

Governance of misinformation by public service bots

The potential of automated social networking site accounts operated by public service media for correcting misinformation

STEFANO PEDRAZZI^I & FRANZISKA OEHMER-PEDRAZZI^{II}

^I DEPARTMENT OF COMMUNICATION AND MEDIA RESEARCH, UNIVERSITY OF FRIBOURG, SWITZERLAND

^{II} INSTITUTE FOR MULTIMEDIA PRODUCTION, UNIVERSITY OF APPLIED SCIENCES OF THE GRISONS, SWITZERLAND

ABSTRACT

The rapid dissemination of reliable information from credible sources is of particular relevance in combating misinformation. Such information can be provided by public service media (PSM). This chapter addresses the potential of automated social media accounts operated by PSM for the correction of misinformation. The legitimacy of this governance option is examined through a legal analysis and its effectiveness tested through an experimental survey ($N = 113$). The results of the legal analysis show that a legitimate claim can be derived for the use of automated communication agents by PSM within social networking sites, which can also be used in the fight against the spread of misinformation. The results from the experiment indicate that accounts operated by PSM enjoy higher credibility and generate more interactions than accounts operated by social peers. However, labelling accounts as automated tends to limit the effectiveness of accounts operated by PSM that provide corrective information.

KEYWORDS: correction of misinformation, platform governance, social bots, mixed-methods design

Pedrazzi, S., & Oehmer-Pedrazzi, F. (2023). Governance of misinformation by public service bots: The potential of automated social networking site accounts operated by public service media for correcting misinformation. In M. Puppis, & C. Ali (Eds.), *Public service media's contribution to society: RIPE@2021* (pp. 193–218). Nordicom, University of Gothenburg.
<https://doi.org/10.48335/9789188855756-10>

Introduction

Social networking sites (SNS) allow users to access and exchange information from various sources and participate in discussions (Valenzuela et al., 2019). However, they are also considered to be a place where problematic content such as misinformation and defamation is spread (Vosoughi et al., 2018). Since SNS serve as a primary source of information for a notable number of people, the dissemination, exposure, and use of misinformation through these channels is associated with a wide range of far-reaching societal consequences: They have been linked to propagating conspiracies (Bessi et al., 2015; Wood, 2018), compromising public health (Loomba et al., 2021; Roozenbeek et al., 2020; Sharma et al., 2020), inciting violence (Jolley & Paterson, 2020), affecting trust in mass media and political authorities (Ognyanova et al., 2020; Shin & Thorson, 2017; Vosoughi et al., 2018), and undermining democracy (Clayton et al., 2019; Lewandowsky & Van Der Linden, 2021).

Automated software agents such as social bots can be employed to reinforce the diffusion of mis- and disinformation due to their potentially unlimited communication and networking capabilities. As a consequence, a distorted online discourse can be observed, as studies on the influence of social bots on elections, popular votes, and polarised debates with regard to political and other socially relevant topics such as vaccinations or the emergence and spread of diseases show (Bessi & Ferrara, 2016; Broniatowski et al., 2018; Ferrara, 2017; Howard & Kollanyi, 2016; Swire-Thompson & Lazer, 2020).

Unsurprisingly, policy-makers are increasingly discussing or even implementing measures aimed at countering the spread of mis- and disinformation (by social bots). The debate focuses on an interplay of governance options by regulators, platforms themselves, and users, including measures such as bot disclosure regulations, automated and manual content moderation, and the strengthening of users' media literacy and digital skills (for an overview, see Oehmer & Pedrazzi, 2020).

We argue that the rapid and widespread dissemination of reliable information from credible sources such as public service media (PSM) is of particular relevance in combating misinformation (Hjarvard, 2018; Horowitz et al., 2021). Using a combination of legal analysis and experimental survey, in this chapter we aim to examine the legality, suitability, and effects of corrective information to misinformation provided to users of SNS by means of automated accounts operated by PSM. In order to investigate this, we take a three-step approach, in which the first two research questions serve as the premises to test the third:

- RQ1. Can the use of SNS accounts operated by a public service media organisation for the purpose of correcting misinformation in SNS be considered legal?

- RQ2. Based on the credibility perception of SNS users, can accounts operated by a public service media organisation be considered suitable for correcting misinformation in SNS?
- RQ3. What are the effects of labelling an SNS account (operated by a public service media organisation) as automated that provides corrective information on misinformation on 1) users' assessment of its credibility; 2) users' intention to like, share, or use the corrective information provided; and 3) opinion?

We first discuss problems that arise from the spread of mis- and disinformation, followed by a discussion of available governance options with a focus on corrective interventions. Subsequently, we address the special role of PSM and their suitability as originators of corrective information on SNS from a legal standpoint, using the case of Switzerland. Our experiment was theoretically based on the Elaboration Likelihood Model (Petty & Cacioppo, 1986), an information-processing model allowing for assumptions related to attitude change by persuasive messages depending on several factors, such as the source of a message and the individual relevance of its content. Participants were exposed to a corrective message in response to misinformation posted on SNS. The account from which the corrective message originated was presented as affiliated with the PSM or not, and marked as an automated account or not. In addition to the source's credibility, the effects of the corrective intervention on users' intentions to interact with it (e.g., like, share) and on opinion were examined.

This chapter contributes to the research area in three respects: first, to the hitherto underresearched area of the effectiveness of different sources as originators of corrective information; second, to the effects of labelling SNS accounts as automated, as required by bot disclosure regulations; and third, to the legitimacy and suitability of automated accounts operated by PSM in combating the spread of misinformation in SNS.

Challenges posed by misinformation on social networking sites

Various terms such as fake news, misinformation, disinformation, propaganda, and so on, are often used inconsistently to describe different forms of false and misleading information (Tandoc et al., 2018; Vraga & Bode, 2020). Thorson (2016) understood misinformation to be information that is initially presented as accurate but is later shown to be false, while Vraga and Bode (2020) have emphasised the contradiction of misinformation with the best expert evidence available at the time of its publication. Wardle and Derakhshan (2017) defined misinformation as false information that is not created for the purpose of causing harm; disinformation as content that is

created with an inherent harmful intent; and malinformation as content that is not false but is intended to cause harm. And Zimmermann and Kohring (2018) defined disinformation as the “communication of knowingly and empirically false information about new and relevant facts with a claim to truth [translated]”. The differentiation between mis- and disinformation often takes place at the level of the originator of the information, who is assumed to have malicious intent or at least to act knowingly. Whether a deceptive or malicious intent can be assumed from the (often unknown) originator of content is not clear-cut in most cases. Accordingly, the term disinformation lacks discriminatory power for empirical investigation, which is why in this chapter we use the term misinformation when referring to false or deceptive content, regardless of the intention behind it.

The spread of misinformation poses several challenges to individuals, institutions, and society. Accordingly, measures are being discussed and implemented at various levels (state, organisations, users, technology) to curb the spread of misinformation and its negative effects (Saurwein & Spencer-Smith, 2020). To this end, new legislation has been introduced or is being discussed in various countries, which orders the removal of harmful or illegal content and the prevention of its further dissemination or requires the labelling of automatically operated accounts by SNS operators. However, misinformation is not illegal in itself, and determining the legality and distinguishing between content that is actually punishable or merely undesirable often cannot be done beyond doubt, and it is not uniformly regulated in international contexts (Oehmer & Pedrazzi, 2020). SNS operators are also making adjustments to usage policies and socio-technical arrangements, such as the network architecture; verification of a profile’s authenticity; design options that allow users to report illegal or misleading content; the way in which engagement metrics are displayed; the limitation of high-speed liking or sharing; algorithmic recommendations; content moderation; and flagging, refuting, or correcting controversial content through references, fact-checks, or links to verified journalistic sources (Avram et al., 2020; Clayton et al., 2019; Gillespie, 2014, 2018b; Grinberg et al., 2019; Hameleers et al., 2020; Hameleers & Van Der Meer, 2020; Lazer et al., 2018; Li, 2020; Napoli, 2015; Pennycook & Rand, 2021; Shao et al., 2018; Thorson, 2016). However, SNS operators are neither suited for nor able to combat misinformation (spread by social bots) on their own (Gillespie, 2018a). These forms of governance are not only prone to error, but they also privatise decisions with respect to freedom of expression (Helberger et al., 2018; Langvardt, 2018; Pennycook & Rand, 2021). Moreover, the business model of platforms is not conducive to effective self-regulation (Puppis, 2020; van Dijck et al., 2018). Further, SNS users can also play an important role in preventing the spread of misinformation. However, the prerequisite for this is that users recognise misinformation and behave in such a way that they do not support the spread of such content (Pedrazzi & Oehmer, 2021).

The speed at which misinformation spreads on SNS – and the often inconclusive and time- and resource-intensive processes for assessing its factuality and for rebutting it – mean that it is often impossible to prevent an initial wave of misinformation from spreading, possibly additionally supported by social bots. Once spread, false information and beliefs can persist among users despite correction (Lewandowsky et al., 2012) and continue to exert an influence in the form of so-called “belief echoes” (Thorson, 2016). We therefore argue in this chapter that the tradition and expertise of legacy media, especially PSM, in providing independent, verified, and trustworthy corrective information could thus be of great value in dealing with misinformation spread via SNS.

The correction of misinformation

One of the most discussed instruments for combating misinformation is the corrective intervention, for example, of fact-checking organisations or news media. Studies on when and how corrective content has an impact or success show mixed results (Carey et al., 2020; Chan et al., 2017; Clayton et al., 2019; Lewandowsky et al., 2012; Mena, 2020; Pennycook et al., 2020; Walter et al., 2020b; Walter & Murphy, 2018; Walter & Tukachinsky, 2019): Although corrective interventions show an overall positive influence in various (meta-)analyses, unintended side effects have also been observed, such as that they can adversely affect the credibility attribution of other (correct) content or implicitly validate misinformation that is not labelled; moreover, a continued influence effect of misinformation has also been demonstrated.

Researchers have examined how the type and content of a corrective intervention – which can range from simple warnings that challenge the correctness of a piece to detail-enriched rebuttals – are related to its effectiveness. For example, detail-enriched, coherent, fact-based representations that offer an alternative narrative were found to be more likely to be successful than simple rebuttals (Bode & Vraga, 2018; Chan et al., 2017; Lewandowsky et al., 2012; Walter & Tukachinsky, 2019). Studies of misinformation related to health topics, however, have not found significant differences between these types of corrective messages, nor have differences been found when correction is purely textual or in combination with images (Walter et al., 2020a).

The topic on which misinformation is circulated is also related to the potential for corrective intervention. It was found that misinformation related to crime or health topics can be corrected more easily than misinformation related to political topics, which is attributed to the fact that preconceptions and political identities or ideologies are more strongly anchored (Walter & Murphy, 2018). The duration of time until users are confronted with warnings or corrective content is also considered a critical situational factor for correction. On the one hand, misinformation can be better remembered with

longer duration, and on the other, there is also the possibility of complex contagion processes in SNS (Mønsted et al., 2017). That is, if users (also supported by social bot activities) encounter the same misinformation several times, its reception is further promoted, a phenomenon also referred to as “illusory truth effect” (Hassan & Barber, 2021). No significant differences were found between studies analysing corrective interventions on Facebook or Twitter (Walter et al., 2020a).

A key element for the success of a corrective intervention lies in the receptivity to it. Lewandowsky and colleagues (2012) have identified cognitive parameters that, in a neutral scenario (i.e., when individuals have no intrinsic motivation to resist correction), may explain the persistence of false beliefs following exposure to misinformation, despite correction having occurred:

- The corrective intervention may create an uncomfortable gap in mental models about events or topics that is most easily remedied by ignoring the retraction.
- Memory may fail and there may be confusion at recall about which information or source was actually false.
- A sense of familiarity with the misinformation and possibly fluency in processing may also cause it to persist.
- Corrective interventions may trigger reactance that leads to rejection of the correction.

In addition, ideologically or politically motivated reasoning can lead to identity-protective actions, such that information that is consistent with one’s worldview is overly believed, with related research providing evidence that individuals are less accurate in judging the veracity of content (Pennycook & Rand, 2021). In general, people who are more reflective or have higher political knowledge or media literacy are better at discerning false from true news, irrespective of whether the information is consistent with their worldview (Pennycook & Rand, 2021).

The role of (automated) sources in correcting misinformation

Another element in the correction of misinformation that has been investigated relatively rarely is the source of the corrective intervention. The credibility and expertise of a source of information is considered essential for the persuasiveness of a message (Eagly & Chaiken, 1993; Petty & Briñol, 2010; Petty & Cacioppo, 1984): The more credible a source, the more likely people are to be persuaded by the arguments. This is also true for recipients of a message who lack the motivation, ability, or expertise to carefully examine or process the content of the message, who then tend to transfer the credibility

of the source to the content of the message. This heuristic-based form of information-processing was found to be the dominant form in online environments (Metzger et al., 2010), thus including exposure to misinformation and corrective information in SNS. A few studies have investigated which sources or source characteristics are conducive to corrective intervention: In particular, in the area of health, it has been shown that official authorities or experts can correct misinformation more effectively than unknown users in SNS (Vraga & Bode, 2017; Walter et al., 2020a; Wang, 2021); that adding a link to an external, credible source increases the effectiveness of the intervention (Vraga & Bode, 2018); that corrections from users with whom one has a strong social connection are more effective (Margolin et al., 2018); while social endorsements (e.g., number of likes) showed no influence on the effectiveness of a source of a corrective intervention (Bode et al., 2021), but on that of a corrective message (Wang, 2021). Of particular interest to this study is the finding that news organisations can be an effective source of corrective messages (Van Der Meer & Jin, 2020), with the reputation of a news organisation being a determining factor in whether users rely on it as a source of information (Winter & Krämer, 2014). In addition to the potential suitability, the legality to engage in automated communication activities by PSM serves as a premise for being a corrective source to misinformation. Our first two research questions follow from these arguments.

Still largely unexplored is the effectiveness of corrections provided by automated accounts in SNS. Bode and Vraga (2018) have examined the effectiveness of corrective information suggested by Facebook in an automated way, and their results show that it does not differ in effectiveness from corrections by social peers. A key factor in the effectiveness of the correction, even if it comes from an automated account, is its perception and assessment by users. For bots, studies in the tradition of the Computers Are Social Actors approach (Reeves & Nass, 1996), which investigates interactions with computers or other automated agents, show that individuals unconsciously apply the same social rules and heuristics as when interacting with human users. For example, Edwards, Edwards, Spence, and Shelton (2014) showed in an experiment that bots are perceived as credible, attractive, and competent in communication, and that there was an equally strong intention to interact with a bot or a human user on Twitter. Users also exhibit similar levels of effort and motivation in information-seeking and cognitive processing when receiving information from a bot or a human user (Edwards et al., 2016). As yet, the consequences of labelling an SNS account as automated, as required by bot disclosure regulation, are unknown. It is quite conceivable that labelling may lead to more scrutiny of the content by the user, if the motivation and capacity are available (Petty & Cacioppo, 1986).

Little is also known about whether and how the source of a corrective message can influence the behaviour of users, more precisely in terms of the intention to like or share the corrective message in SNS and to seek further

information about it. Research on this has shown that the reasons for sharing true (but also false) information in SNS can vary widely; these can include reputational considerations, signalling group affiliation, or political motives, among others (Chadwick & Vaccari, 2019; Pennycook et al., 2020). The perceived accuracy of a message (Pennycook et al., 2020) and the identification of a source as true (Bauer & Clemm Von Hohenberg, 2020) tend to lead to a greater intention to share a post on SNS. The effect of warning labels, on the other hand, is ambiguous, as they were able to reduce the intention to share misinformation in the studies of Pennycook and colleagues (2020) and Mena (2020), whereas they failed to show any effect in the study of Oeldorf-Hirsch and colleagues (2020) in this respect, and also with regard to seeking further information. Our third research questions results from the fact that no clear predictions can be made about how users perceive and interact with automated SNS accounts of PSM that provide corrective information about misinformation, and labelled as such.

Legal analysis: Purpose and legal basis of the (online) activities of the Swiss public service media

To investigate whether and to what extent the use of automated communication services by the Swiss PSM as a governance tool to correct false information in SNS can be considered legal (RQ1), we conducted an analysis of all legal documents relevant to the Swiss public service remit. “Public service” refers to services in the infrastructure sector which, due to their particular social relevance, cannot be entrusted to the “free play” of the market and are thus regulated by law. The aim is to guarantee the population permanent access to the relevant goods and services at an affordable price. Due to their special importance for public opinion-forming processes, the media have also been granted the status of public service since 1984, according to the Swiss Federal Constitution (Swiss Confederation, 1874, 1999; Saxer & Brunner, 2018). This is to ensure that the population can be provided with diverse and unbiased media content. However, due to the technical circumstances prevailing at the time of the drafting of the constitutional article (e.g., scarcity of the frequencies), only broadcasting – namely radio and television – was addressed (Dumermuth, 2016; Saxer, 2017; Saxer & Brunner, 2018): There was no need for regulation of the press due to its comparatively unlimited availability and thus also the expected diversity of content. The Internet and social media and their importance as a source of information could not yet be foreseen in the 1980s. Accordingly, the Swiss Federal Constitution (1999: Art. 93, para. 2 BV) commits only broadcasters to “contribute to education and cultural development, to the free shaping of opinion and to entertainment”. Furthermore they “shall present events accurately and allow a diversity of opinions to be expressed appropriately”.

The Swiss PSM is SRG SSR. In the Federal Act on Radio and Television (RTVA), its programme mandate is specified as follows: “The SRG SSR contributes to free public opinion-forming through comprehensive, diverse and accurate information especially regarding political, economic and social matter” (Swiss Confederation, 2006: Art. 24, para. 4). Despite the restriction to broadcasting in the constitution, the public service of the SRG also covers online activities according to its charter. However, these activities must have a reference to a programme or, in the case of news, sports reports, or local information not in the Romansh language, a limited number of characters in order to not compete with private media content providers (Swiss Confederation, 2020).

Looking at these regulations, we can certainly derive a claim to a public service in the online sector, and thus to online activities of the SRG SSR in Switzerland. This is for the following reasons:

1. The purpose of Article 93 of the Swiss Federal Constitution is to provide the population with access to undistorted and accurate media content that serves as the basis for unbiased opinion-forming processes. Radio and television were identified as particularly influential media types during the drafting of the constitutional article and were accordingly named in the constitution as public service providers. However, in view of media change and shifting patterns of media use, a restriction to these types of media must be regarded as too narrow and thus as not purposeful (Dumermuth, 2016). In modern democracies, broadcasting is no longer the only important source of information. Online media are increasingly being used, particularly by young people (Newman et al., 2019; Schwaiger, 2020). The main function of the public service in the online sector would then be to provide trustworthy and correct information and thus ensure an undistorted opinion-forming process. This broad reading of Article 93, paragraph 2 is also supported if one considers the reasoning of the paragraph immediately preceding it that grants the Confederation regulatory competence not only for radio and television, but also for “other forms of public broadcasting of features and information” (Swiss Confederation, 1999: Art. 93, para. 1 BV).
2. The more specific norms for the regulation of the public service (RTVA, SRG charter) already explicitly approve – notwithstanding the provisions of the constitution – an engagement on SRG’s online platforms.

However, to be able to assert this claim to online activities of the public service, two conditions must be met: The communication must be public and relevant, and thus also suitable for influencing public opinion- and will-forming processes (Dumermuth, 2016). If these conditions are met, then not only

can the online engagement of the public service be justified, but an activity on social media can also be considered necessary and meaningful. After all, SNS such as Twitter, YouTube, and Instagram are also increasingly advancing as important sources of information. At the same time, however, they are also the place where misinformation is spread, often also via automated social bots or semi-automated accounts, such as trolls. The participation of PSM as a reliable provider of factually correct information about political, economic, and social issues – also in an automated way – on these online platforms thus seems legitimate and necessary.

Experimental method and design

After the legal analysis revealed the legitimacy of automated SNS accounts operated by the Swiss PSM SRG SSR, we investigated, with an experimental online survey, the potential of PSM to combat misinformation by promoting independent and trustworthy content in a timely manner through automated agents. A 2 x 2 between-subjects factorial design was used (i.e., account labelled as automated: yes vs. no; and account owner: Swiss PSM vs. unknown peer). Participants were randomly assigned to the experimental conditions.

Sample

A total of 113 persons fully completed the experimental online questionnaire. Participants were recruited in courses we taught at our respective universities as well as in our personal environment. The mean age of the participants ranged from 17 to 60 years old ($M = 29.70$, $SD = 9.10$). Approximately 58.6 per cent of the participants were female and almost 33.6 per cent had a university degree. Controls were in place to ensure that participants either resided in Switzerland or were familiar with the Swiss PSM from their daily lives, so as not to compromise the account holder manipulation. Data collection took place 29 March–8 April 2021.

Procedure and stimuli

After the participants had given their consent to participate in the study, they were asked some questions about their use of media and SNS. This was followed by an instruction that a conversation on Twitter would follow, in which a labelling of automated accounts by Twitter would be tested by means of a symbol depicted in the instruction. To do this, we created a symbol specifically for this study that mimicked the Twitter symbol for official accounts in both design and positioning, but instead of the checkmark, showed a robot symbol inside the blue ring. Participants were then asked to look at the depicted Twitter feed in a way that corresponds to their normal way of using SNS.

Subsequently, the participants were shown the Twitter feed in the form of a screenshot. This consisted of a post, either from the Swiss National Bank, which informed about a successful test with a digital currency, or from the Swiss Federal Supreme Court, which announced a ruling to increase the maximum permissible return on rental properties. Both events were real. This initial post was followed by five comments. The first comment was from an author marked as automated, who advertised Bitcoin as an investment opportunity immediately after the post was published. This served the purpose of ensuring that all participants also encountered the symbol for labelling automated accounts. This was followed by two comments from accounts not marked as automated on the content of the tweet according to the different experimental conditions, including a reference to the fact that the Canadian National Bank had conducted a similar test with digital currencies and that the Highest Court in Australia had made a similar ruling. Then came a comment from another, unknown account that was not marked as automated, which contained the misinformation. In the tweet about the National Bank's test, the comment said that the goal of the test was to create the transparent, controlled citizen, and in the tweet on the Federal Court ruling, that rent increases would only be very small, with a false fictitious calculation example.

The fifth comment was the corrective intervention, which rejected the preceding comment as incorrect and included a link to an online article from the Swiss PSM that could be consulted for verification. In the various conditions, the originator of this comment was manipulated so that it was either posted by an unknown user (marked as automated in one condition), by the official account of the Swiss PSM (account name: SRF News), or by an account of the Swiss PSM marked as automated (account name: SRF Newsbot). In total, there were four different authors of an identical corrective message. All account names and account images of the previous commenters, temporal aspects, engagement metrics, and content were kept constant across conditions.

After the participants had viewed the Twitter feed, they were asked questions about the correcting post, namely whether they would interact with the content of the post, about the credibility of the author of the correcting post, and about their opinion, prior knowledge, and relevance assessment of the topic of the Twitter feed. This was followed by questions on digital media skills and news elaboration before they gave an assessment of the journalistic services of the Swiss PSM. The questionnaire ended with questions on sociodemographics and a debriefing regarding the test of labelling accounts and the manipulation of certain content.

Measures

- Credibility of the source of the corrective post: Participants rated the credibility of the source of the corrective post with a single question using a 7-point Likert scale ($M = 3.81$, $SD = 1.83$).
- User engagement intention with the corrective post: Participants could indicate their intention to interact with the post or its content using yes/no. 15 per cent of participants responded that they would like the post, 1 per cent would share the post, and 2 per cent would post a comment themselves. 50 per cent of participants said they would click on the article linked in the post and 14 per cent would recommend the article to another person on another channel (e.g., via messenger service).
- Opinion on the topic of the Twitter feed: Participants expressed their opinion on the topic of the Twitter feed using a 10-point Likert scale to the following questions: “Are you in favour of or against the replacement of the Swiss franc by a digital currency by the Swiss National Bank?” ($M = 3.83$, $SD = 2.00$); and “Are you in favour of or against the change in the calculation of the return on rental properties introduced by the Federal Court ruling?” ($M = 4.87$, $SD = 1.76$).
- Prior knowledge on the topic of the Twitter feed: Using a 7-point Likert scale, participants expressed their assessment of their prior knowledge related to the topic of the Twitter feed: digital currencies ($M = 2.39$, $SD = 1.34$); and maximum allowable rental return ($M = 2.26$, $SD = 1.29$).
- Issue relevance: Issue relevance was assessed using three items combined into an index on a 7-point Likert scale that asked about relevance to oneself, to one’s personal environment, and to society (Kaiser et al., 2018) ($M = 4.22$, $SD = 1.38$, Cronbach’s $\alpha = .79$).
- Digital media skills: Using a 7-point Likert scale, participants answered questions about their information-seeking and privacy behaviour when browsing the Internet and SNS deriving from the Information Literacy (4 items) and Privacy Literacy (3 items) scales (Hoffmann et al., 2019). Due to insufficient scale reliability coefficients (both Cronbach’s $\alpha < .45$), digital media skills were excluded from the analysis.
- News elaboration: This variable measures the intensity with which individuals cognitively engage with news by two items (Shahin et al., 2020) using a 7-point Likert scale (e.g., “I often find myself thinking about what I’ve encountered in the news”), which were combined into an index ($M = 5.00$, $SD = 1.08$, $r = .59$).

- Trust in the journalistic services of the Swiss PSM: Trust in the journalistic services of the Swiss PSM was operationalised with the help of five items based on the dimensions of the “Trust in News Media” scale (Kohring & Matthes, 2007). Using a 7-point Likert scale, participants could indicate, for example, whether SRG SSR reports in detail and correctly, covers the important topics and events, or pays attention to balance in its reporting ($M = 5.70$, $SD = .75$, Cronbach’s $\alpha = .90$).

Check of internal validity

In order to verify that the different experimental conditions did not produce unexpected effects, we conducted various checks. No differences resulted in the assessment of one’s own news elaboration or the journalistic offering of Swiss PSM depending on the topic, the labelling of the originator as automated, or the institutional affiliation to the Swiss PSM of the account. Only with regard to the issue relevance assessment did a difference result as a function of the topic: “ruling to adjust the return on rental properties” ($M = 4.98$, $SD = 1.21$) was perceived as more relevant than “introduction of a digital currency” ($M = 3.54$, $SD = 1.16$; $t(111) = -6.44$, $p < .001$). In contrast, the overall low level of prior knowledge did not differ between the two topics.

Results

Credibility perception of source providing correction

To test the suitability of PSM as providers of corrections to misinformation in SNS (RQ2), we conducted a t-test independent samples for source credibility. An account that could be assigned to the Swiss PSM ($M = 4.75$, $SD = 1.72$) was perceived as significantly more credible ($t(111) = 6.23$, $p < .001$) than an account of an unknown user ($M = 2.89$, $SD = 1.44$). The mean credibility of an account that could be attributed to the Swiss PSM was also significantly higher than the scale mean ($t(55) = 3.27$, $p = .002$), overall indicating a suitability of PSM for providing corrective interventions to misinformation in SNS.

Effect of labelling on source credibility

To investigate whether and what effect labelling as automated exerts on the credibility of an SNS account that provides corrective information about misinformation (RQ3.1), we conducted a factorial between-subjects ANCOVA with the perceived credibility of the source of the corrective post as the dependent variable, the experimental factors and their interaction as independent variables, and prior knowledge, issue relevance, news elaboration, and

trust in Swiss PSM’s journalistic services as covariates. The model proved significant and explained 46 per cent of the variance ($F(7, 105) = 12.65$, $p < .001$, $R^2 = .46$, $R^2_{\text{adjusted}} = .42$).

As shown in Table 10.1, both main effects – labelling the account as automated and account affiliation with the Swiss PSM – proved to be significant, such that labelling negatively affected perceptions of source credibility, whereas accounts identified as belonging to the Swiss PSM were perceived as more credible than accounts from unknown social peers. Moreover, the interaction of labelling as automated and affiliation with the Swiss PSM also proved significant. This suggests that the effect of labelling on perceived credibility differs depending on whether the account belongs to an unknown social peer or to the Swiss PSM: The official account from the Swiss PSM was rated most credible ($M = 5.68$, $SD = 1.33$), followed by the account from the Swiss PSM labelled as automated ($M = 3.82$, $SD = 1.56$), the account from the unknown social peer ($M = 3.14$, $SD = 1.71$), and finally the unknown social peer labelled as automated ($M = 2.66$, $SD = 1.08$). A post hoc test with Bonferroni correction shows that the unlabelled account of the Swiss PSM was perceived as significantly more credible than all other accounts; moreover, the Swiss PSM account labelled as automated was also perceived as more credible than the account of the unknown user labelled as automated. In addition, we found a significant relationship between prior knowledge as well as between news elaboration and perceived source credibility, such that participants who had greater prior knowledge tended to rate the credibility of the source of the corrective post more positively, whereas participants who were more cognitively engaged with news tended to rate its credibility more critically. Somewhat surprisingly, there was no link between trust in the Swiss PSM’s journalistic services and the perception of source credibility, although in all conditions, reference was made to an external article by the Swiss PSM.

Table 10.1 Analysis of covariance in the perceived credibility of the source of the corrective post

Source of variation		Type III SS	F(1,105)	p	η_p^2
Source Bot		42.39	21.87	< .001	.17
Source PSM		81.75	42.18	< .001	.29
Source Bot × Source PSM		16.91	8.73	.004	.08
Covariates	Prior knowledge	8.38	4.33	.04	.04
	Issue relevance	0.80	0.41	.52	.00
	News elaboration	13.32	6.87	.01	.06
	Trust in PSM	0.01	0.01	.94	.00

Comments: $N = 113$; $R^2 = .46$ ($R^2_{\text{adjusted}} = .42$)

Effect of labelling on intention to engage with the corrective information

To study the effect of labelling an SNS account that provides corrective information about misinformation as automated on intended interactions by users (RQ3.2), we again performed a factorial between-subjects ANCOVA. To this end, we consider the intention for social endorsements and for information-seeking activities first individually and then in the aggregate, since in SNS interactions are not limited to one and interactions can affect algorithmic selection processes. Again, the experimental factors and their interaction acted as independent variables and prior knowledge, issue relevance, news elaboration, and trust in the journalistic services offered by the Swiss PSM acted as covariates.

Since only single participants expressed the intention to share or comment on the correcting post, a further analysis of this form of interaction was omitted. As for the intention to like, the model proved significant and explained a total of 14 per cent of the variance ($F(7, 105) = 2.35, p = .03, R^2 = .14, R^2_{\text{adjusted}} = .08$). As shown in Table 10.2, the main effect of the account's affiliation with the Swiss PSM proved significant: There was a greater intention to like corrective interventions by accounts from the Swiss PSM than by accounts from unknown social peers. In contrast, there was no significant main effect associated with labelling the account as automated, meaning that the intention to like a corrective post didn't vary between sources marked as automated or not. The interaction of being labelled as automated and belonging to the Swiss PSM was significant: A post hoc test with Bonferroni correction reveals that the intention to like was significantly higher for the corrective intervention by the official account of the Swiss PSM than for all other accounts. No significant difference resulted for the other comparisons. No statistical correlation was found between the covariates examined in the model and the intention to like the post.

Table 10.2 Analysis of covariance in the intention to like the corrective post

Source of variation		Type III SS	F(1,105)	p	η_p^2
Source Bot		0.24	1.99	.16	.02
Source PSM		0.75	6.31	.01	.06
Source Bot × Source PSM		0.78	6.52	.01	.06
Covariates	Prior knowledge	0.01	0.09	.76	.00
	Issue relevance	0.04	0.37	.54	.00
	News elaboration	0.15	1.30	.26	.01
	Trust in PSM	0.06	0.52	.47	.01

Comments: $N = 113; R^2 = .14 (R^2_{\text{adjusted}} = .08)$

As for the intention of information-seeking, the model proved significant, explaining a total of 13 per cent of the variance ($F(7, 105) = 2.26, p = .035, R^2 = .13, R^2_{\text{adjusted}} = .07$). Only the main effect of account affiliation with the Swiss PSM ($F(1, 105) = 8.24, p = .005, \eta_p^2 = .07$) proved significant, such that the intention to obtain more detailed information on the topic by clicking on the linked article was greater for corrective interventions by accounts from the Swiss PSM than for accounts from unknown social peers (64% vs. 37%). No significant main effect was found in connection with the labelling of the account as automated, meaning that the intention to search information didn't vary between sources marked as automated or not. The interaction of being labelled as automated and affiliation with Swiss PSM also proved to be non-significant. We did not find any statistical association between the covariates examined in the model and the intention to search for information.

Finally, we tested the effect on participants' overall intention to interact with the corrective information. The model proved significant, explaining a total of 16 per cent of the variance ($F(7, 105) = 2.77, p = .011, R^2 = .16, R^2_{\text{adjusted}} = .10$). As shown in Table 10.3, the main effect of account affiliation with the Swiss PSM proved significant. There was a greater intention to engage with the corrective information provided by the accounts affiliated with the Swiss PSM compared with the accounts from unknown social peers. In contrast, there was no significant main effect associated with labelling the account as automated. The interaction of labelling as automated and affiliation with the Swiss PSM proved marginally significant. Labelling accounts as automated tended to equalise the intention to engage in a user interaction with corrective information provided, whereas for accounts not labelled as automated, affiliation with Swiss PSM ($M = 1.36, SD = 1.10$) led to significantly more user interactions than did the account of an unknown social peer ($M = .39, SD = .50$). No statistical relationship was found between the covariates examined in the model and the intention to interact.

Table 10.3 Analysis of covariance in the intention to interact with the corrective post

Source of variation		Type III SS	$F(1,105)$	p	η_p^2
Source Bot		0.25	0.28	.60	.00
Source PSM		9.78	11.21	.001	.10
Source Bot $\times$ Source PSM		3.33	3.82	.053	.04
Covariates	Prior knowledge	1.29	1.48	.23	.01
	Issue relevance	0.09	0.10	.75	.00
	News elaboration	0.52	0.60	.44	.01
	Trust in PSM	1.40	1.60	.21	.02

Comments: $N = 113; R^2 = .16 (R^2_{\text{adjusted}} = .10)$

Effect of labelling on opinion

To test the effect of labelling on opinion (RQ3.3), the mean values of the opinions on the two topics were recoded so that positive values represented a deviation in the sense of the corrective intervention. To do this, first the mean on the topic of the federal court ruling was inverted, then both means were mean centred and merged into one variable. Subsequently, a factorial between-subjects ANCOVA was performed again, with opinion as the dependent variable, the experimental factors and their interaction as independent variables, and prior knowledge, issue relevance, news elaboration, and trust in the journalistic services of the Swiss PSM as covariates. The model proved significant and explained 18 per cent of the variance ($F(7, 105) = 3.20, p = .004, R^2 = .18, R^2_{\text{adjusted}} = .12$). Both main effects and the interaction proved non-significant, only prior knowledge was positively related to opinion ($F(1, 105) = 8.51, p = .004, \eta_p^2 = .08$).

To further elaborate on the role of trust in the journalistic services offered by the Swiss PSM on opinion, we additionally performed a moderation analysis using PROCESS (Model 3), in which the mean centred trust in the Swiss PSM was included as an additional moderator in the model. The model proved significant ($F(10, 102) = 3.59, p = .001, R^2 = .26, R^2_{\text{adjusted}} = .19$). As shown in Table 10.4, prior knowledge remained significant. Additionally, trust in Swiss PSM's journalistic offerings also proved significant, as did the interactions between accounts labelled as automated and trust in Swiss PSM, the interaction between account affiliation with Swiss PSM and trust in Swiss PSM, and the interaction between an account labelled as automated, account affiliation with Swiss PSM, and trust in Swiss PSM. The Johnson-Neyman significance value for the mean centred trust in the Swiss PSM is .24.

Table 10.4 Moderation analysis of effects on opinion

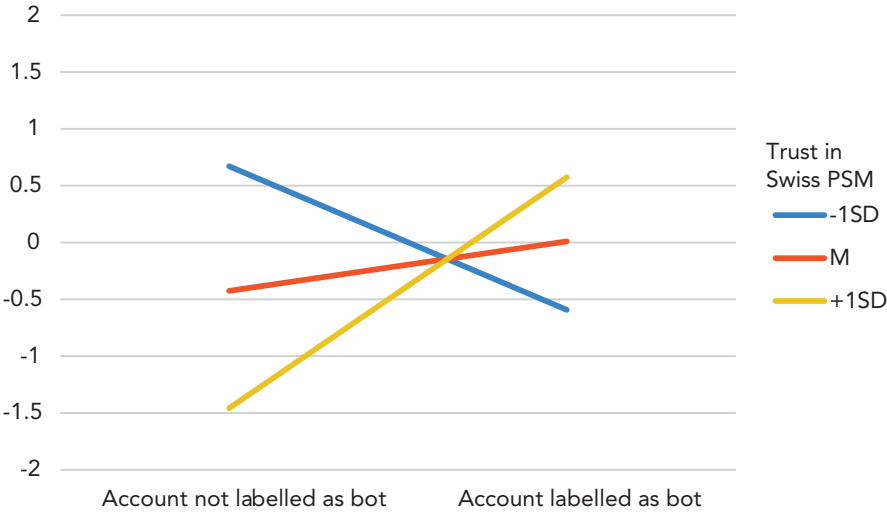
Effect	Estimate	SE	95% CI		p
			LL	UL	
Intercept	-0.28	0.33	-0.93	0.36	.39
Source Bot ^a	0.21	0.46	-0.69	1.12	.64
Source PSM ^a	0.71	0.47	-0.21	1.64	.13
Trust in PSM ^b	-1.37	0.55	-2.46	-0.28	.01
Source Bot × Source PSM	-0.77	0.68	-2.11	0.57	.26
Source Bot × Trust in PSM	2.12	0.65	0.83	3.42	.002
Source PSM × Trust in PSM	2	0.69	0.623	3.36	.005
Source Bot × Source PSM × Trust in PSM	-2.71	0.94	-4.58	-0.83	.005
Prior knowledge ^b	0.43	0.13	0.18	0.68	.001
Issue relevance ^b	0.13	0.12	-0.11	0.38	.28
News elaboration ^b	0.27	0.16	-0.05	0.59	.10

Comments: $N = 113$; $R^2 = .26, R^2_{\text{adjusted}} = .19$; CI = confidence interval; LL = lower limit; UL = upper limit.

^a0 = no, 1 = yes; ^b mean centred

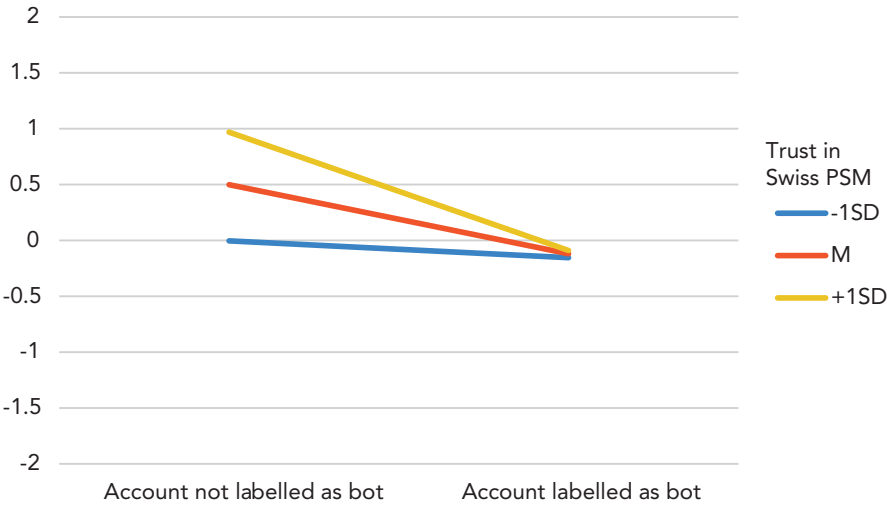
The analysis reveals a more complex picture. As shown in Figures 10.1 and 10.2, high trust in the Swiss PSM journalistic service had a positive corrective effect on opinion when the corrective information came from an account not marked as automated or affiliated to the Swiss PSM, while for unknown peers, high trust tended to have a positive effect when the corrective information was provided by an account marked as automated. One possible explanation for this is that the corrective information from an unknown peer marked as automated was examined more thoroughly, so that it was also more likely to be noticed that the corrective information included an external link to an article from the Swiss PSM.

Figure 10.1 *Opinion after receiving corrective information from the account of an unknown peer*



Comments: N = 113

Figure 10.2 *Opinion after receiving corrective information from an account affiliated with Swiss public service media*



Comments: N = 113

Conclusion

Misinformation and disinformation spread quickly and persistently online and on SNS, making its correction a major challenge. Therefore, “we need to pursue multiple avenues – many of them political – to contain misinformation and redesign the information architecture that facilitates its dissemination” (Lewandowsky & Van Der Linden, 2021: 374).

We have addressed this concern in this chapter by examining the potential of PSM as promoters of timely and trustworthy content through accounts operated by them in an automated way, as a possible governance option in the fight against misinformation. Thus, we have explored one possible way in which PSM can make a contribution to society with regard to a highly relevant and urgent challenge such as the spreading of misinformation on digital communication channels, while also accounting for changing media usage habits. From the legal analysis of the documents relevant to the Swiss PSM, a legitimate claim can be derived for the use of automated communication agents by PSM on SNS, which can thereby also be active with corrective interventions against the spread of misinformation. Despite the similarities of the legal and regulatory conditions, it is the subject of future research to determine the legality of such activities performed by PSM in other countries.

Moreover, whether PSM could also make such a contribution to society in other countries not only depends on the regulatory conditions, but our study also indicates that the credibility attributed to PSM as a provider of journalistic content among the population may influence how successfully it

can perform such a task. The findings from the experimental survey suggest that the high credibility of PSM contributes to users interacting more with corrective information provided by PSM on SNS (via likes and clicks) and that this corrective information can also be effective in terms of opinion – combine this with the fact that higher user engagement can also influence the algorithm-based dissemination of corrective information. Thus, the findings of this study are in line with other studies on the role of sources, including news organisations (Van Der Meer & Jin, 2020), in correcting misinformation, where the credibility and expertise of the source and linking to an external, credible source positively impacted the effectiveness of a corrective intervention (Bode & Vraga, 2018; Walter et al., 2020a). However, the results of our experiment also show that labelling automated accounts, as required by bot disclosure regulations (assuming an indubitable identification could be achieved) could be accompanied by several potential implications (e.g., lower credibility attribution, more intense scrutiny of content) or even side effects, possibly similar to those observed with misinformation labelling (e.g., implicit validation of unlabelled content) (Pennycook et al., 2020).

Previous research on the spread and correction of misinformation has identified another critical factor that can help or hinder the persistence of misinformation, namely the duration of time until a correction is encountered. The temporal sequence that misinformation precedes correction is inherent, but as the duration of time increases until a corrective intervention is made, misinformation can become embedded in memory more strongly; moreover, on SNS there is also the risk of encountering the same misinformation several times, also facilitated by social bots. Therefore, real-time rejections and timely debunking are also considered to be promising responses (Van Der Linden, 2019), which can only be achieved through automation. Users with the appropriate digital media skills could be a valuable resource in the rapid identification of misinformation (Pedrazzi & Oehmer, 2021). This is also the subject of future research.

At this point, it is necessary to highlight the limitations of the study and its results and interpretation. First, the legal analysis is limited to Swiss media regulation. Although similar to many (Western) countries, there may be differences that should be examined in order to determine the legality of engaging in such activities by PSM outside Switzerland. Second, it is necessary to point out the limited ecological validity of the experimental survey, which used an unknown bot disclosure label that the participants encountered for the first time. Third, the artificial SNS environment did not correspond with their account and was represented by a screenshot. Fourth, two rather unfamiliar topics were selected: Research on misinformation correction in particular has shown that misperceptions about more familiar, polarising topics are more difficult to correct. In addition, it is also important to note the small, non-representative sample with which the experimental online survey was conducted.

Despite the limitations, our study was able to show that PSM (through the operation of automated accounts) can be a valuable resource in combating misinformation on SNS. Automation would allow misinformation to be corrected in a timely manner by credible sources such as PSM, which can have an additional positive impact on their effectiveness. Platform operators are under pressure in connection with the spread of misinformation through their networks, as are PSM, whose legitimacy has been challenged in recent times by political actors in various countries (for a discussion of the legitimacy of PSM, see Puppis & Ali, Chapter 1; Fehlmann, Chapter 2 in this volume). Both SNS operators and PSM could thus benefit from cooperating in the fight against misinformation. However, the form that cooperation between platforms and PSM could take in this regard has yet to be defined. Initial media policy regulatory proposals are moving in the direction of online prominence regimes, which oblige platforms to give content from PSM and other legacy media more relevance through their algorithms. But platform operators are also taking action themselves, for example, as part of a collaboration with news agencies Reuters and AP, Twitter is testing the possibility of disseminating information from credible sources about significant events quickly and prioritised. As we have tried to show in this chapter, there are many possible avenues for curtailing misinformation. Future research should accordingly focus on identifying further governance options for dealing with misinformation and testing their effectiveness.

References

- Avram, M., Micallef, N., Patil, S., & Menczer, F. (2020). Exposure to social engagement metrics increases vulnerability to misinformation. *Harvard Kennedy School (HKS) Misinformation Review*, 1(5). <https://doi.org/10.37016/mr-2020-033>
- Bauer, P. C., & Clemm Von Hohenberg, B. (2020). Believing and sharing information by fake sources: An experiment. *Political Communication*, 38(6), 647–671. <https://doi.org/10.1080/10584609.2020.1840462>
- Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Science vs conspiracy: Collective narratives in the age of misinformation. *PLOS ONE*, 10(2), e0118093. <https://doi.org/10.1371/journal.pone.0118093>
- Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 US presidential election online discussion. *First Monday*, 21(11). <https://doi.org/10.5210/fm.v21i11.7090>
- Bode, L., & Vraga, E. K. (2018). See something, say something: Correction of global health misinformation on social media. *Health Communication*, 33(9), 1131–1140. <https://doi.org/10.1080/10410236.2017.1331312>
- Bode, L., Vraga, E. K., & Tully, M. (2021). Correcting misperceptions about genetically modified food on social media: Examining the impact of experts, social media heuristics, and the gateway belief model. *Science Communication*, 43(2), 225–251. <https://doi.org/10.1177/1075547020981375>
- Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., Quinn, S. C., & Dredze, M. (2018). Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health*, 108(10), 1378–1384. <https://doi.org/10.2105/ajph.2018.304567>
- Carey, J. M., Chi, V., Flynn, D. J., Nyhan, B., & Zeitzoff, T. (2020). The effects of corrective information about disease epidemics and outbreaks: Evidence from Zika and yellow fever in Brazil. *Science Advances*, 6(5), eaaw7449. <https://doi.org/10.1126/sciadv.aaw7449>

- Chadwick, A., & Vaccari, C. (2019). *News sharing on UK social media: Misinformation, disinformation, and correction*. Loughborough University.
- Chan, M.-p. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, 28(11), 1531–1546. <https://doi.org/10.1177/0956797617714579>
- Clayton, K., Blair, S., Busam, J. A., Forstner, S., Gance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A. T., Wolff, A. G., Zhou, A., & Nyhan, B. (2019). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, 42, 1073–1095. <https://doi.org/10.1007/s11109-019-09533-0>
- Dumermuth, M. (2016). Die Zuständigkeit des Bundes im Bereich der elektronischen Medien nach Art. 93 BV [The competence of the Confederation in the area of electronic media according to Art. 93 SR]. *Schweizerisches Zentralblatt für Staats- und Verwaltungsrecht*, 7(2016), 335–368.
- Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Harcourt Brace Jovanovich College Publishers.
- Edwards, C., Beattie, A. J., Edwards, A., & Spence, P. R. (2016). Differences in perceptions of communication quality between a Twitterbot and human agent for information seeking and learning. *Computers in Human Behavior*, 65, 666–671. <https://doi.org/10.1016/j.chb.2016.07.003>
- Edwards, C., Edwards, A., Spence, P. R., & Shelton, A. K. (2014). Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter. *Computers in Human Behavior*, 33, 372–376. <https://doi.org/10.1016/j.chb.2013.08.013>
- Ferrara, E. (2017). Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday*, 22(8). <https://doi.org/10.5210/fm.v22i8.8005>
- Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society* (pp. 167–194). MIT Press.
- Gillespie, T. (2018a). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
- Gillespie, T. (2018b). Governance of and by platforms. In J. Burgess, A. Marwick, & T. Poell (Eds.), *The Sage handbook of social media* (pp. 254–278). Sage. <https://dx.doi.org/10.4135/9781473984066>
- Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 US presidential election. *Science*, 363(6425), 374–378. <https://doi.org/10.1126/science.aau2706>
- Hameleers, M., Powell, T. E., Van Der Meer, T. G. L. A., & Bos, L. (2020). A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication*, 37(2), 281–301. <https://doi.org/10.1080/10584609.2019.1674979>
- Hameleers, M., & Van Der Meer, T. G. L. A. (2020). Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers? *Communication Research*, 47(2), 227–250. <https://doi.org/10.1177/0093650218819671>
- Hassan, A., & Barber, S. J. (2021). The effects of repetition frequency on the illusory truth effect. *Cognitive Research: Principles and Implications*, 6(1). <https://doi.org/10.1186/s41235-021-00301-5>
- Helberger, N., Pierson, J., & Poell, T. (2018). Governing online platforms: From contested to cooperative responsibility. *The Information Society*, 34(1), 1–14. <https://doi.org/10.1080/01972243.2017.1391913>
- Hjarvard, S. (2018). Public service in the age of social network media. In G. F. Lowe, H. Van den Bulck, & K. Donders (Eds.), *Public service media in the networked society: RIPE@2017* (pp. 59–74). Nordicom, University of Gothenburg. <http://urn.kb.se/resolve?urn=urn:nbn:se:norden:org:diva-5291>
- Hoffmann, C. P., Weber, J., Zepic, R., Greger, V., & Krcmar, H. (2019). Dimensionen digitaler Mündigkeit und politische Beteiligung im Netz [Dimensions of digital citizenship and

- online political participation]. In I. Engelmann, M. Legrand, & H. Marzinkowski (Eds.), *Politische Partizipation im Medienwandel* [Political Participation in Changing Media] (pp. 79–99). Gesis. <https://doi.org/10.17174/dcr.v6.4>
- Horowitz, M., Cushion, S., Dragomir, M., Gutiérrez Manjón, S., & Pantti, M. (2021). A framework for assessing the role of public service media organizations in countering disinformation. *Digital Journalism*, 10(5), 843–865. <https://doi.org/10.1080/21670811.2021.1987948>
- Howard, P. N., & Kollanyi, B. (2016, June 20). Bots, #StrongerIn, and #Brexit: Computational propaganda during the UK-EU referendum. SSRN. <https://doi.org/10.2139/ssrn.2798311>
- Jolley, D., & Paterson, J. L. (2020). Pylons ablaze: Examining the role of 5G Covid-19 conspiracy beliefs and support for violence. *British Journal of Social Psychology*, 59(3), 628–640. <https://doi.org/10.1111/bjso.12394>
- Kaiser, J., Keller, T. R., & Kleinen-von Königslöw, K. (2018). Incidental news exposure on Facebook as a social experience: The influence of recommender and media cues on news selection. *Communication Research*, 48(1), 77–99. <https://doi.org/10.1177/0093650218803529>
- Kohring, M., & Matthes, J. (2007). Trust in news media. *Communication Research*, 34(2), 231–252. <https://doi.org/10.1177/0093650206298071>
- Langvardt, K. (2018). A new deal for the online public sphere. *George Mason Law Review*, 26(1), 341–394. <https://ssrn.com/abstract=3149513>
- Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380), 1094–1096. <https://doi.org/10.1126/science.aao2998>
- Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction. *Psychological Science in the Public Interest*, 13(3), 106–131. <https://doi.org/10.1177/1529100612451018>
- Lewandowsky, S., & Van Der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, 32(2), 348–384. <https://doi.org/10.1080/10463283.2021.1876983>
- Li, J. (2020). Toward a research agenda on political misinformation and corrective information. *Political Communication*, 37(1), 125–135. <https://doi.org/10.1080/10584609.2020.1716499>
- Loomba, S., De Figueiredo, A., Piatek, S. J., De Graaf, K., & Larson, H. J. (2021). Measuring the impact of Covid-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, 5(3), 337–348. <https://doi.org/10.1038/s41562-021-01056-1>
- Margolin, D. B., Hannak, A., & Weber, I. (2018). Political fact-checking on Twitter: When do corrections have an effect? *Political Communication*, 35(2), 196–219. <https://doi.org/10.1080/10584609.2017.1334018>
- Mena, P. (2020). Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. *Policy & Internet*, 12(2), 165–183. <https://doi.org/10.1002/poi3.214>
- Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of Communication*, 60(3), 413–439. <https://doi.org/10.1111/j.1460-2466.2010.01488.x>
- Mønsted, B., Sapiezynski, P., Ferrara, E., & Lehmann, S. (2017). Evidence of complex contagion of information in social media: An experiment using Twitter bots. *PLOS ONE*, 12(9), e0184148. <https://doi.org/10.1371/journal.pone.0184148>
- Napoli, P. M. (2015). Social media and the public interest: Governance of news platforms in the realm of individual and algorithmic gatekeepers. *Telecommunications Policy*, 39(9), 751–760. <https://doi.org/10.1016/j.telpol.2014.12.003>
- Newman, N., Fletcher, R., Kalogeropoulos, A., & Nielsen, R. (2019). *Reuters Institute digital news report 2019*. <https://reutersinstitute.politics.ox.ac.uk/our-research/digital-news-report-2019>
- Oehmer, F., & Pedrazzi, S. (2020). Was schützt (vor) Social Bots? Vorschläge zur Governance von computergenerierten Softwareagenten im Internet [What protects (against) social bots? Proposals for the governance of computer-generated software agents on the internet]. *UFITA*, 84(1), 7–35. <https://doi.org/10.5771/2568-9185-2020-1-7>

- Oeldorf-Hirsch, A., Schmierbach, M., Appelman, A., & Boyle, M. P. (2020). The ineffectiveness of fact-checking labels on news memes and articles. *Mass Communication and Society*, 23(5), 682–704. <https://doi.org/10.1080/15205436.2020.1733613>
- Ognyanova, K., Lazer, D., Robertson, R. E., & Wilson, C. (2020). Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School (HKS) Misinformation Review*, 1(4). <https://doi.org/10.37016/mr-2020-024>
- Pedrazzi, S., & Oehmer, F. (2021). Bot or no bot? Digital Media Literacy als Governance-Instrument für die Identifikation von und den Umgang mit Social Bots [Digital media literacy as a governance tool for identifying and dealing with social bots]. In G. Keel, & W. Weber (Eds.), *Media literacy* (pp. 189–206). Nomos. <https://doi.org/10.5771/9783748920656-189>
- Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, 66(11), 4944–4957. <https://doi.org/10.1287/mnsc.2019.3478>
- Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, 25(5), 388–402. <https://doi.org/10.1016/j.tics.2021.02.007>
- Petty, R. E., & Briñol, P. (2010). Attitude change. In R. F. Baumeister, & E. J. Finkel (Eds.), *Advanced social psychology: The state of the science* (pp. 217–259). Oxford University Press.
- Petty, R. E., & Cacioppo, J. T. (1984). Source factors and the elaboration likelihood model of persuasion. *ACR North American Advances*, 11(1), 668–672.
- Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In R. E. Petty, & J. T. Cacioppo (Eds.), *Communication and persuasion: Central and peripheral routes to attitude change* (pp. 1–24). Springer. <https://doi.org/10.1007/978-1-4612-4964-1>
- Puppis, M. (2020). Governance der Vermittlung öffentlicher Kommunikation [Governance of the mediation of public communication]. In O. Jarren, & C. Neuburger (Eds.), *Gesellschaftliche Vermittlung in der Krise: Medien und Plattformen als Intermediäre* [Social mediation in crisis: Media and platforms as intermediaries] (pp. 189–224). Nomos.
- Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
- Rozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L. J., Recchia, G., Van Der Bles, A. M., & Van Der Linden, S. (2020). Susceptibility to misinformation about Covid-19 around the world. *Royal Society Open Science*, 7(10), 201199. <https://doi.org/10.1098/rsos.201199>
- Saurwein, F., & Spencer-Smith, C. (2020). Combating disinformation on social media: Multilevel governance and distributed accountability in Europe. *Digital Journalism*, 8(6), 820–841. <https://doi.org/10.1080/21670811.2020.1765401>
- Saxer, U. (2017). Die Online-Zuständigkeiten des Bundes [The online responsibilities of the confederation]. *Aktuelle juristische Praxis: AJP*, 3(2017), 334–350.
- Saxer, U., & Brunner, F. (2018). Der Service public, die digitale Revolution und die Medienverfassung [The public service, the digital revolution and the constitution of the media]. *Aktuelle juristische Praxis: AJP*, 1(2018), 22–41.
- Schwaiger, L. (2020). Mediale Lebenswelten junger Schweizerinnen und Schweizer [Media life worlds of young Swiss]. In *Qualität der Medien. Schweiz – Suisse – Svizzera. Jahrbuch 2020* [Quality of the media. Switzerland. Yearbook 2020] (pp. 89–100). Schwabe.
- Shahin, S., Saldaña, M., & Gil De Zúñiga, H. (2020). Peripheral elaboration model: The impact of incidental news exposure on political participation. *Journal of Information Technology & Politics*, 18(2), 148–163. <https://doi.org/10.1080/19331681.2020.1832012>
- Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, 9(1), 4787. <https://doi.org/10.1038/s41467-018-06930-7>
- Sharma, K., Seo, S., Meng, C., Rambhatla, S., & Liu, Y. (2020). Covid-19 on social media: Analyzing misinformation in Twitter conversations. *arXiv preprint arXiv:2003.12309*.
- Shin, J., & Thorson, K. (2017). Partisan selective sharing: The biased diffusion of fact-checking messages on social media. *Journal of Communication*, 67(2), 233–255. <https://doi.org/10.1111/jcom.12284>

- Swire-Thompson, B., & Lazer, D. (2020). Public health and online misinformation: Challenges and recommendations. *Annual Review of Public Health*, 41(1), 433–451. <https://doi.org/10.1146/annurev-publhealth-040119-094127>
- Swiss Confederation. (1874). *Swiss Federal Constitution, Art. 55bis, para. 2 aBV*. https://www.fedlex.admin.ch/eli/cc/1/1_1_1/de
- Swiss Confederation. (1999). *Swiss Federal Constitution, Art. 93, para. 2 BV*. <https://www.fedlex.admin.ch/eli/cc/1999/404/de>
- Swiss Confederation (2006). *Federal Act on Radio and Television, Article 24, para. 4*. <https://www.fedlex.admin.ch/eli/cc/2007/150/en>
- Swiss Confederation. (2020). *SRG SSR charter*. <https://www.bakom.admin.ch/bakom/de/home/elektronische-medien/informationen-ueber-radio-und-fernsehveranstalter/srg-ssr/konzessionierung-und-technik-srg-ssr.html>
- Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining “fake news”. *Digital Journalism*, 6(2), 137–153. <https://doi.org/10.1080/21670811.2017.1360143>
- Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, 33(3), 460–480. <https://doi.org/10.1080/10584609.2015.1102187>
- Valenzuela, S., Halpern, D., Katz, J. E., & Miranda, J. P. (2019). The paradox of participation versus misinformation: Social media, political engagement, and the spread of misinformation. *Digital Journalism*, 7(6), 802–823. <https://doi.org/10.1080/21670811.2019.1623701>
- Van Der Linden, S. (2019). Countering science denial. *Nature Human Behaviour*, 3(9), 889–890. <https://doi.org/10.1038/s41562-019-0631-5>
- Van Der Meer, T. G. L. A., & Jin, Y. (2020). Seeking formula for misinformation treatment in public health crises: The effects of corrective information type and source. *Health Communication*, 35(5), 560–575. <https://doi.org/10.1080/10410236.2019.1573295>
- van Dijk, J., Poell, T., & de Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. <https://doi.org/10.1126/science.aap9559>
- Vraga, E. K., & Bode, L. (2017). Using expert sources to correct health misinformation in social media. *Science Communication*, 39(5), 621–645. <https://doi.org/10.1177/1075547017731776>
- Vraga, E. K., & Bode, L. (2018). I do not believe you: How providing a source corrects health misperceptions across social media platforms. *Information, Communication & Society*, 21(10), 1337–1353. <https://doi.org/10.1080/1369118x.2017.1313883>
- Vraga, E. K., & Bode, L. (2020). Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation. *Political Communication*, 37(1), 136–144. <https://doi.org/10.1080/10584609.2020.1716500>
- Walter, N., Brooks, J. J., Saucier, C. J., & Suresh, S. (2020a). Evaluating the impact of attempts to correct health misinformation on social media: A meta-analysis. *Health Communication*, 36(13), 1776–1784. <https://doi.org/10.1080/10410236.2020.1794553>
- Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020b). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, 37(3), 350–375. <https://doi.org/10.1080/10584609.2019.1668894>
- Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs*, 85(3), 423–441. <https://doi.org/10.1080/03637751.2018.1467564>
- Walter, N., & Tukachinsky, R. (2019). A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it? *Communication Research*, 47(2), 155–177. <https://doi.org/10.1177/0093650219854600>
- Wang, Y. (2021). Debunking misinformation about genetically modified food safety on social media: Can heuristic cues mitigate biased assimilation? *Science Communication*, 43(4), 460–485. <https://doi.org/10.1177/10755470211022024>
- Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe Report*, 27.
- Winter, S., & Krämer, N. C. (2014). A question of credibility – Effects of source cues and recommendations on information selection on news sites and blogs. *Communications*, 39(4), 435–456. <https://doi.org/10.1515/commun-2014-0020>

- Wood, M. J. (2018). Propagating and debunking conspiracy theories on Twitter during the 2015–2016 Zika virus outbreak. *Cyberpsychology, Behavior, and Social Networking*, 21(8), 485–490. <https://doi.org/10.1089/cyber.2017.0669>
- Zimmermann, F., & Kohring, M. (2018). „Fake news“ als aktuelle Desinformation. Systematische Bestimmung eines heterogenen Begriffs [“Fake news” as current disinformation: Systematic definition of a heterogeneous term]. *M&K Medien & Kommunikationswissenschaft*, 66(4), 526–541. <https://doi.org/10.5771/1615-634X-2018-4-526>