearson, 2021). As a result, we chose to create our stimuli following the Facebook layout because it was the most popular social media to read news on. Moreover, the use of Facebook-like posts aligned with recent studies on online social conformity (e.g., Colliander, 2019; Wijenayake et al., 2020). To accomplish this, we used an online Facebook post generator to mimic the layout of a Facebook post, including the comment section.

To maximize ecological validity, we chose to use actual online news headlines posted on Facebook by well-known online newspapers. To create the stimuli, we searched for reputable online outlets that routinely posted news on Facebook and that had a broad audience. This would also ensure that these outlets would better represent what people encounter in their everyday life on social media. To collect online news, we chose well-known newspapers such as the New York Times, CNN, The Guardian, and Independent, among others. These online newspapers were selected because they received high trustworthiness scores from professional fact-checkers and are thought to have higher editorial standards than other untrustworthy hyperpartisan outlets (Pennycook et al., 2021).

After selecting online newspapers that regularly share news on Facebook, we had to decide which topics to use. These should have been well-known, debatable topics on which people may disagree, rather than extremely polarizing themes like politics and religion. Indeed, past research has shown that, when confronted with these extremely polarizing topics, individuals may be motivated to update their opinions differently than on other topics in other to maintain their identity and group affiliation (Anglin, 2019). We opted to focus on three controversial contemporary issues: climate change, vaccination and

veganism. These are very renowned topics that have sparked continuing global discussions in recent years and are relevant for current socioeconomic and environmental challenges. Therefore, news headlines about these themes were likely to elicit a variety of opinions from participants in the experiments. After determining which topics to employ, we searched the online Facebook pages of these newspapers for relevant news articles concerning the three selected issues. We aimed at finding news headlines that highlighted states of belief rather than factual events, as the first would make it easier to provoke thoughts and generate opinions about them. We gathered 9 suitable news headlines: 3 for climate change, 3 for vaccination, and 3 for veganism.

The next step was to collect real comments from Facebook users about these 9 news headlines. Since we picked large and popular online outlets, we were able to access a wide range of comments for each news headline, ranging from argumentative to subjective and from civil to uncivil. We chose to collect only argumentative, civil and very clear comments, as they have been demonstrated to be more persuasive for the users than subjective comments (Fabian et al., 2018; Winter et al., 2015). The relevant, suitable comments were sorted into two categories: supporting and opposing the specific news headline. For each news headline, we collected 4 supporting and 4 opposing comments. We balanced the lengths of the supporting and opposing comments to ensure that they were regarded as equally compelling, since longer comments may be perceived as more argumentative and therefore more persuasive (Wood et al., 1985). However, in editing the length of the comments, we did not modify any text; instead, we ensured that supportive and opposing comments would have roughly similar length.

After gathering all the necessary comments for each of the 9 news headlines, we created fake Facebook posts using an online Facebook post generator (https://generatestatus.com/fake-facebook-post-generator/) that provides the standard Facebook post layout, while allowing the customization of the image, text and comments to the post. This tool allowed us to control and create stimuli that were consistently similar to one another, reducing the need for extensive editing to remove user reactions or other users' responses to the selected comments. Each stimulus had the layout of a snack news, such as a picture, the news headline (below the picture), and a brief description of the article written by the online newspaper (above the article). Moreover, for each of the

9 news headlines, we crafted multiple comment sections with 4 comments below the post as the main experimental manipulation (see **Fig. 1** for an example of the stimuli created).

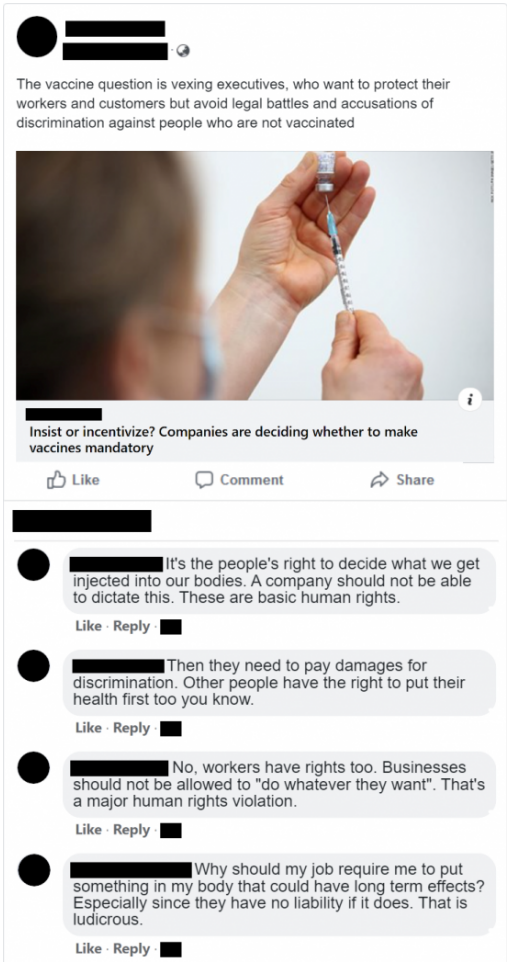


Fig. 1: Example of the stimuli created for the new paradigm.

After creating the Facebook posts, to control for possible confounding effects, we proceed to hide any information that was not relevant for our study. We covered the name and logo of the online newspaper, as well as the name and picture of each Facebook user who commented on that post, to avoid source credibility bias (Hohenberg & Guess, 2022; Nadarevic et al., 2020) and gender and race biases (Hawkins et al., 2023). Moreover, we removed the numbers of likes and shares for each post, as previous studies showed mixed evidence about the effect that likes have on how people perceive online posts. For instance, some studies found that posts with more likes were judged to be more reflective of the public opinion and could more easily trigger the “Bandwagon Effect” (Kim 2018; Lee & Oh, 2017; Xu 2013). Other studies, instead, found that likes were not perceived as a

clear representation of the public climate, because the interpretation of the number of likes is subjective and context-dependent (Lee & Jang, 2010; Neubaum & Krämer, 2016). In total, for each of the 9 news headlines, we developed 4 different versions: one with 4 supportive comments (for the "supportive condition"), one with 4 opposing comments (for the "opposing condition"), and two different versions with 2 supportive and 2 opposing comments ("mixed condition").

2.2.2 Pilot I: Assessing the Valence of Comments

In order to validate the valence of the comments in reflecting the intended manipulation, we conducted a first pilot study where participants rated the valence of the stimuli we collected. The pilot study was implemented using the platform Qualtrics (Qualtrics, Provo, UT) and participants were recruited through the platform Amazon MTurk in May 2021. MTurk is a popular crowdfunding platform that has been highly involved recently in scientific data collection due to its speed and accessibility. However, there are some potential issues with its use that we attempted to address by implementing different measures. For example, due to the anonymous nature of the platform, it may be more challenging (compared to lab settings) to ensure that participants complete tasks accurately. To address these possible concerns and increase the validity and reliability of our data, we used several quality control measures, such as attention checks and exclusion criteria. Moreover, we restricted participation to MTurk workers from the United States who had completed at least 1000 tasks successfully on the platform and had a minimum 98% approval rating. A total of 41 participants were recruited, with 4 being subsequently excluded based on predefined exclusion criteria (final sample: N = 37, 13 female, M_{age} = 34.5, SD_{age} = 12.9). The exclusion criteria included going through the survey at an unreasonable pace and rating the comments' valence inconsistently with the actual sentiment of the comment (for example, rating as "opposing" a comment that supported the news headline).

At the beginning of the task, participants were presented with a news headline accompanied by 4 comments. They were instructed to rate the valence of the 4 comments with a scale from -7 to +7, where -7 indicated that the 4 comments were considered strongly opposing to the content of the news headline, while +7 signified strong support. A rating of 0 indicated that the 4 comments were evenly balanced between supportive and

critical towards the news headline (for example, with two opposing and two supportive comments). The pilot study lasted 15 minutes and participants were compensated 3$ and the possibility of an extra dollar for good performance. Good performance was defined as, for example, rating as positive a supportive comment and rating as negative an opposing comment, demonstrating that participants really read the content of the comments presented. Results from Pilot I confirmed that participants were able to distinguish the valence of the provided comments. Supportive comments received positive ratings (M = 4.6; SD = 3.2), opposing comments received negative ratings (M = -5.2; SD = 2.5), and mixed comments fell in between (M = -0.4; SD = 2.4).

2.2.3 Pilot II: Choosing the Task Designs

Following stimuli validation, we conducted another pilot study to assess what study design would be best suited to use. The goal of the final behavioral task was to measure the impact of other people's opinions expressed in the comments on one's own opinion formation and updating. We developed two distinct study designs to compare and contrast, since each had its own set of advantages and disadvantages. We called them *Sequential Design* and *Block Design*. On one hand, in the *Sequential Design*, participants would first rate their opinion to a single news item without the comments, immediately followed by a rating on the same news headline with comments. On the other hand, in the *Block Design*, participants would rate their opinions on all of the gathered news headlines without comments ("news headlines block") before moving on to the block that displayed all of the news headlines now paired with the comments ("comments block"). One advantage of the Sequential Design is its ecological validity, since individuals in real life view one news headline and then instantly the comment section of that same news item. Because there is more time between the first and second opinion ratings, the Block Design, on the other hand, has the advantage of reducing the *anchoring effect* (Tversky & Kahneman, 1974) which occurs when participants tend to anchor their second opinion rating to their first one, as well as the *need for consistency* (Festinger, 1957) which is the tendency to behave consistently and be resistant to change. However, one disadvantage of the Block Design is that it not only has lower ecological validity, but it also adds a lot of information between the first and second opinion ratings (such as all the other news headlines). This would make it difficult to discern the effect of the comments from that of

the other news headlines. Therefore, the second pilot study was conducted to assess which of the two designs would lead to stronger behavioral effects.

On the platform MTurk, 27 participants from the United States (9 female, M_{age} = 38.5, SD_{age} = 12.9) were recruited for the Sequential Design and 26 participants (7 female, M_{age} = 39.8, SD_{age} = 14) for the Block Design. Two participants were removed from the analyses of the Sequential Design and one from the Block Design due to the exclusion criteria (same as in Pilot I). The Sequential Design demonstrated superiority in evoking social influence both after supportive comments (V = 561.5, p = 0.001) and after opposing comments (V = 103.5, p = 0.01), as opposed to only after supportive comments in the Block Design (V = 19, p = 0.02). Consequently, the Sequential Design was selected due to its greater ecological validity and its ability to elicit a stronger behavioral effect.

2.3 Behavioral Study I and II: Validation of the New Social Influence Paradigm

The first behavioral study was performed in order to validate the new social influence paradigm with a wider participants sample. Before collecting data, we pre-registered our planned study design, hypotheses and statistical analyses at OSF (https://osf.io/5dm7h, date of pre-registration: September 20, 2021). In the results section, we stated any variations from the pre-registered statistical analyses, as well as any exploratory analyses that were not pre-registered. The second behavioral study was conducted as part of the DIGYMATEX project. In this second study, our first goal was to assess whether we could reproduce the previous results with a local sample from Germany and observe its robustness across populations with different characteristics. In fact, sampling from different countries to evaluate a task can improve external validity and generalizability of the findings, since the increased heterogeneity (Demerouti & Rispens, 2014). The second objective was to investigate the relationship between social influence and the construct of digital maturity, as measured by the adult-adapted version of the Digital Maturity Inventory (DIMI) questionnaire (Laaber et al., 2023).

2.3.1 Participants

For the first behavioral study, we recruited 240 American participants from Amazon MTurk, who completed the experiment through their web browsers between September and October 2021. We used a power analysis in G*Power 3 (Faul et al. 2007) to calculate

the necessary sample size of 200 participants. Using the typical 0.05 alpha error probability, we aimed to detect a medium effect size of 0.24 with 0.95 power. The effect sized was computed in G*Power 3 using a Wilcoxon signed-rank test (matched pairs) and comparing the means of the opinions given by participants of the pilot study before and after reading comments from other people (both supportive and critical comments were used). As a result, we determined a minimal sample size of 200 participants for a within-subject design. On MTurk, we increased our recruitment sample size to 240 since some researchers have found suspicious responses that might be from bots or non-serious participants (Paolacci, Chandler, & Ipeirotis, 2010). Based on our manipulation checks and exclusion criteria, we excluded 46 datasets, leaving n = 194 participants (80 female; 9 between 18-25 years old, 33 between 26-30 years old; 91 between 31-40 years old, 61 above 40 years old) in the statistical analyses.

For the second behavioral study, we recruited 221 individuals (143 female; 147 between 18-25 years old; 53 between 26-30 years old; 21 above 31 years old) through the database of the BonnEconLab (University of Bonn, Germany) for the same online experiment. Registration in this database is voluntary and the pool is mainly made of University of Bonn students and staff. Participants were at least 18 years old and fluent in the English language (approximately B2 in the Common European Framework of Reference).

2.3.2 Study Design and Experimental Procedure

The two behavioral studies shared the same study design and experimental procedure, with the only exceptions that the second behavioral study also involved the use of the DIMI questionnaire. We employed 3 out of the 9 possible news headlines that we developed and validated during the first pilot study, one for each topic of interest. The stimuli were Facebook posts of news headlines with four comments written by other Facebook users (see Section 2.2.1 for further information on the stimuli). Each news headline was accompanied by one of the three variations of the comment section: four supportive comments, four opposing comments, or two opposing and two supportive comments. The valence of the comments (i.e. supportive, opposing, mixed) constituted the primary experimental manipulation. On Qualtrics, participants started the online social influence task by reading the instructions which explained the purpose of the study and the exclusion criteria. Participants were excluded if they: moved through the survey at an

unrealistic pace (e.g. if they took less than 3 seconds to read the comments); failed to properly answer the attention checks within the task; filled in the attitudinal questionnaire inconsistently (e.g. always giving the same answer even to reverse items). Easy attention checks were also included in the instructions, allowing the platform Qualtrics to automatically exclude participants who failed to correctly answer them. After reading the instructions, participants completed three validated attitudinal questionnaires about the three contemporary topics used in the task: climate change (Christensen & Knezek, 2015), vaccination (Martin & Petrie, 2017) and veganism (Paslakis et al., 2020). This was done to distinguish between participants who held extreme viewpoints and those who held moderate or weak positions on the topics of the news headlines. Furthermore, the scores obtained from these three questionnaires were used to classify participants as either supporting or opposing the topic of the following news headlines. Specifically, if a participant obtained a score of at least 50% of the possible total score, they were classified as supportive of a certain topic. For example, if a questionnaire had ten items with a four-point Likert scale, the highest possible score would be 40. As a result, an individual who had a score of 20 or higher on that questionnaire would be considered supportive of that specific topic. These were considered “pre-existing attitude scores” towards the three topics. In the second behavioral study, participants additionally filled in the DIMI, a validated questionnaire on the use of digital devices and the internet (Laaber et al., 2023).

The three pre-existing attitude scores were subsequently used to determine what variation of the comment section would be displayed to participants below the news headlines. The variations of the comment section could be: four supportive comments towards the news headline, four opposing comments towards the news headline, two supportive and two opposing comments towards the news headline. We designed the study such that, based on the scores of the attitudinal questionnaires, all participants would be confronted: i) one time with 4 comments that were congruent to their initial attitude towards the topic, ii) one time with 4 comments that were incongruent to their initial attitude towards the topic, and finally iii) one time with 2 congruent comments and 2 incongruent comments. This approach was adopted to prevent participants from potentially being exposed only to comments they would agree or disagree with, or only to mixed comments. Therefore, a partial-randomization of the comments was implemented in Qualtrics to ensure that each

participant would encounter, at least once, comments in agreement with their attitude based on the attitude questionnaire, comments in disagreement, and mixed comments.

At the beginning of the task, a news headline without the comment section was displayed. Participants were asked to rate their opinion about the content of the news headline on a slider ranging from - 7 (= strongly oppose) to + 7 (= strongly support). This rating was coded as R_1, indicating participants' prior opinion on the news headline. Participants were then asked to rate their level of confidence in their previous rating on a scale from 0 (= not confident at all) to 100 (= absolutely confident). The first confidence rating was coded as C_1. After the two ratings were provided, four comments appeared below the same news headline. Participants were instructed to carefully read the comments and then give the opinion and confidence ratings a second time (coded as R_2 and C_2 respectively) (see **Fig. 2**). The names for these ratings are inspired from De Martino et al., 2017. This procedure was repeated for the other two news headlines. At the end of the task, we gathered demographics data such as sex and gender, age, education, political affiliations, Facebook and social media general usage. The study lasted between 20 and 25 minutes and participants were compensated with 6€ for their time.

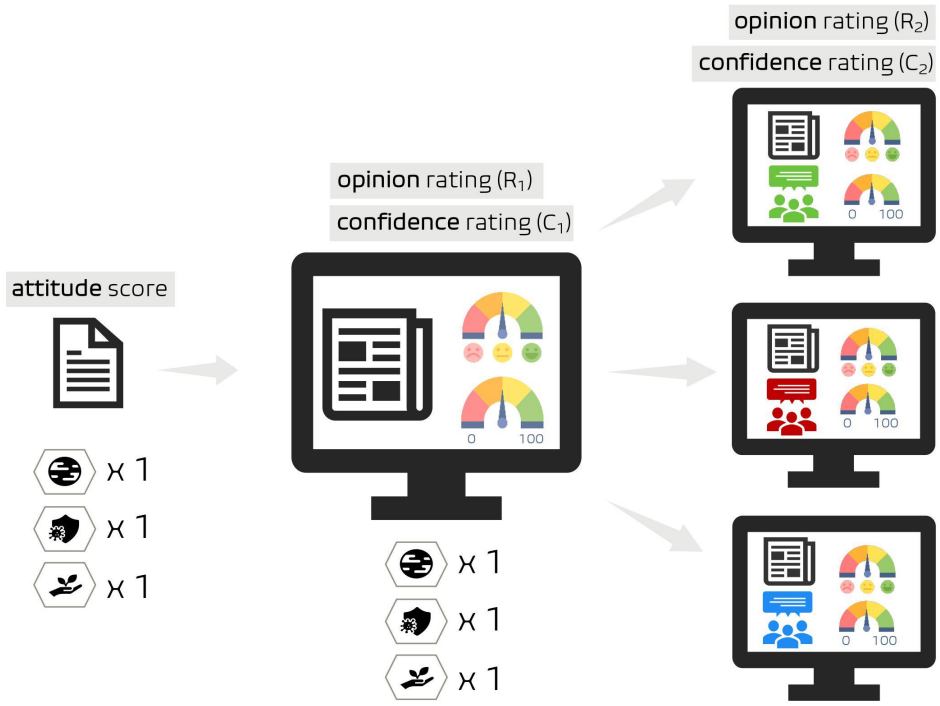


Fig. 2: Illustration of the experimental procedure. **(left)** Participants filled out three attitudinal questionnaires about the three topics. **(center)** At the beginning of the task, they read a news headline and rated their opinion and confidence on two rating scales. **(right)** They then read four comments from other users and rated their opinion and confidence for a second time. The procedure was repeated three times for the three topics.

2.3.3 Statistical Analyses

Statistical analyses for both behavioral studies were conducted with R language (version 4.3.2) and RStudio (version 2023.12.1.402, Posit Team, 2024), as well as the software Jasp (version 0.18.2, JASP Team, 2023) for correlation matrices. The R packages employed for data cleaning, analysis and visualization included: *BayesFactor, cowplot, ggpubr, gmodels, Hmisc, kableExtra, lme4, nlme, plotrix, readxl, reshape2, see, tidyverse.* All statistical analyses were conducted using a two-tailed test and a significance threshold of $\alpha \leq 0.05$.

Deviation from the pre-registration (Study I)

As stated in the pre-registration of Study I, we planned to solely employ one-way ANOVA tests to analyze the main hypotheses of the behavioral study. However, for both

behavioral studies, we ultimately opted to run Mixed-Effects Linear Model analyses, due to the non-normal distribution of our data and because of the robustness of these models against assumption violations. This was a more detailed and flexible approach to modeling the relationship between dependent and independent variables, because it allowed analyzing single-trial data and including both fixed and random effects, resulting in a more accurate and comprehensive interpretation of the data. Therefore, these models were considered most suitable to address our research objectives.

Effects of comments valence (Study I and II)

To determine the presence of a significant difference in opinion ratings before and after presenting the comments, as well as to assess the effectiveness of presenting mixed comments as a control condition, we performed three Paired Wilcoxon tests (one for each level of the explanatory variable: supportive, opposing, mixed) between the first and the second opinion ratings. Subsequently, to examine whether the valence of the comments could predict the direction of the opinion shifts, opinion adjustments were computed (i.e., $R_2 - R_1$) and a Mixed-Effects Linear Regression Model was performed using the function *lme* from the R package *lme4*. In this model, the opinion adjustment ($R_2 - R_1$) was employed as dependent variable, the valence of the comments (i.e. supportive, opposing, mixed) as the explanatory variable, and random intercepts were included to account for individual differences in the average opinion ratings (see **Eq. 1**):

(Eq. 1). $\textit{Opinion adjustment}_{ij} = \beta_0 + \beta_1 \textit{Comment valence}_{ij} + u_j + \epsilon_{ij}$

The subscript *j* indicated the specific participants, instead the subscript *i* indicated the individual observations per participant. $\textit{Opinion adjustment}_{ij}$ represents the dependent variable and $\textit{Comment valence}_{ij}$ represents the independent variable with three levels. u_j indicates the participant-specific random intercept and ϵ_{ij} the residual.

Effects of congruence between initial opinion and comments' valence on opinion updating (Study I and II)

In this pre-registered analysis, our objective was to evaluate the effect of the congruence between participant initial opinion rating and the opinion expressed in the comments

presented below the news headline on opinion updating. Specifically, when participants' first opinion rating (R_1) and the opinion expressed in the comments were either both supportive or both opposing towards the news item, we coded this state as "*congruent condition*". On the contrary, an incongruent "participant-comments" opinion was classified as "*incongruent condition*". However, we decided for a deviation from the pre-registration regarding the levels of the variable congruence. Initially, we had planned to compute a binary index with only two levels (i.e. *congruent* and *incongruent*), leaving out the state where participants were exposed to the control condition. Subsequently, we made the decision to introduce a third level called "*mixed*" in order to create an ordinal variable. This decision was made to better capture the whole range of participants' responses and also to include data from the control condition in this analysis. Therefore, also in this analysis, we decided to define the control condition as "*mixed*", since participants read two supportive comments and two opposing comments. As a result, deviating from the pre-registration, we computed the congruence index as an ordinal variable with three levels: *congruent*, *mixed* and *incongruent*. To investigate the influence of the congruence between participants' initial opinion and the opinion in the comments on opinion updating, we run a Mixed-Effects Linear Regression Model. In this model, the absolute value of opinion updating (i.e., $|R_2 - R_1|$) was utilized as a dependent variable, and the congruence between participants' initial opinion and the opinion in the comments was employed as an explanatory ordinal variable. Random intercepts were included to account for individual differences in the average opinion ratings (see **Eq. 2**):

(Eq. 2). $\textit{Opinion updating}_{ij} = \beta_0 + \beta_1 \textit{Congruence}_{ij} + u_j + \epsilon_{ij}$

Effects of congruence between initial opinion and comments' valence on confidence (Study I and II)

Here, we aimed at assessing the effect of the congruence between the first opinion rating and the subsequent comments on participants' confidence in their opinion. To measure confidence shifts after participants read others' opinions, we computed confidence adjustments as the difference between the second and the first confidence ratings (i.e. $C_2 - C_1$). We then run a Mixed-Effects Linear Model with the confidence adjustment as the

dependent variable and the congruence between the initial opinion rating and the comments' valence as the independent ordinal variable with three levels: *congruent*, *mixed*, *incongruent*. We used random intercepts to control for individual differences in the average confidence ratings (see **Eq. 3**):

(Eq. 3). $Confidence_{ij} = \beta_0 + \beta_1 Congruence_{ij} + u_j + \epsilon_{ij}$

Effects of pre-existing attitudes towards the topics of the news headlines (Study I and II)

Here, we investigated the relationship between pre-existing attitudes towards the three topics of the news headlines and the amount of opinion updating using a Linear Regression analysis. Participants' attitudes towards these topics were assessed using the three attitudinal questionnaires filled out prior the beginning of the task. Due to the variations in the number of items and the different Likert scales across the three questionnaires, the resulting scores were standardized into z-scores to obtain a unique index of pre-existing attitudes that ranged from 0 (weak attitude towards the topic) to 50 (strong attitude towards the topic). We performed a Linear Regression, where we included pre-existing attitude as the exploratory variable and opinion updating ($|R_2 - R_1|$) as dependent variable (see **Eq. 4**):

(Eq. 4). $Opinion\ updating_{ij} = \beta_0 + \beta_1 Attitude\ score_{ij} + \epsilon_{ij}$

Effects of digital maturity on opinion updating (Study II)

Here, our objective was to observe the relationship between opinion updating and digital maturity, assessed through the use of the DIMI questionnaire. Specifically, we were interested in assessing whether individuals with more digital maturity would be less susceptible to the influence of others' opinions. We run Correlation analyses between opinion updating ($|R_2 - R_1|$) and the DIMI with all its sub-categories, such as: *Autonomy of choice, Autonomy within digital contests, Digital literacy, Risk awareness, Controlling negative emotions, Controlling aggressive emotions, Support, Respect, Citizenship*.

Assessing data sensitivity with Bayesian hypothesis testing (Study I and II)

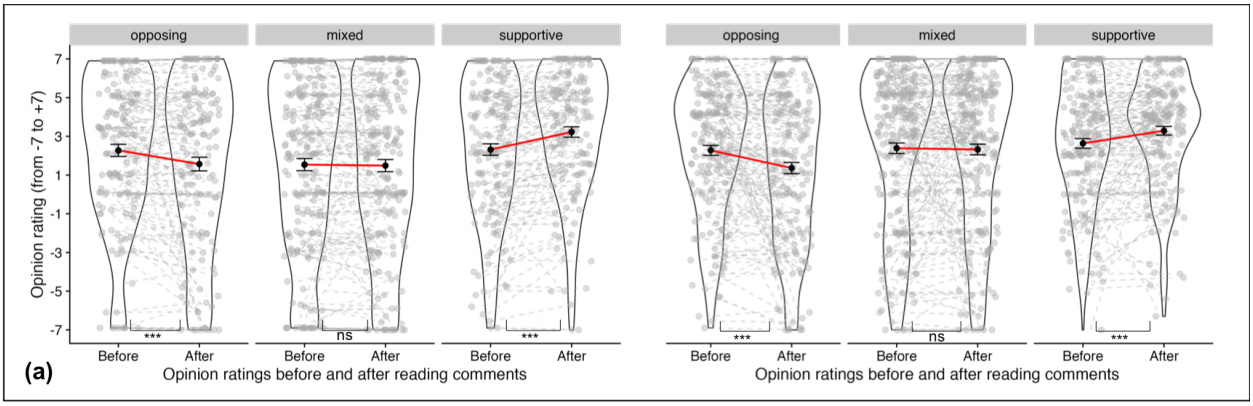
In our behavioral analyses, we also employed Bayesian hypothesis testing to investigate the effects of the above explanatory variables on opinion updating. In this sub-section, we summarize statistical concepts from Dienes (2014). Traditional frequentist methods, such as the null-hypothesis significance testing (NHST), rely on p-values to determine the likelihood of observing the data if the null hypothesis (H0) is true, being able to only provide evidence against H0. If this probability is below a certain threshold (usually 0.05), the alternative hypothesis H1 is accepted over the null hypothesis H0. However, the frequentist approach does not clarify whether non-significant results support H0 or simply reflect data insensitivity. In contrast, Bayesian hypothesis testing measures and evaluates evidence for both H0 and H1. For this reason, in our hypotheses testing, we additionally employed the Bayes Factor (BF) to assess the level of evidence for the alternative hypothesis over the null hypothesis. Specifically, following an often-used convention attributed to Harold Jeffreys (see Dienes 2014), a BF greater than 3 indicated that there was significant evidence for H1 over H0, while a BF less than 1/3 indicated that there was significant evidence for H0 over H1. A BF between 1/3 and 3 indicated that the data was insensitive to distinguish between H1 and H0, so that no firm conclusion could be drawn.

2.3.4 Results

Effects of comments valence (Study I and II)

In both studies, participants significantly adjusted their second rating to conform with the other users' opinions (see **Figure 3a** left and right for Study I and II, respectively). Specifically, participants shifted their opinions in the direction of the social information both after reading supportive (Study I: $V = 3395$, $p < .001$, $BF > 100$; Study II: $V = 4812$, $p < .001$, $BF = 41.88$) and opposing comments (Study I: $V = 9880$, $p = .001$, $BF = 56.48$; Study II: $V = 11538$, $p < .001$, $BF > 100$). Additionally, as we were expecting from the control condition, participants did not statistically change their opinion after reading mixed comments (Study I: $V = 7573$, $p = .8$, $BF = 0.08$; Study II: $V = 7923$, $p = .4$, $BF = 0.078$). Next, we computed the opinion adjustment index as the difference between the second and first opinion ratings (i.e. $R_2 - R_1$). In both studies, we observed a negative adjustment following exposure to opposing comments (Study I: $M = -0.69$, $SD = 2.59$, 95% *CI* [-1.06,

-0.32]; Study II: *M* = -0.93, *SD* = 3.48, 95% *CI* [-1.38, -0.47]) and a positive adjustment when exposed to supportive comments (Study I: *M* = 0.84, *SD* = 1.97, 95% *CI* [0.56, 1.12]; Study II: *M* = 0.62, *SD* = 2.52, 95% *CI* [0.28, 0.95]). However, the rating remained largely unchanged after being exposed to mixed comments (Study I: *M* = -0.05, *SD* = 2.16, 95% *CI* [-0.35, 0.25]; Study II: *M* = -0.6, *SD* = 3.32, 95% *CI* [-0.50, 0.37]) (see **Fig. 3b** left and right for Study I and II, respectively). Indeed, the valence of the comments that participants were exposed to significantly predicted the opinion adjustments in both studies (Study I: $F_{(386)}$ = 23.86, p < .0001, *BF* > 100; Study II: $F_{(440)}$ = 13.76, p < .0001, *BF* > 100), such that: opposing comments towards the news item significantly predicted subsequent negative opinion shifts (Study I: β = -0.65, *SE* = 0.22, 95 % *CI* [-1.09, -0.21], $t_{(386)}$ = -2.929, p = .003; Study II: β = -0.87, *SE* = 0.29, 95 % *CI* [-1.44, -0.29], $t_{(440)}$ = -2.949, p = .003), whereas supportive comments significantly predicted subsequent positive opinion shifts (Study I: β = 0.91, *SE* = 0.22, 95 % *CI* [0.47, 1.35], $t_{(386)}$ = 4.081, p = .0001; Study II: β = 0.69, *SE* = 0.29, 95 % *CI* [0.11, 1.27], $t_{(440)}$ = 2.33, p = .02), compared to mixed comments.



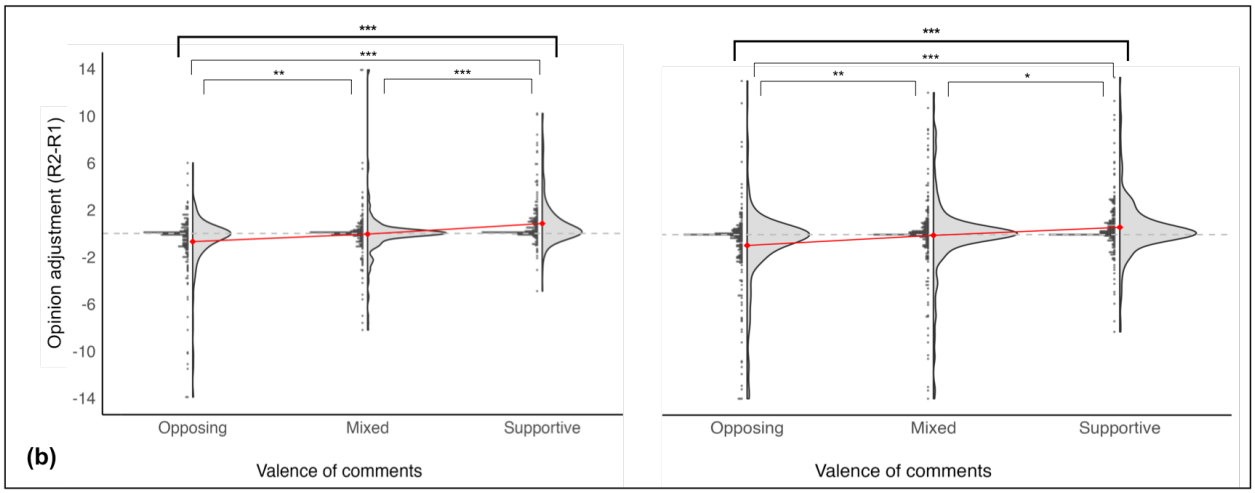


Fig. 3: (a) Effect of the comments' valence on participants' second opinion rating in Study I (left) and Study II (right). The black dots are the mean values across participants, the grey dots are the individual mean values and the error bars represent the standard error of the mean, and the contours represent Kernel density plots **(b)** Opinion adjustments computed as the difference between the first and second opinion ratings in Study I (left) and Study II (right). The red dots are the mean values across participants and the contours represent Kernel density plots.

Effects of congruence between initial opinion and comments' valence on opinion updating (Study I and II)

In this analysis, our objective was to evaluate whether participants would exhibit a greater inclination to update their opinions when exposed to comments that did not align with their first opinion rating, as opposed to comments that aligned. When computing the index for opinion updating (i.e. $|R_2 - R_1|$), we observed a gradual pattern of opinion updating across the levels of the dependent variable. In both studies, we found that the magnitude of opinion updating was the smallest following exposure to congruent comments (Study I: *M* = 0.831, *SD* = 1.04, 95% *CI* [0.67, 0.98]; Study II: *M* = 1.25, *SD* = 1.77, 95% *CI* [1.02, 1.48]), the largest following exposure to incongruent comments (Study I: *M* = 1.62, *SD* = 2.63, 95% *CI* [1.26, 1.99] ; Study II: *M* = 2.10, *SD* = 3.29, 95% *CI* [1.66, 2.54]), and intermediate following exposure to mixed comments (Study I: *M* = 0.99, *SD* = 1.92, 95% *CI* [0.71, 1.26]; Study II: *M* = 1.76, *SD* = 2.81, 95% *CI* [1.39, 2.13]) (see **Fig. 4**). Indeed, the congruence between the participants initial opinion and the comments predicted the subsequent rating behavior in both studies (Study I: $F_{(386)} = 9.54$, $p < .0001$; Study II: $F_{(440)} = 6.33$, $p = .0002$), such that: incongruent information significantly increased the magnitude of the updating behavior compared to congruent information (Study I: $\beta = 0.82$, *SE* = 0.20, 95 % *CI* [0.43, 1.22], $t_{(386)} = 4.09$, $p < .001$; Study II: $\beta = 0.833$, *SE* = 0.24, , 95 % *CI* [0.37, 1.29], $t_{(440)} = 3.52$, $p < .001$). The magnitude of updating significantly differed between congruent and mixed information in Study II but not in Study I (Study I: $\beta = 0.17$, *SE* = 0.20, 95 % *CI* [-0.22, 0.57], $t_{(386)} = 0.8{,}7$ $p = .39$; Study II: $\beta = 0.50$, *SE* = 0.22, 95 % *CI* [0.07, 0.94], $t_{(440)} = 2.26$, $p = .024$); and the difference between incongruent and mixed information significantly differed in Study I but not Study II (Study I: $\beta = 0.65$, *SE* = 0.19, 95 % *CI* [0.27, 1.03], $t_{(386)} = 3.34$, $p < .001$; Study II: $\beta = 0.33$, *SE* = 0.22, 95 % *CI* [-0.11, 0.77], $t_{(440)} = 1.48$, $p = .138$).

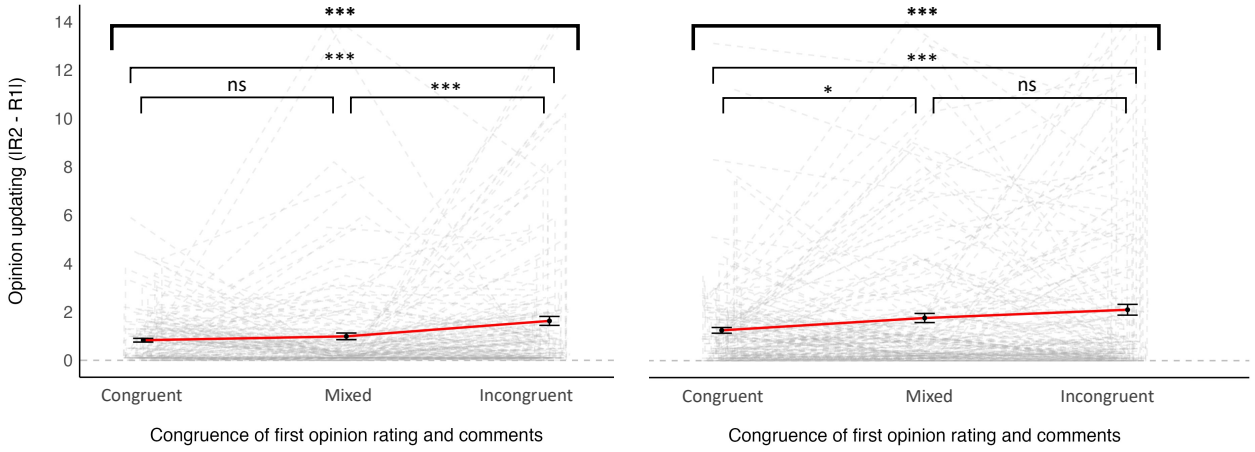


Fig. 4: Effect of congruence of first opinion rating and valence of comments on opinion updating in Study I (left) and Study II (right). The black dots are the mean values across participants, error bars represent the standard error of the mean. The red lines connect the mean values, the grey dotted lines represent the individual participants. ns, not significant.

Effects of congruence between initial opinion and comments' valence on confidence (Study I and II)

In this analysis, our objective was to observe how participants' confidence would change considering the type of comments they read. Specifically, we computed a confidence adjustment index (i.e. $|C_2 - C_1|$) to assess whether participants' confidence would increase with information that aligned with their opinion compared to information that did not align with it. Indeed, we found that in both studies the congruence of the first opinion rating and the comments' valence predicted the confidence participants had in their second judgments (Study I: $F_{(386)} = 5.94$, $p = .003$; Study II: $F_{(440)} = 8.55$, $p < .001$). Specifically, confidence decreased after incongruent information compared to congruent information (Study I: $\beta = -4.44$, $SE = 1.29$, 95 % CI [-6.99, -1.90], $t_{(386)} = -3.427$, $p < .001$; Study II: $\beta = -5.85$, $SE = 1.52$, 95 % CI [-8.82, -2.88], $t_{(440)} = -3.86$, $p < .001$). In Study I but not in Study II, there was a significant decrease in confidence from mixed information to incongruent information (Study I: $\beta = -2.54$, $SE = 1.27$, 95 % CI [-5.03, -0.05], $t_{(386)} = -2.00$, $p = .046$; Study II: $\beta = -1.01$, $SE = 1.52$, 95 % CI [-3.99, 1.98], $t_{(386)} = -0.66$, $p = .81$), and in Study II but not in Study I there was a significant increment in confidence from mixed information to congruent information (Study I: $\beta = 1.90$, $SE = 1.30$, 95 % CI [-0.65, 4.46], $t_{(386)} = 1.46$,

p = .144; Study II: β = 4.84, SE = 1.51, 95 % CI [1.88, 7.80], $t_{(440)}$ = 3.20, p = .001) (see **Fig. 5**).

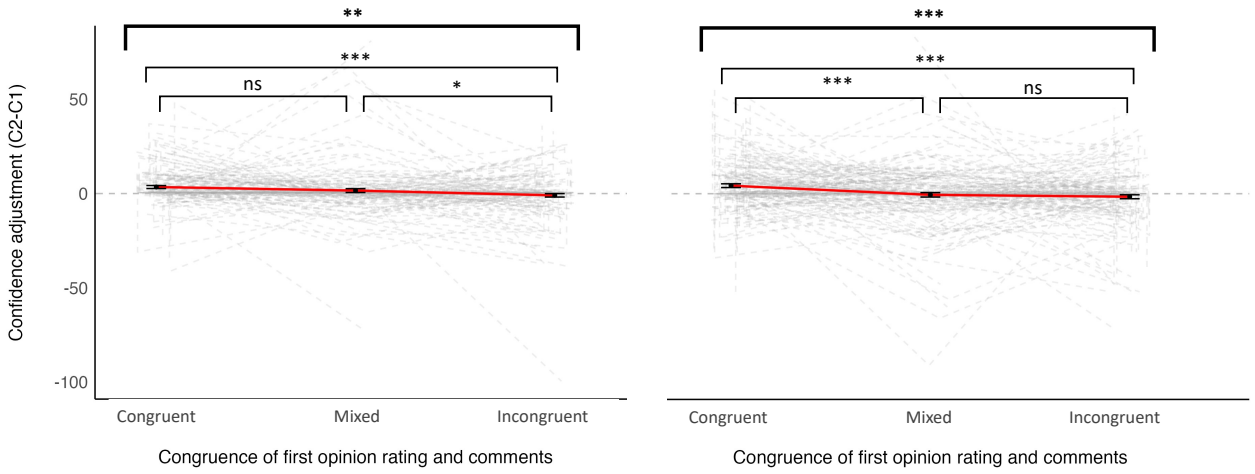


Fig. 5: Effect of congruence of first opinion rating and valence of comments on confidence in Study I (left) and Study II (right). The black dots are the mean values across participants, error bars represent the standard error of the mean. The red lines connect the mean values, the grey dotted lines represent the individual participants. ns, not significant.

Effects of pre-existing attitudes towards the topics of the news headlines (Study I and II)

In this analysis, we aimed at assessing whether having a stronger pre-existing attitude towards the three topics of the task would decrease the magnitude of participants' opinion updating. In both studies, we found that climate change was the topic where people held stronger pre-existing attitudes (Study I: M = 33.82, SD = 13.608; Study II: M = 38.52, SD = 11.138), followed by vaccination (Study I: M = 27.33, SD = 14.98; Study II: M = 22.41, SD = 12.452), and then by veganism (Study I: M = 11.68, SD = 9.245; Study II: M = 12.37, SD = 8.551). We then performed Linear Regression analyses to assess whether the magnitude of pre-existing attitudes towards the topics would have an effect on opinion updating. These analyses did not yield a significant result (Study I: β = -0.004, SE = 0.005, 95 % CI [-0.001, 0.007], $t_{(580)}$ = -0.92, p = 0.35; Study II: β = -0.002, SE = 0.006, 95 % CI [-0.01, 0.01], $t_{(440)}$ = -0.31, p = 0.76). Because the non-significant result supported the null-hypothesis H0, we performed a Bayesian analysis to discern whether the non-significant result provided evidence for the null-hypothesis or indicated data insensitivity. The BF of 0.14 and 0.09 (Study I and II, respectively) provided strong evidence for the null-

hypothesis, thus corroborating the result coming from the frequentist approach. Although we could not predict the amount of opinion updating depending on the topic pre-existing attitudes, in a following exploratory analysis, we found a negative correlation between the two variables, such that participants updated their opinions less when they had stronger pre-existing attitudes compared to when they had weaker pre-existing attitudes towards the topics (*Nonparametric Kendall's tau (τ) correlation test*. Study I: R = -0.07, p = .015; Study II: R = -0.09, p < .001). We performed an additional exploratory analysis to assess whether people would be more prone to opinion resistance in one of the three topics employed in the task. In both studies, we did not find a statistical difference in the amount of opinion updating considering which topic the news headlines were about (Study I: $F_{(386)}$ = 1.37, p > .05; Study II: $F_{(440)}$ = 0.01, p > .05). However, in Study I, although not statistically significant, the topic of vaccination had the stronger resistance to opinion change compared to the other topics.

Effects of initial confidence on opinion updating (Study I and II)

In this exploratory analysis, we were interested in assessing if the amount of confidence participants had in their first opinion rating could influence the subsequence updating behavior. Indeed, in both studies we found a correlation between the two variables, such that participants with higher confidence in their initial opinions were statistically less likely to update their opinion after reading other people's opinions compared to when the initial confidence was lower (*Nonparametric Kendall's tau (τ) correlation test*. Study I: R = -0.2, p < .001; Study II: R = -0.14, p < .001).

Effects of digital maturity on opinion updating (Study II)

In Study II, an additional aim was to assess whether people with more digital maturity were less susceptible to the influence of others' opinions. Overall, we found that participants updated their opinion in the direction of the social information more when their score in the DIMI was lower compared to higher (*Nonparametric Kendall's tau (τ) correlation test*: R = -0.1, p < .001). Three DIMI sub-dimensions mainly drove the effect: *Autonomy in choice* (R = -0.072; p = .01), *Autonomy within digital contexts* (R = -0.076; p = .008), and *Risk awareness* (R = -0.07; p = .01). See the full correlation matrix in **Tab. 1**.

Kendall's Tau Correlations

Variable	Opinion Updating	Autonomy of choice	Autonomy within	Literacy	Growth	Risk	Negative Emotion control	Aggressive Emotion control	Support	Respect	Citizenship
1. Opinion Updating	—										
	—										
2. Autonomy of choice	−0.072*	—									
	0.010	—									
3. Autonomy within	−0.076**	0.176***	—								
	0.008	< .001	—								
4. Literacy	−0.003	0.058*	0.086**	—							
	0.912	0.046	0.004	—							
5. Growth	−0.032	−0.042	0.092**	0.103***	—						
	0.268	0.152	0.002	< .001	—						
6. Risk	−0.070*	0.152***	0.130***	0.110***	0.025	—					
	0.013	< .001	< .001	< .001	0.401	—					
7. Negative Emotion control	−0.042	0.271***	0.169***	0.141***	0.010	0.008	—				
	0.138	< .001	< .001	< .001	0.736	0.782	—				
8. Aggressive Emotion control	−0.044	−0.163***	−0.084**	−0.136***	0.021	−0.059*	−0.141***	—			
	0.126	< .001	0.006	< .001	0.490	0.046	< .001	—			
9. Support	−0.022	−0.009	−0.067*	−0.078**	0.084**	0.152***	−0.125***	0.025	—		
	0.427	0.753	0.021	0.007	0.003	< .001	< .001	0.387	—		
10. Respect	−0.029	0.081**	0.065*	0.065*	0.038	0.186***	0.114***	−0.269***	0.197***	—	
	0.303	0.005	0.029	0.028	0.200	< .001	< .001	< .001	< .001	—	
11. Citizenship	−0.019	−0.146***	−0.057	−0.021	0.098***	0.042	−0.124***	0.056	0.106***	0.099***	—
	0.498	< .001	0.056	0.475	< .001	0.149	< .001	0.059	< .001	< .001	—
12. DYMI total score	−0.100***	0.281***	0.262***	0.264***	0.278***	0.410***	0.269***	0.102***	0.254***	0.258***	0.186***
	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001	< .001

* p < .05, ** p < .01, *** p < .001

Tab. 1: Correlation matrix between opinion updating and all the sub-scales of the digital maturity construct.

Demographics of study group (Study I and II)

In this exploratory investigation, our objective was to examine the potential relationships between different demographic variables, such as *age*, *sex*, *education*, *political affiliation* and *social media usage*, and updating one's own opinion. We conducted a series of non-parametric correlation analyses, using Kendall's Tau Correlations, of the demographic variables together with opinion updating. In Study I, we found significant correlations only with the variable *politic*, measured on a scale between 1 (= Liberal) and 7 (= Conservative), and the amount of Facebook use. Specifically, participants with more conservative political views updated their opinion more compared to those with a more liberal political ideology (R = 0.100, p = 0.001), and participants who engaged more frequently with Facebook in their everyday life were more susceptible to other people comments compared to those who used Facebook less frequently (R = 0.096, p = 0.004). In Study II, we found that only engaging with the comment section was positively correlated to opinion updating (R = 0.067, p = .026). However, a closer look at the plots in **Fig. 6** revealed another trend (although not statistically significant) in the amount of technology used and opinion updating. Specifically, participants who checked more often social media seemed more susceptible to other people's opinions.

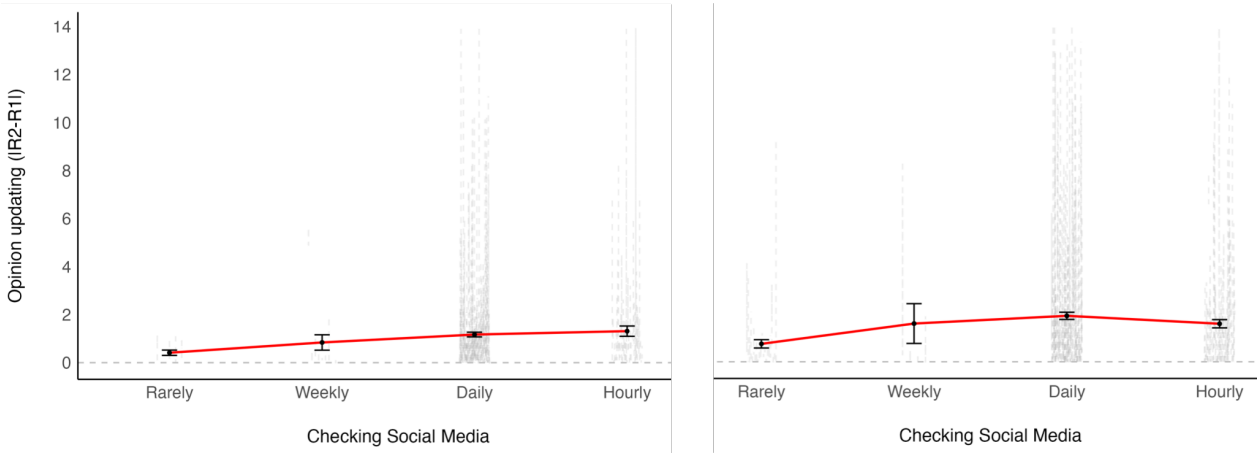


Fig. 6: Opinion updating considering how often people check social media in Study I (left) and Study II (right).

2.4 Discussion

Effects of comments on personal opinion

The goal of this project was to design a new behavioral task that could address some of the limitations of previous studies in the field and contribute to the literature regarding the influence of user-generated comments on personal opinion. We were also interested in assessing whether our novel paradigm could replicate previous findings of social influence on social media. Indeed, results from both our behavioral studies corroborated previous findings that individuals tend to be influenced by user-generated comments and adjust their opinions in the direction of the group opinion (Anderson et al., 2014; Anglin, 2019; Colliander, 2019; Lee & Jang, 2010; Shi et al., 2018; Wijenayake et al., 2020; William & Hsieh, 2021). Specifically, both after reading supportive and opposing comments towards the news headlines, participants significantly shifted their personal opinion to conform to the opinion expressed in the comments. However, participants did not significantly shift their opinion after reading messages without a majority (i.e. mixed comments). This is consistent with previous research demonstrating the proportion of the group opinion having an effect on how people update their opinions (Wijenayake et al., 2020). Indeed, in our studies, reading comments where there was a majority led to more pronounced conforming behaviors, and reading an equal number of supporting or opposing comments did not significantly influence people's opinions.

Previous studies found that negative comments consistently decrease participants' perception and opinion toward the news, and that supportive comments do not have an equivalent effect on participants (Waddell, 2018; Winter et al., 2015). However, in our behavioral studies we found significant social influence effects following both negative and positive comments. As we opted for argumentative, civil comments, it is possible that these supportive comments were compelling as much as the negative ones, thus successfully influencing participants' opinions. Indeed, Williams and Hsiech (2021) found that positive, technical comments about a scientific article influenced more positively participants towards the study's methods and findings than low-quality comments. Winter and colleagues (2015) found influence effects following both argumentative and subjective negative comments, but especially after argumentative comments. It seems indeed that argumentative, civil comments have a strong effect in influencing other people's perception and opinion about the news posts, probably signaling a higher expertise about the topic (Williams & Hsiech, 2021). As online comments do influence users' perception of news articles, it is advisable for social media platforms to actively promote the visibility of civil, argumentative comments (Williams & Hsiech, 2021). By doing so, these platforms can cultivate environments for constructive discourse while simultaneously avoid to undermine the credibility of the articles being discussed.

Effect of comments' congruence

In both our studies we found that participants were more susceptible to the opinion of others when they were exposed to comments that did not align with their initial opinion about the news headline, compared to when the opinions aligned. We found an upward trend of opinion updating, with congruent comments to participants' initial opinion leading to the smallest opinion updating, incongruent comments leading to the biggest updating, and mixed comments falling between the two. This is consistent with previous studies that found that participants tended to shift their opinion more when the group opinion was more challenging to their beliefs compared to when it supported it (Anglin, 2019). An explanation for this phenomenon could be found in the Cognitive Dissonance Theory (Festinger, 1957). It is possible that participants might have experienced cognitive dissonance when confronted with opinions that diverged from their own, which in turn might have motivated them to seek ways to alleviate the discomfort by conforming more to the group's view.

Effects of pre-existing attitudes about the topics

The amount of opinion updating was not only mediated by the discrepancy between personal and group opinions, but also by the strength of pre-existing attitudes towards the specific topics encountered. Indeed, in both studies we found a significant negative correlation between the strength of pre-existing attitudes towards the three topics of the news headlines and the magnitude of opinion updating. Specifically, stronger pre-existing attitudes towards the topics led to smaller opinion updating. Previous research found similar trends of belief perseverance particularly when individuals were holding strong beliefs about the topics (Anglin, 2019; Shi et al., 2018). These findings can be explained in light of the Social Identity Theory, where individuals are motivated to maintain a sense of identity in order to feel part of their social group (Tajfel & Turner, 1979). In the context of the current findings, it is possible that participants with stronger pre-existing attitudes were perceiving the topics as part of their identity, leading them to be more resistance to opinion change. Indeed, for some participants, the topics chosen for the task may have likely primed a strong sense of identity based on deep beliefs and values, such as environmental awareness, animal welfare, and health and ethical principles.

In a subsequent exploratory analysis, we found no statistical difference in the degree of opinion updating across the different topics used in the task. While climate change was the topic with the strongest pre-existing attitudes in both studies, the news headlines regarding the topic of vaccination (in Study I but not in Study II) were the ones with the most resistance to social influence, showing the smallest amount of opinion updating, although not significant. The difference in resistance to opinion change between the two studies could potentially come from the temporal context in which the data were acquired: data for Study I were collected during the COVID-19 pandemic in 2021, while data for Study II were collected one year later in 2022, thus towards the end of the pandemic. This temporal difference could have increased the emotional resonance towards news headlines and other people's comments concerning vaccination, making vaccination-related information particularly vivid and potentially more subject to motivated reasoning. Indeed, when looking at the attitude scores towards this topic, the mean attitude score towards vaccination decreased from Study I to Study II, although the attitude scores of the other two topics increased from Study I to Study II. This decrement from Study I to Study

II suggests that the topic of vaccination might have been perceived as more salient and emotionally charged during data collection of Study I compared to Study II. However, this is a speculative explanation and demographic variables of the two samples should also be taken into account.

Effects of personal confidence

The degree of confidence participants had in their first opinion rating also mediated the magnitude of opinion updating, such as, participants who were less confident in their opinion rating were also more likely to be influenced by what they read in the comments. Indeed, in both studies, participants who had more confidence in their opinions updated significantly less their opinions compared to when their confidence was lower. This aligns with previous research which found initial confidence modulating the degree of influenceability (Wijenayake et al., 2020; De Martino et al., 2017). However, participants confidence in their opinions decreased when they encountered comments that were discordant with their personal opinions. In Study I, we observed a significant decline in participants' confidence when they read comments that were incongruent compared to mixed comments, but there was no significant difference in confidence between congruent and mixed comments. Conversely, in Study II, we found the opposite pattern: confidence decreased significantly between congruent and mixed comments, but there was not a significant decrease between mixed and incongruent comments. However, in both studies, we found a general significant decrease in participants' confidence the more their opinion was discordant with the group opinion.

Effects of Demographics

When taking into account demographics variables such as age, sex, education, political affiliation and general social media usage, we obtained different outcomes from the two studies. In Study I, we identified a positive correlation between individuals' frequency of Facebook usage and the inclination to update their opinions, such that, the more people used Facebook the more they were susceptible to be influenced by the comments. This finding aligned with our findings from Study II, where susceptibility to other people's opinions was positively correlated with their level of engagement in online comment sections, like reading and writing online comments. We also observed a positive

correlation between social media usage and opinion updating, although not statistically significant: the more people checked their social media, the more they were influenced by other people's comments. These findings suggest that the extent to which technology is integrated in our daily life can affect how much we are influenced by the opinions expressed through social media comments. This trend could be attributed to the fact that individuals who spend more time using technology, particularly by engaging with social media comment sections, might be more susceptible to the social cues provided online compared to those who spend less time on these online platforms. Indeed, the attribution of importance of online content could be assumed by the fact that non-social media users tend to valuate online discussions on social media as a waste of time, compared to social media users (Springer et al., 2015). It is indeed reasonable to assume that people who use these platforms daily and spend more time interacting with them would find online discussions more relevant, and thus could display a heightened susceptibility to the social cues embedded within these platforms. Interestingly, in Study I, social influence was particularly pronounced among people who identified as holding a conservative political viewpoint compared to people with a more liberal political viewpoint. One explanation for this finding could be that conservative people tend to be the ones who are more engaged with social media comment sections (Köcher, 2016; Spinger et al., 2015), thus they might be more responsive to what they read online.

Effects of Digital Maturity on opinion updating

Overall, these findings highlight the potential risk that heavy consumption of social media usage might pose to individuals, potentially compromising their ability to independently shape their opinions and resist the influences and manipulations exerted within these platforms. We corroborated the idea that high level of digital maturity acts as protecting factor in resisting social influence exerted by other people's opinions on social media (Laaber et al., 2023). Indeed, in Study II, we found a negative correlation between digital maturity and influenceability. The more individuals were digitally mature the less they were influenced by online comments. Specifically, three dimensions of the DIMI predominantly drove these results: *Autonomy of choice*, *Autonomy within digital contexts*, and *Risk awareness*. Within the DIMI, *Autonomy in choice* is conceptualized as the conscious selection of mobile device usage, which comes from personal preference rather than

compulsion or obligation. *Autonomy within digital contexts* refers to the intentional navigation of digital environments, choosing to engage with the content that more resonate and stimulate the interest. *Risk awareness* denotes the vigilance of individuals regarding the use of mobile device and the awareness of potential risks and influences in online interactions. Based on the descriptions of these dimensions, it is evident that individuals who are aware of the content they consume online, as well as the risk associated with that consumption, exhibit lower susceptibility to online influences. These findings underline the importance of digital maturity, which seems to act as a protective factor against pervasive social influences on social media (Laaber et al., 2023). Indeed, it is reasonable to believe that individuals with higher levels of digital maturity are better equipped to critically evaluate the information they find online (Laaber et al., 2023), such as, for instance, other people's opinions in the comment sections.

2.5 Limitations and Future Directions for Research

This study, however, is not without limitations. While effort was made to design a behavioral task that would operationalize opinion updating by reducing confounding variables as much as possible, the highly controlled experimental setting may have limited the capacity to capture the complexity of real-world online interactions. The generalizability of our findings might also be constrained by the specific topics chosen for the task, which are familiar and polarizing to participants. Future research should explore the effects of online comments across a broader range of less known and polarizing topics. Another potential limitation is that we only considered pre-existing attitudes towards the topics as main modulating variable; however, other individual and psychological factors should be considered to understand why some people are more prone to be influenced by online comments. Moreover, the lack of longitudinal data collection may impede our ability to assess the long-lasting effects and persistence of opinion changes over time, highlighting the need for future research to investigate the stability of social influence from online comments.

One last potential limitation is the explicit nature of our main question within the task, which asked participants "*How much do you personally support or oppose the content of the news headline?*". Such explicit question may trigger the "demand characteristics effect"

(Orne, 1962), where participants become aware of the study's purpose and adjust their responses to align with what they believe the researcher expects or desires. This effect is particularly pronounced in lab settings, where participants physically meet the researcher and might feel more pressure to conform to perceived expectations (McCambridge et al., 2021). However, in our project, this concern is mitigated by the fact that both experiments were conducted online, which may reduce the potential for demand characteristics in several ways. First, the absence of face-to-face interactions means that participants are less likely to feel the social pressure to conform to the researcher's expectations (Mummolo & Peterson, 2019). Second, participants perform the study in a more naturalistic and relaxed environment (typically their home), which might make more genuine their responses. Finally, the anonymous nature of online setting might further reduce perceived expectations, potentially making participants answering more honestly since their identity remains unknown (Esposito et al., 1984). Future studies could try to incorporate more implicit questions or measures to capture opinion change following the presentation of other people opinions in form of online comments.

3. Neural Correlates of the New Social Influence Paradigm

3.1 Introduction

3.1.1 Social Media and Theory of Mind

It is not surprising that over the last decades research on social media has exponentially increased, involving a multitude of disciplines such as psychology, economics, communications, marketing and sociology (Meshi et. al., 2015). However, a gap in our understanding still remains when it comes to the neuroscience of social media and online social interactions. A review by Meshi and colleagues (2015) highlighted the primary neural networks involved in the cognitive processes triggered on social media, namely: self-referential cognition, social reward and mentalizing. Specifically focusing on mentalizing, Frith and Frith (2003) explained that mentalizing is the capacity to represent and understand mental states, desires and beliefs of others in order to predict their intentions and behavior. Given our interest in the neural mechanisms activated when individuals are confronted with other people's opinions under online news articles posted on social media, we decided to narrow our focus specifically on the theory of mind network. Indeed, when someone reads a news posted on social media, they might initially consider their own opinion about the topic. However, upon reading comments from other users, individuals might engage in complex cognitive processes to understand the beliefs and attitudes expressed by other people in the comments. They might evaluate whether these external opinions align or contradict their own view and consider the validity and reasonableness of integrating these perspectives into their beliefs. All of these cognitive processes fall under the umbrella of mentalizing. Indeed, "mentalizing" or "Theory of Mind" (ToM) (Premack & Woodruff, 1978) is a form of social cognition required for reasoning about one's own mental states and for mentally representing the attitudes and beliefs of others (Gallup, 1985; Monticelli et al., 2021). Some scholars suggest that mentalizing seems to be heavily involved during interpersonal communication in online context, as it is specifically required to navigate online social environments that lack physical and verbal cues presented in face-to-face interactions (Doheny & Lighthall, 2023). Indeed, it is possible that in the absence of these physical social signs individuals must solely rely on what others write online to infer their opinions and intentions.

Neurally, the core ToM system involves key brain regions like the medial prefrontal cortex (mPFC), bilateral temporoparietal junction (TPJ), and the precuneus (PC) (Overwalle, 2009; Schurz et al., 2014). Additionally, research indicates the involvement of other brain regions like superior temporal gyrus (STG), inferior frontal gyrus (IFG) and temporal poles during cognitive and affective mentalizing processes (Molenberghs et al., 2016). While the activation of core mentalizing regions remains consistent across different ToM tasks, difference in activation of these secondary regions emerge depending on the nature of the tasks and stimuli used (Molenberghs et al., 2016). For instance, mentalizing tasks involving emotional, visual, implicit stimuli tend to engage more ventral brain regions such as ventral mPFC, anterior insula, IFG, orbitofrontal cortex (OFC), whereas the employment of cognitive, verbal, explicit stimuli seem to activate more dorsal regions like dorsal mPFC, precuneus and TPJ (Molenberghs et al., 2016). Given that our task primarily involves reading explicit, argumentative, civil comments, we would expect activations of mentalizing regions mainly associated with cognitive reasoning.

In the neuroimaging literature, the core mentalizing regions exhibit heightened activation when reasoning about the mental states of others (see review from Monticelli et al., 2021; Kim et al., 2020), when generating alternative explanations for other's opinions and beliefs (Kim et al., 2020; Kliemann et al., 2008), and when inferring and integrating mental states of others for decision-making (Young & Saxe, 2008; Young & Saxe, 2009). Mentalizing regions are also involved during social influence processes, when individuals are actively analyzing and elaborating attitudes from others and deciding whether to integrate these attitudes with their own, especially when others might hold discrepant viewpoints (Welborn et al., 2016). For instance, Cascio and colleagues (2015) found activation within mentalizing regions, specifically the TPJ, among individuals who were more influenceable by online recommendations from other users. Interestingly, they found variation in TPJ activation, with an increased activity correlating with higher susceptibility to social influence. TPJ and PC were also more active when participants were presented with divergent recommendations compared to reinforcing group recommendation. The authors concluded that TPJ is involved not only in reasoning about mental states of others, but also in actively incorporating the social information provided into one's decision-making process (Cascio et al., 2015). Core regions associated with mentalizing also exhibit increased activation when individuals encounter and integrate social information that is

more surprising compared to less surprising. In a study by Kim and colleagues (2021), activation of the mPFC and TPJ was observed when participants adjusted their opinions of others in response to newly acquired information about these people's behavior. Notably, these activations intensified when the new information contradicted strong prior beliefs about the characters, compared to weak prior belief. The authors suggested that this stronger ToM activity might signify an effort to generate alternative explanations for unexpected social information learned (Kim et al., 2021). In our fMRI study, we expect the process of generating explanations for others' beliefs, and thus ToM network activity, to increase more when individuals engage with comments incongruent with their initial opinions, potentially reflecting cognitive effort to understand other users' divergent viewpoints.

ToM regions come also into play when encoding and integrating mental representations of other people's beliefs for moral judgments. Specifically, bilateral TPJ and PC were active during encoding relevant beliefs of others (Kliemann et al., 2008; Young and Saxe, 2008), while mPFC, bilateral TPJ, and PC were active during the integration of this new information to judge the moral status of these people's attitudes and actions (Young and Saxe, 2008). Interestingly, when participants had to integrate the new social information to make a moral judgment, mPFC exhibited increased activity following negative beliefs about a potential outcome of the character's actions compared to neutral beliefs (Young and Saxe, 2008; Young & Saxe, 2009). The authors argued that these regions support the spontaneous inference of other people's mental states to evaluate the morality of their intentions, even when the moral intent of the agent is not explicitly stated (Young & Saxe, 2009). Consequently, given the potential ethical and moral implications inherent in the topics presented in our task, such as climate change, vaccination and veganism, engagement of the mentalizing regions might also signify potential moral judgment processes, especially while facing other people's discrepant opinions.

3.1.2 Aim of the Project

Neuroimaging research about social media is still in its infancy. To our knowledge, prior investigations have not yet examined the underlying neural mechanisms that lead individuals to shift their opinions regarding important contemporary news posted on social media in response to user comments. Thus, the primary aim of this project is to explore

whether regions associated with mentalizing are involved when individuals engage with comments from other users regarding news headlines and integrate this new social information into their own opinion. Specifically, we hypothesized that activation within the core ToM network will be observed when participants read comments related to the news headlines, compared to comments that are irrelevant (as part of a control condition explained in the method section). Additionally, following prior evidence that divergent viewpoints stimulate increased mentalizing processes, reflecting an effort to understand other people's divergent beliefs, we hypothesized that user comments incongruent to participants' initial opinion will elicit greater activation within these brain regions compared to user comments congruent with participants' initial opinions about the news headlines. It is important to note that we do not suggest that reading online comments about news is only mediated by mentalizing processes, rather we view it as a reasonable starting point for this investigation.

3.2 Designing fMRI-compatible Version of the Behavioral Task

3.2.1 Creation of the Stimuli

In order to create a task compatible with fMRI, our first goal was to increase the pool of stimuli so that we could obtain more samples of the BOLD response evoked during our task, and thus enhance our ability to capture the expected hemodynamic response. Additionally, as we use a novel und untested fMRI paradigm, having more stimuli would help us to increase sensitivity to better detect smaller effects within the neural responses. The methodological approach to stimulus creation was the same as employed in the behavioral studies. We navigated the Facebook pages of the same journals we used in the prior investigations and searched for novel and more recent news headlines related to the three contemporary topics. We used the same criteria described in Section 2.2.1 and collected an additional set of 9 news headlines (three per topic), so that, combined with the previous stimuli collected, we had a total pool of 18 news headlines for the fMRI task. Then, we proceeded to search suitable comments for these newly acquired stimuli. As for the behavioral studies, we collected 4 supportive comments and 4 opposing comments written by real Facebook users for each of the 9 news headlines. Then, to better isolate the BOLD response to participants' reading of these comments, we decided

to introduce a control condition to employ for contrasts in the subsequent imaging analyses. This time, differently from the control condition utilized in the behavioral studies (which employed a mixture of supporting and opposing comments), we opted for comments that were unrelated to the news headlines. This strategy was motivated by increasing the separation of BOLD signals coming from supportive versus opposing comments, as this new control condition did not include any of them. In this second iteration of stimuli creation, we collected 24 unrelated comments for the control condition and other 72 related comments (half supportive and half opposing) for the main manipulation. This yielded a pool of 18 news headlines and 168 comments to be employed in the fMRI task. We then used the same Facebook post generator to create the news headline posts and the comments. This time, however, we decided to present only one of the four comments at a time (instead of presenting the four comments together) in order to increase temporal separation between each comment. This allowed us to better isolate BOLD responses to each comment, improving signal-to-noise ratio and statistical power. For this reason, we created single images for each of the 168 comments, so that we could present single comments separated by a fixation cross (more details about the study design in Section 3.3.2).

3.2.2 Pilot Studies

In order to validate the new stimuli and the fMRI-compatible task, we carried out three behavioral and one fMRI pilot studies. The first pilot study was used to validate the new stimuli collected, the second one to assess the stability of the new task programmed in Psychopy, and the third one to assess whether this fMRI-compatible task would elicit a clear behavioral shift in participants' opinions even by presenting the four comments one by one and not altogether. The fMRI pilot was carried out to assess whether the new fMRI-compatible task would elicit brain activations relevant to the task.

In Pilot I, we recruited 21 participants from Amazon MTurk to assess the valence of the new comments collected. Participants were presented with the new 9 news headlines and with 8 comments per news headline. Their task was to indicate on a slider from -7 (= Strongly opposing) to +7 (= Strongly supportive) whether each comment was supportive or opposing the news headline. Each participant rated a total of 72 comments. The task lasted around 20-25 minutes and participants were remunerated with $3 plus a

performance bonus of $1 (following the same criteria of the pilots in Chapter 2). Pilot I confirmed that participants were able to discern the valence of the comments: supportive comments received positive ratings (M = 4.24; SD = 1.24) and opposing comments received negative ratings (M = -4.05; SD = 1.22).

Pilot II was performed to test the stability of the new task developed in Psychopy and to eventually optimise it after participants' feedback. To this end, we recruited 10 volunteers from the University of Bonn to test the new fMRI-compatible task on a laptop. Results from Pilot II confirmed the stability of the task on Psychopy and a preliminary analysis of the results further confirming the robustness of the task to capture opinion updating, although not as strong as in the previous main behavioural studies. Indeed, we found a significant difference between the first and the second opinion rating both after supportive (V = 2777, p = 0.001) and opposing comments (V = 6146, p = 0.009). However, when computing the opinion adjustments, we found no significant difference between the positive adjustment following the supportive comments and the negative adjustment following the opposing comments (F = 2.41, p = 0.12). These results could have been due to the small sample collected. For this reason, we decided to perform a third pilot on a wider pool of participants to solely assess the behavioural effect of opinion updating with this new design. Specifically, we were interested in assessing whether presenting the four comments one by one, rather than all together, would significantly impair the process of opinion updating and thus our behavioural findings.

For Pilot III, we recruited 41 participants through the Bonn EconLab database for an online experiment, of which 38 were included for the statistical analyses. As for the previous pilot studies, participants had to rate their opinions of 9 out of the 18 news headlines before and after reading 4 comments, but this time the four comments were displayed one at a time. The task lasted around 20-25 minutes and participants were remunerated 6€ for their participation. Results from Pilot III corroborated the previous behavioural results, confirming the ability for this new fMRI-compatible task to capture social influence. Indeed, the second opinion rating was more positive following supportive comments (β = 0.86, SE = 0.31, 95 % CI [0.25, 1.48], $t_{(302)}$ = 2.77, p = .006) and more negative following opposing comments (β = -1.18, SE = 0.31, 95 % CI [-1.80, -0.57], $t_{(302)}$ = -3.77, p < .001) compared to unrelated comments. The last step of the piloting phase was to run the new task in the

MRI scanner and to assess whether we could find significant brain activations related to the task.

For the fMRI pilot, we recruited 3 participants to assess whether the new task would be properly compatible with MRI scanner and to assess whether it would activate brain regions relevant to the task. We had to exclude the imaging data from participant 1 due to technical problems during the scanning session. For participant 2, we found activation of the putamen, insula, left superior temporal gyrus and left postcentral gyrus (*p*-uncorrected < .001, k = 10) when reading comments congruent to their opinion compared to irrelevant comments. Only activation of the putamen was left when reading comments related to the news headlines compared to unrelated comments. For participant 3, we found significant activation of the left inferior frontal gyrus and supramarginal gyrus (*p*-uncorrected < .001, k = 10) both while reading related comments to the news headlines versus unrelated comments, and when reading incongruent comments versus unrelated comments. Overall, the brain regions identified in the fMRI pilot included the regions we expected to be engaged by our novel task, as they underlay semantic and language processes (i.e. left inferior frontal gyrus, left superior temporal gyrus, supramarginal gyrus), emotional processes (i.e. putamen, insula) and mentalizing processes (i.e. left inferior frontal gyrus, supramarginal gyrus); therefore, we proceeded with data collection.

3.3 fMRI Study

We carried out the fMRI study in compliance with the most recent revision of the Declaration of Helsinki and it received approval from the ethics committee of the Medical Faculty of the University of Bonn (Ref: 419/21). The study design and planned statistical analyses were pre-registered on OSF after data collection began but before analyses were conducted (https://osf.io/94kyg, date of pre-registration: January 31, 2023). Variations from the pre-registered analyses, as well as any non-registered exploratory analyses, are stated in the results section.

3.3.1 Participants

We chose to acquire 40 datasets in order to achieve a reasonable balance between power and use of financial resources. For cognitive tasks, this is expected to yield a Pearson product-moment correlation coefficient between two independent replications of non-

thresholded fMRI statistical parametric maps of 0.7 (Bossier et al., 2020). We recruited participants between September 2022 and December 2022 via social media and flyers in cafeterias and common spaces of the University of Bonn. Interested participants completed an online screening survey on the platform Qualtrics and if eligible they would provide us with their email. Eligible participants had no prior history or neurological and mental disorders, they were at least 18 years old, they had normal or corrected-to-normal vision, and had good proficiency in the English language to be able to understand the news headlines and the user-generated comments displayed during the study. Candidates for participation were called via phone to book the session and to respond to any questions they might have about MRI safeties and their participation. We limited information regarding the study's hypotheses to a minimum in order to prevent biases that might influence participants' expectations and their opinion ratings. The participants were remunerated for their time at an hourly rate of €15/hour. 43 participants took part in our fMRI study. However, two participants' data had to be excluded from the neuroimaging analyses due to technical problems with the response grips, and two other datasets were excluded due to excessive movements during the scanning sessions. Thus, we included n = 41 (22 female; M_{age} = 24.37 years; SD_{age} = 4.17) datasets in our behavioral statistical analyses and n = 39 (21 female; M_{age}: 24.26 years; SD_{age}: 4.25) datasets in our neuroimaging statistical analyses.

3.3.2 Study Design

This study was designed to quantify opinion updating by social information, while observing the neural correlates associated with encountering consensus/disagreement with one's opinions. The stimuli were the screenshots of 18 news headlines posted on social media, and the social information were the screenshots of written comments made by other users posted together with the news headline (see Behavioral studies above). As in our behavioral studies, the topics of the news headlines were: climate change, vaccination and veganism. The comments accompanying the news headline could be: opposing, supportive or unrelated (see **Fig. 7**). The stimuli were presented using PsychoPy version 2021.2.3 through a screen placed at the back of the participants. During the task, each participant was presented with news headlines and was asked to rate their opinion about the content of the news headline on a scale from "Strongly oppose (the

content of the news headline)" (-7) to "Strongly support (the content of the news headline)" (+7). After their first opinion rating (R_1), participants read four comments under the news headline written by real users and presented one by one. After reading all the four comments presented, participants were asked again to rate their opinion about the news headline (R_2). The presentation order of the headlines was determined randomly for each participant. The type of comments following each news headline was presented pseudo-randomly, such that each participant was exposed randomly to an equal amount of supportive, opposing and unrelated comments towards the 18 news headlines, as part of the within-subject design.

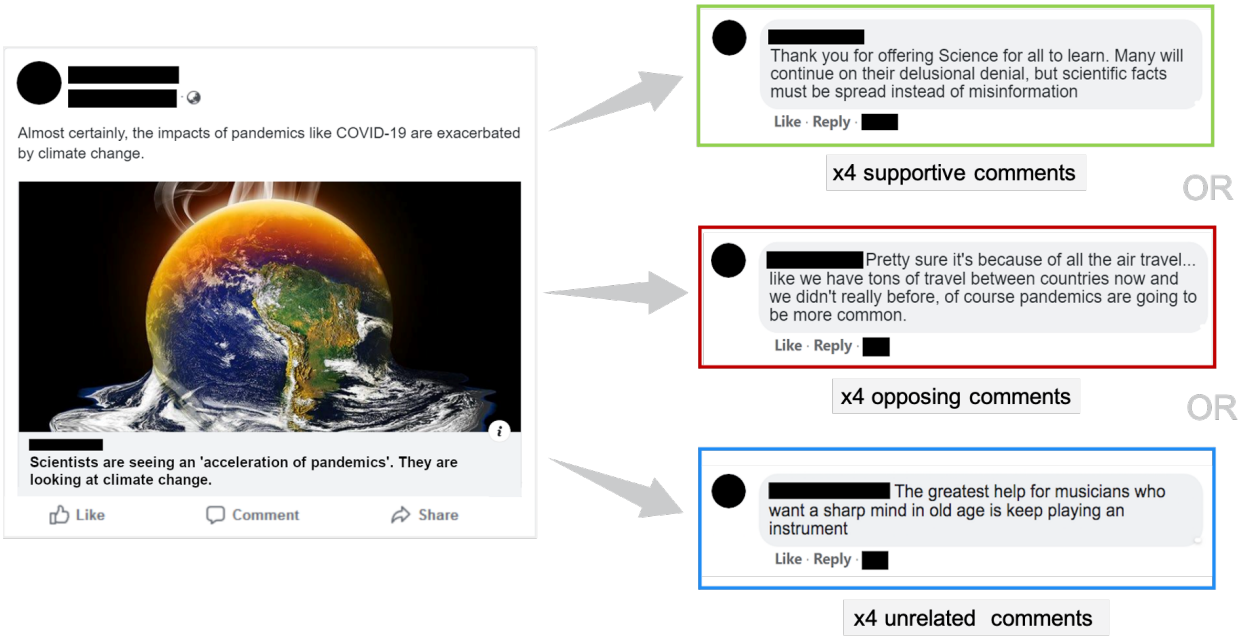


Fig. 7: Stimuli used in the fMRI task. On the left side, an example of one of the 18 news headline that was presented to participants. On the right side, an example of one comment that was presented below the news headline per each experimental condition (respectively: supportive, opposing and unrelated).

Specifically, as shown in **Fig. 8**, each trial started with a jittered fixation cross that lasted between 2 and 3 seconds, followed by the presentation of one news headline along with a rating scale. Participants had at most 50 seconds to carefully read the news headline and rate their opinion about it. Once participants gave their first opinion rating (R_1), a fixation cross was automatically displayed for 2 to 3 seconds. The trial continued with the presentation of the first of the four comments. By using the grips, participants could read

the comment below the news headline and move to the next comment at their own pace, having at most 40 seconds before the next comment would be displayed automatically. Once again, a fixation cross separated each of the four comments. Once participants went through all the four comments, the same news headline along with the rating scale was again displayed. Here participants would rate their opinion about the news headline for a second time (R_2). This whole sequence was repeated for each of the 18 news headlines. Our main manipulation variable was the valence of the user-generated comments about the news headlines displayed to participants. This variable had three levels: supportive, opposing, and unrelated comments.

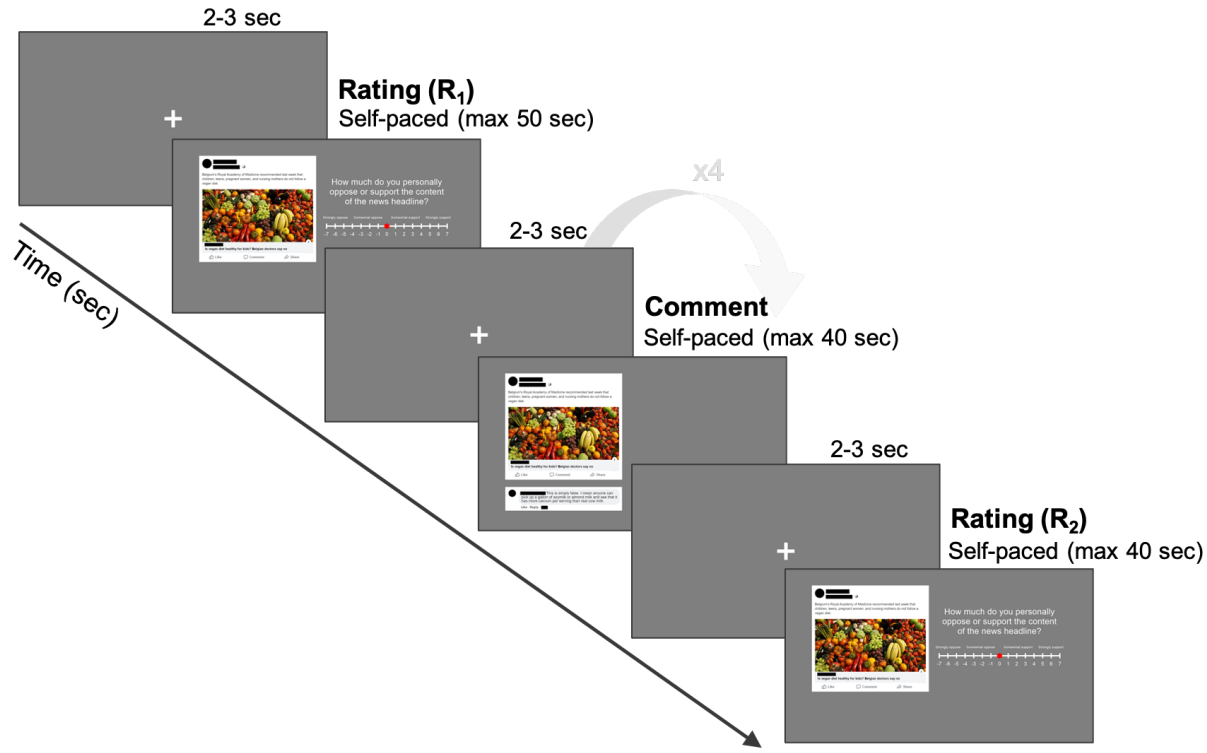


Fig. 8: Sequence of the Online Social Influence Task. After reading the instructions on the screen, participants saw a fixation cross that lasted between 2 and 3 seconds. They were then presented with a news headline that they had to carefully read and rate their opinion about it on a scale between -7 (= Strongly oppose) and 7 (= Strongly support). This was the first opinion rating (R_1). After another fixation cross, they saw one comment below the news headline. When participants read the comment, they could proceed to the next one, until they read all 4 comments presented. After another fixation cross, participants saw the news headline again with the same rating scale and could rate their opinion for a second time (R_2).

3.3.3 Experimental Procedure

The fMRI study was conducted from October 2022 to January 2023 at the MRI Core Facility (Life & Brain Research Center) of the Medical Faculty of the University of Bonn. Once arrived at the facility, participants were welcomed and signed documentations related to their consent to participate in the study, data storage and MRI safety information. Before entering the MRI scanner, participants filled out three pre-exposure attitudinal questionnaires about the three contemporary topics used in the study, the DIMI questionnaire, and completed a desktop tutorial version of the task. Prior to scanning, participants were given instructions on how to use the emergency ball and the controllers in their hands. They were also provided with earplugs to protect their ears against the loud noises of the MRI scanner. While inside the scanner, participants viewed the instructions and the fMRI task via a mirror-system attached to the head coil, which was individually adjusted for clear visibility of the screen at the back of the scanner. Responses were recorded with controllers in participants' hands (Nordic NeuroLab, Bergen, Norway). The scanning procedure included a functional scan, a gradient field map (GFM), and T1-weighted structural images. Specifically, the functional scanning session where participants performed the online social influence task session was composed of 3 runs, each of them included the presentation of 6 out of 18 news headlines. Between each run, participants could take a break of few minutes while still remaining inside the MRI scanner and they were asked to inform the team when they were ready to start a new run. Each run lasted between 10 and 15 minutes, depending on the pace of the participant. The functional session lasted between 30 and 45 minutes. After that, we acquired a gradient field map and T1-weighted structural images of participants' brains with a sequence that lasted approximately 6 minutes. During this time, participants could relax and close their eyes, while still trying to remain as still as possible inside the MRI scanner. After the scanning session, participants were debriefed and could ask questions about the nature and the hypotheses of the study and received their compensation. The overall session lasted approximately 1 hour and 30 minutes (see **Fig. 9** for a scheme of the procedure).

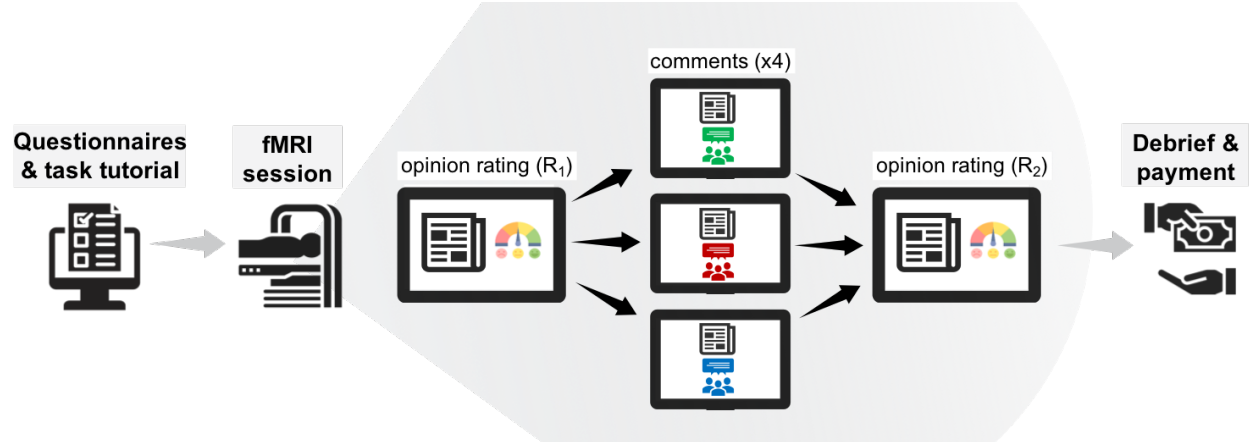


Fig. 9: Experimental procedure of the fMRI online social influence study. Participants underwent a pre-scanning phase on a computer, where they filled in attitude questionnaires on the topics used in the task and then performed a tutorial of the task that they would perform inside the scanner. Inside the scanner, participants had to rate their opinion on 18 news headlines (6 per run), before and after reading 4 comments from other users (presented sequentially). The comments could be either supportive towards the news headline, opposing towards the news headline, or not related to the news headline (control condition).

3.3.4 Image Acquisition

We acquired functional and structural MRI data on a 3T Siemens TRIO MRI scanner with a Siemens 32-channel head coil. We acquired functional images with a T2* echo-planar imaging (EPI) BOLD sequence with the following settings: TR = 2500 ms, TE = 30 ms, flip angle = 90°, field of view = 192 x 192 mm, voxel size (x,y,z) = 2 x 2 x 3 mm^3, and 37 slices acquired axially in ascending order. The number of images varied per participant due to the self-paced nature of the task. We acquired all volumes within 3 different runs per experimental session. Structural images were acquired with a T1-weighted using the following parameters: TR = 1660 ms, TE = 2540 ms, TI = 850 ms, flip angle = 9°, field of view = 256 x 256 mm, voxel size (x,y,z) = 0.8 x 0.8 x 0.8 mm^3. Each participants had 208 images taken in a sagittal direction.

3.3.5 fMRI Data Pre-processing

We used SPM12 (Wellcome Department of Imaging Neuroscience, London, UK) software package based on MATLAB R2023b to pre-process our fMRI data. The pre-processing steps included: realignment, slice-time correction, co-registration, segmentation, normalization, and smoothing. With realignment, data was first corrected for head motion

that could happen during the fMRI scanning procedure by aligning each volume to a reference image. The realignment parameters were visually evaluated, and any participant who moved more than 3 mm during the scanning session was eliminated from the neuroimaging analysis. During the slice-time correction, the images were corrected and aligned properly to the first functional image as the reference timepoint via temporal interpolation. This step helps in dealing with delays of the acquisition of different brain slices. During the co-registration, the functional images were aligned with the individual high-resolution T1-weighted anatomical image, ensuring spatial correspondence between structural and functional images. Afterwards, during segmentation, the different tissues were separated in gray matter, white matter and cerebrospinal fluid in other to create a tissue probability map. Normalization then transformed the images into a standardized coordinate system, specifically the Montreal Neurological Institute (MNI) space. This involved warping the images to align them with a standard brain template. Lastly, a spatial filtering was applied which smoothed the images with an 8 mm 3D Gaussian kernel. This process helps to reduce noise and increase true signal, making subtle brain activity more detectable.

3.3.6 Behavioral Data Analysis

Behavioral data analyses were entirely performed with R language (version 4.3.2) and RStudio (version 2023.12.1.402). The R packages we utilized for data cleaning, analysis, and visualization were: *BayesFactor, cowplot, ggpubr, gmodels, Hmisc, kableExtra, lme4, nlme, plotrix, readxl, reshape2, see, tidyverse.*

Effects of comments valence

To assess whether there was a significant difference in opinion rating before and after presenting the comments and to assess if presenting unrelated comments successfully worked as control condition, we performed three Paired Wilcoxon tests (one for each level of the explanatory variable: supportive, opposing, unrelated) between the first and the second opinion ratings. Subsequently, as pre-registered, to assess whether the valence of the comments would predict the direction of the opinion shifts, we computed the opinion adjustments (i.e., $R_2 - R_1$) and performed a Mixed-Effects Linear Regression Model using the function *lme* from the R package *lme4*. The opinion adjustment ($R_2 - R_1$) was used as

dependent variable, the valence of the comments (i.e. supportive, opposing, unrelated) as the explanatory variable. Random intercepts were employed to control for individual differences in the average opinion ratings.

Effects of congruence between initial opinion and comments valence

In this exploratory analysis, we were interested in assessing the effect of the congruence between participant initial opinion rating (R_1) and the opinion stated in the comments shown below the news headline. Specifically, congruent "participant-comments" opinions occurred when participants' first opinion rating (R_1) and the opinion expressed in the comments were either both supportive or both opposing towards the news headline. This state was coded as "*congruent condition*". On the contrary, the occurrence of diverging "participant-comments" opinions was coded as "*incongruent condition*". Finally, the state where participants were presented with unrelated comments towards the news headline was coded as "*unrelated condition*". To assess whether the congruence between participants' initial opinion and opinion in the comments had an effect on opinion updating, we run a Mixed-Effects Linear Regression Model, where the absolute value of opinion updating (i.e., $|R_2 - R_1|$) was used a dependent variable and the congruence between participants' initial opinion and the opinion in the comments was used as explanatory variable. We allowed random intercepts to account for individual differences in the average opinion ratings.

Effects of pre-existing attitudes towards the topics of the news headlines

Here, we explored the association between having pre-existing attitudes towards the topic of the news headlines and the amount of opinion updating using a Linear Regression analysis. Attitude towards the three topics were assessed with three attitudinal questionnaires that participants filled before entering in the MRI scanner. Because of the different number of items and the different Likert scales in the three questionnaires, the scores resulting from them were transformed into z-scores to obtain a unique index of pre-existing attitudes that ranged from 0 (weak attitude towards the topic) to 50 (strong attitude towards the topic). In the Linear Regression, we included pre-existing attitude as the exploratory variable and opinion updating ($|R_2 - R_1|$) as dependent variable.

Effects of digital maturity

In this analysis, our objective was to investigate the association between increased digital maturity and the degree of opinion updating. Participants' digital maturity was assessed with the use of DIMI questionnaire. To this end, correlation analyses were conducted: the first investigated the link between the overall digital maturity score and opinion updating, while the others explored the associations between each sub-dimension of digital maturity and opinion updating.

Response times

In these exploratory analyses, we aimed to investigate how response times differed across different conditions, specifically regarding the amount of time participants spent reading specific comments. Firstly, we conducted a Kruskal-Wallis rank sum test to compare the time participants spent reading comments that were in agreement with their initial opinion of the news headline (*congruent condition*) with those that were in disagreement (*incongruent condition*). Secondly, we conducted a correlation analysis to examine the association between the time spent reading the comments and the amount of opinion updating.

Assessing data sensitivity with Bayesian hypothesis testing

As in our main behavioral studies described in the previous chapter (Section 2.3.3), we additionally employed Bayesian hypothesis testing to assess the effects of interest.

3.3.7 fMRI Data Analysis

We performed fMRI data analyses using SPM12 software package based on MATLAB R2023b. We used a two-steps univariate analysis, consisting of a first- and a second-level analysis. In the first-level analysis, we specified a statistical model that represented the experimental design along with the expected brain activation. In this statistical model, we included the onsets and durations of different regressors which indicated our experimental conditions, which were convolved with the hemodynamic response function to model the brain's response to each condition. Afterwards, we estimated parameters β for the different regressors, which represent the estimated contribution of each regressor in our statistical model to the observed BOLD data. Using contrasts of parameter estimates, we

defined specific comparisons of interest between regressors. This allowed us to check the effects of our experimental manipulation by calculating the differences in brain activation between conditions, and thus testing our hypotheses. These steps were performed individually for each participant dataset.

In the second-level analysis, we combined the results obtained in the first level analysis in order to observe group-level effects. This analysis allows us to assess what brain regions are consistently activated in each condition in the whole group of participants, allowing us to draw meaningful conclusions about responses in the population of participants we tested. One-sample t-tests were employed to assess if the responses identified by each contrast of parameter estimates were significantly different from zero across all participants. As the statistical tests are calculated over each voxel of the brain, we used different correction methods to decrease the likelihood of false-positive results (Type I error).

During the first-level analysis, to assess the effects of the comments on neural activity, we specified a first GLM containing 11 regressors per run: 1) congruent comments; 2) incongruent comments; 3) unrelated comments (control condition); 4) headline before social information (H_{pre}); 5) headline after social information (H_{post}); 6) six rigid body-movement regressors (three for translation, three for rotation) as confound variables. These 11 regressors were repeated for the second and third run. Every regressor was used to model the responses of specific events, from their onset until their offset. For example, the first regressor (i.e. congruent comments in run 1) modelled the response to each comment of the congruent condition displayed on the screen, with the onset at the appearance of one comment and the offset when that comment would disappear from the screen. As pre-registered, after creating the design matrix, we calculated the contrasts between regressors of interest. The contrasts were the following: i) *Congruent* > *Unrelated*, to assess the neural correlates of being confronted with consensus to one's personal opinion about the news headlines; ii) *Incongruent* > *Unrelated*, to assess the neural correlates of being confronted with disagreement with one's personal opinion about the news headlines; iii) *Congruent* + *Incongruent* > *Unrelated*, to assess the neural correlates of being confronted with comments that are relevant for one's personal opinion

about the news headline; iv) *Incongruent* > *Congruent*, to identify regions more active in the incongruent compared to the congruent condition.

To assess the neural correlates of opinion updating, we specified a second GLM containing 12 regressors per run: 1) when participant updated their opinion; 2) when participant did not update their opinion; 3) headline before social information (H_{pre}); 4) congruent comments; 5) incongruent comments; 6) unrelated comments (control condition); 7) six rigid body-movement regressors (three for translation, three for rotation) as confound variables. These 12 regressors were repeated for the second and third run. Every regressor was used to model the responses of specific events, from their onset until their offset. For example, the first regressor (i.e. opinion updated in run 1) modelled the response obtained when participants rated for the second time their opinion after reading the 4 comments. The onset of each event was the appearance of the news headline with the second rating scale and the offset of the event was when participant rated their opinion a second time. After creating the design matrix, we calculated the following contrasts between regressors of interest: i) *Updated* > *Not updated*, to assess the neural correlates involved in updating one's personal opinion; ii) *Not updated* > *Updated*, to assess the neural correlates involved in resisting to update one's personal opinion.

In the second-level analysis, we analyzed at a group level the effects that our conditions had on neural activity. For the first GLM, we conducted a one-sample t-test over the mean parameter estimates across participants for the contrast about comment congruence performed in the first-level analysis. For the second GLM, we conducted a one-sample t-test over the mean parameter estimates across participants for the contrast about opinion updating performed in the first-level analysis. To limit the number of independent tests in the first GLM, we applied a Small Volume Correction (SVC) to the whole-brain group-level activations, followed by a family-wise error (FWE) correction for multiple comparisons at the cluster-level; this was applied to an uncorrected threshold of $p = .001$ at the voxel-level. To perform the SVC, we created an anatomical ROI using the MarsBar tool in SPM (Brett et al., 2002), based on the results of a functional localizer task identifying brain regions associated with Theory of Mind (*Neurovault*: https://neurovault.org/images/25863/). In the results section, we also chose to present results at a more liberal threshold ($p_{uncorrected} < .001$) due to the innovative and explorative

nature of the paradigm. This practice is commonly employed in fMRI research in the attempt to reduce the possibilities of missing true effects (i.e. Type II error) and enhance the likelihood of uncovering neural landmarks that could guide future investigations (Lieberman & Cunningham, 2009). For this reason, we stated every non-preregistered analysis as exploratory analysis in the results section below.

3.4 Results

3.4.1 Behavioral Results

Effects of comments valence

Participants rated their opinions about the news headlines on a scale between -7 to 7. As shown in the figure below (**Fig. 10a**), and replicating the results from the behavioral studies, participants significantly adjusted their opinions in the direction of the social information, both after reading supportive comments (Wilcoxon signed rank test: V = 23932, p < .001, BF > 100) and after reading opposing comments (Wilcoxon signed rank test: V = 86412, p < .001, BF > 100). After reading unrelated comments, participants did not significantly update their opinion (Wilcoxon signed rank test: V = 12170, p = .4, BF = .04). Thus, participants adjusted their opinion in the direction of the comments' valence. When computing the index for opinion adjustments (i.e. $R_2 - R_1$), participants' second rating was more negative after reading the opposing comments (M = -0.57, SD = 2.25, 95% CI [-0.71, -0.43]) and more positive after reading the supportive comments (M = 0.45, SD = 1.75, 95% CI [0.33, 0.56]), and did not change after reading unrelated comments (M = 0.03, SD = 1.59, 95% CI [-0.13, 0.07]) (**Fig. 10b**). Indeed, the opinions adjustments were significantly predicted by the valence of the comments that participants read (F = 76, p < .001, BF > 100). Specifically, comments supporting the news headline significantly predicted positive opinion adjustments (β = 0.47, SE = 0.082, 95 % CI [0.3081, 0.6304], $t_{(2777)}$ = 5.708, p < .001), as well as comments opposing the news headline significantly predicted negative opinion adjustments (β = -0.54, SE = 0.081, 95 % CI [-0.7041, -0.3843], $t_{(2777)}$ = -6.671, p < .001). We also found that critical comments had a significantly stronger impact on opinion updating (computed as the absolute value of the second opinion rating minus the first opinion rating) compared to supportive comments (β = 0.22, SE = 0.07, 95 % CI [0.081, 0.358], $t_{(2777)}$ = 3.112, p = .002).

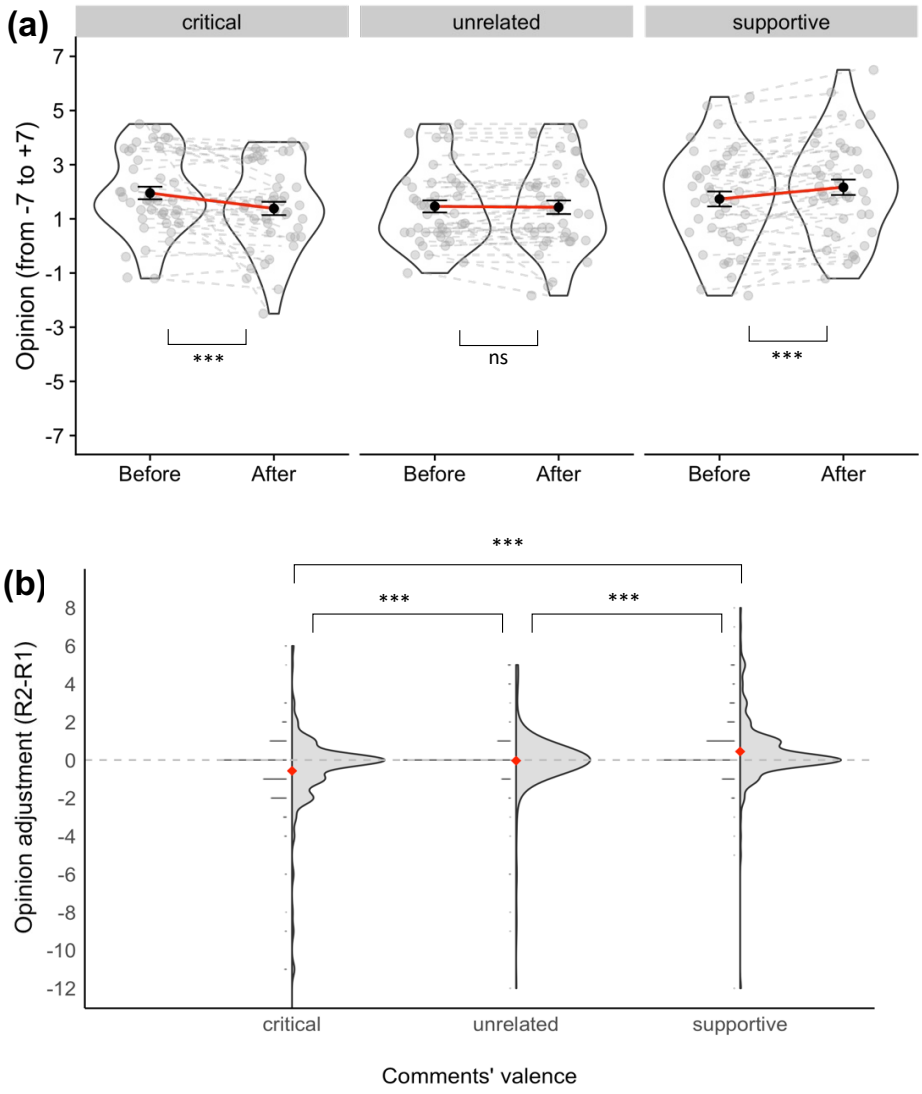


Fig. 10: (a) Effect of the comments' valence on participants' second opinion rating. The black dots are the mean values across participants, the grey dots are the individual mean values and the error bars represent the standard error of the mean. The contours represent Kernel density plots **(b)** Opinion adjustment computed as the difference between the first and the second opinion rating. The red dots are the mean values across participants. The contours represent Kernel density plots.

Effects of congruence between initial opinion and comments valence

In this exploratory analysis, we computed an index of the congruence between participants' initial opinion about the news headline and the opinions expressed in the comments. This index had three levels: *congruent*, *incongruent*, *unrelated*. An opinion updating measure was also calculated for each news headline as the absolute value of

the difference between the second rating and the first rating (i.e. $|R_2-R_1|$). As shown in the figure below (**Fig. 11**), participants updated their opinion more when confronted with a group opinion that was incongruent with their initial opinion (B = 0.671, SE = 0.068, 95 % CI [0.537, 0.805], $t_{(2777)} = 9.82$, $p < .001$), compared to when the initial opinion and the comments were congruent (B = 0.256, SE = 0.073, 95 % CI [0.113, 0.34], $t_{(2777)} = 3.49$, $p < .001$).

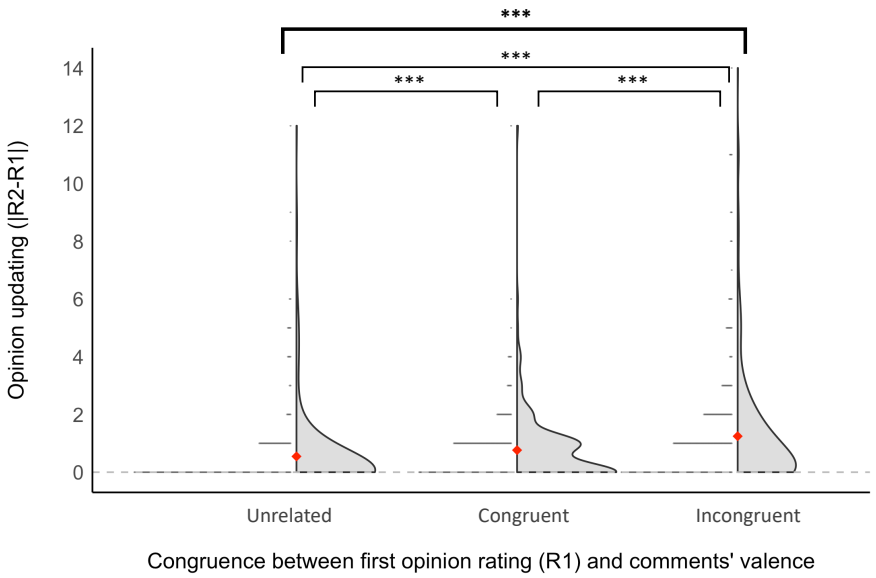


Fig. 11: Effect of congruence of first opinion rating and comments' valence on opinion updating. The red dots are the mean values across participants.

To further evaluate the relationship between the congruence of initial opinions and the valence of subsequent comments in influencing how participants adjusted their ratings, we calculated a polarisation index with three categorical levels: *polarisation*, *depolarisation*, and *no change*. In this exploratory analysis, "polarization" was coded as an increase in magnitude in the same direction from the first opinion rating (R_1) to the second opinion rating (R_2). "Depolarisation" was calculated as a decrease in magnitude from the initial opinion rating (R_1) to the subsequent one (R_2). Within the depolarisation category, we included opinion shifts that moved from one pole of the rating scale to the opposite pole (for instance, if the first opinion rating was one of support for the news headline but the second rating was opposing, or vice versa). "No change" was coded when participants maintained their first opinion rating from the first to the second rating. As

shown in **Fig. 12**, polarisation was significantly higher than depolarisation following congruent comments ($\chi 2$= 10, *df* = 1, *p* = .001). In contrast, depolarisation was significantly higher than polarisation following incongruent comments ($\chi 2$= 4.5, *df* = 1, *p* = .03). There was no significant difference between polarisation and depolarisation following unrelated comments ($\chi 2$= 0.4, *df* = 1, *p* = .5). Interestingly, both congruent and incongruent comments led to opinion polarization, with no significant difference between these two conditions ($\chi 2$= 0.2, *df* = 1, *p* = .7).

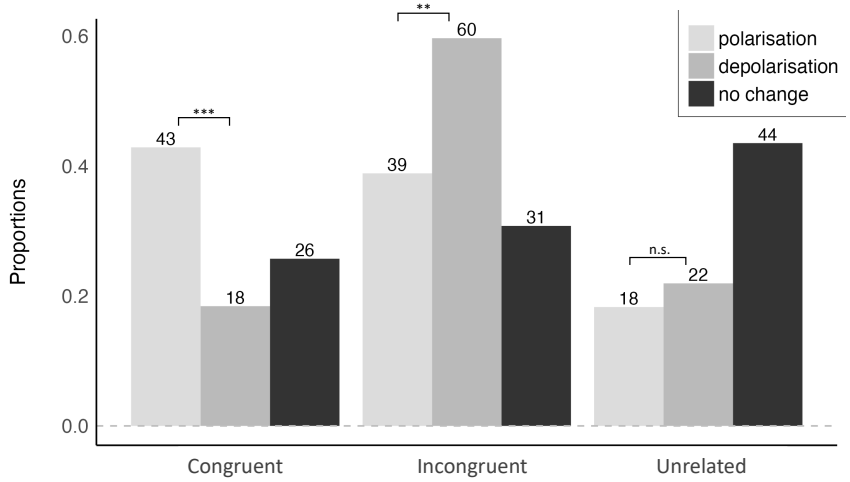


Fig. 12: Bar plot showing the polarization index considering the congruence between first opinion rating and the comments below the news headline. The numbers above the bars represent the specific proportion of polarization in % of the number of trials in the three different levels and in different congruence conditions.

Effects of pre-existing attitudes towards the topics of the news headlines

In this preliminary analysis, we aimed to examine the impact of pre-existing attitudes towards the three topics of the news headlines on how much participants updated their opinions. Linear Regression analysis revealed that pre-existing attitudes towards the topics of the news headlines did not explain the change in opinions after reading the comments below the news headlines (β = 0.003, SE = 0.003, 95 % CI [-0.001, 0.007], $t_{(2818)}$ = 1.43, p = 0.15). Since the null result supported the null-hypothesis over the experimental hypothesis, we conducted a Bayesian analysis to distinguish whether the non-significant result provided evidence for H_0 or indicated data insensitivity. The BF of 0.116 indicates strong evidence for the null-hypothesis and confirmed the result from the

frequentist approach. We then decided to run an additional exploratory Linear Regression analysis between the variable opinion updating and the strength of participants' initial opinion (R_1) towards the news headline (which was coded as the absolute value of their first opinion rating, i.e. $|R_1|$). Here, the Linear Regression showed that participants updated their opinion significantly less when they had a stronger initial opinion towards the news headline compared to when they had a weaker initial opinion (β = -0.128, *SE* = 0.02, 95 % *CI* [-0.16, 0.09], $t_{(2818)}$ = -7.63, $p < .001$, *BF* > 100). Thus, the strength of the first opinion rating predicted the amount of opinion updating.

Effects of digital maturity

Before entering the MRI scanner, participants filled in a questionnaire about their digital maturity (the DIMI questionnaire) that yielded a digital maturity score and sub-scores for the different aspects of digital maturity. We ran a Shapiro-Wilk test for bivariate normality on all the variable pairs to be assessed. This test assesses whether variable pairs are normally distributed. As the test was significant for some pairs, we resorted to using the non-parametric Kendall's tau (τ) correlation test for our analyses. The results of the exploratory analyses show that participants updated their opinion slightly less when they had higher digital maturity compared to participants who scored as less digitally mature, although this correlation did not reach statistical significance (*R* = -0.03, *p* = .058). However, some of the sub-scores of the DIMI had a significant correlation with opinion updating (**Fig. 13**). We found significant positive correlations between opinion updating and digital risk prevention (DIMI_Risk) (*R* = 0.069, $p < .001$), adequate negative emotion regulation (DIMI_EmotionNeg) (*R* = 0.077, $p < .001$), autonomy within digital contexts (DIMI_AutonomyWithin) (*R* = 0.077, $p < .001$) and digital citizenship (DIMI_Citizenship) (*R* = 0.051, $p < .001$). In addition, we found negative correlations between opinion updating and regulation of aggressive impulses in digital contexts (DIMI_EmotionAgg) (*R* = -0.088, $p < .001$), respect towards others in digital contexts (DIMI_Respect) (*R* = -0.062, $p < .001$), autonomy of choice to use mobile devices (DIMI_AutonomyChoice) (*R* = -0.143, $p < .001$) and digital literacy (DIMI_Literacy) (*R* = -0.065, $p < .001$).

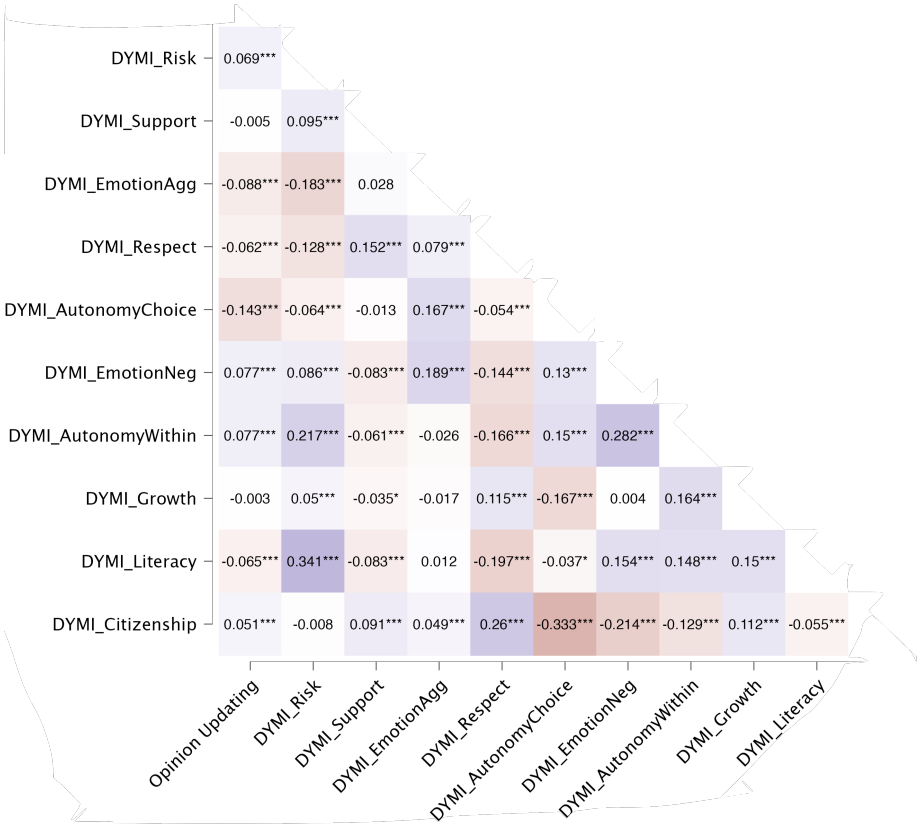


Fig. 13: Heatmap of Kendall's tau-b correlations between the DIMI sub-scores and the dependent variable opinion updating. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Response times

We conducted exploratory analyses to assess how response times varied across different conditions. We found that, on average, participants spent more time reading the comments when these were incongruent with their initial opinion about the new headline compared to when the comments were congruent (Kruskal-Wallis rank sum test: $W = 9$, $p = .003$, $BF > 100$) (see **Fig. 14a**). Interestingly, we also found that the time spent reading the comments was positively correlated with the amount of opinion updating, such that the more participants spent time reading comments of other people, the more they updated their opinion (Nonparametric Kendall's tau (τ) correlation test: $R = 0.17$, $p < .001$) (see **Fig. 14b**).

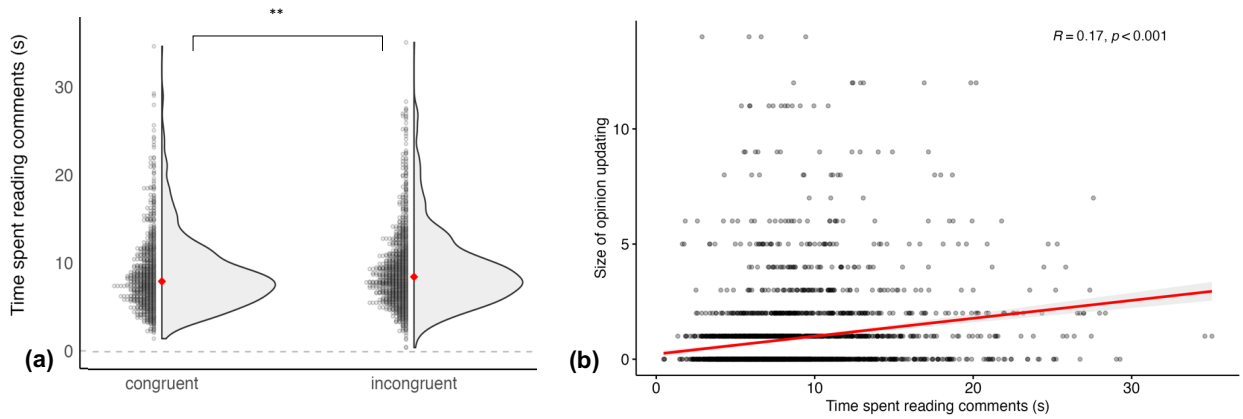


Fig. 14: (a) Half-violin plot showing time spent reading comments that were congruent or incongruent with participants' initial opinion about the news headline. The red dots are the median values across participants. **(b)** Correlation plot showing a positive association between time spent reading comments and the subsequent amount of opinion update.

3.4.2 fMRI Results

Effects of comments valence on brain activation

We tested whether the valence of the comments displayed below the news headlines increased activity in parts of the brain previously linked to mentalizing.

Congruent > Unrelated comments: As pre-registered, we performed a whole-brain analysis that showed significant activation in the left angular gyrus ([x, y, z] = [-45, -58, 35], k = 91, *T*-value = 5.02) at $p < .05$, corrected for FWE at the cluster level using an uncorrected height-threshold of $p < .001$ to define clusters. Next, we performed a SVC using the ROI mask of mentalizing regions (https://neurovault.org/images/25863/) and we found significant activation not only in the left angular gyrus ([x, y, z] = [-45, -58, 35], k = 69, *T*-value = 5.02, $p_{FWE} < .05$), but also in the left precuneus ([x, y, z] = [-3, -58, 38], k = 49, *T*-value = 4.65, $p_{FWE} < .05$) (**Fig. 15**). Given the exploratory nature of the study, we also decided to report findings using a more liberal threshold ($p_{uncorrected} < .001$, k > 0) that additionally yielded brain activations of the left middle frontal gyrus ([x, y, z] = [-39, 17, 53], k = 37, *T*-value = 5.17, $p_{uncorrected} < .001$), the left middle temporal gyrus ([x, y, z] = [-51, -40, 2], k = 40, *T*-value = 4.66, $p_{uncorrected} < .001$) and the left superior frontal gyrus ([x, y, z] = [-12, 26, 59], k = 43, *T*-value = 4.17, $p_{uncorrected} < .001$).

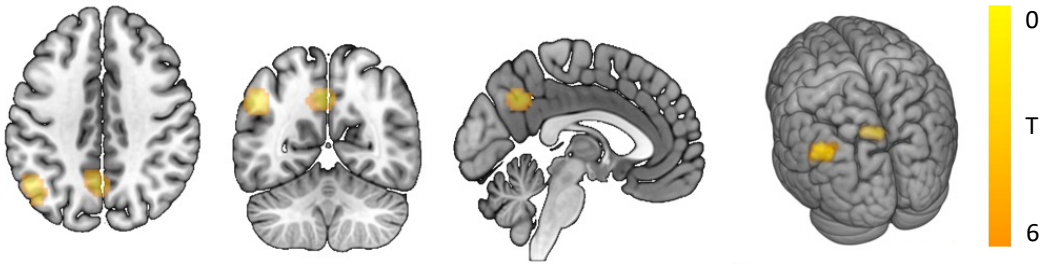


Fig. 15: Significant brain activations resulting from the contrast *Congruent* > *Unrelated comments*. SVC was performed to restrict search area to mentalising regions. SVC performed using FWE correction with a threshold of $p < 0.05$.

Incongruent > *Unrelated comments*: As-preregistered, we performed a whole-brain analysis that showed significant activation in the left angular gyrus ([x, y, z] = [-45, -61, 38], k = 118, *T*-value = 6.12) at $p < .05$, corrected for FWE at the cluster level using an uncorrected height-threshold of $p < .001$ to define clusters. Next, we performed a SVC using the ROI mask of mentalizing regions and we did not find additional significant activation than the left angular gyrus ([x, y, z] = [-45, -61, 38], k = 64, *T*-value = 5.08, $p_{FWE} < .05$) (**Fig. 16**). A more liberal threshold ($p_{uncorrected} < .001$, k > 0) additionally yielded brain activations of the left middle frontal gyrus ([x, y, z] = [-39, 17, 47], k = 28, *T*-value = 4.72, $p_{uncorrected} < .001$), the left precuneus ([x, y, z] = [-6, -58, 35], k = 36, *T*-value = 4.32, $p_{uncorrected} < .001$), the left superior frontal gyrus ([x, y, z] = [-9, 29, 56], k = 4, *T*-value = 3.92, $p_{uncorrected} < .001$), the left middle temporal gyrus ([x, y, z] = [-51, -37, -1], k = 6, *T*-value = 3.88, $p_{uncorrected} < .001$) and the right angular gyrus ([x, y, z] = [51, -61, 35], k = 2, *T*-value = 3.40, $p_{uncorrected} < .001$).

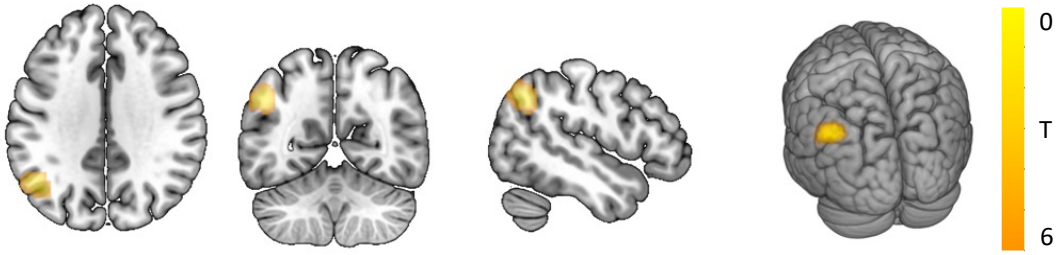


Fig. 16: Significant brain activations resulting from the contrast *Incongruent* > *Unrelated comments*. SVC was performed to restrict search area to mentalising regions. SVC performed using FWE correction with a threshold of $p < 0.05$.

Congruent + Incongruent > Unrelated comments: As pre-registered, we performed a whole-brain analysis that showed significant activation in the left angular gyrus ([x, y, z] = [-45, -58, 35], k = 130, *T*-value = 5.97) at $p < .05$, corrected for FWE at the cluster level using an uncorrected height-threshold of $p < .001$ to define clusters. Next, we performed a SVC using the ROI mask of mentalizing regions, we found significant activation not only in the left angular gyrus ([x, y, z] = [-45, -58, 35], k = 80, *T*-value = 5.97, $p_{FWE} < .05$), but also in the left middle frontal gyrus ([x, y, z] = [-39, 17, 50], k = 42, *T*-value = 5.24, $p_{FWE} < .05$) and in the left precuneus ([x, y, z] = [-6, -55, 35], k = 54, *T*-value = 4.79, $p_{FWE} < .05$) (**Fig. 17**). A more liberal threshold ($p_{uncorrected} < .001$, k > 0) additionally yielded brain activations of the left superior frontal gyrus ([x, y, z] = [-12, 26, 59], k = 43, *T*-value = 4.56, $p_{uncorrected} < .001$), the left middle temporal gyrus ([x, y, z] = [-51, -40, -1], k = 41, *T*-value = 4.60, $p_{uncorrected} < .001$) and the right angular gyrus ([x, y, z] = [51, -61, 35], k = 8, *T*-value = 3.70, $p_{uncorrected} < .001$).

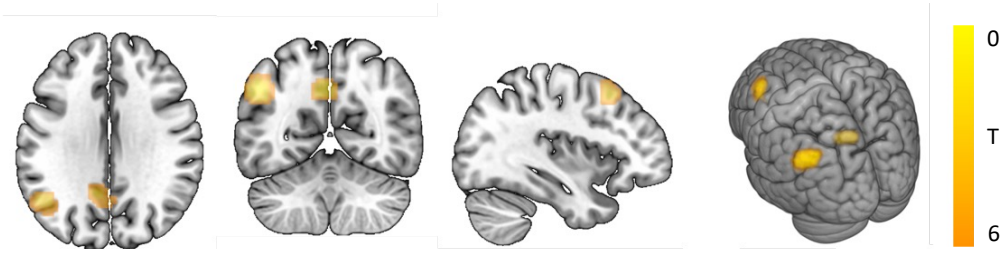


Fig. 17: Significant brain activations resulting from the contrast *Congruent + Incongruent > Unrelated comments*. SVC was performed to restrict search area to mentalising regions. SVC performed using FWE correction with a threshold of $p < 0.05$.

Incongruent > Congruent: We performed a whole-brain analysis that did not yield to significant activations at $p < .05$, corrected for FWE at the cluster level using an uncorrected height-threshold of $p < .001$ to define clusters. Also using a more liberal threshold ($p_{uncorrected} < .001$, k > 0) did not yield any significant brain activity. To further assess and compare the brain activations during the congruent and incongruent conditions, we conducted a Paired Samples T-Test between the average beta of the contrast *Congruent > Unrelated* and the average beta of the contrast *Incongruent > Unrelated* extracted using the ROI of the ToM network (previously implemented in the SVC). As shown in **Fig. 18**, the brain activation of the ToM network in the congruent condition ($M_{cong} = 0.089$, $SD_{cong} = 0.739$) was significantly higher ($t_{(2679)} = 4.721$, $p < .001$) than the one in the incongruent condition ($M_{incong} = 0.045$, $SD_{incong} = 0.015$).

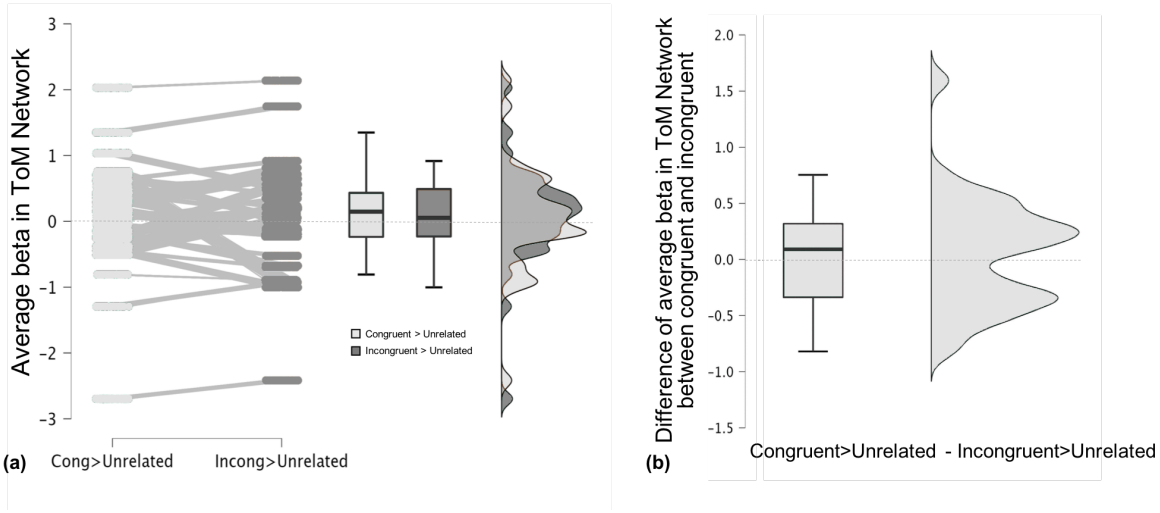


Fig. 18: Raincloud plots displaying **(a)** the average beta estimate of the ToM network for the *Congruent* > *Unrelated* contrast and for the *Incongruent* > *Unrelated* contrast and **(b)** the difference of the average beta estimates of the ToM Network between the *Congruent* > *Unrelated* and *Incongruent* > *Unrelated* contrasts.

After obtaining these results, which contradicted our original hypothesis, our objective was to investigate whether incongruent comments reduced the activity in the ToM network. We hypothesized that participants would avoid direct interaction with divergent opinions to maintain a positive self-concept about their viewpoints. Therefore, an exploratory correlation analysis was conducted using the average betas of the ToM network (when reading comments related to the news headline) and the attitudinal scores obtained from three questionnaires assessing attitudes towards the three contemporary topics used in the task. The result showed a decrease in the ToM network's activity among participants with stronger pre-existing attitudes towards the three contemporary topics, compared to those with weaker pre-existing attitudes (R = -0.03, p = .01) (**Fig. 19**)

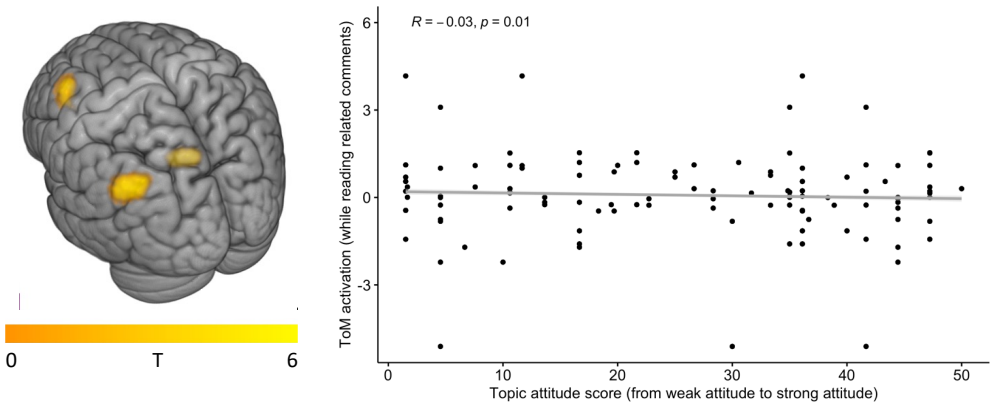


Fig. 19: Effect of pre-existing attitudes on brain activation. **(left)** Significant ToM activation

when reading comments related to the news headlines at $p_{FWE} < .05$, small volume corrected. **(right)** Stronger pre-existing attitudes were associated with decreased ToM activity when reading comments related to the news headlines.

Neural correlates of opinion updating

In these exploratory analyses, we aimed to assess the neural correlates of opinion updating after reading comments written by other people vs. when the second opinion was not updated.

Updated > Not updated: A whole-brain analysis did not show any significant activation at $p < .05$, corrected for FWE at the cluster level using an uncorrected height-threshold of $p < .001$ to define clusters. However, with a more liberal threshold ($p_{uncorrected} < .001$, $k \geq 5$), we found a significant increased BOLD signal in a region close to the right hippocampus ([x, y, z] = [32, -28, 0], k = 7, *T*-value = 3.99, $p_{uncorrected} < .001$) and in the right caudate ([x, y, z] = [15, 2, 20], k = 5, *T*-value = 3.50, $p_{uncorrected} < .001$).

Not updated > Updated: We performed a whole-brain analysis that did not yield to significant activations at $p < .05$, corrected for FWE at the cluster level using an uncorrected height-threshold of $p < .001$ to define clusters. Also using a more liberal threshold ($p_{uncorrected} < .001$, $k > 0$) did not yield any significant brain activation in the not updated conditions compared to the updated ones.

3.5 Discussion

Neuroimaging research on user-generated comments on social media is still in its early stages. Using a novel paradigm, our fMRI study aimed to advance this emerging field by examining the impact of online comments on opinion formation and the corresponding neural activity when individuals are exposed to these online opinion cues. In this study, participants read real news headlines posted on Facebook and provided their initial opinions. Subsequently, unlike our previous behavioral studies, participants were exposed to four comments presented one at a time and then rated their opinions a second time, all while their brain activity was recorded. In the following section, we will briefly address behavioral results covered in the previous discussion section and we will focus and explore novel findings in greater detail.

3.5.1 Behavioral Results

Effect of the valence of comments

Similar to our previous behavioral studies, we found that participants significantly updated their opinions in line with the social information provided, both after reading supportive and opposing comments towards the news headlines. Unlike before, we did not use mixed comments as a control condition; instead, we presented comments unrelated to the news headlines. In this case as well, participants did not change their opinions after reading these irrelevant comments. This finding, alongside our previous behavioral results about mixed comments, suggests that a clear majority of relevant comments is crucial for significantly influencing people's opinion (Wijenayake et al., 2020), which otherwise would remain more aligned with their initial view.

In line with previous research, our findings also indicated that user comments significantly influence opinion change, particularly when comments are critical of the news headlines. This pattern supports the well-documented phenomenon of negativity bias, where negative information has a stronger impact on individuals' perception compared to supportive information (Rozin & Royzman, 2001). This bias seems to be present also in social media interactions, were critical comments significantly influence readers' attitude and opinions negatively (Waddell, 2018; Williams & Hsieh, 2021; Winter et al., 2015). The impact of negativity bias is especially concerning regarding articles on well-established scientific findings, such as climate change and vaccination. Indeed, during the COVID-19 pandemic, critical comments on social media platforms likely shaped public perception of vaccine safety and efficacy, fueling skepticism and undermining public health efforts (Cascini et al., 2022). It is possible that when individuals encounter critical comments, they are primed to question more the validity of what they are reading, making them more susceptible to these comments than to supportive ones, which do not elicit the same level of vigilance.

Effect of comments' congruence

The above-mentioned negativity bias might be similar to what occurs when individuals encounter information that contradicts their own opinions, since such contradiction might prompt individuals to question more their position and trigger a desire to resolve the

conflict between their own views and those of the group (Festinger, 1957). Indeed, we found that comments incongruent with participants' initial opinion significantly influenced them to update their opinions more than congruent comments. As discussed in previous sections, the increased susceptibility to social influence might be interpreted as an effort to reduce cognitive dissonance and the discomfort of holding an opinion that differs from the group consensus (Festinger, 1957). Interestingly, since here we measured reaction times, we also found that participants spent more time reading incongruent comments compared to congruent ones. Furthermore, spending more time reading comments was positively correlated with greater opinion change. One possibility is that the need to reconcile the discrepant group opinion with their own opinions drove participants to spend more time reading the comments to understand the different viewpoints. Consequently, spending more time reading comments might make individuals more receptive to the content of the comments and increase their persuasive impact, leading to a greater opinion change and conformity with the perceived social norm (in this case, the majority opinion expressed in written comments).

Another intriguing and potentially more important effect of being confronted with incongruent comments is opinion depolarization. Although this thesis did not aim to directly address opinion polarization, in an exploratory analysis we computed a polarization index with *polarization*, *depolarization* and *no change* as the three variable's levels. We found that opinion depolarization was primarily driven by exposure to comments discrepant from individuals' personal opinion. This finding is consistent with prior research demonstrating depolarization following exposure to divergent viewpoints (Kubin & Sikorski, 2021). It is plausible that encountering discrepant viewpoints might prompt individuals to engage in more critical thinking and re-evaluate their own positions. Therefore, exposure to divergent opinions may be an effective strategy for mitigating opinion polarization on social media. Conversely, polarization was a consequence of both congruent and incongruent comments. It is possible that two parallels processes may act simultaneously and drive this last result. On one hand, individuals might naturally polarize their view after reading comments that align with their initial viewpoints. This is likely the process in place driving opinion polarization following recommendations algorithm on social media, where exposure to like-minded views strengthen and further polarize opinions (Santos, Lelkes & Levin., 2021). On the other hand, we also found that

incongruent comments may prompt participants to polarize their initial opinions. This phenomenon, known as the backlash effect, describes the tendency to distance oneself from divergent beliefs and opinions by reinforcing one's original position (Zhou, 2016). This reaction could lead to increased resistance to change or stronger adherence to pre-existing views.

Effect of pre-existing attitudes

This resistance to change one's own opinion also informed our hypothesis that strong pre-existing attitudes towards the topics of the study would make participants less susceptible to social influence. However, contrary to our previous behavioral studies, we did not find a negative correlation between the strength of pre-existing attitudes towards the three topics and the amount of opinion updating. Given that we observed a significant correlation in our previous behavioral studies but not in this fMRI study, it is plausible that this result could be due to the small power of the effect, as larger samples previously yielded significant results. Nevertheless, in an exploratory analysis, we found a negative correlation between the strength of the initial opinion rating and the amount of opinion updating, such that, stronger initial opinions towards the news headlines led to smaller opinion updates. These results are consistent with prior research demonstrating that individuals holding strong opinions or beliefs are more resistant to changing their views (Anglin, 2019; Shi et al., 2018).

Effect of digital maturity

Other interesting findings came from the construct of digital maturity. Although the negative correlation between the overall digital maturity index and opinion updating did not reach statistical significance (p = .058), we identified several correlations within the sub-scores of the index. Specifically, participants who showed smaller opinion updates tended to have: stronger regulation of impulses in digital context (Regulation of Aggressive Emotions), better respect towards others in digital contexts (Respect Towards Others), greater autonomy in choosing when to use their mobile devices (Autonomy of choice), and higher digital literacy (Digital Literacy). It is possible that individuals who are better at regulating aggressive emotions, who exercise better control over digital contexts and who have higher digital literacy, might be less reactive and therefore have more stable

opinions. However, contrary to our previous behavioral studies, we also found positive correlation with other sub-scales of the DIMI. Indeed, participants updated their opinions more when they had: higher awareness of the risks associated with digital environments (Risk Awareness), better regulation of negative emotions (Regulation of Negative Emotions), greater autonomy within digital contexts (Autonomy Within), and higher scores of digital citizenship (Digital Citizenship). The positive correlations between certain sub-dimensions of digital maturity and opinion updating could suggest that these dimensions may reflect autonomous and active engagement within digital contexts, which in turn could lead to a greater emphasis on online contents. Specifically, the ability to deliberately select and pursue goals in online context, along with active online social engagement, may enable individuals to critically evaluate what they read online (Laaber et al., 2023). As a result, when individuals encounter well-reasoned, argumentative comments, they may consciously decide to actively engage with the content and revise their opinions based on sound arguments (Winter et al., 2015). Given that we only used civil, argumentative comments, it is possible that higher levels of digital maturity enabled participants to autonomously decide to integrate the information from the comments into their own opinion. However, the above-mentioned interpretations are highly speculative, and the possibility of insufficient power for these correlations should also be considered when drawing conclusions.

3.5.2 fMRI Results

Neural correlates of processing relevant comments

The present fMRI study aimed to explore the neural mechanisms underlying the processing of user-generated comments below news headlines posted on social media. Our findings revealed a significant increase in activation within the core mentalizing regions, specifically the left PC, left TPJ, and left dlPFC, when participants read comments related to the news headlines compared to irrelevant comments. Given the cognitive nature of the task and the use of verbal, explicit stimuli, we primarily observed activation in regions associated with cognitive mentalizing processes (i.e., PFC, TPJ and PC), rather than affective mentalizing processes (e.g., vmPFC, IFG, insula, OFC). This finding aligns with the distinction in mentalizing area activations based on the nature of the task and stimuli used (Molenberghs et al., 2016). Moreover, we observed predominant activation

of this network in the left hemisphere, supporting previous studies on the left hemisphere's role in various aspects of language processing (Vigneau et al., 2006). For example, the left dlPFC is not only involved in working memory (Andrews et al., 2011; Petrides, 2000) and decision-making (Heekeren et al., 2006), but also in language comprehension and production (Klaus & Schutter, 2018). The significant activation of these brain regions when participants read comments related to the news headlines, compared to unrelated comments, could indicate that ToM regions are specifically engaged in encoding and reasoning about others' opinions, rather than responding to any irrelevant comment. This suggests that mentalizing regions are crucial for understanding and reasoning about socially relevant information, particularly regarding the beliefs and attitudes of others. Indeed, our findings support previous studies showing the involvement of mentalizing regions in encoding others' attitudes and opinions (Kliemann et al., 2008; Young and Saxe, 2008) and in reasoning about mental states of others (Monticelli et al., 2021; Kim et al., 2020). Young and Saxe (2009) observed activation in core ToM regions, particularly the right TPJ, when participants were presented with implicit morally relevant information, rather than any information, suggesting a spontaneous attribution of beliefs and intentions to others engaged in morally relevant actions. Although our study did not utilize explicitly moral information, the identity-related nature of the topics used in our task may have involved moral judgment processes linked to mentalizing regions. For example, reading comments attributing the causes of climate change to natural rather than anthropogenic factors may have triggered judgment processes, leading participants to judge the person who wrote the comment as a climate change denier.

Due to the exploratory nature of this new fMRI task, we conducted an additional whole-brain analysis with a more liberal threshold ($p < .001$, uncorrected) to further investigate brain regions beyond the mentalizing network and gain insights for future studies. When participants read comments related to the news headlines compared to unrelated comments, we observed significant activation of the left superior frontal gyrus (SFG), the left middle temporal gyrus (MTG) and the right angular gyrus (AG). Patients with lesions in the left SFG show impairments in working memory functions, suggesting the role of this area for short-term maintenance of relevant information (Boisgueheneuc et al., 2006). Additionally, the SFG is implicated with emotion regulation, such as reducing emotional reactions to morally and non-morally charged stimuli (Harenski & Hamann, 2006).

Therefore, it should not be excluded that reading comments related to controversial contemporary topics like climate change, vaccination, and veganism may engage emotional regulation processes compared to reading irrelevant comments. Activation in the left MTG, instead, is linked to: sematic and conceptual processing (Wei et al., 2012), encoding meaningful (vs meaningless) verbal materials, and arousal following novel information (see review from Martin, 1999). These findings are consistent with the nature of the comments provided in our task, which were semantic, meaningful and potential novel compared to comments unrelated to the news headlines. Finally, the right AG, which is part of the right TPJ, is not only engaged when accessing mental representations (as part of the ToM network), but also in complex language functions, abstract thinking, and moral judgment (see review from Seghier, 2013; Saxe & Young, 2008). Overall, these findings suggest that ToM regions are specifically involved in processing relevant information, whether moral or emotional, highlighting the importance of a meaningful context in eliciting reasoning about others' beliefs or intentions. Indeed, ToM may have an adaptive role in navigating and interpreting the social world, where understanding mental states and predicting behaviors of others are crucial for effective decision-making (Baron-Cohen, 1999).

Neural correlates of congruent and incongruent comments

After investigating the neural processes associated with reading comments related to the news headlines versus irrelevant comments, we were interested in separately examining the neural correlates of encountering comments congruent and incongruent to participants' initial opinions. We found significant activations of core mentalizing regions, specifically the left AG and left PC, when participants read comments congruent to their first opinion, and the left AG alone after incongruent comments. Whole-brain analyses using more liberal thresholds ($p < .001$, uncorrected) revealed additional activations in the left middle frontal gyrus (MFG), left MTG and left SFG for congruent comments. For incongruent comments, activations were observed in these regions as well as in the right AG. Previous research found activation of the right AG (part of the right TPJ) in moral judgments tasks involving explicit and implicit statements about others' beliefs (Kliemann et al., 2008; Young & Saxe, 2008; Young & Saxe, 2009). It is plausible that reading comments incongruent with one's opinions may engage more extensive moral judgment

processes than reading comments that align with personal beliefs. Additionally, prior research indicated that activity in the right TPJ decreases when maintaining positive impressions of ingroup members and increases when updating these impressions negatively (Kim et al., 2020; Park et al., 2020). Although our study did not explicitly differentiate between ingroup and outgroup conditions, due to the anonymous presentation of the comments, activation in this area may reflect social prediction errors when encountering discrepant opinions (Park et al., 2020), irrespective of group membership. Indeed, Decety and Lamm (2007) proposed a similar theory in their meta-analysis, suggesting that the right TPJ may be involved in comparing internal predictions with external incongruent outcomes during social cognition. However, given the exploratory nature of this analysis, this interpretation remains speculative.

Contrary to our pre-registered hypothesis, contrasting incongruent and congruent comments - to assess whether incongruent comments would activate ToM regions more than congruent comments - did not yield statistically significant differences in ToM activation. Unexpectedly, the beta average values in ToM regions indicated greater activation when participants read congruent comments compared to incongruent ones. This result contradicts previous research suggesting increased ToM activity when processing unexpected social information, likely due to the cognitive effort required to generate alternative explanations for others' beliefs and actions (Kim et al., 2020; Kliemann et al., 2008). Different factors may explain this discrepancy. First, Kim and colleagues (2020) showed increased ToM network activity when strong (vs. weak) prior beliefs about others' mental states and intentions were violated. In our study, participants had no prior beliefs about the commenters. Thus, it is possible that incongruent comments may not have violated any prior expectations about the commenters, leading to no increased ToM activation. This difference in task designs could account for the non-replication of previous results. Another possibility is that, contrary to our expectation that participants would mentalize more with incongruent comments, participants might have actually not actively engaged with these comments in order to protect their existing beliefs. This phenomenon, known as motivated reasoning, involves dismissing contradictory information to one's belief and overvaluing supporting evidence (Kunda, 1990). When facing discordant opinions, avoidance of unwanted information has been linked to decreased activation in brain regions like posterior medial frontal cortex (pMFC) (Kappes

et al., 2020), lateral PFC, dorsal anterior cingulate (dACC) and others (Hughes et al., 2017). In our study, it is possible that some participants might have passively discarded undesirable information, thereby reducing activation in mentalizing regions. Kim and colleagues (2020) also observed reduced ToM activation when participants were exposed to inconsistent information, suggesting disengagement from mentalizing when participants didn't feel the need to reconcile this new information with prior beliefs. Following the motivated reasoning hypothesis, an exploratory analysis of our data revealed that ToM activity decreased as participants' pre-existing attitudes about the topics strengthened, suggesting possible engagement in motivated reasoning and reduced consideration of alternative viewpoints. Moreover, the topics used in our study, while not explicitly political, may have triggered motivated reasoning due to their association with personal and political identities (Bolsen et al., 2014; Druckman & McGrath, 2019), particularly in the polarizing context of social media. However, these interpretations remain speculative and future studies should further investigate the neural correlates of motivated reasoning.

Neural correlates of opinion updating

Finally, we investigated the neural correlates of opinion updating by analyzing mental activation during the second opinion ratings. This non-pre-registered analysis did not yield significant results with a conservative threshold ($p < .05$, FWE-corrected). However, a more liberal threshold ($p < .001$, uncorrected) showed significant activation in the right hippocampus and right caudate when participants updated their opinions compared to when they did not. The hippocampus is traditionally involved in encoding and retrieving episodic and emotional information (Fanselow, 2010), and has also been implicated in learning new information within motivationally relevant contexts (Schriber & Guyer, 2016). Interestingly, recent research highlights the hippocampus' role in cognitive flexibility and social behaviors, essential for facilitating flexible use of novel information in social contexts, such as constructing, manipulating and updating mental representations (Rubin et al., 2014). Damage to the hippocampus can produce behavioral inflexibility, impairing the formation, integration and flexible use of information (see a review from Rubin et al., 2014). Cognitive flexibility seems essential in order to update one's opinion in face of new information, like in social contexts, and this process seems to rely on the hippocampal-

dependent memory system (Rubin et al., 2014). Therefore, in our study, hippocampal activation may signify the real-time cognitive flexibility necessary to gather, integrate and use new and familiar information for opinion updating.

We also found activation of the right caudate, part of the dorsal striatum, when participants updated their opinions. Like the hippocampus, the caudate supports cognitive flexibility and value updating over habitual stability (An et al., 2024). Similarly to hippocampal lesions, patients with traumatic brain injuries in the caudate exhibit strong impairments in executive functions, including information processing speed and cognitive flexibility (Xu et al., 2022), highlighting its importance for flexible behaviors. Both hippocampus and striatum are crucial for flexible spatial navigation and adaptation to environmental changes in both rodents and humans (see review from Gahnstrom & Spiers, 2020). Anatomically, strong connectivity between the hippocampus and subcortical regions, such as the amygdala and striatal regions, is associated with greater belief changes following exposure to favorable information (Moutsiana et al., 2015). Indeed, it seems that the interplay between hippocampus and striatal regions (of which the caudate is a component) supports behavioral flexibility, which is necessary for opinion updating. Overall, these findings suggest the possibility that opinion updating is mediated by cognitive flexibility processes, reflected in the activation of the hippocampus and striatal areas.

3.5.3 Limitations and Future Directions for Research

There are several limitations in this fMRI study. Firstly, this study represents an initial effort to measure the neural correlates of reading comments below news headlines on social media. To isolate the BOLD signals of different comment types more easily, we created distinct conditions with sets of four comments of the same type. However, in real online settings, user-generated comments are typically mixed in nature. Future studies should try to implement a mixed condition with both supportive and opposing comments to better capture neural correlates that more closely reflect real online environments. Another interesting research direction involves varying the source of the comments, for example differentiating between in-group and out-group, to observe whether social information is processed differently. On social media like Facebook, commenters usually have identifiable names and pictures, leading users to judge both the content of the comments and the perceived credibility of the commenter. Investigating these additional social cues

may provide deeper insights into the behavioral and neural mechanisms involved in social media interactions. A second limitation is our sample. First, we only tested people over 18 years old. Adolescents, who are major users of social media and appear to be more susceptible to social cues (see review from Ciranka & van den Bos, 2019), may exhibit different and potentially stronger effects both in influenceability and in their neural processing. Therefore, examining how online comments are represented in adolescents' brains would be valuable. Second, our sample size may have been insufficient to detect smaller effects, particularly for the non-pre-registered analysis of the neural correlates of opinion updating. This measure was calculated post hoc, and since not all participants updated their opinions, the analysis was conducted with a smaller dataset. Consequently, our sample may have been underpowered, limiting our ability to detect significant brain activations related to opinion updating with a stringent and corrected threshold. Future research should address these limitations by employing a larger and more diverse sample in order to increase robustness and generalizability of the findings. A third limitation concerns the different duration of comments' presentation across different stimuli and conditions. Our task was self-paced with a fixed maximum duration for each stimulus, allowing participants to read both comments and news headlines at their own pace. This variability in reading times could have introduced confounding effects in the BOLD signal. Specifically, different stimulus durations may have led to inconsistencies in the BOLD signals, complicating the disengagement of the content-related neural activity from the duration-related effects (Mumford et al., 2024). Therefore, we cannot exclude that the observed differences in neural activity between congruent and incongruent comments might be due to task design artifacts rather than effective differences in neural processing. To mitigate these potential confounding effects, future studies should try to standardize and optimize the duration of the comments and use reaction times as possible explanatory parameters (Mumford et al., 2024).

4. General Discussion and Conclusion

In an era where our lives are inevitably interconnected in online environments, social media pose a great challenge due to their pervasive influence and easy accessibility. While social media offer unparallelled connectivity, they might also pose threats for societal decision-making and democracies processes (Saunders, 2020). This PhD project started during the first year of COVID-19 pandemic, a period during which social media played a pivotal role in accelerating the spread of fake news and conspiracy theories about vaccine safety and in influencing public compliance with government recommendations, such as mask-wearing and social distancing (Cascini et al., 2022; Van Raemdonck, 2019). The goal of this PhD thesis was to contribute to shedding light into the cognitive and neural mechanisms by which social media comments influence individuals' perceptions of important contemporary topics. To achieve this, we first developed and tested a new paradigm aimed at isolating the influence of social media comments on personal opinions about news headlines. Subsequently, we investigated the neural correlates of being confronted with other people's opinion about these topics.

Our studies consistently demonstrated that user-generated comments significantly influenced participants' opinions on these critical contemporary issues. We corroborated previous behavioral findings that individuals are susceptible to be influenced, specifically when they have weaker prior attitudes about the topics and when their confidence in their initial opinions is lower. Additionally, we found that participants were more likely to be influenced by comments that were incongruent with their initial opinions. These findings all highlight the susceptibility to social influence exerted by user-generated comments. On the other hand, we also observed possible effects of being motivated to maintain one's belief, particularly when participants held stronger prior attitudes about the topics. Similarly, we also found a tendency to polarize one's initial opinion after reading incongruent comments, as a possible mechanism aiming at distancing oneself with the divergent group opinion (i.e. the backlash effect, Zhou, 2016). These two phenomena - being easily influenced and being more motivated to keep own's belief and even polarize initial opinions - are two complementary consequences of reading online comments. However, it is worth noting that neither of these outcomes is inherently negative. Sometimes it may be beneficial to be able to recognize and distance oneself from poor-

quality or extreme online information, just as it can be rational to change one's opinion in response to well-argued, sound viewpoints. The aim of this research is not to evaluate whether opinion change is inherently good or bad, but rather to raise awareness of how social media comments can influence our perceptions and opinions, even with a single exposure with anonymous comments in a laboratory setting.

Findings from this doctoral thesis and previous research highlight the need for improved regulation and moderation of social media platforms, as current measures remain highly insufficient. There are two primary ways through which positive change can be achieved. Firstly, social media platforms must be held accountable for their lack of transparency regarding recommendation algorithms and their ineffective content moderation mechanisms (Barbu, 2016). As demonstrated by this thesis, exposure to content that only aligns with existing beliefs strengthens those opinions and could lead to a distorted perception of reality. Indeed, recommendation algorithms and echo chambers on social media seem to significantly contribute to opinion polarization and partisan divisions (Schmidt et al., 2018). Social media platforms should mitigate the strong presence of recommendation algorithms and provide broader visibility to more heterogeneous content. By highlighting diverse content and encouraging users to engage with a variety of viewpoints, platforms might foster users critical thinking and facilitate mutual understanding, potentially leading to an array of less polarized and extreme viewpoints. Additionally, social media platforms should implement stricter regulations and moderations for posted content. Since the present doctoral thesis and previous research demonstrated that users are generally influenced by what they read on social media, uncivil, hateful, and misleading comments should be heavily moderated, while constructive and civil dialogue in the comment sections should be promoted (Williams & Hsieh, 2021). Secondly, another crucial responsibility is to increase people's digital literacy. Institutions should aim to make citizens more informed and less susceptible to negative influences on social media. This effort should begin in schools as children and adolescents are not only chronically online, but also are the most susceptible to social influences (Ahmed et al., 2020). By educating the younger generations about the potential dangers of social media and teaching them how to best navigate and critically evaluate the vast sea of online material, we can cultivate a more resilient population as active agents and not passive victims of online manipulations.

5. Abstract

Social media has become central in our daily life, serving both as a mean of interpersonal connection and as a primary source of news consumption. This doctoral dissertation investigates the behavioral and neuroimaging effects of online comments written below news headlines posted on social media platforms. We developed a novel behavioral paradigm to address limitations in previous research. This behavioral paradigm was validated using a sample of around 440 individuals from U.S. and Germany. Results indicated that participants consistently updated their opinions in accordance with the sentiment expressed in the comments. The degree of opinion change was significantly greater when participants had both strong pre-existing attitudes toward the topics and when they had high confidence in their initial opinions about the news headlines. Susceptibility to social influence was mitigated by the level of digital maturity. Subsequently, the behavioral task was adapted for fMRI compatibility. In a fMRI study with 41 participants, we corroborated our previous behavioral findings showing participants updating their opinions following the sentiment expressed in the comments. The update was significantly greater both when comments were incongruent with participants' initial opinions and when their initial opinion was weak. Comment congruence with initial opinions also led to increased opinion polarization. Neuroimaging data revealed activation in the theory of mind network when participants read comments related to the news headlines compared to irrelevant comments. The activation was smaller for comments incongruent with participants' initial opinions, possibly signifying a motivation not to engage with divergent viewpoints. In conclusion, this doctoral thesis demonstrates the powerful influence of online comments on social media in shaping public opinion on important contemporary issues, with potential implication for societal decision-making.

6. List of figures

7. List of tables

8. References

Ahmed, S., Foulkes, L., Leung, J. T., Griffin, C., Sakhardande, A., Bennett, M., ... & Blakemore, S. J. (2020). Susceptibility to prosocial and antisocial influence in adolescence. *Journal of Adolescence*, *84*, 56-68

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, *31*(2), 211-236

An, S. Y., Hwang, S. H., Lee, K., & Kim, H. F. (2024). Distinct representation of cognitive flexibility and habitual stability in the primate putamen, caudate, and ventral striatum. *bioRxiv*, 2024-02

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2014). The “nasty effect:” Online incivility and risk perceptions of emerging technologies. *Journal of computer-mediated communication*, *19*(3), 373-387

Andrews, S. C., Hoy, K. E., Enticott, P. G., Daskalakis, Z. J., & Fitzgerald, P. B. (2011). Improving working memory: the effect of combining cognitive activity and anodal transcranial direct current stimulation to the left dorsolateral prefrontal cortex. *Brain stimulation*, *4*(2), 84-89

Anglin, S. M. (2019). Do beliefs yield to evidence? Examining belief perseverance vs. change in response to congruent empirical findings. *Journal of Experimental Social Psychology*, *82*, 176-199

Axsom, D., Yates, S., & Chaiken, S. (1987). Audience response as a heuristic cue in persuasion. *Journal of personality and social psychology*, *53*(1), 30

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130-1132

Baptista, J. P., & Gradim, A. (2020). Understanding fake news consumption: A review. *Social Sciences*, *9*(10), 185

Barbu, C. M. (2016). Increasing the Trustworthiness of Recommendations by Exploiting Social Media Sources. In *Proceedings of the 10th ACM Conference on Recommender Systems* (pp. 447-450).

Baron-Cohen, S. (1999). *The evolution of a theory of mind* (pp. 261-277)

Basole, R. C. (2004). The value and impact of mobile information and communication technologies. In *Proceedings of the IFAC Symposium on Analysis, Modeling & Evaluation of Human-Machine Systems* (9), 1-7

Berger, J., & Milkman, K. L. (2012). What makes online content viral?. *Journal of marketing research*, *49*(2), 192-205

Boisgueheneuc, F. D., Levy, R., Volle, E., Seassau, M., Duffau, H., Kinkingnehun, S., ... & Dubois, B. (2006). Functions of the left superior frontal gyrus in humans: a lesion study. *Brain*, *129*(12), 3315-3328

Bolsen, T., Druckman, J. N., & Cook, F. L. (2014). The influence of partisan motivated reasoning on public opinion. *Political Behavior*, *36*, 235-262

Bossier, H., Roels, S. P., Seurinck, R., Banaschewski, T., Barker, G. J., Bokde, A. L., Quinlan, E. B., Desrivières, S., Flor, H., Grigis, A., Garavan, H., Gowland, P., Heinz, A., Ittermann, B., Martinot, J., Artiges, E., Nees, F., Orfanos, D. P., Poustka, L., . . . Moerkerke, B. (2020). The empirical replicability of task-based fMRI as a function of sample size. *NeuroImage*, *212*, 116601

Brett, M., Anton, J. L., Valabregue, R., & Poline, J. B. (2002). Region of interest analysis using the MarsBar toolbox for SPM 99. *Neuroimage*, *16*(2), S497

Cascini, F., Pantovic, A., Al-Ajlouni, Y. A., Failla, G., Puleo, V., Melnyk, A., ... & Ricciardi, W. (2022). Social media and attitudes towards a COVID-19 vaccination: A systematic review of the literature. *EClinicalMedicine*, *48*

Cascio, C. N., O'Donnell, M. B., Bayer, J., Tinney Jr, F. J., & Falk, E. B. (2015). Neural correlates of susceptibility to group opinions in online word-of-mouth recommendations. *Journal of Marketing Research*, *52*(4), 559-575

Christensen, R., & Knezek, G. (2015). The climate change attitude survey: Measuring middle school student beliefs and intentions to enact positive environmental change. *International Journal of Environmental and Science Education, 10*(5), 773-788.

Ciranka, S., & van den Bos, W. (2019). Social influence in adolescent decision-making: A formal framework. *Frontiers in Psychology*, *10*

Colliander, J. (2019). “This is fake news”: Investigating the role of conformity to other users’ views when commenting on and spreading disinformation in social media. *Computers in Human Behavior*, *97*, 202-215

DataReportal, & Meltwater, & We Are Social. (January 31, 2024). Number of internet and social media users worldwide as of January 2024 (in billions) [Graph]. In *Statista*. Retrieved from https://www.statista.com/statistics/617136/digital-population-worldwide/

Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition. *The neuroscientist*, *13*(6), 580-593

Demerouti, E., & Rispens, S. (2014). Improving the image of student-recruited samples: A commentary. *Journal of Occupational and Organizational Psychology*, *87*(1), 34-41

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in psychology*, *5*, 781

Doheny, M. M., & Lighthall, N. R. (2023). Social cognitive neuroscience in the digital age. *Frontiers in Human Neuroscience*

Druckman, J. N., & McGrath, M. C. (2019). The evidence for motivated reasoning in climate change preference formation. *Nature Climate Change*, *9*(2), 111-119

Dutceac Segesten, A., Bossetta, M., Holmberg, N., & Niehorster, D. (2022). The cueing power of comments on social media: how disagreement in Facebook comments affects user engagement with news. *Information, Communication & Society*, *25*(8), 1115-1134

Dvir-Gvirsman, S. (2019). I like what I see: Studying the influence of popularity cues on attention allocation and news selection. *Information, Communication & Society*, *22*(2), 286-305

Esposito, J. L., Agard, E., & Rosnow, R. L. (1984). Can confidentiality of data pay off?. *Personality and Individual Differences*, *5*(4), 477-480.

Fabian Prochazka, Patrick Weber, and Wolfgang Schweiger. 2018. Effects of Civility and Reasoning in User Comments on Perceived Journalistic Quality. *Journalism Studies 19*, 1 (2018), 62–78

Fanselow, M. S., & Dong, H. W. (2010). Are the dorsal and ventral hippocampus functionally distinct structures?. *Neuron*, *65*(1), 7-19

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175-191

Festinger, L. (1957). A theory of cognitive dissonance. Evanston, IL: Row, Peterson

Festinger, L. (1957). Social comparison theory. *Selective Exposure Theory*, *16*, 401

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 358*(1431), 459-473.

Gahnstrom, C. J., & Spiers, H. J. (2020). Striatal and hippocampal contributions to flexible navigation in rats and humans. *Brain and Neuroscience Advances*, *4*, 2398212820979772

Gallup Jr, G. G. (1985). Do minds exist in species other than our own?. *Neuroscience & Biobehavioral Reviews*, *9*(4), 631-641

Guo, B., Ding, Y., Sun, Y., Ma, S., Li, K., & Yu, Z. (2021). The mass, fake news, and cognition security. *Frontiers of Computer Science*, *15*, 1-13

Harenski, C. L., & Hamann, S. (2006). Neural correlates of regulating negative emotions related to moral violations. *Neuroimage*, *30*(1), 313-324

Hawkins, I., Roden, J., Attal, M., & Aqel, H. (2023). Race and gender intertwined: why intersecting identities matter for perceptions of incivility and content moderation on social media. *Journal of Communication*, *73*(6), 539-551

Heekeren, H. R., Marrett, S., Ruff, D. A., Bandettini, P. A., & Ungerleider, L. G. (2006). Involvement of human left dorsolateral prefrontal cortex in perceptual decision making is independent of response modality. *Proceedings of the National Academy of Sciences*, *103*(26), 10023-10028.

Hughes, B. L., Zaki, J., & Ambady, N. (2017). Motivation alters impression formation and related neural systems. *Social Cognitive and Affective Neuroscience*, *12*(1), 49-60

Jeffreys, H. (1998). *The theory of probability*. OuP Oxford

Jones, R., Colusso, L., Reinecke, K., & Hsieh, G. (2019, May). r/science: Challenges and opportunities in online science communication. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1-14)

Jordan, T. (2013). *Internet, society and culture: Communicative practices before and after the internet*. Bloomsbury Publishing USA.

Kaiser, J., & Rauchfleisch, A. (2018). Unite the right? How YouTube's recommendation algorithm connects the US far-right. *Data & Society Media Manipulation*, *11*

Kappes, A., Harvey, A. H., Lohrenz, T., Montague, P. R., & Sharot, T. (2020). Confirmation bias in the utilization of others' opinion strength. *Nature neuroscience*, *23*(1), 130-137

Kim, J. W. (2018). They liked and shared: Effects of social media virality metrics on perceptions of message influence and behavioral intentions. *Computers in Human Behavior*, *84*, 153-161

Kim, M. J., Mende-Siedlecki, P., Anzellotti, S., & Young, L. (2021). Theory of mind following the violation of strong and weak prior beliefs. *Cerebral Cortex*, *31*(2), 884-898

Klaus, J., & Schutter, D. J. (2018). The role of left dorsolateral prefrontal cortex in language processing. *Neuroscience*, *377*, 197-205

Kliemann, D., Young, L., Scholz, J., & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia, 46*(12), 2949-2957

Köcher, R. (2016). "Flüchtlingszustrom: Auswirkungen eines gesellschaftlichen Aufregungszyklus auf politisches Interesse und Mediennutzung." *AWA 2016*. Retrieved from www.ifd-allensbach.de/fileadmin/AWA/AWA_Praesentationen/2016/AWA_2016_Koecher_Fluechtlingskrise_Medien.pdf

Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest*, *21*(3), 103-156

Ksiazek, T. B., & Springer, N. (2018). User comments in digital journalism: Current research and future directions. *The Routledge handbook of developments in digital journalism studies*, 475-486

Kubin, E., & Von Sikorski, C. (2021). The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, *45*(3), 188-206

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, *108*(3), 480.

Laaber, F., Florack, A., Koch, T., & Hubert, M. (2023). Digital maturity: Development and validation of the Digital Maturity Inventory (DIMI). *Computers in Human Behavior*, *143*, 107709

Lang, A. (2000). The limited capacity model of mediated message processing. *Journal of communication*, *50*(1), 46-70

Lee, E. J., & Jang, Y. J. (2010). What do others' reactions to news on internet portal sites tell us? Effects of presentation format and readers' need for cognition on reality perception. *Communication research*, *37*(6), 825-846

Lee, E. J., & Tandoc Jr, E. C. (2017). When news meets the audience: How audience feedback online affects news production and consumption. *Human communication research*, *43*(4), 436-449

Lee, E. J., Jang, Y. J., & Chung, M. (2021). When and how user comments affect news readers' personal opinion: perceived public opinion and perceived news position as mediators. *Digital Journalism*, *9*(1), 42-63

Lee, H., & Oh, H. J. (2017). Normative mechanism of rumor dissemination on Twitter. *Cyberpsychology, Behavior, and Social Networking*, *20*(3), 164-171

Lelisho, M. E., Pandey, D., Alemu, B. D., Pandey, B. K., & Tareke, S. A. (2023). The negative impact of social media during COVID-19 pandemic. *Trends in Psychology*, *31*(1), 123-142

Lieberman, M. D., & Cunningham, W. A. (2009). Type I and Type II error concerns in fMRI research: re-balancing the scale. *Social cognitive and affective neuroscience*, *4*(4), 423-428

Livingstone, S., Mascheroni, G., & Staksrud, E. (2018). European research on children's internet use: Assessing the past and anticipating the future. *New Media & Society, 20*(3), 1103-1122

Martin, A. (1999). Automatic activation of the medial temporal lobe during encoding: Lateralized influences of meaning and novelty. *Hippocampus*, *9*(1), 62-70

Martin, L. R., & Petrie, K. J. (2017). Understanding the dimensions of anti-vaccination attitudes: The vaccination attitudes examination (VAX) scale. *Annals of Behavioral Medicine, 51*(5), 652-660.

McCambridge, J., de Bruin, M., & Witton, J. (2012). The effects of demand characteristics on research participant behaviours in non-laboratory settings: a systematic review.

Meshi, D., Tamir, D. I., & Heekeren, H. R. (2015). The emerging neuroscience of social media. *Trends in cognitive sciences*, *19*(12), 771-782

Molleman, L., Kanngiesser, P., & van den Bos, W. (2019). Social information use in adolescents: The impact of adults, peers and household composition. *PloS one*, *14*(11), e0225498

Molenberghs, P., Johnson, H., Henry, J. D., & Mattingley, J. B. (2016). Understanding the minds of others: A neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews*, *65*, 276-291

Monticelli, M., Zeppa, P., Mammi, M., Penner, F., Melcarne, A., Zenga, F., & Garbossa, D. (2021). Where we mentalize: Main cortical areas involved in mentalization. *Frontiers in Neurology*, *12*, 712532

Moutsiana, C., Charpentier, C. J., Garrett, N., Cohen, M. X., & Sharot, T. (2015). Human frontal–subcortical circuit and asymmetric belief updating. *Journal of Neuroscience*, *35*(42), 14077-14085

Mumford, J. A., Bissett, P. G., Jones, H. M., Shim, S., Rios, J. A. H., & Poldrack, R. A. (2024). The response time paradox in functional magnetic resonance imaging analyses. *Nature Human Behaviour*, *8*(2), 349-360

Mummolo, J., & Peterson, E. (2019). Demand effects in survey experiments: An empirical assessment. *American Political Science Review*, *113*(2), 517-529.

Nadarevic, L., Reber, R., Helmecke, A. J., & Köse, D. (2020). Perceived truth of statements and simulated social media postings: an experimental investigation of source credibility, repeated exposure, and presentation format. *Cognitive Research: Principles and Implications*, *5*(1), 1-16

Neubaum, G., & Krämer, N. C. (2017). Monitoring the opinion of the crowd: Psychological mechanisms underlying public opinion perceptions on social media. *Media psychology*, *20*(3), 502-531

Noelle-Neumann, E. (1974). The spiral of silence a theory of public opinion. *Journal of communication*, *24*(2), 43-51

Orne, M. T. (2017). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. In *Sociological methods* (pp. 279-299). Routledge

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, *5*(5), 411-419

Papacharissi, Z. (2002). The virtual sphere: The internet as a public sphere. *New media & society*, *4*(1), 9-27

Park, B., Fareri, D., Delgado, M., & Young, L. (2021). The role of right temporoparietal junction in processing social prediction error across relationship contexts. *Social Cognitive and Affective Neuroscience*, *16*(8), 772-781

Park, C. (2001). News Media Exposure and Self-Perceived Knowledge: The Illusion of Knowing. *International Journal of Public Opinion Research*, 13(4), 419-425

Paslakis, G., Richardson, C., Nöhre, M., Brähler, E., Holzapfel, C., Hilbert, A., & de Zwaan, M. (2020). Prevalence and psychopathology of vegetarians and vegans–Results from a representative survey in Germany. *Scientific reports, 10*(1), 6840.

Pearson, G. (2021). Sources on social media: Information context collapse and volume of content as predictors of source blindness. *New Media & Society*, *23*(5), 1181-1199

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590-595

Petrides, M. (2000). The role of the mid-dorsolateral prefrontal cortex in working memory. *Experimental brain research*, *133*, 44-54

Posit team (2024). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind?. *Behavioral and brain sciences*, *1*(4), 515-526.

Prochazka, F., Weber, P., & Schweiger, W. (2018). Effects of civility and reasoning in user comments on perceived journalistic quality. *Journalism studies*, *19*(1), 62-78

Rauchfleisch, A., & Kaiser, J. (2020). The German far-right on YouTube: An analysis of user overlap and user comments. *Journal of Broadcasting & Electronic Media*, *64*(3), 373-396

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and social psychology review*, *5*(4), 296-320

Rubin, R. D., Watson, P. D., Duff, M. C., & Cohen, N. J. (2014). The role of the hippocampus in flexible cognition and social behavior. *Frontiers in human neuroscience*, *8*, 742

Santos, F. P., Lelkes, Y., & Levin, S. A. (2021). Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences*, *118*(50), e2102141118

Saunders, J. (2020). Dark advertising and the democratic process. *Big Data and Democracy*, 73-85.

Schäfer, S. (2020). Illusion of knowledge through Facebook news? Effects of snack news in a news feed on perceived knowledge, attitude strength, and willingness for discussions. *Computers in human behavior*, *103*, 1-12

Schäfer, S., Sülflow, M., & Müller, P. (2017). The special taste of snack news: An application of niche theory to understand the appeal of Facebook as a news source. *First Monday*

Schmidt, A. L., Zollo, F., Scala, A., Betsch, C., & Quattrociocchi, W. (2018). Polarization of the vaccination debate on Facebook. *Vaccine*, *36*(25), 3606-3612

Schriber, R. A., & Guyer, A. E. (2016). Adolescent neurobiological susceptibility to social context. *Developmental cognitive neuroscience*, *19*, 1-18

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *42*, 9-34

Seghier, M. L. (2013). The angular gyrus: multiple functions and multiple subdivisions. *The Neuroscientist*, *19*(1), 43-61

Shearer, E., & Grieco, E. (2019). *Americans are Wary of the Role Social Media Sites Play in Delivering the News: Getting News from Social Media is an Increasingly Common Experience, Nearly Three-in-ten US Adults Do So Often*. Pew Research Center

Shi, R., Messaris, P., & Cappella, J. N. (2014). Effects of online comments on smokers' perception of antismoking public service announcements. *Journal of Computer-Mediated Communication*, *19*(4), 975-990

Springer, N., Engelmann, I., & Pfaffinger, C. (2015). User comments: Motives and inhibitors to write and read. *Information, Communication & Society*, *18*(7), 798-815

Steinfeld, N., Samuel-Azran, T., & Lev-On, A. (2016). User comments and public opinion: Findings from an eye-tracking experiment. *Computers in human behavior*, *61*, 63-72

Stroud, N. J., Van Duyn, E., & Peacock, C. (2016). News commenters and news comment readers. *Engaging News Project*, *21*

Sude, D. J., Knobloch-Westerwick, S., Robinson, M. J., & Westerwick, A. (2019). "Pick and choose" opinion climate: How browsing of political messages shapes public opinion perceptions and attitudes. *Communication Monographs*, *86*(4), 457-478

Sülflow, M., Schäfer, S., & Winter, S. (2019). Selective attention in the news feed: An eye-tracking study on the perception and selection of political news posts on Facebook. *new media & society*, *21*(1), 168-190

Sundar, S. S., Oeldorf-Hirsch, A., & Xu, Q. (2008). The bandwagon effect of collaborative filtering technology. In *CHI'08 extended abstracts on Human factors in computing systems* (pp. 3453-3458)

Tajfel, H., Turner, J. C., Austin, W. G., & Worchel, S. (1979). An integrative theory of intergroup conflict. *Organizational identity: A reader*, *56*(65), 9780203505984-16.

Tamir, D. I., & Ward, A. F. (2015). Old desires, new media. *The psychology of desire*, 432-455

Trott, V., Li, N., Fordyce, R., & Andrejevic, M. (2021). Shedding light on 'dark'ads. *Continuum*, *35*(5), 761-774

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, *185*(4157), 1124-1131

Van Overwalle, F. (2009). Social cognition and the brain: a meta-analysis. *Human brain mapping*, *30*(3), 829-858

Van Raemdonck, N. (2019). The echo chamber of anti-vaccination conspiracies: Mechanisms of radicalization on Facebook and Reddit. *Institute for Policy, Advocacy and Governance (IPAG) Knowledge Series, Forthcoming*

Vigneau, M., Beaucousin, V., Hervé, P. Y., Duffau, H., Crivello, F., Houde, O., ... & Tzourio-Mazoyer, N. (2006). Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. *Neuroimage*, *30*(4), 1414-1432

Von Hohenberg, B. C., & Guess, A. M. (2023). When do sources persuade? The effect of source credibility on opinion change. *Journal of Experimental Political Science*, *10*(3), 328-342

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *science*, *359*(6380), 1146-1151

Waddell, T. F. (2018). What does the crowd think? How online comments and popularity metrics affect news credibility and issue importance. *New Media & Society*, *20*(8), 3068-3083

Waddell, T. F. (2019). When comments and quotes collide: How exemplars and prior attitudes affect news credibility. *Journalism Studies*, *20*(11), 1598-1616

Weber, P. (2014). Discussions in the comments section: Factors influencing participation and interactivity in online newspapers' reader comments. *New media & society*, *16*(6), 941-957

Wei, T., Liang, X., He, Y., Zang, Y., Han, Z., Caramazza, A., & Bi, Y. (2012). Predicting conceptual processing capacity from spontaneous neuronal activity of the left middle temporal gyrus. *Journal of Neuroscience*, *32*(2), 481-489

Welborn, B. L., Lieberman, M. D., Goldenberg, D., Fuligni, A. J., Galvan, A., & Telzer, E. H. (2016). Neural mechanisms of social influence in adolescence. *Social cognitive and affective neuroscience*, *11*(1), 100-109

Wijenayake, S., Hettiachchi, D., Hosio, S., Kostakos, V., & Goncalves, J. (2020). Effect of conformity on perceived trustworthiness of news in social media. *IEEE Internet Computing*, *25*(1), 12-19

Williams, S., & Hsieh, G. (2021). The effects of user comments on science news engagement. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1-29

Winter, S., Brückner, C., & Krämer, N. C. (2015). They came, they liked, they commented: Social influence on Facebook news channels. *Cyberpsychology, Behavior, and Social Networking*, *18*(8), 431-436

Wood, W., Kallgren, C. A., & Preisler, R. M. (1985). Access to attitude-relevant information in memory as a determinant of persuasion: The role of message attributes. *Journal of Experimental Social Psychology*, *21*(1), 73-85

Xu, H., Zhang, X., & Bai, G. (2022). Abnormal dorsal caudate activation mediated impaired cognitive flexibility in mild traumatic brain injury. *Journal of Clinical Medicine*, *11*(9), 2484

Xu, Q. (2013). Social recommendation, source credibility, and recency: Effects of news cues in a social bookmarking website. *Journalism & Mass Communication Quarterly*, *90*(4), 757-775

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *neuroimage*, *40*(4), 1912-1920

Young, L., & Saxe, R. (2009). An FMRI investigation of spontaneous mental state inference for moral judgment. *Journal of cognitive neuroscience, 21*(7), 1396-1405

Zhou, J. (2016). Boomerangs versus javelins: How polarization constrains communication on climate change. *Environmental politics*, *25*(5), 788-811

Zillmann, D., & Brosius, H. B. (2012). *Exemplification in communication: The influence of case reports on the perception of issues*. Routledge

9. Acknowledgements

I want to express my gratitude to everyone who contributed to this doctoral thesis.

Thank you, Johannes, for your invaluable time and support, and for fostering such a friendly and encouraging atmosphere at CENs. Thank you, Prof. Silke Lux, Prof. Wouter van den Bos, and Prof. Bernd Weber, I am deeply grateful for your guidance and the insight you shared on this journey. Thank you, Angela, for being such a wonderful friend, always by my side. Thank you, Dominik, for being a great friend and for your generosity and help in any matter. Thank you, Omar, for your constant assistance with IT and the fun time we shared in Ghent. Thank you, Holger, for your relentless pursuit of quality and for your insightful feedback. Thank you, Qëndresa, for being always ready to help, the great conversations, and fun time together. Thank you, Daniela, for our lovely chats, your help, and the nice empowerment workshop. Thank you, Jana, for your assistance, fun moments together, and the great conversations. Thank you, Nahid, for your support and nice conversations. Thank you, Aline, for the insight into my project and the fun chats. To Nayara, Neha, Emir, Lennard, and Xenia, thank you for your research support and the great moments we shared. Finally, a special thanks to my close student collaborators over these years, Olimpia, Annkathrin, Nicole, and especially Jan; you are great contributors of this thesis and made the work much more enjoyable.

Beyond Academia, my heart is full with gratitude for those who offered me constant emotional end motivational support.

Thank you, Silvia, for being a true friend and sharing countless fun and deep moments together. Thank you, Mariachiara, for being my lost sister and my genial friend. Thank you, AnnaPaola, for always lighting my mood with funny and chaotic conversations. And lastly, Stefan, thank you for being the life partner I never imagined but am eternally grateful to have found. This doctoral thesis is dedicated to my wonderful family; Emilia, Antonio and Alessio, you are the greatest certainty of my life; everything I am and achieved is because of you.

Disclosure: during the preparation of this work the author used ChatGPT-3.5 in order to improve the language of the manuscript. After using this tool, the author reviewed and edited the content as needed and take full responsibility for the content of the dissertation.