Behavior Change Interventions Combating Online Misinformation: A Scoping Review

Loukas Konstantinou
lukekonsta@gmail.com
Cyprus University of Technology
Limassol, Cyprus

Evangelos Karapanos
evangelos.karapanos@cut.ac.cy
Cyprus University of Technology
Limassol, Cyprus

Abstract

It is increasingly acknowledged that simply presenting users with corrective information is unlikely to produce the desired effects against misinformation. As such, the need for systematic use of behavioral theory is increasingly acknowledged, and behavioral interventions against misinformation are rising. This paper presents a scoping review of digital behavioral interventions countering misinformation, inquiring into their behavioral objectives, theoretical foundations, design and evaluation practices, and the factors that were empirically proven, or speculated, to contribute to interventions' failure. Among others, we identify 17 distinct behavioral objectives, organized into three stages of the online news cycle: composition, amplification and consumption, 24 theoretical frameworks employed in designing these interventions, and nine reasons of failure. We synthesize the findings into a set of design cards with the goal of guiding intervention designers during concept ideation and refinement, and highlight areas for future research.

CCS Concepts

• **Human-centered computing** → **HCI theory, concepts and models**.

Keywords

Online Misinformation, Behavior Change, Attitude Change, Scoping Review

ACM Reference Format:
Loukas Konstantinou and Evangelos Karapanos. 2025. Behavior Change Interventions Combating Online Misinformation: A Scoping Review. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3706598.3713127

1 Introduction

Three-quarters of online news consumers encounter at least one instance of misinformation daily [91], while an alarming number of users have come to embrace conspiracy theories due to *"a variety of alternative worldviews"* [86]. Misinformation, defined as *false or misleading information shared without malicious intent* [126], can easily spread on digital platforms [125]. While misinformation is far from a new phenomenon, internet technology has contributed to the problem due to low entry costs, permitting journalism competitors to enter the sphere and undermine information credibility [76], and due to the pervasiveness of bots contributing to misinformation propagation [22, 44].

Figure 1: We synthesize the findings of this review into a set of design cards, the *Behavioral Responses to Misinformation* (BRM) cards. The BRM cards support designers by outlining 17 behavioral objectives, 24 theoretical frameworks, and 9 reasons of failure for behavioral interventions against misinformation. The cards can be downloaded at: http://ekarapanos.com/BRMcards/

Given the prevalence of misinformation, it is understandable the research community has spent great efforts inventing ways to counteract this phenomenon. While the majority of initial attempts had focused on technical challenges, such as misinformation detection (see Fernandez and Alani [44] for a review), the role of human behavior is increasingly acknowledged [125] and behavioral interventions against misinformation are rising. With many early user-facing interventions focused on warnings and labels, researchers are calling for a systematic use of behavioral theory, acknowledging that simply presenting users with corrective information is more likely to fail in changing their salient beliefs and opinions, and may even backfire, prompting them to reinforce existing beliefs despite contradictory evidence [44, 62].

Given that a considerable body of work on behavioral interventions against misinformation now exists, it is a good time to step

CHI '25, April 26-May 1, 2025, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713127

back and reflect on the different approaches taken and spot areas that need greater attention. This paper presents a scoping review of digital behavioral interventions against misinformation[1], inquiring into the objectives of interventions, the behavioral theories they rely upon, the design and evaluation methodologies, as well as the reasons for their failure.

The paper makes five respective contributions to the field. First, it identifies 17 distinct behavioral objectives of the interventions, classified into three stages of the *"News Process"* cycle: composition, amplification and consumption. Second, it identifies 24 theoretical frameworks employed in designing interventions, classified into five broad categories: theories of persuasion, decision-making, motivation, socially situated theories, and finally, descriptive theories & models. Third, it examines the design process of behavioral interventions, identifying three categories of design approaches, ones rooted in user-centered design methodologies, ones rooted in participatory design, and ones rooted in behavioral science frameworks. Fourth, the review highlights the prevalence of empirical evaluation through controlled lab experiments and field trials and investigates the type of study designs (control, multiple or single arms), as well as the infrequent employment of post-removal phases, among others. Lastly, the review expands upon nine reasons that were either speculated or proven empirically to lead to the behavioral failure of interventions.

2 Method

As our objective was to explore and map the breadth of behavioral interventions for misinformation mitigation, we decided that a scoping review was the most appropriate, which typically seeks to present an overview of the literature on a broad topic, to identify emerging themes and to highlight literature gaps [23, 90].

2.1 Research questions

Our review aimed to address the following research questions:

RQ1. Which behavioral objectives were the interventions designed to pursue? Behavioral interventions may adopt diverse objectives to mitigate misinformation, from instilling reflective content composition to increasing exposure to fact-checking. To identify and structure the different objectives employed by the interventions in our sample, we employed the *"News Process"* model [122], a semi-formal model that identifies five stages - *create*, *edit*, *publish*, *amplify* and *consume* - in the circulation of online information and news. In the *creation* stage, online content is generated by users and news sources. The *editing* stage refers to quality evaluation assessments of content and language; processes that are mostly obsolete as authors have also become editors. *Publishing* refers to the virtual space where content is put out and the relevant policies that approve or reject its publication. *Amplification* is about the activities performed by users and platforms to promote specific news pieces; sharing and recommending content to others. Last, *consumption* refers to the process by which consumers decide to allocate their limited attention to an online news piece.

[1]Digital behavioral interventions against misinformation are defined as *tools designed to induce behavior change and provide tailored support and advice for users* [82, 97], *through web and mobile platforms* [5], *in the context of misinformation*

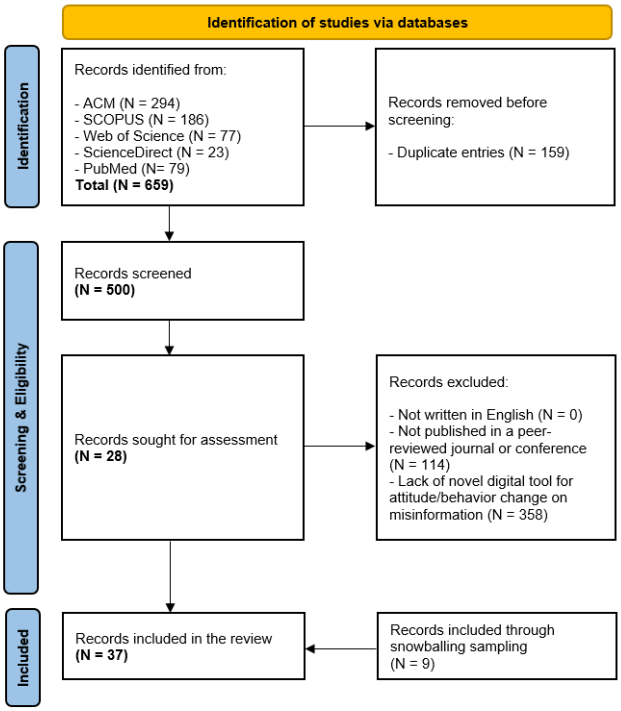

Figure 2: The PRISMA-ScR flowchart for publication selection.

These stages may suggest different approaches for tackling online misinformation. In the *creation* stage, tools might discourage the generation of misinformative content. In the *editing* stage, interventions might scale content-editing activities to minimize the likelihood of misinformative content being published. In the *publication* stage, (digital) platform policies and processes may be implemented to ban misinformation or related non-credible news outlets. In the *amplification* stage, tools can minimize the spread of misinformative content, either by digital platforms or by users. For instance, credibility labels can be put in place to warn users when they are about to share misinformation. Last, in the *consumption* stage, tools may encourage users to assess the quality of the information they view, providing fact-checking services that cultivate media literacy skills and critical thinking.

Using a modified version of the *"News Process"* model, we first characterized the stage of online news production each intervention was designed to tackle: a) composition, b) amplification, and c) consumption, and second, we identified the different objectives in each stage. While the original framework [122] differentiates between the *creation* and *editing* stages, we consolidated them into a single stage - *composition* - as in the modern era, formal news editing processes have mostly disappeared, with authors also adopting the role of the editor [122]. Moreover, we decided to omit the *publication* stage, as it primarily refers to practices of denying hosting, listing users, banning content or blocking Internet access [122]; actions violating users' autonomy and not being in line with the ethical practice of behavior change [117]. Charting the objectives across

these stages provides a collection of actions to be targeted when designing interventions aiming to tackle online misinformation.

RQ2. What theories of attitude or behavior change were interventions grounded upon? Critical to the success of any behavioral intervention is an accurate *logic model - "a bespoke map of what the intervention is designed to change and how it will work"* [1]. For instance, the Information-Motivation-Behavioral Skills (IMB) model [47] suggests that people may be more likely to exhibit the desired behavior if they are *informed*, have sufficient *motivation*, and the *behavioral skills* required for the behavior to take place. Identifying which behavioral antecedents will be targeted is critical to the success of the intervention. However, doing so can be challenging, as designers have to choose from a wide range of theories. A recent review [31] identified 83 theories of behavior and behaviour change, and it has been noted that "*designers and researchers are having a hard time deciding with confidence which of the theories and techniques to use in their design and research"* [80].

With this research question, we aim to explore the theories of attitude and behavior change that are most frequently used in the design of behavioral interventions for misinformation-resilient users, but also the ones that have received less attention. Davis et al. [31], for instance, noted an uneven distribution in the use of theories, with some of the frequently used theories having received weak empirical support, while other prominent theories received little attention from intervention designers.

RQ3. How prevalent is stakeholder involvement in the design of interventions? The design of behavioral interventions is approached by many disciplines, such as Behavioral Science and Human-Computer Interaction (HCI). Each discipline brings unique strengths to intervention design; Behavioral Science emphasizes theory and evidence-based intervention design [59, 82], while HCI, with its User-Centered Design (UCD) methodology, emphasizes the need for iterative design and evaluation cycles to understand better users' requirements and come up with interventions that are easy to use. Both disciplines stress the need for user (and other stakeholder) involvement early in the design process. For instance, recent attempts in the behavioral sciences [133] have argued that theory and evidence-based approaches to intervention design (*e.g.*, the Intervention Mapping Framework [14]) must be coupled with qualitative inquiry aimed at understanding and accommodating the perspectives of the people who will engage with the intervention (see Person-Based Approach [133]).

However, one may note that each discipline brings different values and practices to intervention design. In Behavioral Science, intervention design is still more tuned to a sequential model rooted in pharmacological research [20], which separates the development of the intervention from its effectiveness evaluation. The Person-Based Approach [133], for instance, clearly distinguishes between Planning, Optimization and Evaluation, while Prototyping is a one-off event that takes place once the Planning phase has been completed. UCD, on the other hand, draws on design processes where problem definition and solutions co-evolve [37], and problem reframing is a significant objective in the design process. A variety of methods have been proposed to help designers think divergently, or *outside the box*, while *parallel design* - creating multiple alternative solutions to an identified problem - is seen as superior to the design of a single solution, as it may help designers to discover unseen constraints and opportunities [39] and overcome design fixation [64]. UCD has known for long that just asking people what they want is insufficient, and employs ethnographic and other field methods that combine observation with contextually situated interviewing to gain insight and draw inspiration for design [60, 83].

With this research question, we inquire into the frameworks and practices of stakeholder involvement during the design and development of the interventions. What frameworks of intervention design did researchers rely upon, if at all, and what stakeholders did they involve in the design process? To what extent did they engage in parallel design?

RQ4. To what extent were interventions evaluated to assess their behavioral or attitudinal effects? Empirical studies are critical to developing knowledge of what behavior change strategies work, when and for whom, in mitigating online misinformation. However, such studies have produced mixed findings. For instance, in a recent review of behavioral interventions aimed at reducing vaccine hesitancy, Ruggeri et al. [106] found that while accuracy prompts have shown promise in reducing the spread of misinformation, subsequent replication studies have documented significantly smaller effect sizes, casting doubt on prompts' efficacy. Klasnja et al. [72] argue that measuring behavior change as a result of intervention impact - especially in early-stage research - is challenging as behavior change is a complex and long-term process with high relapse rates, influenced by circumstantial nuances, internal factors and limited human self-control. They suggest complementing *final outcome* measures that reflect the extent to which intervention objectives are met, with *intermediate* measures that capture the extent to which a change in the conditions affecting behavior change has occurred [55], and thus are assumed to capture if progress towards the final outcomes has been made [127]. Similarly, Müller et al. [89] have argued for the measurement of *distal outcomes* reflecting the long-term intervention goals, and *proximal* ones that mediate the effect of interventions on distal outcomes.

Previous work has also raised concerns about the quality of empirical evidence on behavioral interventions. For instance, Caraban et al. [26] who reviewed the use of *nudging*[2] in HCI literature (*i.e.*, a widely recognized key behavioral concept in influencing behavioral decision-making [69]), found 65% of studies to have a duration of a day or less, and only 19% of studies to last longer than a month. They also found only 14% of studies to inquire whether the effects of the nudge hold after its removal. Given the high prevalence of side effects and backfires in nudging, they suggested that our knowledge of the effectiveness of nudging is limited. Similarly, Bergram et al. [18] found only 15% of behavior change interventions to be evaluated through field trials, with the majority being evaluated through controlled experiments that may suffer from limited ecological validity.

We investigate how many of the interventions have undergone empirical evaluation, what type of evaluations have been conducted (lab studies or field trials), the study duration, whether studies explored the impact of interventions post-removal (returning to a non-treatment condition to evaluate the long-term effects of the

[2]A nudge is defined as "*any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives*" (p. 6) [117].

intervention), the intermediate and final outcomes employed, as well as intervention effectiveness.

RQ5. What are the reasons for interventions' failure? Behavioral interventions may not produce the aimed results, may lead to side effects, or can even backfire. Understanding those reasons for failure is critical to advancing our capability to design effective behavioral interventions against misinformation. We aim to identify the factors that were either found or speculated to negatively impact the effectiveness of the behavioral interventions in our sample. In doing so, we build on the taxonomy of reasons for the failure of nudges of Sunstein [114] and Caraban et al. [26]. Sunstein [114] suggested that nudges may fail due to the provision of complex and confusing information, losing resonance over time due to repeated intervention exposure or producing compensating behaviors (*e.g.*, successfully encouraging people to be more active and energetic, yet triggering rebound effects like eating more). Similarly, Caraban et al. [26] identified seven reasons for failure in producing behavior change, such as lack of educational gains, the inability to sustain behavioral effects over time, or backfiring effects and compensating behaviors.

2.2 Information sources

Publications were retrieved from the following databases: the Association for Computing Machinery (ACM) Guide to Computing Literature, SCOPUS, Web of Science, ScienceDirect and PubMed.

2.3 Search strategy

We searched for articles by applying the following query on publication titles, abstracts, or author keywords: ***("behavior change" OR "behaviour change") AND ("misinformation" OR "fake news")***. Fake news - *false content that is intentionally and verifiably false and can mislead readers* [6, 96] - and misinformation are often used interchangeably in research discourse [70]. As such, including both terms allowed us to capture a broader volume of relevant articles. The following databases - ACM, Web of Science and PubMed - did not permit a simultaneous search within all three sections. For those, we applied the query to the full publication text. All publications were retrieved between the 1st and 15th of June 2024.

2.4 Eligibility criteria

Publications were eligible for inclusion when they were a) written in English, b) published in a peer-reviewed journal or conference, and c) presented one or more novel digital interventions addressing misinformation through behavior or attitude change. For the scope of this review, drawing on Alkhaldi et al. [5], we define digital interventions as programs delivered through digital platforms (*e.g.*, mobile apps, websites). All interventions should demonstrate a clear attempt to influence human behaviors or attitudes, and establish a link to theories and principles of behavior or attitude change (see Davis et al. [31], Vogel and Wanke [124]). Online misinformation is defined as *false or misleading information shared without malicious intent* [63, 126], in the form of online messages, ads, news articles, and social media posts [96].

2.5 Procedure

The review was conducted according to the PRISMA-ScR guidelines [120], and consisted of the following steps (see Figure 2):

- **Identification:** In total, 659 publications were identified by applying the query to the five databases. Necessary publication metadata (publication title, authors, year, journal or conference of publication, database where publication was found) were extracted in a spreadsheet. Duplicate entries (N=159) were removed from the publication pool.
- **Screening & Eligibility:** A total of 500 publications were screened against the eligibility criteria. All publications were written in English. 114 publications were excluded due to not being published in a peer-reviewed journal or conference (92 were conference proceedings collections, seven were books, four book chapters, four workshop reports, three abstracts, one thesis, one editorial report, one letter and another general report).
  For the remaining 386 publications, we examined their abstract, and when needed the full text, to assess whether they fulfilled the third eligibility criterion: presenting one or more novel digital interventions addressing misinformation through behavior or attitude change. The vast majority (N=257) were excluded for not presenting a digital intervention. Their contributions ranged from frameworks for modelling human actions in epidemic simulations [32], providing systematic reviews on topics like social media fatigue [136], or attempting to interpret users' digital behaviors like the propagation of COVID-19 checks from unreliable sources [28]. Also, 51 publications were excluded for not addressing misinformation. For example, one publication presented the design and development of an educational game for increasing resilience and improving community readiness in case of wildfires [66]. 50 more publications were excluded for not presenting interventions aiming for behavior or attitude change. As an example, exploring the potential of the Internet of Things objects to convey journalism and media-related information to individuals [84]. Based on these three criteria, 358 publications were removed from the pool.
- **Included:** Overall, 28 publications fulfilled all three inclusion criteria and were incorporated in the review. This list was expanded with nine more publications retrieved through snowballing sampling, resulting in 37 publications.

3 Findings

Out of the 37 publications, we identified 67 digital interventions. In this section, we present the findings regarding each research question.

3.1 Which behavioral objectives were the interventions designed to pursue?

All in all, we identified 17 distinct objectives (see Table 1) and classified them into the three stages of the online *"News Process"* model: composition, amplification, and consumption.

3.1.1 Composition. For the composition stage, behavioral interventions intend to discourage the creation of online misinformation. We identified four distinct objectives from five interventions.

One of the most popular objectives for the consumption stage is to *throttle misinformation composition activities right when they take place*; discouraging the creation of false claims before being published. For example, Konstantinou et al. [75] proposed four concepts that share this objective. *Await* introduces a ten-second countdown when users attempt to post a non-credible hyperlink on social media. During the countdown, users may choose to cancel the publication, wait for the countdown to be finished and for the post to be published, or bypass the waiting time by clicking on the *"post now"* button. Differently, *Aureole* intervenes as users compose the social media post by applying color-coding based on the credibility status of an attached link - pink when the link is fact-checked as non-credible, green when the link is fact-checked and verified as true, grey when it cannot be assessed, and white when no link is included. A relevant short text below informs the user (*e.g., "The link you included in the tweet has been flagged as misinformative. Click here to check why."*). Differently from *Aurole*, *Noticeable* informs the post author of the implications of including a misinformative link. Posts including links that are fact-checked and verified as true are promoted, reaching double the audience, while posts containing misinformative links only reach half of the intended audience. Finally, *Socially* discourages users from generating misinformative posts by reminding them of people in their network who may perceive their post as misinformative [75]. This nudge leverages the spotlight effect: people's tendency to overestimate the extent to which their actions are noticeable to others [50]. All nudges attempt to throttle mindless activity (see Caraban et al. [26], Konstantinou et al. [75]) and promote reflecting decision-making. They do so by tapping on regret aversion: the tendency to become more cautious decision-makers when a certain level of risk is identified [108].

Interventions may *reinforce fact-checking during content composition*, encouraging users to verify and address the credibility of a post they are creating. *Aureole* and *Await* strategically place a link (*e.g., "check here why"*) alongside the color-coding and countdown, respectively, encouraging users to participate in content verification before submission, gain deeper insights on the post subject and be more careful content creators [75]. Also, with color changing, *Aureole* and *Noticeable* attract reflective processing and prompt users to rethink ways for changing the color from red to green [138].

A third objective is to *retract generated misinformation*: reexamine content credibility and remove or issue corrections in case of misinformation. For instance, Tsipursky et al. [121] proposed the *Pro-Truth Pledge*, whereas users agree to sign and abide by 12 behaviors intended to counteract believing and sharing online misinformation. When signing the digital pledge, users may share about it on social media, putting an indicator in their profile photos showing they took the pledge [121]. One of the agreed behaviors encourages users to honour the truth: reevaluate if the credibility of their generated content is challenged, and retract when they cannot verify it [121]. The *Pro-Truth Pledge* leverages public commitment [26]; having users publicly commit to a truthful behavior taps into their tendency to be true to their words [113], thus adhering to the pledge behavior.

A similar objective is for interventions to encourage users to *communicate with others for editing or retracting generated misinformation* and celebrate those who do so. Continuing from the previous example, the *Pro-Truth Pledge* prompts another behavior: users asking mutual contacts to retract or correct unreliable information they may have generated [121]. Taking advantage of users' desire to conform with actions expected from them [26] through the profile indicator, the *Pro-Truth Pledge* motivates users to stick to the consented pledge action.

3.1.2 Amplification. In the amplification stage, behavioral interventions aim to minimize the spread of misinformation. We identified five distinct objectives from 12 interventions.

The most popular objective is to *throttle misinformation propagation activities right when they take place*, targeting actions such as sharing, liking and commenting on posts containing false claims. For instance, a nudge introduced the following message above an article thumbnail on social media, encouraging general caution to avoid propagating misinformation: *"There is a lot of misleading and false information online. Most responsible people think twice before sharing content with their friends and followers"* [7]. This nudge employs the throttling mindless activity technique [26], attempting to interrupt users from mindlessly sharing online (mis)information and to encourage more deliberate content examination before doing so. Another intervention warns users how many other accounts viewed but didn't share posts with false claims (*e.g., "In your personal Twitter network, X out of X people saw but did not like, reply or retweet this tweet."*) when attempting to share misinformative posts [67].

A second objective is to *promote content fact-checking* before re-posting it; encouraging users to verify information accuracy and source reliability through fact-checking sources. Armeen et al. [10] came up with an implementation intention intervention - an *"if-then"* plan on social media (*"If I want to share a story on social media, then I will check its validity on Snopes.com."*). Also, Konstantinou et al. [75] came up with *Alternative* and *Changy* nudges. When users attempt to share a post containing links to misinformation, the former provides a warning with a fact-checking article debunking the false claims provided in the link, and the latter proposes a switch to an alternative, credible link. Both nudges employ the multiple viewpoints strategy by Caraban et al. [26], which taps into the confirmation bias - the tendency to search for, interpret and favour information that confirms one's pre-existing views and beliefs [92].

A third objective is to *encourage social interactions that reward and defend pro-truth behaviors online*, thus creating a culture where content truthfulness and accuracy are valued, celebrated and actively promoted. For instance, the *Pro-Truth Pledge* encourages users to acknowledge when others share truthful information and defend them when they wrongfully come under attack, even when they disagree otherwise - an attempt to tackle their inner confirmation bias [121].

Propagating credible content is another amplification-oriented objective. It ensures that credible, well-sourced information reaches a wider online audience. For instance, Reuter et al. [102] developed and formulated anti-smoking health messages, posted across social

Table 1: List of 17 behavioral objectives grouped into the three stages of the modified "News Process" model.

Composition	Amplification	Consumption
1. Throttle misinformation composition activities right when they take place [74, 75] 2. Reinforce fact-checking during content composition [74, 75] 3. Retract generated misinformation [121] 4. Communicate with others for editing or retracting generated misinformation [121]	1. Throttle misinformation propagation activities right when they take place [7, 67, 74, 75] 2. Promote content fact-checking [10, 74, 75, 121] 3. Encourage social interactions that reward and defend pro-truth behaviors online [121] 4. Propagating credible content [102] 5. Reporting misinformation [51]	1. Increase knowledge on a topic of concern [2, 8, 13, 17, 21, 24, 25, 35, 36, 48, 61, 100, 103, 115, 118, 134] 2. Boost users' media literacy skills [4, 24, 35, 45, 56, 94, 95, 112] 3. Stimulate critical and reflective thinking [15, 41, 74, 75, 77, 94, 95, 103, 135] 4. Foster collaborative attempts to establish consensus [24, 48, 103, 121] 5. Encourage careful formation of search queries [74, 75, 132] 6. Promote careful browsing [74, 75, 132] 7. Encourage consumption of fact-checked content [74, 75, 102, 134] 8. Discourage the consumption of misinformation [74, 75, 110]

media, as paid ads and unpaid organic results, focusing on the risks of using tobacco products and highlighting negative effects and misinformation. By ensuring accurate and credible information is more visible and accessible, the intervention seeks to positively influence anti-smoking behaviors and increase the pool of credible posts publicly available.

A fifth objective is targeting a specific action: *reporting misinformation*. This objective encourages users to identify and report misinformation, ensuring it is reviewed by moderators or fact-checkers to prevent further spread. Gimpel et al. [51] designed and developed textual messages shown at the beginning of a social media news feed (*"Fake news is increasingly threatening public opinion. It is, therefore, important that our users report such inappropriate content to improve quality for all of us."*) and alongside social media posts (*"This content has been reported by X users."*), in an attempt to encourage users to report misinformation. Both messages employ social comparison [26], tapping into the herd instinct bias, encouraging users to replicate others' actions [46].

3.1.3 Consumption objectives. In the consumption stage, behavioral interventions aimed at promoting critical engagement with online content. We identified eight distinct objectives from 55 interventions. The most popular objective is to *increase knowledge on a topic of concern*. To provide credible information that helps users gain a deeper understanding of the topic at hand while addressing relevant misinformation, thereby fostering informed decision-making. Adib and Orji [2] designed a web application that employed visualizations, interactive game narratives and persuasive principles to raise awareness about climate change while tackling myths and false claims. Data visualizations aim at effectively communicating evidence about climate change and the pivotal role humans play in accelerating it. An interactive game allows users to drive a story by making decisions regarding daily behaviors and simulating their effects at scale, while persuasive principles are employed to promote behaviors that mitigate climate change. In another example, Youngseung et al. [134] designed *ChamberBreaker*, a game-based system that aims at increasing players' awareness and responses to the *echo chamber* effect: users being exposed to information and opinions of like-minded peers, leading to the reinforcement and amplification of such beliefs, thus generating polarization and biased information, while excluding different perspectives [111]. With *ChamberBreaker*, players become anonymous social media users and are asked to continuously share biased posts that are expected to form an echo chamber [134]. With this game, players are expected to learn about the characteristics of echo chambers and educate themselves about the negative consequences of belonging in one. Other interventions employed short, animated, story-based videos to foster knowledge, behavioral intent, and engagement around COVID-19 vaccinations [13, 17, 115].

A second objective is to boost *media literacy skills* - critical thinking skills enabling individuals to decipher information from mass communication channels and empower the development of independent evaluation skills for media content [99]. The focus is on teaching users to assess content critically, enabling them to distinguish between reliable and unreliable sources with the help of various techniques (*e.g.*, lateral reading) [87]. *TrustyTweet* is a browser plugin assisting users in assessing social media posts by providing a set of politically neutral, understandable and intuitive warnings and hints, adhering to media literacy guidelines for identifying misinformation [56]. Such hints include excessive emoticon and punctuation use, the absence of official account verification seals, and consecutive capitalization. Other interventions may target health literacy: the degree to which people can obtain, process and communicate health-related information for taking based health decisions [19]. Examples are two health campaigns among struggling communities, addressing inaccurate and misleading information regarding vaccination, intending to improve health literacy [35]. Both campaigns incorporated communication assets (*e.g.*, media graphics, emails) which were formulated upon particular persuasion tactics, such as the Elaboration Likelihood Model

(ELM) which suggests that attitude building and behavioral persuasion depend on information flow through central or peripheral routes [41, 116]. For instance, emails were used to convey relevant information, emphasizing the importance of being vaccinated, prompting users to reflect critically on such information and take appropriate action [35].

A similar objective is to *stimulate critical and reflective thinking*. To encourage a reflective approach when consuming content, prompting users to think critically about the underlying message and its broader implications, not just its surface-level accuracy. As an example, *ClarifAI* is an automated propaganda detection tool, devised to prompt readers towards critical content consumption by activating the reflective and analytical mode of thinking [135]. *ClarifAI* detects and highlights instances of propaganda in a text, alerting about its potential presence with distinct colors and offers an explanation of the propaganda technique used and why the text may be propagandistic when users hover over a flagged segment [135]. The goal of this tool is to facilitate both quick, intuitive interactions, as well as thorough and reflective thinking through provided explanations. *ClarifAI* relies on just-in-time prompts [26] (*e.g.*, colored highlighting and explanations) as soon as a text containing propaganda comes into view, nudging users to be careful.

Another objective is to *foster collaborative attempts to establish consensus*. This urges users to collaborate to discuss and agree on the most accurate information interpretations, particularly for controversial or complex topics. Through shared dialogue and cooperative fact-checking, users may collectively create a more balanced understanding, reducing polarization and fostering a shared sense of truth. *PandeVITA* is a digital ecosystem that facilitates access to health-related credible information and fosters collaboration and communication between users, health professionals and scientists, to improve relevant decision-making for effectively and efficiently tackling a crisis like COVID-19 [48]. Through collaboration, *PandeVITA* aims to help users mitigate economic, healthcare and social drawbacks associated with COVID-19-related misinformation and generate social awareness and responsible behavior [48].

Objectives may occur in particular contexts, like search engine environments. For one, interventions can *encourage careful formation of search queries*, inspiring users to reflect upon the formulation of queries, to retrieve more credible results. An example of such intervention would be a query auto-completion presenting particular search terms purposefully stimulating critical thinking [132]. The query terms were based on the priming effect: users were presented with particular terms (*e.g.*, comparison, evidence, statistics) aiming to encourage careful search experiences [132]. Another similar objective is to *promote careful browsing*, aiding users in searching, navigating and deciding on the type of search content they consume more mindfully. Continuing from the previous example, priming queries may also populate the retrieved search results with particular words (*e.g.*, verification, research, actual proof) [132].

Another objective is to *encourage consumption of fact-checked content*. The objective involves guiding users toward credible and reliable information, reviewed for accuracy by trusted fact-checkers. A nudge - *Precious* - inserts fact-checked articles in users' timeline feed only for a limited period [75]. By making fact-checked content a scarce resource, *Precious* leverages the scarcity bias: individuals' tendency to appreciate rare objects, as they are more difficult to acquire in the future [30]. These fact-checked articles are perceived as limited and valuable, prompting users to examine them immediately [75].

Lastly, interventions may opt to *discourage the consumption of misinformation*, steering users away from non-credible content. Story-based interstitial warnings (interrupting users' primary tasks) against targeted clickbait may deter users from clicking, viewing and consuming such content [110]. The design of these interventions targets users' reflective engagement and thoughtful decision-making, shifting from their desire to check clickbait to reevaluating their actions through interruption [110]. For example, some of the warning variations tapped into users' emotions by depicting characters who had clicked on clickbait, and as a result, experienced negative emotions like sadness and anxiety. These interventions also count on instigating empathy [26], attempting - through persuasive, emotionally charged narratives - to exploit the affect heuristic and nudge users to avoid clickbait.

3.2 What theories of attitude or behavior change were interventions grounded upon?

Overall, we identified 24 theoretical frameworks employed in the design of interventions. Interestingly, some of the theories featured prominently in the review of Davis et al. [31] were not as evident in our review. For instance, the *Transtheoretical Model of Change* [101], employed in one-third of the interventions in the review of Davis et al. [31], was only used in one intervention design in our sample. Similarly, the *Theory of Planned Behavior*, the second most featured theory in Davis et al. [31], was used in only three intervention designs in the present scoping review. Table 2 outlines the theoretical frameworks organized into five overall categories. As some interventions leveraged more than a single theoretical framework, the percentages displayed below add up to more than 100%.

3.2.1 Theories of Persuasion. A total of 28 (42%) interventions leveraged theories of Persuasion - explaining how and why individuals can be influenced to change their attitudes or behaviors - such as the Inoculation Theory [123], the Elaboration Likelihood Model (ELM), the Systematic-Heuristic Model of Persuasion [40], frameworks describing principles of Persuasion like Cialdini's principles of Influence [29] and the Persuasive Systems Design framework [93], and media consumption theories such as the Affordance Theory [54].

The Inoculation Theory [123] was employed by four intervention designs. As an example, the *Bad News* game [15, 105] aims at building cognitive inoculation against misinformation by pre-exposing users to weakened doses of misinformation. Users take the role of a fake news tycoon who aims to produce and disseminate misinformation and learn to use the tactics of six common misinformation techniques, such as impersonating people, using emotional language, and inducing group polarization.

The Elaboration Likelihood Model (ELM) [98] was employed in the design of three interventions. Ebnali and Kian [41] employed ELM to redesign a health-related web forum, to provoke critical examination of content. To do so, they refrained from adopting

Table 2: Broad categories of theories and theoretical concepts identified within each publication.

Category	Theoretical basis
Theories of Persuasion	Inoculation Theory [15, 61, 74, 75, 134], Elaboration Likelihood Model [35, 41], Persuasive Systems Design framework [2, 52], Other Persuasive Principles [25, 36, 103], Selective Exposure Self-Affect-Management Model [102], Affordance Theory [102], Relevance Theory [110], Continued Influence Effect [74, 75], Systematic-Heuristic Model of Persuasion [112]
Theories of Decision Making	Dual-System Theory of Cognition [135], Dual-Process Theories [77], Heuristics & Cognitive Biases like the Confirmation Bias [56], the Dunning-Kruger Effect [121], Social Norms [7, 8, 51, 67, 94, 95], Priming Effect [132] and others (*e.g.*, Status-Quo Bias, Scarcity Bias, etc.) [74, 75, 134]
Theories of Motivation	Self-Determination Theory [24, 48, 100], Self-Efficacy Theory [45]
Socially Situated Theories	Social Cognitive Theory [24], Social and Behavior Change Theory [4], Social Judgment Theory [35], Social Translucence Theory [74, 75]
Descriptive Theories & Models	Health-Belief Model [13, 17, 115], Diffusion of Innovations Theory [24], Theory of Planned Behavior [10, 13, 17], Steps to Behavior Change Model [118], Extended Parallel Process Model [35], COM-B [21]

peripheral cues, such as the number of likes in a post, or a star system to rank users' credibility, and instead, leveraged central route initiators, such as injecting trusted links with small descriptions related to the topic of a discussion and extensive threads of discussion. The goal of incorporating multiple central route initiators was to trigger more visual and cognitive attention, thus encouraging users to consume health-oriented information with more suspicion and critical thinking [41].

A significant number of interventions (N=13, 20%) employed no theory, but instead relied on principles of persuasion. For instance, Gurgun et al. [52] designed several interventions using seven persuasive principles of the Persuasive Systems Design framework [93]: reduction, suggestion, self-monitoring, recognition, normative influence, tunneling and liking.

Media consumption theories, such as the Selective Exposure Self-and Affect-Management (SESAM) model [73], the Affordance Theory [54], the Relevance Theory [43] and the Continued Influence Effect [27], were employed by eight interventions. For instance, Reuter et al. [102] explored ways to communicate anti-smoking messages on social media. Building upon the Affordance Theory [54], suggesting that user engagement with social media content is influenced by the characteristics of social media platforms, and thus Facebook posts should be made simple and plain as the platform relies on simplicity and immersive interface design, while Instagram posts should leverage visual imagery [102].

3.2.2 Theories of Decision-Making. Theories and concepts of decision-making guided the design of 26 (39%) interventions - examining how individuals assess information and make choices. Two interventions tapped into theories of decision-making and particularly dual-process theories such as the Dual-System Theory of Cognition [68], while another 24 interventions capitalized on the effects of cognitive biases and heuristics, and one also relied on the priming effect [68, 79]. For instance, Tsipursky et al. [121] created the *Pro-Truth Pledge* which encourages signees to abide by twelve behaviors, all intended to counteract different cognitive biases contributing to accepting and spreading online misinformation, such as the confirmation bias - people's tendency to favour information that confirms their prior beliefs and values [92], by asking users to search for evidence that may disprove their initial beliefs, and the Dunning-Kruger effect - people's tendency to have a favourable, inflated perception of their abilities when they lack such knowledge and skills [121] - by encouraging users to *"recognize the opinions of those who have substantially more expertise on a topic than myself as more likely to be accurate in their assessments"* [121].

Yamamoto and Yamamoto [132] redesigned a search engine to promote careful information-seeking and engagement during and after issuing search queries. To do so, they leveraged the ideomotor/priming effect, which suggests that individuals make motions and movements unconsciously or involuntarily due to a thought [68, 79], and reformulated the auto-completion and auto-suggestion keywords in a search engine environment, using words such as comparison, evidence, and research, to prime their respective concepts and, in turn, promote careful information-seeking behaviors [132].

3.2.3 Theories of Motivation. Four interventions (6%) leveraged theories of motivation - exploring internal and external factors driving individuals to act. Three of them employed the Self-Determination Theory (SDT) which suggests the fulfilment of three psychological needs - competence, relatedness and autonomy - to increase individuals' intrinsic motivation for an activity [107]. For example, Buller et al. [24] developed *#4Corners4Health*, a social media campaign to advance media and digital literacy skills for cancer prevention. Among other theories, they employed SDT in formulating messages that foster individuals' perceptions of competence, relatedness and autonomy for accessing, analyzing, reflecting on content, responding to misinformation, and referring to community resources for fact-checking, among others.

One intervention focused on the concept of *self-efficacy* - one's internal assessment of ability to complete a task [12]. In an attempt to boost individuals' self-efficacy beliefs and in turn their media literacy skills, Ferrucci and Hopp [45] developed a social media

intervention providing positive feedback to users, reassuring them about their capabilities, competencies and skills in evaluating the credibility of online content and identifying misinformation.

3.2.4 Socially Situated theories. Five interventions (7%) relied on socially situated theories - considering how the social context shapes behaviors - such as the Social Cognitive Theory [11], the Social and Behavior Change Theory [4] and the Social Translucence Theory [42]. For instance, Buller et al. [24], employed the Social Cognitive Theory which posits that learning occurs in a social context with a dynamic and reciprocal interaction of the person, environment, and behavior [11], when designing a social media campaign that aimed at increasing individuals' digital and media literacy skills and their responses to misinformation [24].

3.2.5 Descriptive Theories & Models. Ten interventions (15%) relied on descriptive theoretical frameworks and models of behavioral determinants - providing descriptions of how individuals behave - such as the Capability-Opportunity-Motivation-Behavior (COM-B) model [128], the Theory of Planned Behavior [3] and the Health-Belief Model [65]. For instance, when designing an audio-visual for health literacy, Bonner et al. [21] employed the COM-B model to identify capability, motivation, or opportunity barriers that users may face and took action such as simplifying the reading level by eliminating complex jargon.

3.3 How prevalent is stakeholder involvement in the design of interventions?

Out of the 67 interventions, the authors provided no design process details for 19 interventions (28%). The remaining 48 interventions and their practices of stakeholder involvement are analyzed in the text below.

3.3.1 What frameworks of intervention design did researchers rely upon? We identified three categories of intervention design frameworks: a) frameworks rooted in User-Centered Design (UCD), b) frameworks of co-creation, co-design, or participatory design, and c) frameworks rooted in the Behavioral Sciences.

UCD frameworks, stressing iterative design and quick feedback cycles, were employed in the case of 6 interventions (13%). Hartwig and Reuter [56], for example, followed a five-step process rooted in a *design science approach* [58]: a) achieving problem awareness, b) suggesting possible solutions, c) implementing solutions, d) evaluating their impact and finally, e) providing concluding comments. Ferrucci and Hopp [45], who designed a textual intervention boosting users' self-efficacy on social media, adopted design principles from prior empirical work regarding users' attention span and engagement with social media interventions and employed these throughout the design process.

Participatory design [88], co-creation [109] and co-design [9] methodologies were used in 35 interventions (73%), involving the active participation of different types of stakeholder in the design process. Stakeholders were involved in different phases of the design process and with varying objectives. For instance, Skipper et al. [112] held workshops with students, social media influencers and academics, to define the topics of the intervention and its possible formats. Shrestha et al. [110] held a structured design session, providing participants with user interface components such as overlays, headers, and navigation buttons, asking them to design alternative implementations of the intervention, while Dhaliwal et al. [36] involved community leaders and health workers in the development of audio-visual content aimed at increasing vaccine acceptance. Others included stakeholders such as researchers, experts and end-users, as equal members of the design team, often involving them throughout the full cycle (*e.g.*, [15, 121]).

Behavioral science frameworks for intervention design and development, such as the Intervention Mapping Framework [14], the Health Information Persuasion Exploration (HIPE) framework [35], and ad-hoc intervention design methodologies developed around a behavioral model, such as the COM-B model [128], were employed in the case of 7 interventions (15%). Desens et al. [35] followed the HIPE framework to detect and analyze harmful health-related misinformation and persuasion tactics, and to design campaigns tackling vaccine hesitancy. They followed a four-step process: a) *detect* - identifying health-related misinformation regarding COVID-19 through customized searches on social media, blogs and newspapers, as well as reports from crowdsourcing services, b) *analyze* - studying the discourse to understand the relevant narratives around COVID-19 vaccination, c) *design* - developing communication assets and response strategies, and d) *evaluate* - assessing the impact of the developed assets [35]. Others, as in the case of Bonner et al. [21], developed an ad-hoc intervention design methodology around a behavioral model, like COM-B [128]. Bonner et al. [21] developed TikTok-style audio-visual interventions to increase knowledge on COVID-19 testing. Their intervention design consisted of four phases, with the first two being national surveys eliciting and coding behavioral barriers to COVID-19 testing using the COM-B framework and estimating their prevalence and health disparities, and the latter two being experiments testing the intervention, health literacy-sensitive information for addressing capability and motivation barriers, first in textual form and then in audio-visual form.

3.3.2 Parallel design. 25 of 48 interventions (52%) were developed using a parallel design approach which involves creating multiple alternatives simultaneously, in an attempt to explore the design space. Parallel design was employed at different phases of the design process: from the initial brainstorming and discovery phase to the ideation phase and, finally, the concept refinement phase. For instance, Hopkins et al. [61] employed parallel design during the initial brainstorming and discovery phase. They provided experts with initial sketches of the *Cranky Uncle* character to receive early feedback and form, based on these, visual mockups and scripts of their game to be used as design probes during their co-design workshops. Most employed parallel design during a structured ideation phase. Skipper et al. [112] and others [100, 110] created and tested several design variations, and prototypes with experts or participants - before deciding on the final intervention. Parallel design was also used during concept refinement; generating intervention variations upon researchers' reflection on those ideas. Konstantinou et al. [75] employed parallel design during ideation and concept refinement, comparing, expanding and grouping ideas that emerged from a design workshop, and mapping them back to their theoretical framework [26].

3.3.3 What types of stakeholders did researchers involve in their design process? For 45 of 67 interventions, researchers reported on stakeholder involvement. The vast majority (N=35, 78%) involved **experts**, from journalists and media researchers (*e.g.,* [15, 24]), to topic experts such as community health workers [36], medical doctors, nurses, epidemiologists and pharmacists [115], to behavioral intervention designers (*e.g.,* [75]). For eight interventions, experts were undefined (*e.g.,* [52, 100]). Ten interventions (22%) were developed with the help of **end-users**. While some of the studies focused on particular populations, such as pregnant women [8], young people aged 16 to 24 years [61], and pupils aged 11–13 [112], others adopted wider inclusion criteria, such as being active social media users [56], or adults, living in the United States or Canada and having a 99% HIT approval to increase the quality of responses in their study.

3.4 To what extent were interventions evaluated to assess their behavioral or attitudinal effectiveness?

Overall, 69% of interventions (46 of 67) were evaluated for their effectiveness in prompting behavioral or attitudinal change. Most interventions (61%, 28 out of 46) were evaluated through a controlled experiment, either online or in a laboratory setup, with the experiment lasting for one hour or less. For instance, Youngseung et al. [134] conducted a between-subjects online experiment that evaluated whether a 15-minute session with *Chamberbreaker* would alter individuals' news consumption behaviors and knowledge about the echo chambers. The remaining interventions (39%, 18 out of 46) were tested through a field trial, lasting from weeks (39%, N=7) to months (56%, N=10), or even more than a year (6%, N=1). For example, Thompson and Harutyunyan [118] launched a multimedia campaign called the *Green Path Campaign for Family Health*, which lasted six months, to boost awareness, acceptance and adoption of modern contraception practices. Following the campaign, a representative sample of 1088 married women were surveyed to evaluate the campaign's impact, which revealed a significant increase in favourable attitudes towards modern contraceptive methods, family planning services and other relevant information-seeking services [118].

3.4.1 Controls, multiple arm and single arm designs. A control condition was adopted in about half of the studies - in 12 out of the 28 (43%) interventions evaluated through experimental studies, and in 9 out of 18 (50%) interventions evaluated through field trials. In some studies, controls consisted of the absence of the intervention. For instance, in a study of the impact of a social norm-based message in reducing the probability of sharing false news, participants in the control condition weren't exposed to any message. In other studies, a baseline intervention was introduced as a control condition. For instance, for assessing *ChamberBreaker*, experimental condition participants were asked to play the game for 15 minutes before completing a post-task questionnaire and those in the control condition were asked to read a definition of the echo chamber effect and examples of scenarios and posts. Similarly, in the study of the *Bad News* game [15], participants in the experimental condition played the game for 15 minutes and participants in the control condition played Tetris - as a control for *Bad News* [15].

While only about half of the interventions were compared to a control condition, the remaining interventions were evaluated through a multiple-arm design, comparing two or more variations of the interventions. This was the case for 13 of the 28 (46%) interventions evaluated through an experimental study and 4 of 18 (22%) assessed through a field trial. For example, Gurgun et al. [52] compared seven interventions, each leveraging a different persuasion strategy. Participants were exposed to all interventions and assessed their persuasiveness in tackling online misinformation. Reuter et al. [102] compared organic (unpaid) to advertised (paid) anti-smoking health messages on social media, tackling false claims regarding tobacco use and monitoring users' engagement with the messages such as liking, sharing, commenting and clicking on the message link directing them to an educational website.

3 of 28 (11%) interventions in lab studies and 5 of 18 (28%) in field trials, did not involve a control condition or any other form of comparison. For instance, Adib and Orji [2] had participants interact with their web app tackling climate change-related misinformation, followed by one-on-one interviews and self-reports on the system's perceived persuasiveness. Similarly, Powell et al. [100] publicly deployed the web-based health intervention addressing vaccine misinformation and hesitancy, for five months, measuring changes in vaccine hesitancy before and after the intervention, with the help of the Vaccine Hesitancy Scale.

3.4.2 Studying interventions post-removal. We found 6 of 67 (9%) interventions to be investigated for their effect after their removal. For instance, Yamamoto and Yamamoto [132] who evaluated the priming effect through a search query auto-completion technique on critical thinking, monitored participants' search actions with and without the auto-completion and auto-suggestion with priming terms, in an attempt to examine the persistence of the query priming effect. In another case, Orosz et al. [94] asked students to compose a letter addressing a close, older family member, in which they provided a summary of six strategies to identify misinformation. Following the letter composition, students were presented with real and fake news headline items. Four weeks later, participants completed a similar follow-up test of real and fake news headlines [94].

3.4.3 Participants and recruitment. The studies' median sample size was 264, with 13% (N=5 out of 40) having less than one hundred participants, 65% (N=26) having more than one hundred and less than one thousand, and 23% (N=9) studies having more than one thousand participants. Recruitment, most often (63%, N=25), took place through crowd-sourcing platforms such as Prolific Academic, Amazon's Mechanical Turk and Lancers, a Japanese crowd-sourcing service. However, this does not imply that the studies were conducted on crowd-sourcing platforms. For instance, Armeen et al. [10] conducted a pre-experiment survey on MTurk to gather qualified participants to join the trials.

3.4.4 Objective versus subjective measures. In assessing behavioral intervention effectiveness, researchers used objective and subjective measures (see Table 3). *Objective measures* ranged from gaze

behavior metrics, such as the total duration of fixation on the intervention, to engagement metrics such as the number of false claims reported by users, and the number of likes, shares, and comments made. *Subjective measures* were further classified into six categories: (a) *content accuracy* measures capturing perceptions of how accurate and trustworthy a piece of information is, (b) measures of *knowledge* assessing how the intervention impacts understanding or awareness of specific topics, (c) measures of *motivation and behavioral intention* regarding target behaviors such as sharing or engaging with (misinformative) content, or receiving a vaccine, (d) measures attempting to establish whether a change in *beliefs, attitudes, or emotional state* was incurred by the intervention, (e) measures of overt behaviors and practices, and (f) evaluative judgments on the intervention, such as perceived persuasiveness and usability.

3.4.5 Intermediate versus final outcome measures. We classified measures into intermediate and final outcome measures. *Intermediate measures* captured progress towards final outcomes [127], reflecting whether changes in conditions affecting behavior change occurred [55]. *Final outcome measures* reflected the extent to which interventions achieved their set objectives [55, 127]. Of the 46 interventions, 33 (72%) were assessed with final outcome measures, 8 (17%) with intermediate, and 5 (11%) with both.

Final outcome measures were classified into seven categories. Most frequently (N=20), interventions were evaluated in terms of their capacity to produce a change in *overt behavior*, such as an uptake in vaccination [35]. Other interventions (N=18) aimed at, and measured changes in *behavioral intentions* such as individuals' willingness to challenge online misinformation [52]. Eight interventions employed measures of *information perception* such as participants' ability to discern fake news headlines [94]. Five interventions were measured on *knowledge outcomes*, such as participants' knowledge about COVID-19 testing [21]. Four interventions aimed at affecting, and measured, individuals' *attitudes* towards certain behavioral practices such as contraception practices [118], while another two aimed at and measured *psychological outcomes*, such as improving anxiety levels among older adults for COVID-19 [115]. Last, two interventions measured the extent to which individuals engaged in *reflective thinking*. For example, Liao [77] evaluated the efficacy of a nudge aimed to engage users in high-level thinking while viewing videos, and employed measures of the extent to which individuals engaged in reflection and critical thinking.

Intermediate measures were classified into three categories. Most frequently (N=4), they reflected individuals' *behavioral intentions*. For example, Shrestha et al. [110] who developed interventions aimed at discouraging users from engaging with clickbait, measured participants' interest and likelihood of clicking on the provided clickbait, as an intermediate outcome of the intervention. The second category (N=3) related to individuals' *perceptions and attitudes* assumed to mediate, or reflect the intervention's capacity to produce the aimed effect, such as the perceived usability [56], usefulness [56], and persuasive capacity of the intervention [2]. The last category of intermediate outcome measures (N=1) related to *cognitive and psychological outcomes*, such as knowledge or self-efficacy beliefs. For example, Dhaliwal et al. [36] who developed an intervention facilitating vaccine acceptance, measured knowledge gains intermediate outcomes to evaluate progress toward this goal, such as improved knowledge of vaccination purposes and side effects.

3.4.6 Behavioral effectiveness. Of the 46 interventions, we could deduct their effectiveness in all but three, either because they were involved in an ongoing study [24, 100] or due to lack of information [8]. The remaining 43 interventions were classified as *successful*, *partially successful*, or *unsuccessful*, based on whether they had a statistically significant effect, in line with researchers' hypotheses, in all, some, or none of the primary outcomes (both intermediate and final outcome measures).

Among the 43 interventions, 23 (54%) were successful. For example, Zavolokina et al. [135] illustrated that a propaganda detection tool, when paired with explanations, received significantly more favourable opinions and effectively raised propaganda awareness (intermediate outcomes), and was also proved highly effective in enhancing critical thinking and increasing reading time (final outcomes). Ten (23%) interventions achieved partial effectiveness. For example, the intervention of Ebnali and Kian [41] which aimed at triggering reflective thinking toward health-oriented content, had no significant effect on intermediate measures such as participants' perception of health information, but a significant effect on the primary final outcome, with participants spending more time viewing health content, which was interpreted as an indicator of the extent to which they engaged in reflective cognitive activity [41]. The remaining 10 (23%) interventions did not meet their objectives. Gurgun et al. [52] evaluated the perceived persuasiveness of seven prototypes on willingness to challenge misinformation and reported that particular prototypes, such as *"sentence openers"* (guiding users through complex experiences), were deemed the least influential.

3.5 What are the reasons for interventions' failure?

Building on the taxonomies of Sunstein [114] and Caraban et al. [26], we identified nine factors of behavioral failure in the context of online misinformation. We begin by discussing six factors speculated by the authors to hinder intervention effectiveness and proceed to three factors that were empirically shown to reduce effectiveness.

3.5.1 Geo-cultural factors. Geo-cultural factors regarding populations being addressed, cultural norms and other regional communication habits, were hypothesized as critical factors for success or failure in several studies. For instance, Tsipursky et al. [121] suggested the success of the *Pro-Truth Pledge* in reducing the propagation of misinformation may have been dependent on participants' location; they suspected the pledge had a stronger, positive behavioral impact on Americans as they are often targeted by malicious foreign actors that spread misinformation, coupled with a lack of meaningful action taken by the US government in addressing misinformation [121].

Similarly, Gurgun et al. [52] observed that most participants considered that existing social media don't provide sufficient tools to challenge misinformation, and found several of the interventions

Table 3: Types of subjective and objective measures.

Subjective measures	**Content Accuracy:** Perceived Accuracy [112], Attitudinal Certainty [15], Perceived Trustworthiness [112], Fake News Evaluation Accuracy [94], Information Accuracy [45], Perceived Bias in a News Article [135], Confidence in Assessing Article [112]
	Knowledge: Coronavirus Knowledge Assessment Form [115], Vaccine-Related Knowledge [17], Article Comprehension [7], Propaganda Awareness [135]
	Motivations & Behavioral Intentions: Motivation to Share an Article [7], Net Promoter Score [135], Intent to be Vaccinated [17], Vaccine Hesitancy Scale [100], Willingness to Share an Article [7], Likelihood of Article Sharing [112], Likelihood to Click on Clickbait [110], Intention to Test for COVID-19 When Being Asymptomatic [21], Intention to Get Vaccinated [25]
	Beliefs, Attitudes and Emotions: Echo Chamber Breaking Questions [134], Interest to Click on Clickbait [110], Geriatric Anxiety Inventory [115]
	Behaviors & Practices: Extent to Which One's Actions Aligned With the Pro-Truth Pledge [121], Contraceptive Practices [118], Vaccination Action [25], Number of Stories Checked for Validity [10], Slow vs Fast thinking [135], Informed Decision-Making for Vaccination [8]
	Evaluations of the Intervention: Perceived System Persuasiveness [2], Intervention Acceptance & Usability [8]
Objective measures	Total Fixation Duration on Intervention [41], Order of Fixation in Areas of Interest of an Intervention [41], Number of False Claims Reported by Users [51], Reading Duration [135], Number of Queries Participants Issued in a Task [132], Number of Search Engine Result Pages Visited in a Task [132], Social Media Campaign Engagement Measures (Number of Likes, Shares and Comments) [24, 102, 103], Engagement (Click-Through Rate) with the Intervention [102], Vaccine Uptake (Percentage of Vaccinated Community Members) [35], Interaction with Misinformation (Number of Likes and Shares of Misinforming Articles) [67]

(*e.g.*, predefined question stickers, private commenting, thinking face reactions) effective in motivating users to challenge misinformation. The authors suggested this could be attributed to the sample, as participants were from the United Kingdom: a (Western) society where people are more open to direct discussions and confrontations, especially for challenging misinformation [52].

3.5.2 Lack of educational effects. Several interventions leveraged nudging principles that tap into the automatic mind. While such interventions can work without depleting users' cognitive resources, their effects are speculated to disappear once the interventions are removed. Konstantinou et al. [75] compared two of their interventions. *Sorted* reorders search results based on the likelihood of containing misinformation (placing credible information at the top and misinformation at the bottom). While the intervention is expected to decrease the likelihood of interaction with misinformation, it does not engage users in reflection on the importance of interacting with truthful content, or the ways to identify true from false information. As such, its effects will not likely be sustained once the intervention is removed. On the contrary, *Await*, which adds a countdown when attempting to post misinformative content, encourages users to reflect on their actions. As such, users are more likely to internalize the behavior and be less dependent on the intervention.

Similarly, Basol et al. [15] argued that inoculation interventions may not produce scalable and generalizable protection against misinformation if they fixate on building resistance against particular instances of misinformation (*e.g.*, certain topics), instead of the techniques underlying misinformation production. The *Bad News* game attempts to alleviate these concerns by preemptively exposing users to weakened instances of techniques underlying misinformation production [15].

3.5.3 Behavioral or attitudinal effects not sustaining over time. A failure to sustain the intervention's effects over time was frequently mentioned in our sample. For instance, Shrestha et al. [110] suggested story-based warnings against clickbait lose their resonance over time due to repeated exposure with users ignoring them. Likewise, Zavolokina et al. [135] highlighted that *ClarifAI* effectiveness may suffer due to prolonged use.

3.5.4 Unexpected effects and backfiring. Interventions are also hypothesized to produce unexpected and unintended outcomes, such as compensatory behaviors. Andı and Akesson [7] warned that nudges, reminding users to be more cautious about the content they share, could also decrease engagement with high-quality articles. Ferrucci and Hopp [45] pointed out that media literacy interventions increase belief scepticism in all types of content, whether they are credible or not. Zavolokina et al. [135] also emphasized the risk that interventions may make errors when detecting misinformation, thereby increasing susceptibility to it. Real-time detection techniques based on Large Language Models (LLMs), identifying and warning of textual misinformation, are prone to hallucinations: generating assessments that are irrelevant, made-up, or inconsistent with the input data, which in turn may reinforce misinformative claims [135].

3.5.5 Intrusiveness and reactance. Some interventions attempt to induce behavior change through friction (*i.e.*, pausing unwanted actions by instilling doubt and hesitancy) [26]. For instance, in an attempt to protect users from clickbait journalism, Shrestha et al. [110] introduced story-based warnings that interrupt users' primary

task of engaging with clickbait and attempt to prompt reflection about their decisions. They speculated that few users may feel the intervention undermines their autonomy and suggested a risk of reactance [110]. Similarly, Konstantinou et al. [75] suggested that nudges like *Await* and *Changy* are highly intrusive as they disrupt users' primary activity and require immediate action; *Await* pauses the submission of a post containing a non-credible hyperlink and prompts a ten-second countdown so users may reconsider their decision, and *Changy* asks users to switch to an alternative link from trustworthy sources when they attempt to share a post containing a link pointing to misinformation. In both cases, users may perceive nudges as disruptive, imposing a forced interaction before allowing users to resume their action [75].

3.5.6 Unattainable spillover effects. Interventions may fail to trigger positive spillovers to other behaviors [49]. Interventions, such as nudges developed by Konstantinou et al. [75], may positively influence decision-making regarding online misinformation (*e.g.*, prompt users to seek valid and truthful information to make more informed decisions), yet might fail to impact subsequent unrelated actions. Mazar and Zhong [78] found that exposure to a priming nudge (*e.g.*, making green products more prominent to promote ethical consumerism) did not activate norms of social responsibility for subsequent unrelated actions (*e.g.*, collecting the correct money users earned by themselves, after playing a game); users were more likely to cheat, stealing $0.48 more from the money envelope. In this case, the licensing effect (individuals acting in self-indulgence after behaving altruistically) [71] might have played a significant role in users acting negatively after doing something positive.

3.5.7 Psychological characteristics. Psychological characteristics, such as personality traits, often play a critical factor in the success of the intervention. For instance, Orosz et al. [94, 95] developed a family-based pro-social intervention asking participants to write a letter to their digitally less experienced relatives, explaining six strategies to identify misinformation. The intervention worked for individuals scoring high in need for cognition (one standard deviation above the mean), but not for those that scored lower.

3.5.8 Strength. The strength of the intervention - the intensity with which it is applied in a situation - was cited in some studies as an empirical reason for failure. For instance, Gimpel et al. [51] created descriptive social norm warnings (*e.g.*, *"This content has been reported by X users"*) motivating users to report misinformation. They found that when the message over-reported the number of users who flagged a particular content, users lost their motivation to flag due to the reduced benefit of the action, as many others had already done so [51].

3.5.9 Strong preferences and established habits. Other interventions failed due to participants' strong preferences towards particular outcomes. In the case of *ChamberBreaker*, conservative users showed significantly less flexibility in changing their attitude regarding echo chamber effects after playing the game [134]. The same was observed by Beleites et al. [17], who evaluated the impact of short story-based videos on knowledge acquisition, intent and engagement regarding COVID-19 vaccination. The results demonstrated that participants who identified as right-wing were least interested in engaging with the video while ending the experiment early [17].

4 Discussion

This scoping review aimed to provide an overview of digital behavioral interventions against misinformation. We reviewed the interventions' behavioral objectives, their theoretical bases, the methods and processes of their design and evaluation, and the reasons for failure.

4.1 Gaps in targeting misinformation at early stages

Overall, we identified 17 distinct **behavioral objectives** of interventions, organized into three distinct stages of the online news cycle: *composition*, *amplification* and *consumption*. We suggest that this framework of objectives and stages can guide intervention design during its early steps, as it may structure the discussion around the intervention scope, and the means to achieve it. Moreover, we found the majority of interventions to target misinformation during the consumption stage (47%), followed by amplification (29%) and composition (24%), highlighting the need for further exploration of behavioral interventions targeting the earlier stages of the news cycle. Other frameworks also demonstrate this limitation. Zhu and Yang [137] mapped out actions for addressing misinformation by reviewing policy documents issued by the US and the Chinese governments, during the COVID-19 pandemic. They identified three main action categories: direct (government agency actions to prevent or control misinformation), expenditure-based (affirmative and negative governmental actions), and information-based (efforts to influence individuals through knowledge transfer, reasoned arguments, and moral persuasion) [137]. While the former two focused on policy measures (*e.g.*, direct moderation), the latter sought to engage individuals constructively, influencing their behavior while respecting their autonomy. This group aligns with our focus on actionable objectives empowering users to make informed choices. Nonetheless, even within this category, the actions identified were limited (N=9) and consumption-oriented, such as promoting media literacy, or raising misinformation awareness [137].

4.2 Addressing the gap of behavioral frameworks gap in misinformation interventions

We identified 24 **theoretical frameworks** that guided the design of interventions, with the most prominent being theories of persuasion (42%) and decision-making (39%), followed by descriptive theories and models (15%), socially situated theories (7%) and theories of motivation (6%). As one would expect, we found limited use of popular, in other domains, theoretical models, such as the *Transtheoretical Model of Change* [101] (see Davis et al. [31]) which was used only by one intervention in our sample. Instead, dual process theories of decision-making, and the use of heuristics and cognitive biases, featured prominently in our sample, in contrast to prior reviews in the behavioral sciences [31]. The same was true for theories of Persuasion, with the Inoculation Theory being one of the most prominent theories in our sample. Future work can further attempt to inquire

into the exact behavior change techniques employed by the interventions, what Michie et al. [81] define as the *"observable, replicable, and irreducible component(s) of an intervention designed to alter or redirect causal processes that regulate behavior"*. Such analysis will help bridge the gap between the high-level (behavioral theories and antecedents), and the low-level (intervention implementation), thus generating scalable knowledge on a repertoire of effective interventions for misinformation. To the best of our knowledge, existing literature on misinformation interventions has mostly focused on techniques, rather than behavioral frameworks or theories underpinning them. For instance, Whitehead et al. [129] conducted a systematic review to identify communications-based intervention strategies for addressing and preventing vaccination-oriented misinformation. Overall, nine approaches were identified, such as scare tactics (*e.g.*, disease images) aiming to shock individuals into changing their attitudes, or humour-based corrections attached to false posts [129]. The review did not explore the theoretical frameworks applied to alter individuals' behavior.

4.3 A broader view of reasons for intervention failure

We identified nine **reasons of interventions' failure**, such as backfires and compensating behaviors, building on the taxonomies of Sunstein [114] and Caraban et al. [26]. One should note that only a few researchers explored ways to counter such behavioral failures. For instance, Armeen et al. [10] speculated that implementation intentions may lose resonance over time and suggested several counteract measures to be tested, such as periodic reminders and self-monitoring to strengthen the likelihood of fact-checking before sharing a post. For the same failure reason, Andı and Akesson [7] suggested deploying interventions for limited periods (*e.g.*, during election campaigns) or occasionally introducing modifications in color and placement. This highlights a noteworthy refinement compared to other frameworks identifying reasons for failure related to misinformation. For example, Gurgun et al. [53] pointed barriers to challenging misinformation by deriving insights from a survey among social media users. They reported demographic factors (*e.g.*, age, gender) influencing the likelihood of challenging misinformation; older users and men are more likely to address misinformation [53]. The survey also highlighted social conformity; 76% of participants reported the fear of provoking aggressive reactions from others as a significant barrier [53]. While such findings shed light on important barriers, they fixate on social media, rather than providing a broader view.

4.4 The 'Behavioral Responses to Misinformation' Design Cards

To assist researchers during the design and development of behavioral intervention against misinformation, we synthesize our findings from three research questions - *objectives*, *theories* and *reasons of failure* - into a set of design cards. Design cards are a widely adopted design support tool in Interaction Design, providing, what Rogers [104] calls *knowledge transfer* (*i.e.,* the translation of research findings from one discipline into another), and deliver several benefits for the design process: making the process visible and less abstract, communicate knowledge between group members and increase creativity and idea generation [34, 130].

The BRM cards comprise 50 double-face cards (see Figure 1). The first 17 cards cover the behavioral objectives that interventions may target, classified into *composition*, *amplification* and *consumption*. The next 24 cards cover the theoretical frameworks employed in the design of interventions, grouped into theories of *persuasion*, *decision-making*, *motivation*, *socially situated theories*, and *descriptive theories & models*. The last nine cards cover common or possible failure reasons, as identified in our review. We suggest that the former two categories of cards - *objectives* and *theories* - can assist researchers and intervention designers in the initial stages of the design process, through *scoping* (*i.e.* crystallizing the intervention objective) and *framing* (*i.e.*, crystallizing the behavioral antecedents the intervention aims), while the latter category - *reasons of failure* - can assist later in the design process, during *concept refinement*.

4.5 The rise of participatory design and parallel design

We found 73% of interventions to be designed with the help of participatory design methods, with the remaining ones relying on methods rooted either in User-Centered Design or Behavioral Science. Yet, despite the surprising adoption of participatory design methods, end-users, the individuals who will eventually be exposed to the intervention, were included in the design process only in 1 out of 4 times, with the remaining interventions involving experts such as journalists, media researchers and healthcare practitioners. We advocate for a stronger inclusion of end-users in the design of behavioral interventions against misinformation, as this will lead to more suitable, acceptable and culturally appropriate interventions [16, 61].

Surprisingly, we found about half of the interventions to come out of a parallel design process, applied both early on, during the initial brainstorming and discovery phase, as well as later in the design process, during concept refinement. The benefits of parallel design - helping design teams uncover unseen constraints, avoid fixation, and derive better and diverse solutions - are well documented in HCI and design literature (see [38, 119]), yet, parallel design has been reported to be infrequently practiced due to small budgets, lack of time and its propensity to lead to decision paralysis [57].

4.6 A growing emphasis on ecological validity

A frequent critique on behavioral interventions, and more broadly empirical research in Psychology, is that most of them take place in isolated laboratory environments, over limited periods, often with participant cohorts composed of university students, thus rendering doubts whether any of the findings will replicate in the real world [85]. In a recent review of nudging in HCI literature, for instance, Caraban et al. [26] found 65% of the empirical studies of nudging to have a duration of a day or less, with only 19% of the studies to last over a month. Moreover, they found only 14% of studies to inquire into whether the effects of the nudge hold after its removal.

Our review revealed a more positive picture. While 61% of the interventions were evaluated through a controlled experiment, either online or in a laboratory setup, with the experiment lasting for

one hour or less, 39% of the interventions were deployed in the real world, through a field trial, lasting from weeks (39%), to months (56%), or even more than a year (6%). About half of the interventions (46%, 21 of 46) were compared against a control condition, while another 37% (17 of 46) of the interventions were evaluated through a multiple-arm design, comparing two or more variations of the intervention. The median sample size was 264 with only 13% of the interventions studied with a sample of less than 100 participants. All combined paint a promising picture, highlighting a growing recognition of the importance of ecological validity. However, we suggest that more attention is needed in studying whether the intervention effects are sustained after their removal, as, similarly to Caraban et al. [26], we found that researchers inquired into this only for 13% of the interventions.

4.7 The importance of intermediate measures in evaluating intervention effectiveness

Evidence on the effectiveness of behavioral interventions against misinformation is considerably limited, and the findings are mixed, with a recent review of behavioral interventions aimed at reducing vaccine hesitancy, finding replication studies to often produce significantly smaller effect sizes, casting doubt on the efficacy of the techniques [106]. While our review did not aim, neither could allow for a systematic inquiry into interventions' efficacy, given the small and divergent sample of interventions in terms of their behavioral objectives, techniques, topics and participants, it portrays a similar picture, with almost half of the interventions in our sample either failing or partially failing to produce a significant effect on their primary outcomes. In line with Klasnja et al. [72], we emphasize the importance of measuring intermediate outcomes. Our review revealed only 20% of the studies included an inquiry of interventions' impact on an intermediate outcome. Such inquiries are critical in HCI research as they contribute to an understanding of which behavior change techniques work, when, and for whom [33], and can shed light on the dynamic, sequential nature of misinformation and its mitigation, from one's initial exposure to the development of misperception, and the formation of attitudes and subsequent behaviors [131].

5 Conclusion

In this paper we reviewed the state of the art of behavioral interventions against misinformation, inquiring into their behavioral objectives, theoretical foundations, design and evaluation practices, and factors that did or may have contributed to their failure. All in all, we identified 17 unique behavioral objectives that interventions may pursue, mapped 24 different theoretical frameworks for prompting behavioral or attitudinal change, noted the rise of participatory design, as well as an uptake in ecological validity studies, and highlighted nine reasons that impact or may play a role in effectiveness. We synthesized the findings into a set of design cards to guide researchers and designers of behavioral interventions during concept ideation and refinement and provided several recommendations for future work, including (a) a need for further exploration of the earlier stages of the news cycle, (b) a need for future work to inquire into the exact behavior change techniques employed by the interventions, bridging the gap between the high and the low level, (c) a need for a stronger inclusion of end-users in the design of behavioral interventions against misinformation, and (d) a need for an increased emphasis on studying interventions' effectiveness post-removal.

Acknowledgments

The study was partially funded by the Co-Inform project (770302), under the Horizon 2020 call "H2020-SC6-COCREATION-2016-2017 (CO-CREATION FOR GROWTH AND INCLUSION)" of the European Commission.

References

[1] Charles Abraham, Sarah Denford, et al. 2020. Design, implementation, and evaluation of behavior change interventions: A ten-task guide. *The handbook of behavior change* (2020), 269–84.
[2] Ashfaq Adib and Rita Orji. 2023. Persuasive System Design for Climate Change Awareness. In *International Conference on Entertainment Computing*. Springer, Springer Nature Singapore, Singapore, 115–129.
[3] Icek Ajzen. 1991. The theory of planned behavior. *Organizational behavior and human decision processes* 50, 2 (1991), 179–211.
[4] Altug Akin, Selin Turkel, and Pinar Umul Unsal. 2023. Infodemic Management for Social and Behavior Change: Youth Mobilization for Combating Disinformation During COVID-19. *Journal of Health Communication* 28, sup2 (2023), 41–48.
[5] Ghadah Alkhaldi, Fiona L Hamilton, Rosa Lau, Rosie Webster, Susan Michie, and Elizabeth Murray. 2016. The effectiveness of prompts to promote engagement with digital interventions: a systematic review. *Journal of medical Internet research* 18, 1 (2016), e6.
[6] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–236.
[7] S Andı and J Akesson. 2021. Nudging away false news: Evidence from a social norms experiment. Digital Journalism, 9 (1), 106-125.
[8] Charlotte Anraad, Pepijn van Empelen, Robert AC Ruiter, Marlies Rijnders, Katja van Groessen, and Hilde M van Keulen. 2024. Promoting informed decision making about maternal pertussis vaccination: the systematic development of an online tailored decision aid and a centering-based group antenatal care intervention. *Frontiers in Public Health* 12 (2024), 1256337.
[9] Matteo Antonini. 2021. An overview of co-design: advantages, challenges and perspectives of users' involvement in the design process. *Journal of Design Thinking* 2, 1 (2021), 45–60.
[10] Inaiya Armeen, Ross Niswanger, and Chuan Annie Tian. 2024. Combating Fake News Using Implementation Intentions. *Information Systems Frontiers* (2024), 1–14.
[11] Albert Bandura. 1999. Social cognitive theory of personality. *Handbook of personality* 2, 1 (1999), 154–196.
[12] Albert Bandura and Sebastian Wessels. 1997. *Self-efficacy*. Cambridge University Press Cambridge.
[13] Sandra Barteit, Violetta Hachaturyan, Ferdinand Beleites, Tilman Kühn, Caterina Favaretti, Maya Adam, and Till Bärnighausen. 2022. The effect of a short, animated story-based video on COVID-19 vaccine hesitancy: A study protocol for an online randomized controlled trial. *Frontiers in Public Health* 10 (2022), 939227.
[14] L Kay Bartholomew, Guy S Parcel, and Gerjo Kok. 1998. Intervention mapping: a process for developing theory and evidence-based health education programs. *Health education & behavior* 25, 5 (1998), 545–563.
[15] Melisa Basol, Jon Roozenbeek, and Sander Van der Linden. 2020. Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of cognition* 3, 1 (2020), 1–9.
[16] Jan M Bauer, Sabine Bietz, Julius Rauber, and Lucia A Reisch. 2021. Nudging healthier food choices in a cafeteria setting: A sequential multi-intervention field study. *Appetite* 160 (2021), 105106.
[17] Ferdinand Beleites, Maya Adam, Caterina Favaretti, Violetta Hachaturyan, Tilman Kühn, Till Bärnighausen, and Sandra Barteit. 2024. Evaluating the impact of short animated videos on COVID-19 vaccine hesitancy: An online randomized controlled trial. *Internet Interventions* 35 (2024), 100694.
[18] Kristoffer Bergram, Marija Djokovic, Valéry Bezençon, and Adrian Holzer. 2022. The digital landscape of nudging: A systematic literature review of empirical research on digital nudges. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA). Association for Computing Machinery, New York, NY, USA, 1–16.
[19] Nancy D Berkman, Terry C Davis, and Lauren McCormack. 2010. Health literacy: what is it? *Journal of health communication* 15, S2 (2010), 9–19.

[20] Ann Blandford, Jo Gibbs, Nikki Newhouse, Olga Perski, Aneesha Singh, and Elizabeth Murray. 2018. Seven lessons for interdisciplinary research on interactive digital health interventions. *Digital health* 4 (2018), 2055207618770325.

[21] Carissa Bonner, Carys Batcup, Erin Cvejic, Julie Ayre, Kristen Pickles, Tessa Copp, Samuel Cornell, Brooke Nickel, Mustafa Dhahir, Kirsten McCaffery, et al. 2023. Addressing behavioral barriers to COVID-19 testing with health literacy–sensitive eHealth interventions: results from 2 national surveys and 2 randomized experiments. *JMIR Public Health and Surveillance* 9, 1 (2023), e40441.

[22] Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. 2011. The socialbot network: when bots socialize for fame and money. In *Proceedings of the 27th annual computer security applications conference* (Orlando, Florida, USA). Association for Computing Machinery, New York, NY, USA, 93–102.

[23] Amy Bucher, E Susanne Blazek, and Christopher T Symons. 2024. How Are Machine Learning and Artificial Intelligence Used in Digital Behavior Change Interventions? A Scoping Review. *Mayo Clinic Proceedings: Digital Health* 2, 3 (2024), 375–404.

[24] David B Buller, Andrew L Sussman, Cynthia A Thomson, Deanna Kepka, Douglas Taren, Kimberly L Henry, Echo L Warner, Barbara J Walkosz, W Gill Woodall, Kayla Nuss, et al. 2024. # 4Corners4Health Social Media Cancer Prevention Campaign for Emerging Adults: Protocol for a Randomized Stepped-Wedge Trial. *JMIR Research Protocols* 13, 1 (2024), e50392.

[25] Maximilian Nicolaus Burger, Matthias Mayer, and Ivo Steimanis. 2022. Repeated information of benefits reduces COVID-19 vaccination hesitancy: Experimental evidence from Germany. *PLoS One* 17, 6 (2022), e0270666.

[26] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (Glasgow, Scotland Uk). Association for Computing Machinery, New York, NY, USA, 1–15.

[27] Ee Pin Chang, Ullrich Ecker, and Andrew Page. 2018. Continued Influence Effect of Misinformation in Rumination. In *45th Annual Conference of the Australasian Society for Experimental Psychology*. Hobart, Tasmania, 1.

[28] Qiang Chen, Yangyi Zhang, Richard Evans, and Chen Min. 2021. Why do citizens share COVID-19 fact-checks posted by Chinese government social media accounts? The elaboration likelihood model. *International Journal of Environmental Research and Public Health* 18, 19 (2021), 10058.

[29] Robert B Cialdini and Robert B Cialdini. 2007. *Influence: The psychology of persuasion*. Vol. 55. Collins New York.

[30] Robert B Cialdini and N Garde. 1987. *Influence (Vol. 3)*.

[31] Rachel Davis, Rona Campbell, Zoe Hildon, Lorna Hobbs, and Susan Michie. 2015. Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. *Health psychology review* 9, 3 (2015), 323–344.

[32] Jan de Mooij, Parantapa Bhattacharya, Davide Dell'Anna, Mehdi Dastani, Brian Logan, and Samarth Swarup. 2023. A framework for modeling human behavior in large-scale agent-based epidemic simulations. *Simulation* 99, 12 (2023), 1183–1211.

[33] Walter Dempsey, Peng Liao, Pedja Klasnja, Inbal Nahum-Shani, and Susan A Murphy. 2015. Randomised trials for the Fitbit generation. *Significance* 12, 6 (2015), 20–23.

[34] Ying Deng, Alissa N Antle, and Carman Neustaedter. 2014. Tango cards: a card-based design tool for informing the design of tangible learning games. In *Proceedings of the 2014 conference on Designing interactive systems* (Vancouver, BC, Canada). Association for Computing Machinery, New York, NY, USA, 695–704.

[35] Linda Desens, Brandon Walling, Anna Fiedor, Vanessa Howard, Zue Lopez Diaz, Katherine Kim, and Denise Scannell. 2023. A Comparative Case Study Analysis: Applying the HIPE Framework to Combat Harmful Health Information and Drive COVID-19 Vaccine Adoption in Underserved Communities. *Vaccines* 11, 6 (2023), 1107.

[36] Baldeep K Dhaliwal, Rajeev Seth, Betty Thankachen, Yawar Qaiyum, Svea Closser, Tyler Best, and Anita Shet. 2023. Leading from the frontlines: community-oriented approaches for strengthening vaccine delivery and acceptance. In *BMC proceedings*, Vol. 17. Springer, 5.

[37] Kees Dorst and Nigel Cross. 2001. Creativity in the design process: co-evolution of problem–solution. *Design studies* 22, 5 (2001), 425–437.

[38] Steven P Dow, Alana Glassco, Jonathan Kass, Melissa Schwarz, Daniel L Schwartz, and Scott R Klemmer. 2010. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Transactions on Computer-Human Interaction (TOCHI)* 17, 4 (2010), 1–24.

[39] Karl Duncker and Lynne S Lees. 1945. On problem-solving. *Psychological monographs* 58, 5 (1945), i.

[40] Alice H Eagly and Shelly Chaiken. 1993. *The psychology of attitudes*. Harcourt brace Jovanovich college publishers.

[41] Mahdi Ebnali and Cyrus Kian. 2020. Nudge users to healthier decisions: A design approach to encounter misinformation in health forums. In *Advances in Human Factors in Communication of Design: Proceedings of the AHFE 2019 International Conference on Human Factors in Communication of Design, July 24-28, 2019, Washington DC, USA 10*. Springer, Springer International Publishing, Cham, 3–12.

[42] Thomas Erickson and Wendy A Kellogg. 2000. Social translucence: an approach to designing systems that support social processes. *ACM transactions on computer-human interaction (TOCHI)* 7, 1 (2000), 59–83.

[43] Jonathan St BT Evans. 2003. In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences* 7, 10 (2003), 454–459.

[44] Miriam Fernandez and Harith Alani. 2018. Online misinformation: Challenges and future directions. In *Companion Proceedings of the The Web Conference 2018* (Lyon, France). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 595–602.

[45] Patrick Ferrucci and Toby Hopp. 2023. Let's intervene: How platforms can combine media literacy and self-efficacy to fight fake news. *Communication and the Public* 8, 4 (2023), 367–389.

[46] Leon Festinger. 1954. A theory of social comparison processes. *Human relations* 7, 2 (1954), 117–140.

[47] Jeffrey D Fisher and William A Fisher. 1992. Changing AIDS-risk behavior. *Psychological bulletin* 111, 3 (1992), 455.

[48] A Gallego, Eugenio Gaeta, Anni Karinsalo, Ville Ollikainen, Pekka Koskela, Lutz Peschke, Frans Folkvord, Eleni Kaldoudi, Timo Jämsä, Francisco Lupiáñez-Villanueva, et al. 2021. Human computer interaction challenges in designing pandemic trace application for the effective knowledge transfer between science and society inside the quadruple helix collaboration. In *International Conference on Human-Computer Interaction*. Springer, Springer International Publishing, Cham, 390–401.

[49] Claus Ghesla, Manuel Grieder, and Jan Schmitz. 2019. Nudge for good? Choice defaults and spillover effects. *Frontiers in psychology* 10 (2019), 178.

[50] Thomas Gilovich, Victoria Husted Medvec, and Kenneth Savitsky. 2000. The spotlight effect in social judgment: an egocentric bias in estimates of the salience of one's own actions and appearance. *Journal of personality and social psychology* 78, 2 (2000), 211.

[51] Henner Gimpel, Sebastian Heger, Christian Olenberger, and Lena Utz. 2021. The effectiveness of social norms in fighting fake news on social media. *Journal of Management Information Systems* 38, 1 (2021), 196–221.

[52] Selin Gurgun, Emily Arden-Close, John McAlaney, Keith Phalp, and Raian Ali. 2023. Can We Re-design Social Media to Persuade People to Challenge Misinformation? An Exploratory Study. In *International Conference on Persuasive Technology*. Springer, Springer Nature Switzerland, Cham, 123–141.

[53] Selin Gurgun, Deniz Cemiloglu, Emily Arden Close, Keith Phalp, Preslav Nakov, and Raian Ali. 2024. Why do we not stand up to misinformation? Factors influencing the likelihood of challenging misinformation on social media and the role of demographics. *Technology in Society* 76 (2024), 102444.

[54] Najmeh Hafezieh and Farjam Eshraghian. 2017. Affordance theory in social media research: systematic review and synthesis of the literature. In *25th European Conference on Information Systems (ECIS 2017)*. Guimarães, Portugal, 1–12.

[55] Patricia Hanrahan and William J Reid. 1984. Choosing effective interventions. *Social service review* 58, 2 (1984), 244–258.

[56] Katrin Hartwig and Christian Reuter. 2019. TrustyTweet: an indicator-based browser-plugin to assist users in dealing with Fake News on Twitter. In *Proceedings of the international conference on wirtschaftsinformatik*. Siegen, Germany, 1844–1855.

[57] Marc Hassenzahl and Rainer Wessler. 2000. Capturing design space from a user perspective: The repertory grid technique revisited. *International Journal of Human-Computer Interaction* 12, 3-4 (2000), 441–459.

[58] Alan R Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. 2004. Design science in information systems research. *MIS quarterly* 28, 1 (2004), 75–105.

[59] Chris Hollis, Richard Morriss, Jennifer Martin, Sarah Amani, Rebecca Cotton, Mike Denis, and Shon Lewis. 2015. Technological innovations in mental healthcare: harnessing the digital revolution. *The British Journal of Psychiatry* 206, 4 (2015), 263–265.

[60] Karen Holtzblatt and Hugh Beyer. 1997. *Contextual design: defining customer-centered systems*. Elsevier.

[61] Kathryn L Hopkins, Chelsey Lepage, Wendy Cook, Angus Thomson, Surangani Abeyesekera, Stacey Knobler, Nicholas Boehman, Brianna Thompson, Peter Waiswa, Jacquellyn Nambi Ssanyu, et al. 2023. Co-Designing a Mobile-Based Game to Improve Misinformation Resistance and Vaccine Knowledge in Uganda, Kenya, and Rwanda. *Journal of Health Communication* 28, sup2 (2023), 49–60.

[62] Benjamin D Horne, Mauricio Gruppi, and Sibel Adali. 2019. Trustworthy misinformation mitigation with soft information nudging. In *2019 first IEEE international conference on trust, privacy and security in intelligent systems and applications (TPS-ISA)*. IEEE, 245–254.

[63] Cherilyn Ireton and Julie Posetti. 2018. *Journalism, fake news & disinformation: handbook for journalism education and training*. Unesco Publishing.

[64] David G Jansson and Steven M Smith. 1991. Design fixation. *Design studies* 12, 1 (1991), 3–11.

[65] Nancy K Janz and Marshall H Becker. 1984. The health belief model: A decade later. *Health education quarterly* 11, 1 (1984), 1–47.

[66] MJ Johns, Emmanuel Chinedum Ezenwa, Seunghyun Lee, Thomas Maiorana, Ciel Wood, Josh D Levano, Rita Aksum Tesfay, Michael Takami, Cameron A

Dodd, Madison Li, et al. 2024. Participatory Design of a Serious Game to Improve Wildfire Preparedness with Community Residents and Experts. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA). Association for Computing Machinery, New York, NY, USA, 1–8.

[67] Christopher M Jones, Daniel Diethei, Johannes Schöning, Rehana Shrestha, Tina Jahnel, and Benjamin Schüz. 2023. Impact of social reference cues on misinformation sharing on social media: Series of experimental studies. *Journal of Medical Internet Research* 25 (2023), e45583.

[68] Daniel Kahneman. 2011. *Thinking, fast and slow.* macmillan.

[69] Randi Karlsen and Anders Andersen. 2019. Recommendations with a nudge. *Technologies* 7, 2 (2019), 45.

[70] Tanveer Khan, Antonis Michalas, and Adnan Akhunzada. 2021. Fake news outbreak 2021: Can we stop the viral spread? *Journal of Network and Computer Applications* 190 (2021), 103112.

[71] Uzma Khan and Ravi Dhar. 2006. Licensing effect in consumer choice. *Journal of marketing research* 43, 2 (2006), 259–266.

[72] Predrag Klasnja, Sunny Consolvo, and Wanda Pratt. 2011. How to evaluate technologies for health behavior change in HCI research. In *Proceedings of the SIGCHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 3063–3072.

[73] Silvia Knobloch-Westerwick. 2015. The selective exposure self-and affect-management (SESAM) model: Applications in the realms of race, politics, and health. *Communication Research* 42, 7 (2015), 959–985.

[74] Loukas Konstantinou and Evangelos Karapanos. 2023. Nudging for Online Misinformation: a Design Inquiry. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA). Association for Computing Machinery, New York, NY, USA, 69–75.

[75] Loukas Konstantinou, Dionysis Panos, and Evangelos Karapanos. 2024. Exploring the Design of Technology-Mediated Nudges for Online Misinformation. *International Journal of Human–Computer Interaction* (2024), 1–28.

[76] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[77] Jingxian Liao. 2023. Restructuring Unstructured Video Resources for Collaborative Learning and Work. In *Companion Proceedings of the 2023 ACM International Conference on Supporting Group Work* (Hilton Head, SC, USA). Association for Computing Machinery, New York, NY, USA, 60–62.

[78] Nina Mazar and Chen-Bo Zhong. 2010. Do green products make us better people? *Psychological science* 21, 4 (2010), 494–498.

[79] David E Melnikoff and John A Bargh. 2018. The mythical number two. *Trends in cognitive sciences* 22, 4 (2018), 280–293.

[80] Susan Michie and Andrew Prestwich. 2010. Are interventions theory-based? Development of a theory coding scheme. *Health psychology* 29, 1 (2010), 1.

[81] Susan Michie, Michelle Richardson, Marie Johnston, Charles Abraham, Jill Francis, Wendy Hardeman, Martin P Eccles, James Cane, and Caroline E Wood. 2013. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Annals of behavioral medicine* 46, 1 (2013), 81–95.

[82] Susan Michie, Maartje M Van Stralen, and Robert West. 2011. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implementation science* 6 (2011), 1–12.

[83] David R Millen. 2000. Rapid ethnography: time deepening strategies for HCI field research. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques* (New York City, New York, USA). Association for Computing Machinery, New York, NY, USA, 280–286.

[84] John Mills, Mark Lochrie, Tom Metcalfe, and Peter Bennett. 2018. NewsThings: exploring interdisciplinary IoT news media opportunities via user-centred design. In *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction* (Stockholm, Sweden). Association for Computing Machinery, New York, NY, USA, 49–56.

[85] Gregory Mitchell. 2012. Revisiting truth or triviality: The external validity of research in the psychological laboratory. *Perspectives on Psychological Science* 7, 2 (2012), 109–117.

[86] Shaheed N Mohammed. 2019. Conspiracy theories and flat-earth videos on YouTube. *The Journal of Social Media in Society* 8, 2 (2019), 84–102.

[87] Ryan C Moore and Jeffrey T Hancock. 2022. A digital media literacy intervention for older adults improves resilience to fake news. *Scientific reports* 12, 1 (2022), 6008.

[88] Aloysius Gonzaga Mubuuke and Brenda Leibowitz. 2013. Participatory action research: The key to successful implementation of innovations in health professions education. *African Journal of Health Professions Education* 5, 1 (2013), 30–33.

[89] Andre Matthias Müller, Ann Blandford, and Lucy Yardley. 2017. The conceptualization of a Just-In-Time Adaptive Intervention (JITAI) for the reduction of sedentary behavior in older adults. *Mhealth* 3 (2017), 1–12.

[90] Zachary Munn, Micah DJ Peters, Cindy Stern, Catalin Tufanaru, Alexa McArthur, and Edoardo Aromataris. 2018. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC medical research methodology* 18 (2018), 1–7.

[91] Thanh Tâm Nguyên. 2019. *Debunking Misinformation on the Web: Detection, Validation.* Technical Report. and Visualisation. Technical Report. EPFL.

[92] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.

[93] Harri Oinas-Kukkonen and Marja Harjumaa. 2009. Persuasive systems design: Key issues, process model, and system features. *Communications of the association for Information Systems* 24, 1 (2009), 28.

[94] Gábor Orosz, Laura Faragó, Benedek Paskuj, and Péter Krekó. 2024. Strategies to combat misinformation: Enduring effects of a 15-minute online intervention on critical-thinking adolescents. *Computers in Human Behavior* 159 (2024), 108338. https://doi.org/10.1016/j.chb.2024.108338

[95] Gábor Orosz, Benedek Paskuj, Laura Faragó, and Péter Krekó. 2023. A prosocial fake news intervention with durable effects. *Scientific Reports* 13, 1 (2023), 3958.

[96] Nathaniel Persily, Joshua A Tucker, and Joshua Aaron Tucker. 2020. *Social media and democracy: The state of the field, prospects for reform.* Cambridge University Press.

[97] Olga Perski, Ann Blandford, Claire Garnett, David Crane, Robert West, and Susan Michie. 2020. A self-report measure of engagement with digital behavior change interventions (DBCIs): development and psychometric evaluation of the "DBCI Engagement Scale". *Translational behavioral medicine* 10, 1 (2020), 267–277.

[98] Richard E Petty, John T Cacioppo, Richard E Petty, and John T Cacioppo. 1986. *The elaboration likelihood model of persuasion.* Springer.

[99] W James Potter. 2010. The state of media literacy. *Journal of broadcasting & electronic media* 54, 4 (2010), 675–696.

[100] Leigh Powell, Radwa Nour, Youness Zidoun, Sreelekshmi Kaladhara, Hanan Al Suwaidi, Nabil Zary, et al. 2022. A web-based public health intervention for addressing vaccine misinformation: protocol for analyzing learner engagement and impacts on the hesitancy to vaccinate. *JMIR Research Protocols* 11, 5 (2022), e38034.

[101] James O Prochaska and Wayne F Velicer. 1997. The transtheoretical model of health behavior change. *American journal of health promotion* 12, 1 (1997), 38–48.

[102] Katja Reuter, Melissa L Wilson, Meghan Moran, NamQuyen Le, Praveen Angyan, Anuja Majmundar, Elsi M Kaiser, Jennifer B Unger, et al. 2021. General audience engagement with antismoking public health messages across multiple social media sites: comparative analysis. *JMIR Public Health and Surveillance* 7, 2 (2021), e24429.

[103] Ashley Z Ritter, Shoshana Aronowitz, Lindsey Leininger, Malia Jones, Jennifer Beam Dowd, Sandra Albrecht, Alison M Buttenheim, Amanda M Simanek, Lauren Hale, and Aparna Kumar. 2021. Dear pandemic: Nurses as key partners in fighting the COVID-19 infodemic. *Public Health Nursing* 38, 4 (2021), 603–609.

[104] Yvonne Rogers. 2004. New theoretical approaches for HCI. *Annual review of information science and technology* 38, 1 (2004), 87–143.

[105] Jon Roozenbeek and Sander Van der Linden. 2019. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications* 5, 1 (2019), 1–10.

[106] Kai Ruggeri, Friederike Stock, S Alexander Haslam, Valerio Capraro, Paulo Boggio, Naomi Ellemers, Aleksandra Cichocka, Karen M Douglas, David G Rand, Sander Van der Linden, et al. 2024. A synthesis of evidence for policy from behavioural science during COVID-19. *Nature* 625, 7993 (2024), 134–147.

[107] RM Ryan. 2017. *Self-determination theory: Basic psychological needs in motivation, development, and wellness.* Guilford Press.

[108] William Samuelson and Richard Zeckhauser. 1988. Status quo bias in decision making. *Journal of risk and uncertainty* 1 (1988), 7–59.

[109] Elizabeth B-N Sanders and Pieter Jan Stappers. 2008. Co-creation and the new landscapes of design. *Co-design* 4, 1 (2008), 5–18.

[110] Ankit Shrestha, Audrey Flood, Saniat Sohrawardi, Matthew Wright, and Mahdi Nasrullah Al-Ameen. 2024. A First Look into Targeted Clickbait and its Countermeasures: The Power of Storytelling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA). Association for Computing Machinery, New York, NY, USA, 1–23.

[111] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter* 19, 1 (2017), 22–36.

[112] Yvonne Skipper, Daniel Jolley, and Joseph Reddington. 2023. 'But wait, that isn't real': A proof-of-concept study evaluating 'Project Real', a co-created intervention that helps young people to spot fake news online. *British Journal of Developmental Psychology* 41, 4 (2023), 371–384.

[113] Barry M Staw. 1981. The escalation of commitment to a course of action. *Academy of management Review* 6, 4 (1981), 577–587.

[114] Cass R Sunstein. 2017. Nudges that fail. *Behavioural public policy* 1, 1 (2017), 4–25.

[115] Sri Susanty, Made Ary Sarasmita, I Wayan Sudarma, Danur Azissah, Jipri Suyanto, SUR Kamil, Budiman Budiman, and Suharjiman Suharjiman. 2023. Animated video development COVID-19 prevention and management for anxiety among older adults in Indonesia. *Geriatric Nursing* 49 (2023), 13–21.
[116] Kar Yan Tam and Shuk Ying Ho. 2005. Web personalization as a persuasion strategy: An elaboration likelihood model perspective. *Information systems research* 16, 3 (2005), 271–291.
[117] Richard H Thaler and Cass R Sunstein. 2009. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.
[118] Michael E Thompson and Tsovinar L Harutyunyan. 2006. Contraceptive practices in Armenia: panel evaluation of an information-education-communication campaign. *Social science & medicine* 63, 11 (2006), 2770–2783.
[119] Maryam Tohidi, William Buxton, Ronald Baecker, and Abigail Sellen. 2006. Getting the right design and the design right. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (Montréal, Québec, Canada). Association for Computing Machinery, New York, NY, USA, 1243–1252.
[120] Andrea C Tricco, Erin Lillie, Wasifa Zarin, Kelly K O'Brien, Heather Colquhoun, Danielle Levac, David Moher, Micah DJ Peters, Tanya Horsley, Laura Weeks, et al. 2018. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Annals of internal medicine* 169, 7 (2018), 467–473.
[121] Gleb Tsipursky, Fabio Votta, and James A Mulick. 2018. A psychological approach to promoting truth in politics: The Pro-Truth Pledge. *Journal of Social and Political Psychology* 6, 2 (2018), 271–290.
[122] Ziga Turk. 2018. Technology as enabler of fake news and a potential tool to combat it.
[123] Sander Van der Linden, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach. 2017. Inoculating the public against misinformation about climate change. *Global challenges* 1, 2 (2017), 1600008.
[124] Tobias Vogel and Michaela Wanke. 2016. *Attitudes and attitude change*. Psychology Press, London.
[125] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science* 359, 6380 (2018), 1146–1151.
[126] Claire Wardle and Hossein Derakhshan. 2017. *Information disorder: Toward an interdisciplinary framework for research and policymaking*. Vol. 27. Council of Europe Strasbourg.
[127] Eric W Welch, M Jae Moon, and Wilson Wong. 2006. 15 What drives global e-government? An exploratory assessment of existing e-government performance. *Public Service Performance: Perspectives on Measurement and Management* (2006), 275.
[128] Robert West and Susan Michie. 2020. A brief introduction to the COM-B Model of behaviour and the PRIME Theory of motivation [v1].
[129] Hannah S Whitehead, Clare E French, Deborah M Caldwell, Louise Letley, and Sandra Mounier-Jack. 2023. A systematic review of communication interventions for countering vaccine misinformation. *Vaccine* 41, 5 (2023), 1018–1034.
[130] Christiane Wölfel and Timothy Merritt. 2013. Method card design dimensions: A survey of card-based design tools. In *Human-Computer Interaction–INTERACT 2013: 14th IFIP TC 13 International Conference, Cape Town, South Africa, September 2-6, 2013, Proceedings, Part I 14*. Springer, Springer Berlin Heidelberg, Berlin, Heidelberg, 479–486.
[131] Shan Xu, Ioana A Coman, Masahiro Yamamoto, and Christina Jimenez Najera. 2023. Exposure effects or confirmation bias? Examining reciprocal dynamics of misinformation, misperceptions, and attitudes toward COVID-19 vaccines. *Health Communication* 38, 10 (2023), 2210–2220.
[132] Yusuke Yamamoto and Takehiro Yamamoto. 2018. Query priming for promoting critical thinking in web search. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Brunswick, NJ, USA). Association for Computing Machinery, New York, NY, USA, 12–21.
[133] Lucy Yardley, Leanne Morrison, Katherine Bradbury, Ingrid Muller, et al. 2015. The person-based approach to intervention development: application to digital health-related behavior change interventions. *Journal of medical Internet research* 17, 1 (2015), e4055.
[134] Jeon Youngseung, Kim Bogoan, Xion Aiping, and Han Kyungsik. 2021. ChamberBreaker: Mitigating Echo Chamber Effects and Supporting Information Hygiene through a Gamified Inoculation System. *ACM Hum.-Comput. Interact* 5, CSCW2 (2021), 472.
[135] Liudmila Zavolokina, Kilian Sprenkamp, Zoya Katashinskaya, Daniel Gordon Jones, and Gerhard Schwabe. 2024. Think fast, think slow, think critical: designing an automated propaganda detection tool. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA). Association for Computing Machinery, New York, NY, USA, 1–24.
[136] Han Zheng and Rich Ling. 2021. Drivers of social media fatigue: A systematic review. *Telematics and Informatics* 64 (2021), 101696.
[137] Xiaohua Zhu and Shengnan Yang. 2023. Toward a Sociotechnical Framework for Misinformation Policy Analysis. In *The Usage and Impact of ICTs during the Covid-19 Pandemic*. Routledge, 11–45.
[138] Verena Zimmermann and Karen Renaud. 2021. The nudge puzzle: matching nudge interventions to cybersecurity decisions. *ACM Transactions on Computer-Human Interaction (TOCHI)* 28, 1 (2021), 1–45.

A Identified publications

The list of the 37 publications for the scoping review.

(1) Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news [15]
(2) Nudge users to healthier decisions: A design approach to encounter misinformation in health forums [41]
(3) Nudging away false news: Evidence from a social norms experiment [7]
(4) A psychological approach to promoting truth in politics: The Pro-Truth Pledge [121]
(5) 'But wait, that isn't real': A proof-of-concept study evaluating 'Project Real', a co-created intervention that helps young people to spot fake news online [112]
(6) Let's intervene: How platforms can combine media literacy and self-efficacy to fight fake news [45]
(7) The effectiveness of social norms in fighting fake news on social media [51]
(8) TrustyTweet: an indicator-based browser-plugin to assist users in dealing with Fake News on Twitter [56]
(9) Think fast, think slow, think critical: designing an automated propaganda detection tool [135]
(10) Query priming for promoting critical thinking in web search [132]
(11) Infodemic Management for Social and Behavior Change: Youth Mobilization for Combating Disinformation During COVID-19 [4]
(12) ChamberBreaker: Mitigating Echo Chamber Effects and Supporting Information Hygiene through a Gamified Inoculation System [134]
(13) Human computer interaction challenges in designing pandemic trace application for the effective knowledge transfer between science and society inside the quadruple helix collaboration [48]
(14) Restructuring Unstructured Video Resources for Collaborative Learning and Work [77]
(15) Co-Designing a Mobile-Based Game to Improve Misinformation Resistance and Vaccine Knowledge in Uganda, Kenya, and Rwanda [61]
(16) Can We Re-design Social Media to Persuade People to Challenge Misinformation? An Exploratory Study [52]
(17) Persuasive System Design for Climate Change Awareness [2]
(18) Dear pandemic: Nurses as key partners in fighting the COVID-19 infodemic [103]
(19) Animated video development COVID-19 prevention and management for anxiety among older adults in Indonesia [115]
(20) A web-based public health intervention for addressing vaccine misinformation: protocol for analyzing learner engagement and impacts on the hesitancy to vaccinate [100]
(21) #4Corners4Health Social Media Cancer Prevention Campaign for Emerging Adults: Protocol for a Randomized Stepped-Wedge Trial [24]
(22) Strategies to combat misinformation: Enduring effects of a 15-minute online intervention on critical-thinking adolescents [94]

(23) A prosocial fake news intervention with durable effects [95]
(24) Evaluating the impact of short animated videos on COVID-19 vaccine hesitancy: An online randomized controlled trial [17]
(25) The effect of a short, animated story-based video on COVID-19 vaccine hesitancy: A study protocol for an online randomized controlled trial [13]
(26) Contraceptive practices in Armenia: panel evaluation of an information-education-communication campaign [118]
(27) General audience engagement with antismoking public health messages across multiple social media sites: comparative analysis [102]
(28) A First Look into Targeted Clickbait and its Countermeasures: The Power of Storytelling [110]
(29) A Comparative Case Study Analysis: Applying the HIPE Framework to Combat Harmful Health Information and Drive COVID-19 Vaccine Adoption in Underserved Communities [35]
(30) Addressing behavioral barriers to COVID-19 testing with health literacy–sensitive eHealth interventions: results from 2 national surveys and 2 randomized experiments [21]
(31) Promoting informed decision making about maternal pertussis vaccination: the systematic development of an online tailored decision aid and a centering-based group antenatal care intervention [8]
(32) Impact of social reference cues on misinformation sharing on social media: Series of experimental studies [67]
(33) Leading from the frontlines: community-oriented approaches for strengthening vaccine delivery and acceptance [36]
(34) Repeated information of benefits reduces COVID-19 vaccination hesitancy: Experimental evidence from Germany [25]
(35) Nudging for Online Misinformation: a Design Inquiry [74]
(36) Exploring the Design of Technology-Mediated Nudges for Online Misinformation [75]
(37) Combating Fake News Using Implementation Intentions [10]