



Popping misinformation bubbles: persona-informed style-based messaging in Facebook ad experiments

JooYoung Lee¹ · Emily Booth¹ · Hany Farid² · Marian-Andrei Rizoiu¹

Received: 20 February 2025 / Accepted: 8 September 2026
© The Author(s) 2026

Abstract

Misinformation is often viewed as an information problem, attributed to people's poor understanding of consensus-based facts. While this is a contributing factor, we propose that misinformation and conspiracy are also communicated differently to their audiences. Can we better reach these audiences by matching the styles used in online communities? We present the findings of a multiple Facebook ad study in which we tailored consensus-based messages to communication styles associated with three persona-informed Facebook audiences. Our results demonstrate that styled messages led to higher engagement among targeted Facebook audiences than the neutral-toned variants. We introduce our concept of the misinformation ecosystem, based on our Australian dataset, as a way of understanding online communities and their communication styles, and develop three personas whose demographic and interest profiles informed the Facebook ad audiences used in our experiment. Our findings suggest that styled messaging about consensus-based information can better engage audiences associated with vulnerable communities otherwise resistant to external influence, providing considerable implications about possible future interventions to counter misinformation.

Keywords Misinformation ecosystem · Styled messages · Vulnerable personas

✉ JooYoung Lee
jooyoung.lee@uts.edu.au

Emily Booth
Emily.Booth@uts.edu.au

Hany Farid
hfarid@berkeley.edu

Marian-Andrei Rizoiu
Marian-Andrei.Rizoiu@uts.edu.au

¹ Data Science Institute, University of Technology Sydney, 15 Broadway, 2007 Ultimo, New South Wales, Australia

² School of Information, University of California, Berkeley, 2000 Carleton Street, 94720 Berkeley, California, USA

1 Introduction

Misinformation has typically been defined as a category of material with factual inaccuracies, “fake news”, and other fabricated narratives, including political propaganda ([1, 2]). This definition exists in opposition to concepts of accurate or consensus-based information (i.e. not misinformation), in that it spotlights the “inaccurate element” of a piece of content in order to categorize it as misinformation. While this is no doubt useful for the purposes of classification, such definitions also carry inherent assumptions that can inhibit attempted interventions. For example, the highly polarised nature of social media, and the ability for opinions and interpretations of virtually any event to be published without adhering to journalistic standards of accuracy and objectivity mean that the classification of content as misinformation is not always straightforward. Likewise, the understanding of misinformation, as defined by the presence of inaccuracies, has lent itself to a series of interventions that, while logical, are not always effective. For example, fact-checking initiatives, while a popular intervention, are when they seek to change the opinions of people who find the facts counter to their preference [3]. Furthermore, visuals are commonly used despite the fact that they inhibit the potential corrective power of fact-checking tools [3]. Additionally, content moderation has had limited impacts on the visibility of misinformation content and pages on platforms such as Facebook [4]. Therefore, it follows that if these strategies are only suitable for those who are already predisposed to believe them then the most vulnerable cohorts—those distrustful or unwilling to engage with authoritative sources like fact-checkers—are beyond the reach of current intervention methods. The challenge, then, is to design new intervention methods that can reach them.

Technological factors, while enabling the spread of misinformation, simultaneously inhibit our ability to reach these communities. Data voids, defined as the outcome of “search terms for which the available relevant data is limited, non-existent, or deeply problematic” ([5], p. 1), provide rich opportunities for everyone from conspiracy theorists to con-artists to profit. By building custom landing pages for oblique search terms that have few results, and then disseminating the search terms online accompanied by encouragement that users “do their own research” and input the term into a search engine, these bad actors can drive people down rabbit holes they might never have otherwise sought out ([5], p. 3). These strategies deliberately misdirect users and manipulate *how* they conduct their fact-checking, which in turn makes it less likely that users will find a credible source to inform them. Likewise, recommendation algorithms, used across social media platforms, have been demonstrated to progressively drive audiences towards more polarizing fringe content that they may never have sought out on their own [6]. This pushes users into increasingly insular communities rather than exposing them to a diversity of content that could provide them with numerous perspectives on a topic. Digital media, more generally, have already had widespread political impacts: Lorenz-Spreen et al. [7] find both correlational and causal evidence that the introduction of digital media into established democracies, such as the United States, contributes to the “declining political trust, increasing populism and growing polarization” among the population, thereby undermining democracy (p.74). Although most people are not totally cut off from

sources that can counter unreliable online content, Booth et al. [8] used testimonies from former members of fringe communities to show that people who were highly invested in the online community severed ties with friends and family, and actively avoided mainstream information sources. This indicates that such people do exist, beyond the reach of existing interventions.

If simply making consensus-based information available does not halt the spread of misinformation, then it follows that misinformation may not be an information problem, but a human problem. Existing research has explored this human side, and looking into the psychological traits and profiles of people who typically believe misinformation and conspiracy theories provides some insight. Uncertainty, insecurity, alienation, and distrust in authorities have been found to commonly occur in believers of conspiracy theories and misinformation [9–13]. Additionally, whether or not someone believes a conspiracy – or generally, a piece of information – depends on pre-existing views on the topic ([14] p.6). However, we argue that this human-centric approach needs to be taken further, and that online misinformation needs to be studied in context. Who is posting and consuming it? What are their apparent motivations or needs for doing so? How is misinformation being shared, so as to engage people effectively? Although the reach and consequences of misinformation are amplified by the increasing use of online social media as a primary source of news and information by many people [15], the existence of misinformation does not guarantee believers. What makes certain misinformation narratives appealing to their believers? And can we appropriate these attributes to design interventions?

In this article, we explore the design of interventions that allow us to reach consumers of misinformation, focusing on the Australian environment. We first investigate why fact-based interventions can fail by identifying the types of people active in these communities, understanding the community structure and communication style, and the perceived barriers to fact-based communication. From this qualitative research, we propose the concept of misinformation ecosystems, in which concentrated networks of misinformation producers and consumers initially form around key topics of concern, but remain part of the network for reasons other than (mis)information about that catalyst topic. We illustrate the core components of a misinformation ecosystem, including the distinctive communication styles used in each group, and provide three personas to represent the typical demographic traits and concerns of the members from some of the ecosystems studied. These personas help us to understand the human factors that fuel people's desire to remain connected to these communities, and how they can gradually become immune to outsider interventions like fact-checking.

Next, we design outreach measures. In a Facebook ad experiment, we demonstrate how adopting the communication styles of these groups and applying them to consensus-based information can be used to target messages towards these users; and that this can lead to greater uptake of interest in evidence-based resources compared to the generic, evidence-based format that is commonly-employed in fact-checking methods, which we also test. Building on this, we also provide the results of deliberate error tests designed to act as controls for our experiment. Although these results do not prove that people's perspectives were changed, they do demonstrate that people were reached because they actively clicked on the ads. From our results, we argue

that understanding the key components of a misinformation ecosystem enables the design of more targeted interventions that can reach users who may have cut themselves off from external influences.

2 Methods

Prior to data collection, the study received full ethical clearance from the Human Research Ethics Committee of the University of Technology (application ID ETH23-8018). This also includes a waiver of consent for the experiment involving the ads, in compliance with the Australian National Statement on Ethical Conduct in Research Involving Humans.

2.1 Qualitative development of the misinformation Ecosystem

At the outset of this study, we first sought to understand the online misinformation landscape in Australia, to distinguish it from other contexts like the United States of America. We began by identifying numerous Facebook groups that we suspected posted or facilitated misinformation about particular topics of concern, such as vaccination skepticism and far-right groups. This was done through a search of Facebook pages for keywords; and after identifying a relevant page, looking into what pages it followed or engaged with on the platform. We manually assembled two lists of Facebook pages for specific ideologies, namely Australian Antivax and Far-right groups. Using CrowdTangle API, we collected all posts from the Facebook groups between January 23, 2021 and February 24, 2023. Through this process, we collected 6017 posts from the Far-right groups and 2969 posts from the Antivax groups. We conducted qualitative online observation of these pages and groups over several months, during which time detailed field notes were recorded, capturing the nature of posts on the page (i.e. topic and writing style), the types of people who appeared to be active on the page based on comments, and links and other content posted to it.

Due to limitations on the team's size, only one member of the team was qualified to conduct the qualitative analysis that developed the style guides and field notes. The field notes served as an observational record documenting recurring demographic characteristics, community concerns, interests, communication patterns, and writing styles observed throughout the ethnographic observation. Rather than being coded themselves, these observations were synthesised to develop the personas and communication style guides used in the subsequent stages of the study. As the qualitative analysis was conducted by a single qualified researcher, formal assessment of inter-coder agreement was not applicable. To address this, we also computationally collected posts from the groups under observation. A total of 8,986 posts were collected. Following the development of the communication style guides, the same researcher manually coded a random sample of 66 de-identified posts against these existing categories as a robustness check. While some posts were only one or two sentences, some were extended paragraphs, thus providing ample data to code and allowing data saturation to be achieved. This process, done across 2 stages, allowed categories to be robustly evaluated, independent of any bias attached to the source.

During this observation period, the research team had weekly meetings to discuss the observations. Questions that were raised about the online communities were recorded, so that any relevant observations that might answer the question could be fed back to the group. Thus, it became an iterative process whereby new observations continually led to new insights, which then guided new observations. During this time, we observed that different groups may share the same topic of concern, yet depending on the community demographic, these topics would be framed in different ways by members. For example, an anti-vaccination group that was dominated by young women would more likely frame their opposition as being a choice to embrace more natural health remedies and lifestyles, whereas opposition to vaccination could arise in far-right communities as part of their larger hostility against doctors and other members of the medical establishment (or “the elite”).

In summary, the qualitative workflow consisted of ethnographic observation of online communities, recording field notes to document recurring characteristics and communication patterns, synthesising these observations into personas and communication style guides, and subsequently coding a random sample of de-identified posts against the resulting style categories as a robustness check.

2.2 Qualitative development of the personas

The concept of a persona as a representative character intentionally designed to embody a group of people within a specific, pre-determined context for research was coined by Alan Cooper in 1999 [16], as a tool for software developers to better understand the needs of their users and design their products accordingly. A persona in this context can be understood as a figure based on empirical data about real people in which the most common features are combined to create a representative character, with supplementary details typically added through elements of fiction that naturally extend from the data. Personas are not stereotypes: a stereotype is a pre-existing assumption about a particular demographic that may be applied to this group regardless of context and can be transmitted at a broader societal level such as in entertainment media; whereas a persona is an evidence-based understanding of a predefined cohort, which is bound to the context it is created in and cannot be used to generalize elsewhere [17]. Personas have been widely used in both research and industry including in the contexts of business, health, IT and software design, and more [18]. Personas are beneficial for researchers working with sensitive data, as they enable the representation of a cohort without ever identifying an individual, thus addressing ethical concerns around identification. Personas are most frequently developed from data gathered through qualitative research methods, such as focus groups, interviews, ethnography, and observation, typically taking at least several months to reach data saturation [19]. Data saturation refers to the point in qualitative research where no new findings are discovered, and the data gathered merely repeats what has already been gathered. At this point, data collection is terminated because it is determined that there is nothing new to discover.

Developing a persona involves first identifying and cohesively drawing together the most relevant and prominent features in the data, such as demographic features, and perspectives expressed like frustrations or hopes. The result is a clearly-defined

character that can be named and envisioned as a real person, from which we can better understand their specific wants, fears, and goals [16]. In this study, due to safety considerations involved with engaging fringe communities directly, we conducted months of online observation of key communities, instead of directly interacting with members, and recorded these observations in detailed field notes. We compiled a list of the most common demographic traits of each group, such as approximate age, gender, and occupation, as well as the most common interests, like raising a family or achieving physical fitness, and fears and hopes. For example, observation of one group containing mostly young men who were concerned about the implementation of 5 G towers in Australia and feared its impacts on their physical health led to the development of the persona of Aaron. Based on the distinctions between communities noted in the above section, we designed three personas that accurately embodied each community as a whole by combining the most common traits into named entities. Personas are deployed through the creation of *scenarios*, which is where Nielsen notes, the persona “comes to life and acts” [17]. A scenario is a story developed around a persona and the related data, with the purpose of imaginatively describing how they might engage with a specific task, goal, routine, or product [17, 20]. For us, scenarios were built around the idea of a regular day in their lives and how their engagement in a misinformation community online might fit into their otherwise ordinary lives. Both personas and scenarios are tools that emphasize the humanity of participants, foregrounding empathy and understanding of a cohort’s behavior rather than stigmatizing it, thus enabling the design of more ethical solutions. This information is typically brought together on a persona card, such as those included at the end of this article [20].

2.3 Facebook A/B testing methods

The Facebook advertisements were designed and deployed through Facebook’s A/B testing program. This allows for two versions of an ad to be tested among the same target demographic, to determine what achieves the best result. Specifically, we use a “multiple-ad study without holdout” implemented by Facebook which has become the most widely adopted approach to collect data on advertising platforms [21]. The personas we created provided both the template for the demographic traits used to run the Facebook ads, and our understanding of the best styles and topics for each cohort. We selected the cohort of users the ad would be shown to based on factors such as gender and age, but also the perceived personal interests of individuals. We specify the persona’s characteristics in the field “People who match” – this determines the audience of the advertisements. For example, to target the audience of Jennifer, we configured “female between the ages of 25 to 40 with interests in painting, health, wellness, yoga or childcare”. We create the textual content of the ads to match the identified best-suited styles for each persona. By “style”, we refer to all aspects of written language beyond the propositional content: tone, readability, emotional framing, directness, question format, and implied concern or urgency. In this study, it is predominantly the non-nouns that convey the style, while the nouns (such as “vaccination”) remain constant. For the purposes of coding, the three communication styles were operationalised according to their defining linguistic characteristics. Intimacy

was characterised by highly personal language, emotional openness, and expressions of affection or vulnerability. Expert was characterised by authoritative, explanatory language that presented claims through references to evidence, reports, or expertise. Empowerment was characterised by motivational and confrontational language intended to encourage action against a perceived adversary. More detailed qualitative descriptions of these styles are provided in Sect. 3.3. We also created ‘errors’ with the style and topic to test the strength of the findings, as presented in Table 1. The styled messages for each topic and each persona are summarised in Table 2.

We sourced copyright-free images to accompany each ad, and these were consistent regardless of the accompanying text. The landing page – the target where the viewer is taken to when clicking the ad – for both versions of the ad was identical: an Australian government-affiliated website providing facts about the topic, to ensure individuals landed on a website with reliable information about the topic of their interest. Because of the constraints on the length of the text that could be used for

Table 1 Summary of Facebook Ad outcome report for the 12 unique paired comparisons included in the primary analysis. The significance is measured by a two-sided chi-squared test. * $p < 0.05/12$, ** $p < 0.01/12$, *** $p < 0.001/12$ using Bonferroni correction. “-” sign indicates the difference was non-significant

Topic	Message Style	Target	Impressions	Link clicks	CTR	Significance (<i>p</i>)
<i>Effect of Style</i>						
5 G	Neutral	Aaron	29888	133	0.44	*** (3.36 × 10 ^{−7})
	Empowerment	Aaron	24203	191	0.79	
Supplements	Neutral	Aaron	14554	102	0.70	− (0.759)
	Empowerment	Aaron	17708	118	0.67	
Vaccines	Neutral	Jennifer	15567	101	0.65	* (0.002)
	Intimacy	Jennifer	13107	129	0.98	
School Education	Neutral	Patrick	13502	148	1.10	*** (7.72 × 10 ^{−100})
	Expert	Patrick	10276	617	6.00	
Climate change	Neutral	Patrick	20654	135	0.65	*** (6.47 × 10 ^{−71})
	Expert	Patrick	10083	334	3.31	
<i>Error of Style</i>						
5 G	Neutral	Aaron	28618	132	0.46	−(0.079)
	Intimacy	Aaron	28238	161	0.57	
School Education	Neutral	Patrick	14619	194	1.33	*** (1.32 × 10 ^{−54})
	Empowerment	Patrick	11235	506	4.50	
Vaccines	Neutral	Jennifer	13200	168	1.27	−(0.025)
	Expert	Jennifer	13173	212	1.61	
5 G	Neutral	Patrick	8939	168	1.88	*** (3.86 × 10 ^{−28})
	Empowerment	Patrick	8861	431	4.86	
<i>Error of Topic</i>						
School Education	Neutral	Jennifer	17495	101	0.58	*** (4.36 × 10 ^{−6})
	Intimacy	Jennifer	11105	119	1.07	
School Education	Neutral	Jennifer	17336	132	0.76	*** (2.93 × 10 ^{−20})
	Empowerment	Jennifer	8639	181	2.10	
5 G	Neutral	Patrick	11340	216	1.90	*** (1.78 × 10 ^{−7})
	Expert	Patrick	10960	328	2.99	

Table 2 The styled messages for each topic and each persona

Persona	Jennifer	Patrick	Aaron	
Style	Intimacy	Expert	Empowerment	Neutral
Topic	Styled messages			
5 G	Are our children safe around 5 G? What do new mothers need to know?	The world changes quickly - Can 5 G help you keep up?	What is 5 G really, and why is it being rolled out around Australia? How will it change our lives?	5 G is the new industry standard for a cheap, energy-efficient future, and is now available around Australia
Supplements	Tired and need energy for two? Could supplements be your answer?	Not as fit as you were? Find the right supplement to get your strength back	Interested in supplements? Get the facts to get the wins	Supplements should only be used with medical advice
Vaccines	Do you know all the effects of childhood vaccinations on your babies?	Natural immunity or modern medicine? Do vaccines help today's kids grow up stronger?	Do childhood vaccinations really improve natural human immunity?	Protect the community and your kids by booking their vaccinations today
School education	What child sexuality education will your children receive at school?	What's included in today's child sexuality education curriculum?	What does Australia's school child sexuality education include?	Get the facts about the new kid's health education
Climate change	How will climate change affect your children's future?	What kind of world will your grandchildren grow up in?	The truth about climate change—check the facts yourself	Australia is taking action against climate change

the ads, we had to carefully balance the topic (message) with the style properties. An example of deployed advertisements is shown in Fig. 2.

The campaigns were run between May 1 and September 30, 2023, with each experiment running for one week. All campaigns used Facebook's Traffic objective, a 7-day click or 1-day view attribution setting, and lifetime ad set budgets of AUD 100. The bid type recorded in the Facebook report was ABSOLUTE_OCPM. Most ads spent the full AUD 100 budget, while one rejected climate-change ad spent AUD 4.29 before cancellation. Within each paired test, the sponsor, image, landing page, campaign objective, attribution setting, and budget were held constant, with only the ad text varying. Facebook's delivery system remained a platform constraint: ad delivery was mediated by Facebook's opaque optimization system, and two climate-change campaigns were cancelled by Facebook, possibly because they were classified as ads related to social issues.

After the ads ran for a given period of time (here, one week), Facebook made available a range of performance indicators, including *Impressions*, *Link clicks*, *Cost per link click*, *Post reactions*, etc. All performance measures provided by Facebook is available in Supplementary information. We selected the Click Through Rate (CTR), as this indicated that a user had taken an active step to find out more about the sample text they were shown. CTRs are the rate at which an ad was clicked, calculated by considering the percentage of people who viewed the ad (its impressions) in relation

to the percentage of people who then click the ad (clicks). A higher CTR indicates that the ad is drawing more attention and action from individuals.

Of the 15 pairs of tests we designed, two climate change pairs were cancelled by Facebook: climate change targeted towards Jennifer (intimacy/neutral style), and climate change targeted towards Patrick (empowerment/neutral style). Of the 13 pairs that ran, one was a repeated Supplements/Aaron pair, which we ran to confirm the stability of Aaron's lower engagement; because it returned similar results, we retained the original run for the primary analysis, leaving 12 unique paired comparisons. One possibility for why these ads may have been prevented from running is that Facebook determined they did not comply with the policy around ads related to social issues (¹). However, given that our initial test of climate change targeted towards Patrick (expert/neutral style) ran without issue, and that our ads mentioned but did not take a stance on the topics featured, we are unable to determine why those two particular campaigns were disabled. Drafts of all ads were submitted to the university Human Research Ethics Committee prior to the experiment commencing.

3 Results

This research unfolded through an explorative and iterative process (fully detailed in the Methods section) with the aim of understanding the online misinformation landscape and users in Australia. The qualitative exploration led to our development of the misinformation ecosystem and the personas and their respective communication styles. Following this, we deployed what we had learned about the Australian online misinformation environment and its users through a series of experiments in which we re-styled consensus-based information into more engaging formats tailored to each cohort. The findings demonstrate that targeted styling produces higher engagement with the content from the user base, but also that custom styles are more effective at engaging users than a generic, authoritative message. We next detail each of these steps.

3.1 Misinformation ecosystem

We propose the concept of a misinformation ecosystem to understand the formation of communities around individual producers who package misinformation like a product for their followers, depicted in Fig. 1. We built the ecosystem based on the qualitative observations, conducted over eight months in 2023, of 23 public Facebook pages and groups expressing critical views about vaccination (anti-vaxxers) and extreme-leaning conservative views and narratives (nationalism and conspiracy theories). This role-based structure differs from the informal spread of misinformation through personal networks or viral content, as a misinformation ecosystem contains a clear internal hierarchy. Producers generate misinformation content, and they are often the page or group owners and administrators. On the other side are the consum-

¹ <https://www.facebook.com/business/help/214754279118974?id=288762101909005>

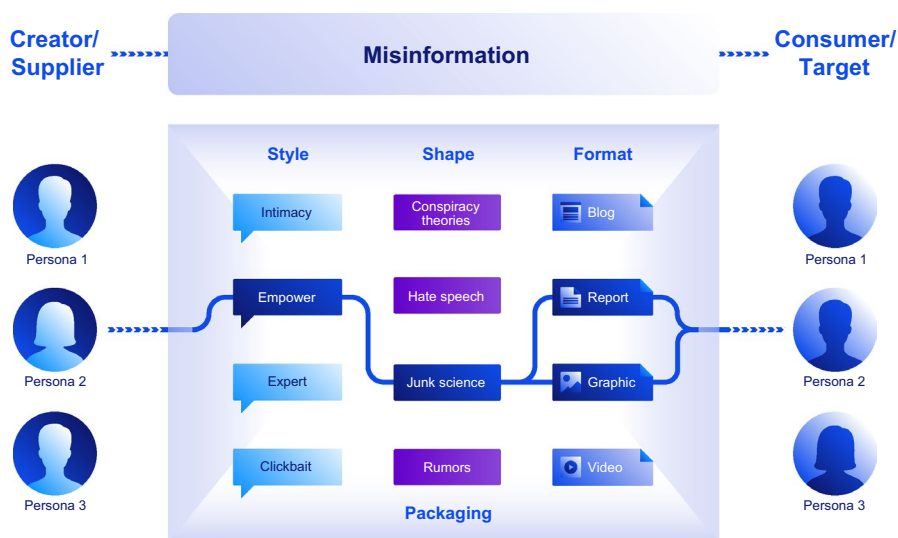


Fig. 1 The misinformation ecosystem: the producers, the misinformation content and the consumers. The content has a writing style, a shape (or genre), and a format – all of which are specific to categories of consumers

ers, who make up the majority of the community and drive most of the engagement with the page or group.

We found that these pages and groups were not just topic-based hubs focused on information exchange, but social communities with their own distinct communication style. When we compared how different pages with different demographics of users discussed the same topics, we found three additional layers to the distribution of misinformation content beyond the topic itself: the writing style, the shape or genre of the content, and the format. These layers of styling, which vary between groups even when the topics are similar, are what we term misinformation packaging. It is this packaging that makes producers valuable, as they not only share content relevant to their audiences; they do so in a particular way that engages their specific audience, such as through their choice of language or content format. Consideration of these additional layers pushes beyond the common confines of examining misinformation based on topic and accuracy, and instead considers how it is communicated. We note that while the examples we use to illustrate our proposed concept of the misinformation ecosystem in this paper are conceptualized based on our observations of Australian online misinformation communities, the concept of a misinformation ecosystem itself could be adapted to other cultural contexts in future research.

3.2 Three vulnerable personas

The personas developed for this research are a synthesis of the months of online observation of specific groups in Australia, as is standard for persona design (see Materials and Methods section for details). We developed three personas to represent these communities without compromising anyone's privacy. First, we identified a

list of the most commonly observed demographic traits for each community (e.g. approximate age, gender, current or former occupation), as well as the interests of members beyond misinformation topics (e.g. healthy eating, family, fitness), and the concerns and hopes they expressed. Next, we developed representative personas for each cohort by assembling the information we gathered into a narrative form—that is, essentially writing the story that brought together all the information we had gleaned during our qualitative observations into one fictional “person”. The traits used to build these personas were based on what was the most common for that cohort. For example, we observed that all anti-vaxxers in one such community we followed were young women raising children who were passionate about health and wellbeing; therefore, these traits were included in the persona of Jennifer. These profiles are then extended through the process of scenario building, as detailed in the Methods section. Without standing in for or identifying any single person, the personas represent the type of person we observed to be vulnerable to misinformation related to specific topics, when presented in specific styles. Our personas are Jennifer, Patrick, and Aaron, and we briefly introduce them here. Jennifer is a young mother experiencing anxiety about her first child’s mandatory vaccinations. Patrick is a retired school teacher concerned about the state of the world for his grandchildren, including what they are taught in schools and whether climate change is a real threat to them. Aaron is an ambitious young man who resents being single and not having what he perceives as the ideal male body, leading to his interest in supplements and new technology. The full personas and their visuals are presented in the Supplementary Materials, and we include our persona cards in Figs. A1 to A3.

3.3 Misinformation packaging

Through a close reading of the user posts from each page and comparing the discussions of the same topic on other pages, we determined distinctive attributes of the writing styles used for each persona. Jennifer’s style, named ‘intimacy’, involves the topics of conversation being expressed through an intensely personal lens, as though posts are addressed to a close friend. Common emotions include joy, gratitude, and a kind of fragile fear or trepidation to enhance the poster’s image of being open and vulnerable with the audience. High levels of affection towards others are expressed, through words or emoji use. This style was suitable for genres like rumors or conspiracy theories, which prompted anxiety, which in turn prompted the need for community members to provide emotional support for each other. Patrick’s style was ‘expert’, which relied on creating the impression of taking a logical approach to the topic by referencing legitimate news articles and even academic sources while stripping them of context and nuance to misrepresent the results (or perhaps, not understanding these factors in the first place). This approach tapped into the preference to come across as a knowledgeable authority on a topic, and is commonly seen in the genre of junk science. Lastly, Aaron’s style of ‘empowerment’ involved writing as though trying to rally the readership to face an enemy, through a consistently angry tone and inflammatory language, which could include slurs or language that implied the speaker and their in-group were being subjected to oppression. This is the type of language commonly seen in hate speech. Although each style has a common shape

or format through which it may be deployed as part of the misinformation packaging process, we hypothesized that this style element could be applied to different genres of text, including those grounded in consensus-based information.

Finally, we compared the three observed writing styles to that of Australian government factsheets about similar topics, to determine how much of a gap lay between government communications, and the style used in each community. We found that these resources used a more authoritative tone, focused primarily on conveying facts and information. This is not inherently a problem, but the distance between this style and the styles used in the online communities was significant. We hypothesized that this gap in style could contribute to certain users being unwilling or unable to access authoritative resources, and that alternative messages conveying information about the same topics in a community's preferred style could potentially increase engagement with official resources. To test this hypothesis, we designed and implemented a Facebook experiment. We created a series of short messages about the topics of concern for our personas, drawing from official information from Australian government-affiliated websites. Thus, although the style was intentionally reminiscent of their familiar online communities, the content users would access would be reliable, consensus-based information rather than misinformation.

3.4 Styled messages receive more engagement

We developed a series of Facebook advertisements targeting our personas to determine the impact of language style on these different communities, about their social issues of concern. We leveraged Facebook's A/B testing facility, which allows for two different versions of the same ad to be tested with the same demographic to determine what achieves the best result. The image, the ad target and the clearly identified sponsor are identical across the two versions, with only the text's styling differing. Our success metric was the Click-through Rate (CTR) on the ad, defined as the percentage of ad viewers who also clicked on the ad. We use CTR because we aim to measure how many individuals were willing to take the active step of clicking an ad, while noting that this does not measure comprehension, trust, retention, or attitude change. All other metrics provided by Facebook's A/B testing are also shown in Supplementary information section. The development of the personas provided us with the necessary insights into user demographics to design and target these Facebook interventions, as factors such as demographics and interests can be added to the interface to refine the targeted audience. First, we tested the impact of custom-styled advertisements for each persona, compared to the neutral style we associated with the Australian government fact sheets. Figure 2 shows an example of a pair of ads targeting Jennifer. Each experiment ran for one week in 2023. When users clicked on the text for either version of the ad, they were redirected to an Australian government or affiliated factsheet on the topic, ensuring no exposure to misinformation. Jennifer's ad used the 'intimacy' style and the topic of child vaccination, Patrick's used the 'expert' style on both climate change and school health and sexuality education, and Aaron's used the 'empowerment' style along with the topics of supplements and 5 G technology. In Table 2, we summarize the styled messages for each topic and per-

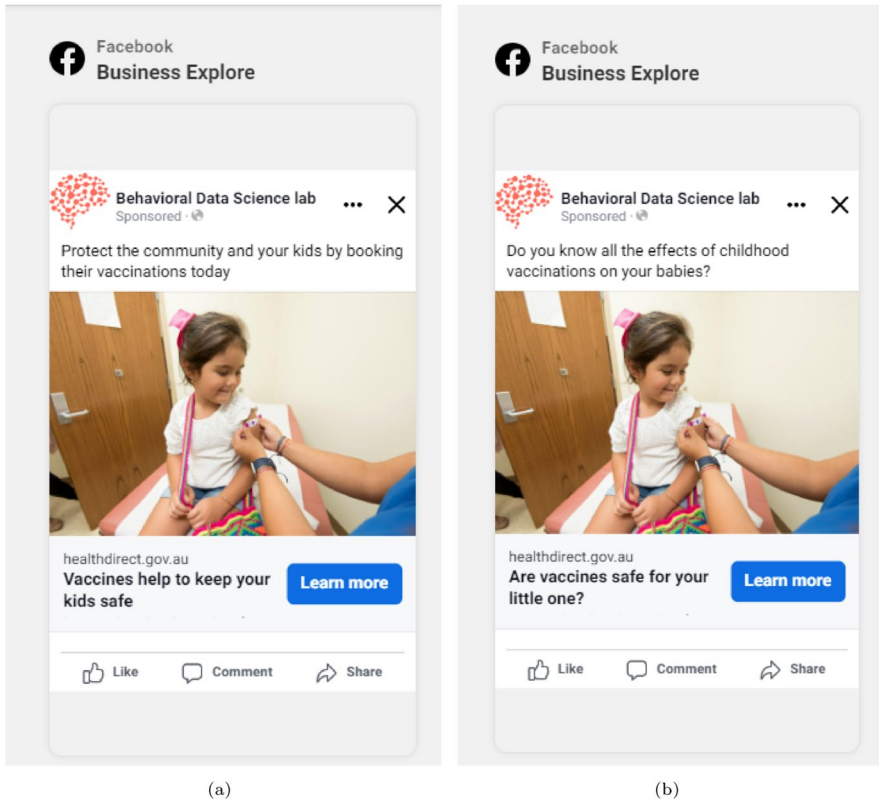


Fig. 2 Screenshots of real-world advertisements about vaccination in Australia. **a** General purpose non-styled version. **b** Intimacy styled (preferred by the Jennifer persona)

sona. We fully detail designing and deploying the Facebook experiment in Materials and Methods.

Figure 3 compares the CTR of the styled and neutral ads tailored for each persona and each topic. The styled ad versions outperform the neutral version in all experiments. We calculate the two-sided chi-squared statistics for each of the 12 unique paired comparisons in the A/B testing result (styled-neutral) to determine whether there was a significant association between group assignment (styled-neutral) and click status (click or non-click); we find that the differences in click rates for all pairs (styled-neutral) are statistically significant when applying a Bonferroni correction ($p < 0.05/12$), except for the Supplements ad targeted for Aaron. We believe that Aaron, being a member of the current Generation Z, is a demographic less likely to be on Facebook and more likely to instead be on platforms like TikTok, and this could account for the lower engagement from his cohort. Overall, the styled versions of ads achieved a higher clickthrough rate, suggesting that appropriately styled texts are more attractive to the demographics represented by our personas than the neutral style of government fact sheets. We show the complete set of performance metrics for each ad, including impressions and clicks, in Table 1, Effect of Style.

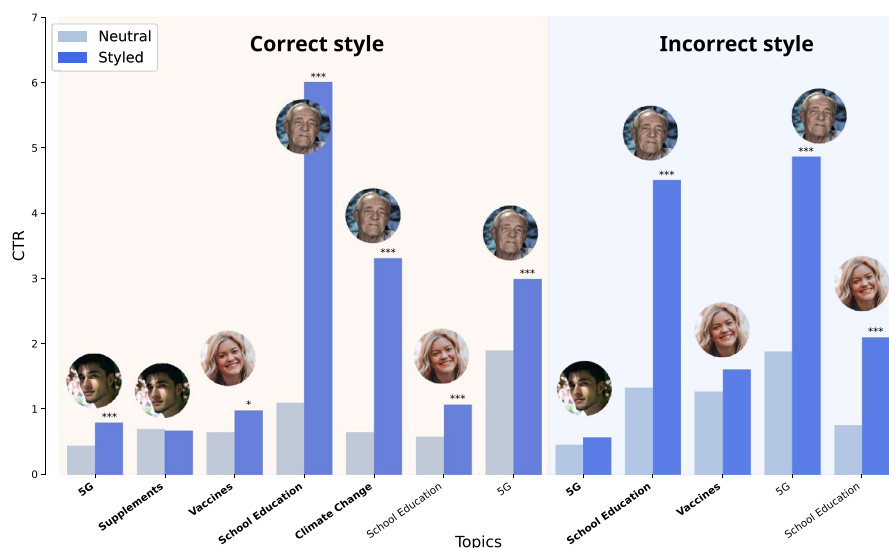


Fig. 3 Engagement of styled messages. We compare the engagement of advertised messages via CTR for the 12 unique paired comparisons included in the primary analysis. The x-axis shows the tested topics; the faces above the bar indicate the targeted persona. The style matches the persona in the first seven bars (orange background); in the next five bars (blue background), the custom-applied style is “incorrect” for the persona. These figure groupings map onto Table 1 by style correctness: the orange group contains the five Effect of Style comparisons plus the two Error of Topic comparisons that use the persona’s correct style (Intimacy/School Education for Jennifer and Expert/5 G for Patrick); the blue group contains the four Error of Style comparisons plus the Error of Topic comparison that uses an incorrect style (Empowerment/School Education for Jennifer). We boldface the tests where the topic matches the persona – e.g., in the seventh group of bars, the topic (5 G) does not match Patrick, but the message is correctly styled. The stars above the columns indicate the strength of the statistical significance measured by a two-sided chi-squared test – * ($p < 0.05/12$), ** ($p < 0.01/12$), *** ($p < 0.001/12$). When correct styles are used, statistically significantly higher CTRs are achieved except when targeting Aaron for supplements. Even when the incorrect styles are applied, the CTR is higher than the neutral style: three out of five styled ads had significantly higher CTRs than their non-styled (neutral) counterparts

3.5 Any style is better than no style

In order to determine the strength of the relationship between a style and its user demographic (represented by our personas), we ran an additional sequence of tests. We targeted the same topics of interest toward our personas in an A/B format, deliberately deploying the wrong style for each persona (represented as Incorrect style in Fig. 3). Although we used an incorrect style that is not preferred by each persona, the styled versions still achieved a higher CTR than the neutral versions. When we used incorrect styles, only the CTR of the ads for Jennifer concerning vaccines, and Aaron about 5 G were not statistically significant. This indicates that any styled communication is better than the neutral style when seeking engagement from these populations. A possible explanation is that all three customized styles are easier to understand than the facts-focused ad and, therefore, are more accessible to our demographics. However, another explanation is that our three demographics distrust the

government (the “establishment”, the “elites”) and automatically ignore any formal, government-sounding communication.

In the final testing phase, we investigated whether the style was more relevant to our cohorts than the topic. In this setting, we used topics that were not identified as topics of particular interest to each cohort. Surprisingly, across all pairs within this setting (represented with non-bold face topics in Fig. 3), the styled versions achieved notably higher CTR values and exhibited statistical significance. This implies that styled text can effectively engage audiences associated with vulnerable demographics that are otherwise resistant to information that might change their minds, even when the topic is not a prominent pre-existing concern in the demographic.

4 Discussion and conclusion

Our proposed misinformation ecosystem provides a new framework with which to think about online misinformation: not solely as untethered falsehoods that go viral online because of their claims, but as content that resonates so deeply with a specific demographic that a social community can spring up around. Within this community, we see people adopting different roles just as they would in benign community groups; and producers and consumers are co-dependent in that both must be present and engaged to keep the ecosystem alive. This also creates a strong shield against outside intervention efforts, as they do not simply threaten individual worldviews, but the existence of a social community with its own norms, including communication preferences. Through the lens of the misinformation ecosystem, we can clearly identify the leaders (producers), who both fuel and profit from the investment of consumer community members.

In creating the personas representative of fringe communities, we aimed to present a holistic and humane portrait of vulnerable individuals, highlighting their genuine concerns related to their topic of interest, as well as how their involvement in fringe communities can have ramifications for their lives more broadly. Scholarship has historically taken a deficit view towards people drawn in by misinformation, alternatively portraying them as digitally illiterate, lacking critical thinking skills, or being defined by a series of negative psychological traits such as high levels of anxiety or a weak self-concept. Whether or not these factors apply to an individual, such an approach does not account for the richness of human experience, which cannot be reduced to a single trait. We hope these personas will benefit future research by providing an accessible and sympathetic portrait of some of the types of people who can be vulnerable to certain topics, in certain circumstances. As we have only compiled a small number due to the scope and feasibility of the research, we also hope that future work will lead to the creation of more personas that better help us to understand these cohorts.

A/B testing on Facebook has inherent limitations due to the platform’s algorithmic complexities. Although researchers can specify target audience demographics, Facebook’s algorithm—a black box mechanism—introduces compounding effects that obscure the full impact of styled messages. This algorithm dynamically selects and presents ads to users based on numerous factors beyond the specified criteria, leading

to variations in exposure that cannot be controlled or fully understood. Consequently, it is challenging to attribute the success of an advertisement solely to its design, such as our styled messages. Several studies, including [21–24], have addressed additional concerns regarding Facebook’s A/B split testing, which can lead to a significant overestimation of the effects of online advertising. These concerns are primarily discussed in the context of commercial advertising, where the primary goal is to maximize click-through rates (CTR) and conversions to sales. While we present our results using CTR, our research only aims to use the A/B testing function to expose the targeted Facebook audiences to new perspectives outside their existing worldviews, so concerns about conversions to sales are beyond the scope of our goals. Furthermore, the compounding of Facebook’s algorithm with the message styling can be seen as an added benefit for our work, as the goal is to better engage audiences associated with insular misinformation bubbles; if we can enlist Facebook’s algorithms to help with it, the better. Finally, although CTR is not a complete metric for success, given that we cannot measure the quality of the traffic, comprehension, trust, retention, or whether accessing the information sources changed worldviews, it provides a clear and direct measure of ad engagement with the tailored messaging. The Facebook campaign reports did not record landing-page views or any other post-click indicators, so no downstream engagement measures were available to report. We also note that the styled and neutral messages differ across a bundle of co-varying textual features (tone, readability, emotional framing, directness, question format); with the sponsor, image, and landing page held constant within each pair, our design establishes the effect of this style bundle as a whole, and disentangling the contribution of individual stylistic features remains future work.

The results of our experiments to test our concepts of the misinformation ecosystem, the packaging of misinformation, and the types of personas vulnerable to certain types of misinformation, provide strong initial evidence for the importance of each of these factors when designing interventions. It is intriguing to observe that styled communication almost always received higher engagement regardless of audience and topics. This is consistent with other studies investigating the effect of communication styles on user engagements utilizing generic style comparisons such as formal versus informal styles [25] and populist style versus self-centered style [26]. Although we could not complete all the paired experiments due to limitations from the Facebook platform which would not permit us to run some of the campaigns, there are some pairs that show interesting insights to consider for the next development of personas. Tests targeted to Patrick showed significantly higher engagements using styled messages, even when the test was performed using incorrect styles on the incorrect topics. Patrick’s resulting engagement using styled messages was significantly higher and also absolutely higher than other tests on other personas. However, because the experimental matrix was incomplete and the number of tested topics differed across personas, this persona-level pattern should be treated as speculative. It may indicate that the demographic Patrick represents is more interested in such online communications and information in general, while Aaron’s lower engagement may suggest that his demographic is less interested in engaging with these topics on the Facebook platform.

It has been shown that matching the content of ads to individuals' psychological characteristics attract significantly more clicks [27] on the Facebook advertising platform. While Matz et al. [27] extracted characteristics of the target audience quantitatively based on data points such as profiles, *likes*, and other interactions on Facebook, we built detailed personas using demographic traits, interests, hopes, and fears that users openly expressed on these public pages and groups as detailed in Materials and Methods. This is to say that our approach is focused on understanding the interiority of the users, as the users themselves express it through their own communication style. This is only possible due to the months of qualitative online observation we conducted. As a result, our personas offer a richer portrait of these users; and the results confirm our approach successfully engaged persona-informed Facebook audiences.

4.1 Ethical considerations for style-based interventions

Adapting communication styles from misinformation ecosystems raises ethical risks that must be taken seriously. We do not argue that researchers or institutions should reproduce harmful rhetoric, deception, hostility, or inflammatory claims in order to increase engagement. Rather, style-based adaptation should be limited to making consensus-based information more accessible, familiar, and emotionally legible to the intended audience, while preserving accuracy, transparency, and respect for users. In our study, safeguards included using clearly identified advertisements, maintaining the same sponsor and landing page across paired ads, directing users only to Australian government or government-affiliated factual resources, and submitting all draft advertisements to our university's Human Research Ethics Committee as part of the institutional ethics approval obtained before deployment. Future uses of this approach should avoid impersonation, fear amplification, stigmatizing language, or the reuse of hateful frames, and should include review processes that evaluate both informational accuracy and potential rhetorical harm. These safeguards are especially important because the same knowledge of audience style could be misused to manipulate vulnerable groups if separated from ethical oversight and a commitment to consensus-based information.

If changes to the style of otherwise consensus-based information can stimulate interest among those for whom standard, neutral information resources are off-putting, then this opens the door for a range of new approaches to misinformation interventions. For example, providing multiple versions of existing resources around vaccines, climate change, and school education, could increase the likelihood of good information getting picked up. In an age where attention spans are being impaired by digital technologies [28, 29], where screen-reading is becoming more common despite its negative impacts on information comprehension [30], and where bad-faith actors are quick to exploit our everyday human anxieties for profit, the styling could give good information a fighting chance at engaging audiences associated with insular communities, and reigniting interest in the world around us.

In summary, while Facebook's A/B testing presents certain theoretical limitations, our study demonstrates that it is possible to leverage the platform's targeting capabilities to achieve significant engagement boost with well-crafted informational

messages. Since higher engagement leads to higher visibility in such platforms, by engaging the initial persona-informed Facebook audiences, we have a better chance at exposing other vulnerable groups to consensus-based information.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s42001-026-00493-0>.

Author contributions J.L., E.B., H.F., and M.A.R. designed research; J.L. and E.B. performed research; J.L. analyzed data; and J.L. and E.B. wrote the paper.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This study is supported by Department of Home Affairs (PRN0006958).

Data availability Full data and the analysis code can be found at <https://code.research.uts.edu.au/159233/popping-filter-bubbles>.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical approval The study received full ethical clearance from the Human Research Ethics Committee of our university (Reference Number: ETH23-8018).

Consent for publication All authors consent to publish.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Vraga, E. K., & Bode, L. (2020). Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation. *Political Communication*, 37(1), 136–144.
2. Tandoc, E. C., Jr., Lim, Z. W., & Ling, R. (2018). Defining “fake news” a typology of scholarly definitions. *Digital journalism*, 6(2), 137–153.
3. Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political communication*, 37(3), 350–375.
4. Johns, A., Bailo, F., Booth, E., & Rizoiu, M.-A. (2024). Labelling, shadow bans and community resistance: did meta's strategy to suppress rather than remove covid misinformation and conspiracy theory on facebook slow the spread? *Media International Australia*. <https://doi.org/10.1177/1329878X241236984>
5. Boyd, D., & Golebiewski, M. (2018). Data voids: Where missing data can easily be exploited

6. Bryant, L. V. (2020). The youtube algorithm and the alt-right filter bubble. *Open Information Science*, 4(1), 85–90.
7. Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2023). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*, 7(1), 74–101.
8. Booth, E., Lee, J., Rizoiu, M.-A., & Farid, H. (2024). Conspiracy, misinformation, radicalisation: understanding the online pathway to indoctrination and opportunities for intervention. *Journal of Sociology*. <https://doi.org/10.1177/14407833241231756>
9. Kruglanski, A. W., Molinaro, E., Ellenberg, M., & Di Cicco, G. (2022). Terrorism and conspiracy theories: A view from the 3n model of radicalization. *Current Opinion in Psychology*, 47, Article 101396.
10. Abalakina-Paap, M., Stephan, W. G., Craig, T., & Gregory, W. L. (1999). Beliefs in conspiracies. *Political Psychology*, 20(3), 637–647.
11. Grzesiak-Feldman, M. (2013). The effect of high-anxiety situations on conspiracy thinking. *Current Psychology*, 32, 100–118.
12. Klausen, J., Campion, S., Needle, N., Nguyen, G., & Libretti, R. (2016). Toward a behavioral model of “homegrown” radicalization trajectories. *Studies in Conflict & Terrorism*, 39(1), 67–83.
13. Winter, D. A., & Feixas, G. (2019). Toward a constructivist model of radicalization and deradicalization: A conceptual and methodological proposal. *Frontiers in Psychology*, 10, Article 419929.
14. Dahlke, R., & Hancock, J. (2022). The effect of online misinformation exposure on false election beliefs. *OSF Preprints*, 17,
15. Shearer, E., & Mitchell, A. News Use Across Social Media Platforms in 2020. <https://doi.org/APO-311092>. <https://apo.org.au/node/311092>
16. Cooper, A. (1999). *The Inmates Are Running the Asylum*. Indianapolis, Indiana: Sams.
17. Nielsen, L. (2013). *Personas-user Focused Design* (Vol. 15). London, UK: Springer.
18. An, J., Kwak, H., Jung, S.-G., Salminen, J., Admad, M., & Jansen, B. (2018). Imaginary people representing real numbers: Generating personas from online social media data. *ACM Transactions on the Web (TWEB)*, 12(4), 1–26.
19. Jansen, B. J., Jung, S.-G., Nielsen, L., Guan, K. W., & Salminen, J. (2022). How to create personas: Three persona creation methodologies with implications for practical employment. *Pacific Asia Journal of the Association for Information Systems*, 14(3), 1.
20. Stephanidis, C., & Salvendy, G. (2024). *User Experience Methods and Tools in Human-Computer Interaction*. Boca Raton, FL: CRC Press.
21. Braun, M., De Langhe, B., Puntoni, S., & Schwartz, E. M. (2024). Leveraging digital advertising platforms for consumer research. *Journal of Consumer Research*, 51(1), 119–128.
22. Orazi, D. C., & Johnston, A. C. (2020). Running field experiments using facebook split test. *Journal of Business Research*, 118, 189–198.
23. Lewis, R. A., Rao, J. M., & Reiley, D. H. (2011). Here, there, and everywhere: correlated online behaviors can lead to overestimates of the effects of advertising. *Proceedings of the 20th International Conference on World Wide Web* (pp. 157–166).
24. Berman, R., & Bulte, C. (2022). False discovery in a/b testing. *Management Science*, 68(9), 6762–6782.
25. Della Giusta, M., Jaworska, S., & Vukadinović Greetham, D. (2021). Expert communication on twitter: Comparing economists’ and scientists’ social networks, topics and communicative styles. *Public Understanding of Science*, 30(1), 75–90.
26. Bene, M., Ceron, A., Fenoll, V., Haßler, J., Kruschinski, S., Larsson, A. O., Magin, M., Schlosser, K., & Wurst, A.-K. (2022). Keep them engaged! investigating the effects of self-centered social media communication style on user engagement in 12 european countries. *Political Communication*, 39(4), 429–453.
27. Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, 114(48), 12714–12719.
28. Vedeckina, M., & Borgonovi, F. (2021). A review of evidence on the role of digital technology in shaping attention and cognitive control in children. *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2021.611155>
29. Firth, J., Torous, J., Stubbs, B., Firth, J. A., Steiner, G. Z., Smith, L., Alvarez-Jimenez, M., Gleeson, J., Vancampfort, D., Armitage, C. J., & Sarris, J. (2019). The “online brain”: how the internet may be changing our cognition. *World Psychiatry*, 18, 119–129. <https://doi.org/10.1002/wps.20617>

-
30. Liao, S., Yu, L., Kruger, J.-L., & Reichle, E. D. (2024). Dynamic reading in a digital age: new insights on cognition. *Trends in Cognitive Sciences*, 28, 43–55. <https://doi.org/10.1016/j.tics.2023.08.002>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.