Cagé, Julia; Gallo, Nathan; Hengel, Moritz; Henry, Emeric; Huang, Yuchen

Working Paper

Fact-Checking and Misinformation: Evidence from the Market Leader

CESifo Working Paper, No. 12319

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Cagé, Julia; Gallo, Nathan; Hengel, Moritz; Henry, Emeric; Huang, Yuchen (2025) : Fact-Checking and Misinformation: Evidence from the Market Leader, CESifo Working Paper, No. 12319, Munich Society for the Promotion of Economic Research - CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/336021

CES ifo Working Papers

12319
2025

December 2025

Fact-Checking and Misinformation: Evidence from the Market Leader

Julia Cagé, Nathan Gallo, Moritz Hengel, Emeric Henry, Yuchen Huang

CES ifo

Imprint:

CESifo Working Papers

ISSN 2364-1428 (digital)

Publisher and distributor: Munich Society for the Promotion
of Economic Research - CESifo GmbH

Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740

Email office@cesifo.de
https://www.cesifo.org

Editor: Clemens Fuest

An electronic version of the paper may be downloaded free of charge
· from the CESifo website: www.ifo.de/en/cesifo/publications/cesifo-working-papers
· from the SSRN website: www.ssrn.com/index.cfm/en/cesifo/
· from the RePEc website: https://ideas.repec.org/s/ces/ceswps.html

Fact-Checking and Misinformation: Evidence from the Market Leader*

Julia Cagé[1], Nathan Gallo[2], Moritz Hengel[3], Emeric Henry[4], and Yuchen Huang[5]

[1,4]Sciences Po Paris and CEPR
[2,3]Sciences Po Paris
[5]Sciences Po Paris and HEC Liège

First version: June 2023. **This version:** November 2025

Abstract

What are the dynamic effects of fact-checking on the behavior of those who circulate misinformation and on the spread of false news? In this paper, we provide causal evidence on these questions, building on a unique partnership with the Agence France Presse (AFP), the world's largest fact-checking organization and a partner of Facebook's Third-Party Fact-Checking Program. Over an 18-month period (December 2021-June 2023), we collected information on the stories proposed by fact-checkers during the daily editorial meetings, some of which were ultimately fact-checked while others, despite being ex ante "similar", were left aside. Using two complementary Difference-in-Differences approaches, one at the story level and the other at the post level (within fact-checked stories), we show that fact-checking reduces the circulation of misinformation on Facebook by approximately 8%, an effect driven entirely by stories rated as "False." Furthermore, we provide evidence of behavioral responses: the publication of a fact-check more than doubles the deletion of posts in the fact-checked stories, and users whose posts appear in fact-checked stories become less likely to share misinformation in the future. While our results clearly confirm the effectiveness of fact-checking, we provide policy recommendations to further strengthen its impact.

Keywords: fact-checking, misinformation, Facebook, Meta, Crowdtangle, fake news, third-party verification, social media.
JEL No: D8, D83, D91, L82, L86, P00.

*We are grateful to Malka Guillot, Nicolas Hervé, Andrea Prat, Koleman Strumpf, and Katia Zhuravskaya, to seminar participants at the Paris School of Economics, Sciences Po Paris, the University of Neuchatel, and the VIDE Digital Economy Seminar, and to conference participants at the CEPR Workshop on Media, Technology, Politics, and Society, the Digital Methods Summer School in Amsterdam, the "Economic and Social Challenge in the Digital Age" conference in Paris, and the Tilec conference in Tilburg for very helpful comments and suggestions. We also thank all the AFP Factuel team, and in particular Grégoire Lemarchand and Pauline Talagrand who made this project possible. We further thank Constance Frohly for her outstanding research assistance. The research leading to this project has received funding from the McCourt Institute and the Paris Region PhD programme. All errors remain our own.

1 Introduction

While the fact-checking industry has grown significantly in recent years in response to global concerns about the spread of fake news (Allcott and Gentzkow, 2017; Allcott et al., 2019), its impact is still under intense scrutiny. In January 2025, Facebook announced the end of its flagship Third-Party Fact-Checking Program in the US, citing concerns over potential threats to free speech. Other critics have questioned whether fact-checking has any meaningful impact on the circulation of misinformation. Based on a unique partnership with a leading fact-checking organization, we provide causal evidence on these questions. We find that fact-checking is effective: it reduces the spread of fake news on social media by approximately 8%. Moreover, we observe significant behavioral responses from users: posts related to fact-checked stories are more likely to be deleted, and users whose posts were rated as false become less likely to share misinformation in the future. These findings suggest that fact-checking not only slows the spread of fake news but also, rather than hindering free speech, nudges users towards more responsible online behavior.

The main empirical challenge to identify the causal effect of a fact-check is the lack of credible counterfactuals. It is not possible to simply compare the circulation of a story rated as false with that of another story, as false stories often have specific characteristics. To overcome this challenge, we rely on a unique partnership with the Agence France Presse (AFP), the third-largest news agency in the world and the world's largest fact-checking organization. A journalist was hired for 18 months (December 2021 - June 2023) to attend the daily editorial meetings of AFP Factuel, the AFP's unit working on fact-checking French-language news.[1] He collected information on all the stories that were discussed during the daily meetings, those that were approved and fact-checked and those that were left aside. He also recorded the reasons for rejection (lack of resources, lack of virality, etc.) based on regular meetings with the AFP Factuel's chief editors. This allows us to identify stories that could have been fact-checked – and hence are "similar" to stories that were ultimately chosen – but were not for reasons such as a lack of time or resources. We are thus in a unique position to define a valid set of counterfactual stories to quantify the causal impact of fact-checking.

AFP is a member of the Third-Party Fact-Checking Program (TPFC program henceforth) set up by Facebook. This gives AFP journalists direct access to the Facebook tool where they can rate posts directly once a fact-check is produced.[2] Importantly, the agreement with Facebook does not provide incentives to systematically flag all the posts that relate to the same fact-checked story. For each story, we recover additional posts, associated with the same misinformation narrative, that were not flagged. This allows us to define, *within fact-checked stories*, a valid set of counterfactual posts to quantify the causal impact of signaling a post as false.

[1] The journalist was hired by the research team but selected in conjunction with AFP Factuel. He was fully involved in the AFP fact-checking unit, producing fact-checks and carrying out the same work as the other journalists. The only difference is that he had the extra task of collecting information on the stories as part of this research project.

[2] It also gives journalists access to the so-called "Facebook claim", which contains a list of suspicious posts automatically detected by Facebook using algorithms (see Section 3 below). Only the stories fact-checked as "False", "Altered Content", "Partly False" or "Missing Context" are rated; the (very few) stories fact-checked by the journalists but ultimately considered to be "True" are not rated.

To investigate quantitatively the causal impact of fact-checking, for each of the stories – fact-checked or not – and posts – flagged or not – we then collect information on all the associated public posts on Facebook, and in particular on their engagement metrics using Crowdtangle, Meta's tool to explore public content on social media. This allows us to study the effect of fact-checking on the engagement with the posts related to a story. Additionally, we also collect data on the most active Facebook accounts and users who posted content related to one of the stories discussed by AFP. For these accounts, we measure their general posting behavior, including on stories not among those discussed by the AFP.

Overall, our dataset includes 944 stories – including information on their origin as well as on their topic – of which 558 were fact-checked by AFP Factuel. For each of the 558 fact-checked stories, we also recover information on the associated flag, and for each of the 386 unchecked ones, information on the reasons for not fact-checking them, based on discussions with the editorial team. This unique data collection offers a singular perspective on the fact-checking process. One of the contributions of the paper is to provide some novel descriptive facts about the way fact-checking is performed. First, we show that the algorithm provided by Facebook to detect fake news is rarely used by journalists who rely more on their own monitoring strategies. Second, the fact-checking process is quite long: the median time between the date of discussion and the first rating is two days. Finally, approximately half of the posts related to a story are flagged by the fact-checkers.[3] All these facts, combined with the empirical analysis of the impact of fact-checking will guide our policy recommendations.

Most importantly, the data collection allows us to build an original identification strategy to identify the causal effect of fact-checking on the circulation of misinformation. We use two approaches, one at the story level and the other at the post level. At the story level, we use a Difference-in-Differences (DiD) approach, comparing stories that were fact-checked to "similar" stories that were initially considered for fact-checking but left aside. The key identifying assumption is that the two types of stories would have had similar trajectories in terms of circulation absent the fact-checking intervention. To ensure the validity of this identifying assumption, we impose four restrictions on the data exploiting the details of the editorial process. First, we exclude the stories that were not fact-checked because of a lack of virality, since the criterion is directly related to our outcome variable. Second, we exclude the stories that were translations of stories fact-checked in another language, since the editorial process is very different due to the much lower cost of fact-checking. Third, we exclude stories that appeared the night just before the editorial meeting, since it is not feasible for us to verify the absence of pre-trends for those "young" stories.Finally, to further improve the balance between control and treated stories, we use propensity score matching using pre-treatment engagements with the stories as covariates.

We provide causal evidence that fact-checking reduces the circulation of posts related to fact-checked stories, with engagement decreasing by approximately 8%. This effect is highly dependent on the type of flag applied. For stories rated as "False," engagement drops by 9%, while for those labeled "Partly False" or "Missing Context," we find no statistically significant reduction in circulation. We further explore key sources of heterogeneity. First, timing matters: when a fact-check is produced within two days, the reduction in engagement reaches 11%. In contrast, fact-checks issued after more than two days

[3]This is an upper bound since the journalist working on our project presumably did not succeed in recovering all the posts related to the stories.

have no statistically significant effect. Second, we find variation by topic. Fact-checks on more recent or less politically entrenched issues – such as the war in Ukraine – are considerably more effective.

The second identification strategy uses only fact-checked stories and, for each story, compares the posts that were rated by the fact-checkers with those that were not. As explained above, the agreement with Facebook does not provide incentives to systematically rate all posts. Rating additional posts linked to the story is up to the willingness of the journalist who fact-checked the story. Working together with the AFP Factuel team allowed us to understand that journalists rate as many posts as they can but not usually all of them due to a lack of time. To further strengthen the identification – since we cannot identify which marginal posts the journalist would have rated if they had pursued their efforts[4] – we use propensity score matching using pre-treatment circulation. As in the story-level analysis, we observe a sharp decrease in engagement for fact-checked stories, and we find that the magnitude of the drop is larger for the stories rated false.

Part of the observed reduction in circulation could be due to Facebook's policy of demoting content shown to be false, although the platform does not disclose the exact mechanisms by which this demotion is implemented. In the second part of the paper, we show that there are clear behavioral responses from users in addition to the demotion effect. First we explore the impact on deletions of posts. We show that the deletion rate doubles for posts related to fact-checked stories. Notably, this effect is driven entirely by stories rated as "False." This suggests that fact-checking triggers users' reputational concerns, prompting them to voluntarily remove flagged content. We then provide a second essential piece of evidence on the behavioral response of users. Based on the tracking of the most active accounts, we show that accounts whose posts appear in a story fact-checked by the AFP decrease the number of posts they produce on other stories in the two weeks following the editorial meeting when the story was considered for fact-checking. Furthermore, we provide evidence that being fact-checked also reduces the probability that an account circulates misinformation in the longer run on Facebook.[5]

We conclude the paper by discussing the policy implications of our findings. The evidence clearly suggests that discontinuing the TPFC program would lead to an increase in the circulation of misinformation. Moreover, our results challenge the notion that fact-checking poses a threat to free speech. Many of the observed effects stem from users' voluntary behavioral responses – such as deleting posts flagged as false or adjusting their future sharing behavior – rather than from content demotion by the platform. In addition, our findings provide a basis for preliminary cost-effectiveness estimates. A back-of-the-envelope calculation suggests that the upper bound on the cost per reduced engagement with misinformation is between 15 and 35 cents. This figure should be compared with the social costs of engagements with misinformation. Moreover it does not account for the spillover effects on the behavior of the users when they interact with different types of content after being flagged.

Our findings also point to concrete ways to improve the fact-checking process. First, on the detection side, the current system relies heavily on manual monitoring of sub-communities within Facebook, due to the absence of effective automated tools. This approach leads to incomplete and path-dependent

[4]This is because we do not observe the order in which journalists see the posts on the "Facebook claim" (to which we do not have access).

[5]We cannot rule out substitution to another platform, however.

detection of misinformation. Second, in terms of production, we show that the speed at which a fact-check is produced is a critical determinant of its effectiveness, highlighting an important trade-off between timeliness and other production constraints. Third, on the rating side, our results underscore the importance of providing appropriate incentives for fact-checkers to comprehensively rate posts once a fact-check has been completed. Improvements across these three dimensions – detection, production, and rating – would likely enhance the overall effectiveness of the fact-checking system at relatively low cost, especially when compared to the broader societal costs of misinformation.

Literature review Our work relates to the growing literature on the impact of fact-checking using evidence from randomized survey experiments. The first strand of papers looks at the effectiveness of fact-checking in correcting false beliefs. Barrera et al. (2020), in the context of presidential elections in France, expose users to false statements from the far-right candidate on the issue of immigration, while some are also randomly shown fact-checks of the statements. The paper shows that fact-checking works to correct purely factual beliefs, but does not change more subjective beliefs. In particular, the voting intentions for the far-right candidate increase by the same amount with or without fact-checking. While similar results are obtained by Swire et al. (2017) and Nyhan et al. (2020) in the context of Trump's 2016 presidential campaign, Wintersieck (2017) – using one political debate from the 2013 New Jersey Gubernatorial race – does find that fact-checks not only affect the respondents' perception of candidates but also their willingness to vote for them.[6]

While it may not be effective in changing people's beliefs, there is a broad consensus, emerging from similar types of randomized survey experiments, that fact-checking can play a role in decreasing the circulation of fact-checked content (Henry et al., 2022; Mena, 2020; Pennycook et al., 2020a; Yaqub et al., 2020). Pennycook et al. (2020a) for instance carried out an online experiment where the participants were shown true and false statements. They find that adding the "false" label to a statement or providing access to a fact-check significantly reduces participants' self-reported intention to share the statement on social media.[7] Guriev et al. (2023) perform a randomized experiment on Twitter during the 2022 US midterm election and 2024 US presidential campaign and show that priming accuracy is the most effective policy for reducing the sharing of false news.[8]

There is, however, a lack of evidence from the field on the impact of fact-checking, in particular dynamic effects, which our paper tries to fill. One exception is Mattozzi et al. (2022) who study how fact-

[6]There is also a large related literature studying more generally the impact of information on political beliefs and behavior (Alesina et al., 2018; Kuziemko et al., 2015; Grigorieff et al., 2016; Cagé, 2016; Cagé, 2020, among others). Our paper also relates to the literature investigating the role played by false news and misinformation in elections (Allcott and Gentzkow, 2017; Munger et al., 2022).

[7]See also the literature on the factors influencing the decision to share (Altay et al., 2020; Guess et al., 2019; Fazio, 2020). Pennycook et al. (2020c) show that the people who share false claims about COVID-19 partly do so because they are not aware that the content is not accurate; consistently, Pennycook et al. (2020b) find that people do not necessarily want to share false news on social media but do so because they fail to implement their preference for accuracy due to attentional constraints. Briole et al. (2025), in a randomized controlled trial, show that exposing young people to high-quality news discourages them from sharing fake news. Chopra et al. (2022), in an online experiment, study the demand for fact-checking; they find that fact-checking increases demand for a newsletter when it features stories from an ideologically non-aligned source (on the demand for fact-checking, see also Assenza et al., 2024).

[8]The paper also shows that fact-checking works to reduce the circulation of false news with the key mechanism being the priming of accuracy-related considerations.

checking affects the behavior of politicians (on the impact of fact-checking on politicians, see also Nyhan and Reifler, 2015; Ma et al., 2023). They partner with an Italian fact-checking company and randomly fact-check middle-range politicians. They find that after being fact-checked, politicians are less likely to make incorrect statements. They also show that it makes them less likely to make statements that are verifiable.[9] Compared to this work, we do not specifically focus on political fact-checks but consider all the dimensions of misinformation,focusing on general circulation on social media and behavioral reactions of users.

Our paper also contributes to the growing literature on fact-checking by providing novel descriptive facts about the way fact-checking is performed in the real world. Using a comprehensive dataset of articles published by six French fact-checkers, Louis-Sidois (2025) documents differences in fact-checkers' political content, which reflect the media outlets' slant. Lim (2018) also studies fact-checkers's performance by comparing the inter-rater reliability of two fact-checkers in the US. While Ribeiro et al. (2022) provide suggestive evidence that the incorporation of online attention signals may help organizations better assess and prioritize their fact-checking efforts, our heterogeneity results – in particular regarding the importance of the speed at which the fact-checks are produced – provide additional evidence supporting such policy recommendations.[10] We complement these papers by presenting an insider's view on the processes adopted by a leading fact-checking organization.

Facebook has ended its TPFC program in the US and moved to a Community Notes model, a crowdsourced fact-checking system. The existing literature has mostly focused on Twitter's Community Notes program. Chuai et al. (2024a) document a very large impact of these notes both on the spread of misleading posts and on the probability that users delete them, a finding consistent with Slaughter et al. (2025) and Gao et al. (2024). Yet the evidence remains somewhat mixed regarding whether Community Notes is an effective intervention to reduce engagement with misinformation on social media (Chuai et al., 2024b).[11] By assessing the effectiveness of fact-checks produced by professional organizations, our findings can inform the debate regarding the potential consequences of the end of Facebook's TPFC program and the transition to a crowdsourced model.

The rest of the article is organized as follows. In Section 2 below we describe the institutional setting and present the unique partnership we have set up with the AFP Factuel team. Section 3 provides novel insights on the fact-checking process grounded in our qualitative approach. In Section 4, we present our empirical strategy, both at the story and at the post level. Section 5 presents our main results and discusses their policy implications. Finally, Section 6 concludes.

[9] On repeated false claims by politicians, see also Larraz et al. (2024).

[10] They are also consistent with the literature indicating that the more rapid spread of false news compared to real news may be attributable to its higher level of novelty (see e.g. Vosoughi et al., 2018).

[11] See also Zhou et al. (2025) who show that while Twitter's community notes received by fact-checked authors can increase audience engagement with these authors in the short term, they decrease audience engagement in the long term.

2 Institutional setting and Data collection

2.1 Institutional Setting

Over the past decade, the fact-checking industry has steadily grown amid increasing concerns about the spread of misinformation. Fact-checking organizations range from small NGOs to mainstream media. The main social media companies started setting up partnerships to verify the content circulating on their platforms. Meta (Facebook), in particular, partners with the International Fact-Checking Network (IFCN) to accredit fact-checking organisations as part of its TFPC program. During our study period, there were about 120 accredited organisations worldwide, for example Reuters and The Associated Press (AP) in the US, or France 24 (a television channel) in France.[12] Following recent developments, the partnership with Meta was suspended in the US in January 2025 and fact-checking activities are generally under threat around the world.

We entered in a partnership (described in more details below) with one of the participants in the TFPC, AFP Factuel. AFP Factuel is the largest fact-checking organization in the world. It was created by the AFP in 2017 and gathers around 130 fact-checkers worldwide.[13] Note that the AFP is a private nonprofit media organization, without shareholders and independent from the State.[14]

We concentrate on the team dealing with the content in French, which fluctuates over the study period at around 10 daily active members, most of whom are based in Paris (see Appendix Figure B.1). Below, we describe the way the fact-checking process is organized. Providing evidence on this process is one of the contributions of the paper.

The daily morning meetings Each weekday morning, journalists gather for an editorial meeting to review news stories proposed for fact-checking.[15] The AFP Factuel team then decides which proposals to pursue, with the editor-in-chief playing a central role in the final decision. When a story is approved, the journalist who proposed it typically writes the fact-check.

Identifying a suitable, fact-checkable story to propose during the morning meeting is critical. Journalists draw on a range of sources to select the false or misleading claims they propose to investigate, with each journalist developing their own search strategies – often shaped by their area of specialization. As discussed in Section 3, these strategies frequently involve monitoring known sources of misinformation across multiple platforms. For the approved stories, the process of writing the fact-check then starts, involving discussion with experts and a careful study of different sources. In a third of cases, the writing

[12] As of July 26th, 2023. For an up-to-date list, consult https://www.facebook.com/formedia/mjp/programs/third-party-fact-checking/partner-map. According to the Duke Reporters' LAB, there are now 160 organizations partnering with Meta around the world.

[13] AFP Factuel published several thousands fact-checks per year (e.g. 8,614 in 2021, of which 792 were in French). As a global agency, the AFP works in several languages, including German, English, Arabic, Spanish and Portuguese.

[14] According to its bylaws, the agency is a *"autonomous organization with a civil status, operating under commercial rules."* Its main goal is to seek "the elements of a complete and objective information service" (see e.g. Cagé et al., 2020).

[15] The meeting is held every weekday morning, i.e. excluding weekends. It is attended by all journalists from the French fact-checking team, the news editors overseeing fact-checking in France, the editor of the Africa desk covering disinformation in the francophone world, and the journalist responsible for disinformation in other French-speaking European countries (Belgium and Switzerland). All the AFP Factuel journalists from the Paris and Brussels desks and the editor of the Africa desk participate in the daily morning meetings, regardless of their location (and thus via Zoom when located abroad).

phase takes less than a day. All fact-checks are published on the AFP Factuel's website and are available for free. In Section 3 we provide a number of key novel facts on this process.

Rating the posts After a fact-check is published, the journalist who authored it is responsible for rating the Facebook posts related to the story.[16] As part of the TPFC program, journalists do the rating directly on the platform, choosing one of the ratings established by Meta. Journalists have the option either to rate individual posts or to rate a link, a video or an image (which then rates every post circulating that content). In its training for participants in the TPFC program, Meta advises journalists to rate posts rather than links or videos to reduce the risk of false positives, for instance in cases where the content includes disclaimers or prior debunking.

The ratings established by Facebook are the following, in order of severity: "False", "Altered", "Partly False" and "Missing Context". Regardless of the rating, the user who posted the content receives a notification and has the possibility to un-share the content. If rated "False", an overlay is put on the post to blur its content which clearly indicates that it was checked by an independent fact-checker (online Appendix Figure B.2). When faced with such a post, users have the choice to view the fact-check or to ignore it and see the post. If users decide to see the post, they are prompted with an additional warning flag (online Appendix Figure B.3). The post is also demoted, though the exact conditions are not clearly communicated by Meta (see below).

If the content is rated "Partly False" (online Appendix Figure B.4) or "Missing Context" (Figure B.5), the post remains visible, but with a banner at the bottom of the page. For the "Partly False" rating there is also a demotion, but it is lighter than for the "False" rating. In all cases, sharing flagged content requires the user to confirm additionally that they want to share content flagged by fact-checks (see the "Share Anyway" button on online Appendix Figure B.3 for a post rated "False", and on Figure B.6 in the case of a story rated "Partly False").

As part of the TPFC program, AFP Factuel is paid a flat rate by Facebook for the production of a fact-check, not for rating the posts.[17] In other words, it is only paid for rating one post related to the story with a link to the fact-check. If additional posts corresponding to the same fact-check are rated, this will not involve additional payments. In fact, journalists also have the option to fact-check stories that circulate on other platforms and websites and that might not be shared on Facebook at the time of the morning meetings. The link to these stories (e.g. a post on X) can be rated on the platform but AFP Factuel will only be paid and a flag will only appear if the link is shared on Facebook. It is therefore up to the journalist to decide how much effort to put into rating additional posts on Facebook that relay a fake news story.

[16] According to Facebook, posts related to statements by politicians cannot be rated. Politicians are defined as candidates for office, current officeholders, and leaders of political parties. We thus do not include the political statement-related stories in our analysis.

[17] The amount paid is not disclosed by Meta.

2.2 Partnership with AFP and data collection

We established a partnership with AFP Factuel in October 2021. Beginning on December 1st, 2021, we recruited a journalist – whom we refer to as the observer – whose role in the project was to attend the daily morning editorial meetings (he was also fully integrated into the AFP Factuel team and wrote fact-checks alongside the other journalists). During these meetings, the observer systematically recorded all news stories considered for potential fact-checking, the content of the discussions, the (anonymized) identity of the proposing journalist, and the final decision on whether or not to proceed with a fact-check. Over the course of the experiment, the observer attended 376 meetings in total.

Following the meeting, for each story that was discussed (regardless of whether it was finally selected), the journalist proposing the story sent links and posts promoting the story's claim to the observer. We refer to these as the seeds of the story. We record in particular the origin of the claim (i.e where it was found by the journalist: Twitter[18], Facebook claim, etc.).

Our observer also had three additional data collection tasks. First, for the rejected stories, he was asked to collect information on the reasons for rejection, following a classification that we designed for the project in collaboration with the AFP team. The categories are (i) lack of resources, (ii) lack of virality, (iii) probably true, (iv) infeasible,[19] and (v) already done. This categorization was performed on a weekly basis with the editor-in-chief, who has the final say in the decision to accept or reject a story and has the most oversight over the team's activities.

Second, the observer was asked to identify other Facebook posts circulating the story. We extracted the main keywords for each story and a search based on these keywords was launched to collect public Facebook posts using Crowdtangle, a tool provided by Meta that collects time-series data on a range of public posts.[20] We refer to this as "phase 1" of the data collection process. Starting in June 2022, for all posts, this exact same search was also done a week after the meeting (which we refer to as "phase 2"). In addition, for the fact-checked posts, the data collection was also performed at the time of the publication of the fact-check, a timing typically occurring between phase 1 and phase 2.

For each post identified by the above process, we extract from Crowdtangle the hourly engagement statistics (shares, comments and likes) as well as the nature of the post (text, video, image) and the nature of the account. This extraction is performed for 14 consecutive days after the post is identified. We can therefore observe whether it is no longer found during that period of 14 days, which could indicate the deletion of the post, as discussed below (see the description of the variables in Section 2.3). Third, the observer was asked to collect information on the fact-check and the ratings on Facebook. For each fact-checked story, he retrieved all topic classifications from the published fact-check. He also collected information on its exact date of publication and its length. Additionally, for stories that were not fact-checked, AFP journalists indicated which topic category a potential fact-check would have aligned with

[18] We use the name Twitter here rather than X given that it was the name of the media at the time of our study.

[19] E.g. a claim that the EU sent diplomatic cables to the Senegalese President Macky Sall asking him to postpone the parliamentary elections – which was unverifiable – or another claim that French President Emmanuel Macron paid extras during his trip to Cameroon in 2022 – a claim that was also unverifiable given that AFP's journalists were not on the scene, and they had no evidence to suggest that this was not the case.

[20] Crowdtangle only collects data for public posts from public pages, groups and verified profiles. It does not collect information on private accounts. Meta shut down Crowdtangle in 2024, but the tool was still in place at the time of our study (2021-2023).

– selected from the set of categories observed in the verified stories (such as climate, COVID, elections, Ukraine, inflation, and others).[21]

Working sample As highlighted above, the partnership with AFP was formalized in October 2021, and our observer began working with the AFP Factuel team on a daily basis on December 1st, 2021. The first period until February 15th, 2022 was considered as a trial period, needed to understand how the AFP Factuel team worked, and the different data collection tasks that needed to be performed. The observer's work ended in June 2023.

Overall, $1,479$ stories were discussed between February 16th, 2022 and June 30th, 2023. The final sample used for the purpose of our analysis is smaller, however, since some stories do not generate usable data, as explained below. First, since Meta does not allow ratings of posts or statements by politicians, these stories cannot be included in our current analysis. This reduces the sample to $1,310$ stories. The second case of stories that are not exploited involve situations where there are no posts identified on Crowdtangle related to the story (this would typically be the case of stories that originate on Twitter but were not be picked up on Facebook) or there are no posts within a -5 days / $+10$ days time window around the time of the discussion.[22] After excluding these stories, we obtain a working sample of 944 stories.

Further data collection: Facebook accounts To analyze whether Facebook accounts modify their behavior after being fact-checked, we also collected data on the activity of a subset of accounts that appeared in our dataset. The data collection process was launched in three waves in August 2022, December 2022 and June 2023, respectively. In each wave, the $2,000$ accounts with the most posts in our dataset (i.e posts associated with stories discussed in the AFP meeting) on that date are tracked.[23] Overall, $3,223$ out of the $8,054$ accounts appearing in our data are tracked this way in at least one wave. We quest Crowdtangle for all the posts created by these accounts since December 2021. A more detailed description of the account data collection process can be found in the online Appendix Section A.1. Appendix Table C.5 report the summary statistics of those accounts.

2.3 Main variables of interest

We describe here the main variables we constructed based on this data collection process. We distinguish story-level, post-level and account-level variables.

[21] As can be seen from Table 1 (and as described in more detail below), the topics of the stories are centered around a small number of issues.

[22] For instance, AFP journalists discussed a video on August 17th 2022 claiming to be a protest by Dutch agricultural workers. The video was in fact an old footage from 2015 filmed in Belgium. Despite the lack of public posts on Facebook, the journalists decided to produce a fact-check due to the virality of the story on Twitter. The fact-check is not included in our working sample given that we do not have data on the circulation of this story on Facebook.

[23] We "only" focus on the top $2,000$ accounts in each wave for data management and storage reasons (the size of the data dump with $2,000$ accounts for these three waves nearly reaches 1TB).

Story-level variables At the story level, we measure four different variables. First, **origin**: the origin of the story as reported by the journalist (e.g. the Facebook claim, Twitter, etc.). Second, **fact-checked**: an indicator variable equal to one if the story is fact-checked, and to zero otherwise. Third, **topic**: the topic of the story considered for fact-check (whether or not it is ultimately fact-checked). Fourth, **reasons**: for the stories that were considered for fact-checking but ultimately not fact-checked, we record the reasons for not fact-checking them. For a given story, several reasons may be invoked.[24]

Post-level variables At the post level, we capture four time-varying variables. First, the **number of engagements** by type (shares, likes, comments) that we computed on an hourly basis using Crowdtangle. Second, we measure for each post whether it disappears from Crowdtangle during the 15-day period where we collect its activity data. There are two possible reasons for disappearance: either the post is deleted or the account becomes private.[25] We cannot distinguish between the two reasons for the disappearance, but later in the text we refer to this variable as **deleted**, since this is the most likely reason for disappearance. Third, we collect from the Facebook tool the information on contents that were **rated** and the type of rating ("False", "Missing Context", etc.). As already highlight, a content could be a link, a video or a photo (in which case all the posts circulating this link or video/photo will be flagged) or could be an individual post. Finally, in addition to the rating information, we also used scrappers to identify whether flags could be found on the individual posts. These scrappers were launched in three different months : September 2022, May 2023 and June 2023. If a post was deleted before the script was run, we cannot recover its flagging status. Thus our main analysis at the post level will rely on the rating information, although flagging will be used in robustness exercises.

Finally, we further aggregate the post-level data at the story level using all posts related to a given story, so that we can obtain, for instance, all the engagement measures at the story level. We also aggregate the type of flag – such as "False" or "Partly false" – at the story level.[26]

Account-level variables At the account level, we measure the **number** of posts published by the account at account $\times$ day level, excluding posts that are in our main dataset (i.e correspond to stories that were discussed during the morning editorial meetings). We then aggregate this variable and calculate the cumulative number of posts published by each account. For each fact-check, we collect information on the cumulative number of posts starting 20 days prior to story's consideration date.

[24] As explained above, we distinguish five possible reasons: (i) a lack of resources, (ii) a lack of virality, (iii) the fact that the story is probably true, (iv) the fact that the fact-check would have been unfeasible (because of a lack of sources or factual elements needed to fact-check the claim) and (v) the fact that the fact-check has already been done. We built these categories after extensive discussions with AFP, and recorded them based on the evaluation of the editor-in-chief. Of course, this evaluation can be considered as "subjective" but it is the most relevant one in our context given that the editor-in-chief is the one who ultimately decides whether or not a story had to be fact-checked.

[25] The Facebook data only allows us to observe public accounts (not private ones), so if an account changes status and becomes private, its post would disappear from Crowdtangle.

[26] In around 10% of cases, a story has multiple types of flags. When this is the case, we take the mode of the type of flags, and if there is a tie, we choose the more severe type of flag.

3 New evidence on the fact-checking process

Our data offers a unique perspective on the fact-checking process, grounded in the decisions of the world's largest fact-checking organization. Below, we summarize several key insights, and provide descriptive statistics in Table 1. Providing descriptive evidence is especially important in this context, where little is known about the actual process behind the production of fact-checks.

The Facebook algorithm is not central in the fact-checking process According to Table 1, where we present descriptive statistics for our working sample, 38% of the stories discussed during the morning meetings originate from Twitter, while 26% come from Facebook sources other than the Facebook's algorithm. In both cases, journalists may either monitor known misinformation-spreading accounts and pages, or screen newly posted content. Additionally, around 14% of the proposed stories are French-language versions of claims that have already been fact-checked by other AFP desks in different languages. Notably, only 16% of the proposed stories originate from the Facebook's Claim system – an algorithm that flags potentially false news for accredited fact-checkers, a key limitation we discuss in Section 5.3.

Regarding the origin of the stories, fact-checked and unchecked stories are similar with two exceptions: unchecked stories are more likely to originate from Twitter and fact-checked stories are more likely to be translations. This makes sense since Twitter stories are less likely to have associated posts on Facebook, which is a criterion for selection given that fact-checkers are paid for flagging at least one post on Facebook, while translations involve very low costs and can be done quickly. This leads us to make further restrictions while setting up our empirical strategy, as described in Section 4.

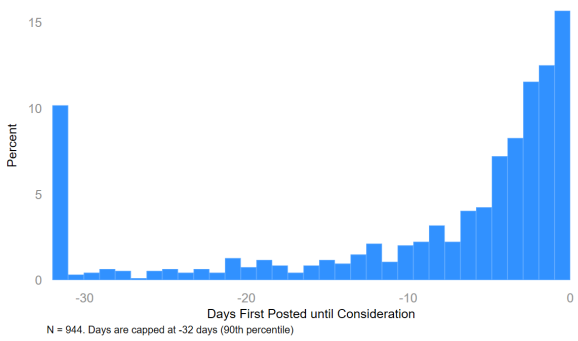
The main reason for not fact-checking a story is that the verification seems infeasible According to Table 1, the most common reason a story is not fact-checked – accounting for 44% of cases – is that the fact-checkers judge the claim to be infeasible to investigate.[27] A sizable part of these claims involve scientific facts difficult to establish on the spot. For example, the story entitled "Bill Gates was disappointed to see that Omicron had immunized people faster than vaccines" is typically unverifiable, as there was no way at the time to accurately estimate the amount of people that contracted the Omicron variant of COVID or estimate its effect in halting the pandemic. Another frequent reason (25% of cases) is that the editorial team considers the story likely to be true. It is AFP policy not to publish fact-checks on stories that are ultimately shown to be accurate.[28] In 19% of cases, the story is rejected due to insufficient virality. Finally, 18% of the stories considered are left aside because of resource constraints. For example, a claim that excess mortality in France and other countries could be linked to vaccination was deemed too resource-intensive to investigate thoroughly, especially given the need to examine data from multiple countries. These reasons will play a key role in shaping our identification strategy in the following section.

[27] Note that the reasons are not mutually exclusive and that the editor could mention several of them as a justification.
[28] This policy is not universally followed across fact-checking organizations.

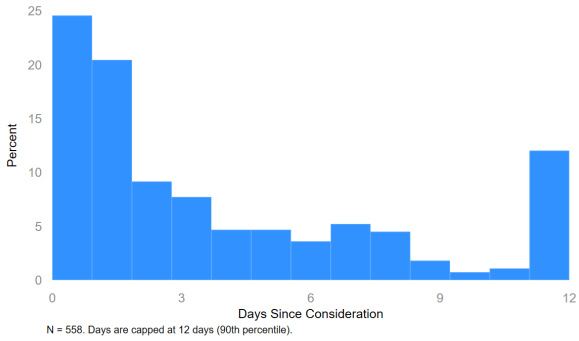
Table 1: Stories considered for fact-checking: descriptive statistics

	Total (N = 944)	Fact-checked (N = 558)	Not fact-checked (N = 386)	Difference	P-value
Story Origin					
Twitter	0.38	0.32	0.47	-0.15	0.00
Facebook Claim	0.16	0.17	0.15	0.01	0.66
Other Facebook	0.26	0.28	0.24	0.04	0.16
WhatsApp	0.06	0.04	0.07	-0.03	0.06
Translation	0.14	0.20	0.06	0.15	0.00
Media	0.01	0.01	0.01	0.00	0.69
TikTok	0.02	0.03	0.01	0.01	0.28
Other Social Media	0.02	0.01	0.02	-0.01	0.26
Topics					
Covid	0.19	0.16	0.24	-0.08	0.00
Ukraine/NATO	0.19	0.18	0.20	-0.02	0.43
Vaccines	0.16	0.13	0.20	-0.08	0.00
Climate	0.10	0.10	0.11	-0.01	0.55
Other	0.23	0.39	0.00	0.39	0.00
Activity at Discussion Date					
N. Active Posts	12.24	13.22	10.81	2.42	0.20
Total Engagement	5831.62	6116.58	5419.67	696.91	0.60
Shares	2036.75	2161.33	1856.65	304.68	0.63
Comments	584.70	643.22	500.11	143.12	0.21
Likes	2423.81	2560.72	2225.90	334.82	0.57
# days btw posting and AFP discussion	46.56	51.69	39.14	12.54	0.41
Flags					
False		0.68			
Altered Media		0.03			
Partially False		0.04			
Missing Context		0.14			
Satire/True		0.11			
Reasons for Not Fact-checking					
Infeasible			0.44		
Probably True			0.25		
Lack of Ressources			0.18		
Lack of Virality			0.19		
Deletion at Discussion Date					
Has Deleted Post	0.19	0.24	0.12	0.13	0.00

Notes: The table provides descriptive statistics on the 944 stories considered for fact-checking and included in our working sample. An observation is a story. Column (1) provides statistics for all the stories, and Column (2) (respectively Column (3)) provides statistics for the stories that are fact-checked (respectively not fact-checked).



(a) Time between first post and consideration



(b) Time between discussion and publication of the fact-check

Notes: The figure describes the length of the fact-checking process. Sub-figure 1a plots the distribution of the time interval between the date of the publication of a story on Facebook and the date it is first discussed by the AFP Factuel team. Each bin is of size one day. Sub-figure 1b plots the distribution of the time interval between the discussion of a story to be fact-checked by the AFP Factuel team and the publication of the fact-check. Each bin is of size one day.

Figure 1: Length of the fact-checking process: Descriptive evidence

The fact-checking process is long There is often a delay between when a story is first posted and when it is discussed by the AFP team. As shown in Figure 1a, while many stories are addressed shortly after they appear, 44% of the stories discussed by AFP Factuel had been circulating for more than four days prior to being considered. Notably, 10% had been circulating on Facebook for over 30 days before being discussed. Even after a story is selected, producing the fact-check takes additional time. In about one-third of cases, the investigation and writing are completed within a day, though the process can take several days. Figure 1b illustrates the distribution of the time interval between a story's discussion during the morning meetings and the publication of the corresponding fact-check. 33% of the fact-checks take more than four days to be written.

Around 60% of the posts related to a fact-checked story are rated As explained above, once the fact-check is written, the journalist in charge of the fact-check searches on Facebook for posts that circulate the fact-checked story and decides how much effort to put into the process.[29] As a result, not all posts that relate to the same fact-checked story are rated. For an illustration, compare online Appendix Figures B.4 and B.7, which both report posts that were published in the same Facebook group on different days with almost identical content, but only one was finally flagged (B.4).

Appendix Table C.2 provides descriptive statistics on the posts, focusing on the posts related to fact-checked stories in the working sample. Roughly 61% of the posts that are associated with a fact-checked story are rated. This will be the basis of our second identification strategy. The posts that are rated are systematically different from the others, including at the time of the first rating; in particular, unrated posts appear to receive more engagement. To deal with this issue, we use propensity score matching in our empirical analysis.

[29] If the source of the story is Facebook, then it is easy for the journalist to find associated posts; if it is not, they usually rely on links or keywords.

Fact-checked accounts are composed of a few superspreaders and a majority of one-time offenders As mentioned above, fact-checkers often monitor accounts known for spreading misinformation when selecting stories to propose during the morning editorial meetings. Appendix Table C.6 lists the 30 accounts most frequently involved in fact-checked stories. Those are often groups and pages dedicated to anti-vaccine or anti-Macron narratives, as their names indicate.[30]

However, those "superspreaders" constitute only a small fraction of the fact-checked accounts. The distribution of the number of stories per account is very long-tailed, as shown in online Appendix Figure B.8. Nearly three-fourths of the accounts in our sample ($5,857$ out of $8,054$) are associated with one single story, and 12% with only two. Appendix Table C.5 provides descriptive statistics on these accounts.

4 Identification and empirical strategy

The objective of our identification strategy is to identify "counter-factuals" for the content that is fact-checked by AFP Factuel, in order to causally estimate the impact of fact-checking on the spread of misinformation and on the behavior of users, in a Difference-in-Differences framework, one at the story-level, the other at the post-level.

4.1 Impact of fact-checking on engagement

4.1.1 Story-level empirical strategy

Difference-in-Differences approach We estimate the following model:

$$\overline{Y}_{st} = \alpha + \beta \text{Fact-checked}_s \times \mathbb{1}_{t>t_{sc}} + \delta_s + \gamma_t + \varepsilon_{st} \tag{1}$$

where s indexes the stories and t the time. In our preferred empirical specification, the dependent variable, $\overline{Y}_{st}$, is the logarithm of the cumulative engagement of the posts p related to story s (which we denote $p \in I(s)$) at time t (i.e. $\overline{Y}_{st} = \sum_{p \in I(s)} Y_{pst}$). The higher $\overline{Y}_{st}$, the higher the engagement with the story. We control for story (δ_s) and time fixed effects (γ_t).[31]

Our main explanatory variable, Fact-checked$_s \times \mathbb{1}_{t>t_{sc}}$, is the interaction between an indicator variable equal to one if the story has been fact-checked and to zero otherwise (Fact-checked$_s$), and an indicator variable equal to one after the story has first been considered by the AFP Factuel team and to zero before ($\mathbb{1}_{t>t_{sc}}$).[32] The coefficient β measures the causal impact of the fact-check on engagement. If fact-checking is effective at reducing the spread of misinformation, β should be negative. We cluster the standard errors at the level of the stories.

[30] For instance, "For the resignation of Emmanuel Macron" or "NO TO THE VACCINATION PASS".

[31] In the main specification, as explained below, engagement always takes strictly positive values and thus a specification in logs is appropriate.

[32] t_{sc} is the time at which the story has first been considered by the AFP factuel team. We use the time of consideration as our benchmark time to define the pre/post treatment period – rather, for example, than the time of the fact-check – given that this time is accurately defined both for the stories that are fact-checked and for those that are not.

The key identification assumption is that the treated and control stories would have followed similar trends absent fact-checking. Note that, when a journalist proposes a story in the morning meetings, they believe and argue the case that it should give rise to a fact-check. This process supports the notion that all proposed stories are, ex ante, relatively comparable. As observed in Table 1, fact-checked and unchecked stories are indeed relatively similar, except that fact-checked stories are less likely to originate on Twitter, more likely to be translations and tend to generate more engagements (though this difference is not statistically significant). Nevertheless, to increase the likelihood that our identification assumption is met, we impose restrictions on both the control and treatment groups. First, we remove from the sample stories that were not fact-checked because of a lack of virality, since the selection was precisely done according to the future trends. Second, we remove stories that are translations, since the process for selecting these is very different: there is very little cost involved in translating and these stories are therefore typically much more likely to be selected. Third, we exclude stories that appeared (i.e had their first engagement) less than 12 hours before the editorial meeting. For such stories, we cannot credibly assess the absence of pre trends because the time window is too short. We show in Appendix Table C.7 that the stories that appear on Facebook the night before the meetings are much more likely to originate on Twitter; the trends on Facebook are therefore hard to predict, making it very difficult to establish why some of these stories were selected and not others. The sample obtained when imposing these restrictions has 589 stories (see online Appendix Table C.4 for descriptive statistics on this set of stories).

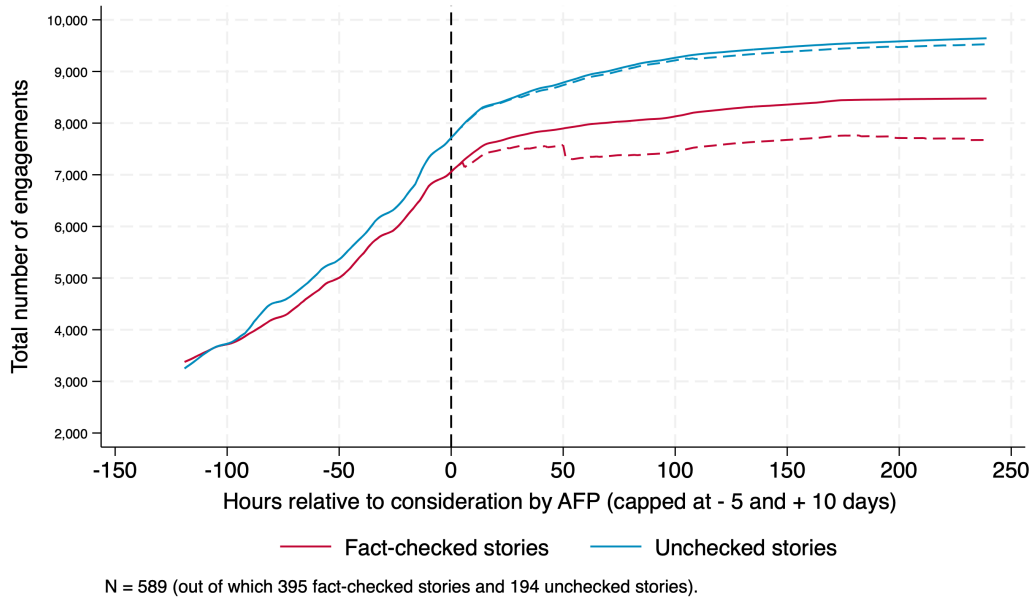
Figure 2 plots the raw trends in engagements with posts associated with fact-checked vs. unchecked stories in this restricted sample. The solid lines represent the total cumulative engagements with the story, while the dashed lines represent the cumulative engagements when setting engagements for a post at zero when the post is deleted. Figure 2 shows that the pre-trends are relatively parallel, but a gap remains. To improve the comparability between treated and control stories prior to estimating treatment effects, we therefore implement a covariate-balancing matching procedure. Specifically, we employ propensity score matching (PSM) based on a set of pre-treatment covariates controlling for engagement pre-treatment.[33] The procedure generates matched samples by reweighting control observations to closely mirror the distribution of covariates in the treated group and, in our main specification, equation (1) is estimated weighting observations accordingly. We consider a number of robustness exercises changing this matching procedure.

Event study We also perform an event-study analysis to provide evidence consistent with the parallel trends assumption. Using four-hour intervals to measure engagement, we estimate the following model:

$$\overline{Y}_{st} = \sum_{h=-12}^{192} \beta_h \text{Fact-checked}_s \times \mathbb{1}_{t-t_{sc}=h} + \delta_s + \gamma_t + \varepsilon_{st} \tag{2}$$

where, as before, t_{sc} is the time at which the story is first considered by the AFP Factuel team. We are interested in the β_h coefficients; the parallel trends assumption requires the coefficients β_{-12} to β_0 to be close to zero and not statistically significant. We expect β_h to gradually decrease over time when $h > 0$

[33] In the baseline specification, we control for engagement at time 0 and twelve hours before consideration.



Notes: Each line plots the average sum of all engagements (shares, comments, posts or other reactions) of posts associated with a story for fact-checked stories (red line) and unchecked stories (blue line) relative to the date of consideration of the story by the AFP Factuel team. The solid lines represent the cumulative engagements of all posts that existed in the story, while the dashed lines represent the engagement of active posts in the story (i.e. deleted posts are considered to have 0 engagement).

Figure 2: Raw trends in engagements: Fact-checked stories vs. unchecked stories

as fact-checks are gradually produced (as shown in Figure 1b above, one third of the fact-checks indeed take more than four days to be produced).

4.1.2 Post-level empirical strategy

Our second empirical strategy exploits quasi-experimental variation at the post level. As previously described, once a story is fact-checked, the journalist responsible for the fact-check also rates several posts associated with that story. However, due to time constraints and the lack of incentives in the agreement with Facebook to rate all posts, journalists typically tend to rate only a share of posts associated with the story. In this second approach, we study the evolution of the engagement with the rated posts (treated posts) compared to the non-rated ones (control posts). Compared to the story-level specification, only the stories that are fact-checked by the AFP Factuel team are included in the sample (within-story identification strategy).

Difference-in-Differences approach We estimate the following Difference-in-Differences model:

$$Y_{p(s)t} = \zeta Rated_{p(s)} \times \mathbb{1}_{t>t_{sr}} + \mu_{p(s)} + \lambda_{s,t} + \varepsilon_{pst} \tag{3}$$

where, as before, p index the posts, s the stories and t the time. The dependent variable of interest, $Y_{p(s)t}$,

is a variable measuring the logarithm of the cumulative engagements with a post p on story s at time t. Only fact-checked stories are included.

Our explanatory variable of interest, $Rated_{p(s)} \times \mathbb{1}_{t>t_{sr}}$, is the interaction between an indicator variable equal to one if the post p on story s has been rated and to zero otherwise ($Rated_{p(s)}$), and an indicator variable equal to one after the first rating on story r and to zero before ($\mathbb{1}_{t>t_{sr}}$). Hence, contrary to the story-level identification strategy, time is taken relative to the date of the first rating within a story rather than the time of consideration. We control for post fixed effects ($\mu_{p(s)}$) as well as for story-time fixed effects ($\lambda_{s,t}$), capturing in a flexible way shocks that can affect the popularity of a story. Standard errors are clustered at the story level.

Our identification assumption is that the last post that is rated is "similar" to the first one that is not, i.e. that the fact-checker decides to stop rating posts at some point for some exogenous reason (e.g. a lack of time or the need to move on to another fact-check), due to a lack of incentives to pursue the effort.[34] Although we cannot directly observe which unrated post was considered last, we adopt the following procedure to construct an appropriate group of control posts. First, as in the story-level analysis, we exclude posts that appeared less than 12 hours before the first rating occurred, given that we cannot verify the absence of pre-trends for those posts. Second, we apply propensity score matching to better balance treated (rated) and control (unrated) posts. We use as covariates to perform the matching engagement at time 0 and -12. In our main specification, matching is performed across stories rather than within them. As a robustness check, we also restrict matching to occur within stories, which further limits the sample size, since successful matching requires sufficient overlap in covariates between treated and control posts within each story.

Event study The corresponding event-study specification is:

$$Y_{p(s)t} = \sum_{h=-12}^{192} \zeta_h Rated_{p(s)} \times \mathbb{1}_{t-t_{pr(s)}=h} + \mu_{p(s)} + \lambda_{s,t} + \varepsilon_{pst} \tag{4}$$

where $t_{pr(s)}$ is the time of the first rating of a post within a fact-checked story. We are interested in the ζ_h coefficients; the parallel trends assumption can be assessed through the pre-treatment coefficients ζ_{-12} to ζ_{-4}. As discussed, we control for story-time fixed effects $\lambda_{s,t}$, which take out the overall trend of a fact-checked story. If rating a post is effective in reducing its spread on Facebook, we expect the coefficient ζ_h to be negative and statistically significant when $h > 0$.

4.2 Behavioral responses

By design, we expect the number of engagements with fact-checked stories and the rated posts to decrease post-treatment, since Meta states that rated content is automatically downgraded, though its ac-

[34] We believe that this assumption is reasonable. First, note that the AFP is only rewarded for the first post fact-checkers rate (so journalists do not have any incentive to maximize the number of posts they rate, nor there is an economic trade-off between rating more posts or considering a novel story). Second, this assumption is consistent with several discussions we had with the members of the AFP Factuel team regarding their decisions on whether to rate additional posts.

tual policy remains non-explicit.[35] We therefore also focus on other outcomes which are clearly identifiable as user strategic reactions.

Deletions As appears in Figure 2, there is a sharp drop in the number of engagements with active posts following a fact-check (dashed lines), suggesting that users appear to react to the treatment by deleting posts. To investigate whether this is indeed the case, we explore whether fact-checking a story influences the rate at which posts are deleted. As explained in Section 2.2, we do not observe deletions directly, but since we follow posts on Crowdtangle for 15 consecutive days from the time they enter our database, we can observe the time at which they disappear from Crowdtangle (if they do so within this 15-day time window).[36] Given this measurement methodology, we cannot observe deletions before the discussion date (by the AFP Factuel team), which is the earliest date on which any post enters our database. We therefore estimate the following model:

$$\overline{D}_{st} = \sum_{h=0}^{192} \beta_h \text{Fact-checked}_s \times \mathbb{1}_{t-t_{sc}=h} + \delta_s + \gamma_t + \varepsilon_{st} \tag{5}$$

where $\overline{D}_{st}$ is the share of posts related to story s, present at date zero, that have been deleted between time 0 and time t. The independent variable of interest, Fact-checked$_s \times \mathbb{1}_{t-t_{sc}=h}$, is the interaction between an indicator variable equal to one if the story has been fact-checked and to zero otherwise (Fact-checked$_s$), and indicator variables defined with respect to the time at which the story has first been considered by the AFP factuel team ($\mathbb{1}_{t-t_{sc}=h}$). We control as before for story (δ_s) and time (γ_t) fixed effects, and cluster the standard errors at the level of the stories.

Account general activity We also explore whether users adapt their subsequent posting activity when one of their posts has been fact-checked. In this case, the unit of observation is the account and we rely on the accounts we tracked as described in Section 2.2. We define date 0 as the first date on which an account was considered for fact-checking within our study period (i.e. a post that the account published was promoting a story considered for fact-checking by the AFP Factuel). This approach allows us to consistently measure the "recidivism rate," referring to the interval between the first observed fact-check of an account's content and its subsequent involvement in a further fact-checked story.[37]

We then estimate a specification similar to the one used in our story-level empirical approach, where the outcome variable is the activity of the account:

[35] See Meta's Transparency Center: "Once a fact-checker rates a piece of content as False, Altered, or Partly False, or we detect it as near identical, it may receive reduced distribution on Facebook, Instagram and Threads. We dramatically reduce the distribution of False and Altered posts, and reduce the distribution of Partly False to a lesser extent. For Missing Context, we focus on surfacing more information from fact checkers. Meta does not suggest content to people once it is rated by a fact-checker, which significantly reduces the number of people who see it. "

[36] As noted above, we interpret this as deletions, although it could also correspond to an account switching its status from public to private.

[37] Note that the number of "repeat offenders" – to use Facebook wording – is not high: as highlighted above, around 72% of the accounts are involved in only one story considered for fact-checking and another 11% of the accounts are involved in only two stories.

$$\overline{Y}_{at} = \sum_{d=-5}^{+15} \beta_d \text{Fact-checked}_a \times \mathbb{1}_{t-t_{ac}=d} + \delta_s + \gamma_t + \varepsilon_{at} \tag{6}$$

a index the accounts and t the time. $\overline{Y}_{at}$ is the logarithm of the cumulative number of posts published by the account and not related to the story discussed by AFP. We initiate this cumulative measure 20 days before the time of consideration and exclude accounts that were entirely inactive between -20 and -5 days before this time, since we do not want accounts exclusively focused on a single story.[38]

The independent variable of interest, Fact-checked$_a \times \mathbb{1}_{t-t_{ac}=d}$, is the interaction between an indicator variable measuring whether the first story in which the account participates in our data – i.e. that is discussed during an editorial meeting – was ultimately fact-checked by the AFP Factuel team (Fact-checked$_a$) and indicator variables for time ($\mathbb{1}_{t-t_{ac}=d}$). As in the previous sections, to strengthen our identification, we apply propensity score matching to better balance treated and control groups. We use as covariates to perform the matching the cumulative number of posts at date 0 and at -5.

Recidivism Finally, we investigate whether the fact of publishing a post in a fact-checked story affects the long-term probability that an account will publish another post in another story considered for fact-checking. To tackle this question, we estimate a Cox proportional-hazard model:

$$h(t|X_a) = h_0(t) \exp(\beta_1 \text{Fact-checked}_a + \mathbf{X'_{at}}\beta_2) \tag{7}$$

where $h(t|X_a)$ represents the hazard function that an account a publishes a content related to a new story considered for fact-checking by AFP Factuel in day t after the first time it has done so. The indicator variable Fact-checked$_a$ takes the value one if the first AFP story in which the account was involved was fact-checked by the AFP, and is equal to zero otherwise. $\mathbf{X'_{at}}$ is a vector of time-varying account controls, including day-of-the-week fixed effects and indicator variables for the journalists present (we control for this set of indicator variables given that certain journalists can have a higher probability of monitoring some Facebook accounts rather than others, which could otherwise affect "by construction" the observed probability of recidivism). We are interested in the coefficient β_1, which represents the effect of the first story being fact-checked on the probability of publishing other posts related to new stories considered for fact-checking ("recidivism").

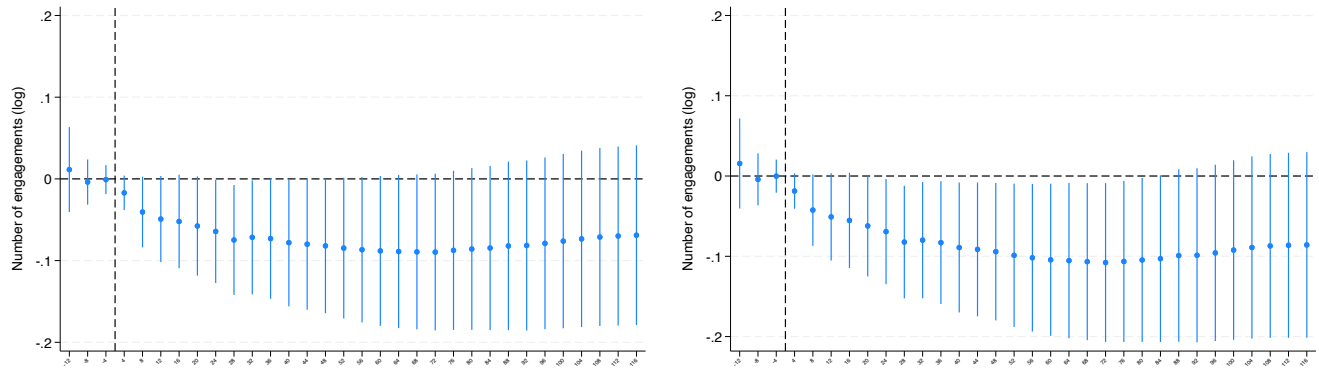
5 Empirical results

5.1 The impact of fact-checking on engagement

5.1.1 Story-level analysis

We first report the event-study estimates (estimation of equation (2)). In the propensity score matching procedure, we use as covariates variables measuring the activity close to the time of consideration by the

[38] This also implies that for all the accounts in our data the cumulative number of posts is always positive, and we can thus use the logarithm of this variable.



(a) All types of flags **(b)** Only stories rated "False"

Notes: The figure reports the results of the estimation of equation (2). An observation is a story. Observations are weighted using weights derived from a propensity score matching procedure using as covariates engagement at consideration date and engagement 12 hours before consideration. The dependent variable is the logarithm of the cumulative number of engagements. Sub-figure 4a includes all the stories considered for fact-checking by the AFP Factuel team. In Sub-figure 4b, only the stories rated "False" are included in the treated group.

Figure 3: Effect of fact-checking on the number of engagements: Story-level analysis, Event-study estimates

AFP Factuel team: the number of engagements at the consideration date and the number of engagements 12 hours before consideration. After matching, we retain 93% of the treated (i.e. fact-checked) stories and 92% of the controls (i.e. unchecked) stories. Standardized differences for all covariates fall below 0.05, indicating strong balance. We drop stories with no common support and proceed with the estimation of equation (2) using the matched sample, weighted by the matching weights.

Figure 3 shows the results, in Panel (a) for all types of flags and in Panel (b) focusing on fact-checked stories rated as false. We first observe the validity of the parallel trends assumption: before the date of consideration, there is no difference between the number of engagements of fact-checked vs. unchecked stories. The coefficients β_{-12} to β_{-4} are indeed close to zero and not statistically significant.

Furthermore, it appears that, compared to the non-fact-checked stories, the popularity of the fact-checked stories decreases following the fact-check. Interestingly, given that we define the pre/post period with respect to the time of the first consideration by the AFP Factuel team, we see that the drop happens gradually (as highlighted in Section 3, the fact-checking process indeed takes time: more than a day in two-thirds of cases). Overall, we see that, compared to the non-fact-checked stories, the publication of the fact-check leads to a drop of 8% on average in the total engagement with posts associated with a story, as reported in Table 2 Columns (1) and (2), where we estimate equation (1).

Furthermore, sub-Figure 4b shows that the results hold when we only consider the stories rated false. In fact, those stories mainly drive the results (see also Columns (3) and (4) of Table 2). For the stories receiving more nuanced ratings ("Partly False" or "Missing Context"), there is a much smaller, not significant decrease in engagement (online Appendix Figure B.9). We return to this point in Section 5.3, when discussing the policy implications of our results.

Table 2: Effect of fact-checking on the number of engagements: Story-level analysis, Difference-in-Differences estimates

	All types of flags		Stories rated False		Fast rating	
	(1)	(2)	(3)	(4)	(5)	(6)
Post * Fact-checked stories	-0.07	-0.08	-0.09*	-0.09*	-0.11**	-0.11**
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Story FEs	✓	✓	✓	✓	✓	✓
Time FEs	✓	✓	✓	✓	✓	✓
Day of the week		✓		✓		✓
Observations	18,227	18,227	15,092	15,092	12,056	12,056
Nb of clusters (stories)	553	553	458	458	366	366
Mean DepVar	6	6	6	6	6	6
Sd DepVar	2.59	2.59	2.56	2.56	2.61	2.61

Notes: * p<0.10, ** p<0.05, *** p<0.01. The Table provides the results of the estimation of equation (1). An observation is a story. Observations are weighted using weights derived from a propensity score matching procedure. The dependent variable is the logarithm of the cumulative number of engagements. Columns (1) and (2) include all the stories considered for fact-checking by the AFP Factuel team. In Columns (3) and (4), only the stories rated "False" are included in the treated group. In Columns (5) and (6), the treatment group includes only the stories that take less than the median time (2 days) to be checked. Odd-numbered columns control for day of the week fixed effects.

Heterogeneity depending on speed Is the impact of fact-checking on engagement affected by how quickly the fact-check is produced? The median production time for a fact-check in our sample is two days. We therefore define "fast" fact-checks as those completed in less than two days. First, we explore what drives the journalists to publish a fact-check more quickly (online Appendix Table C.8). While current engagements with the story are not a significant predictor of faster fact-checking, we see that the topic of the story is important, with fact-checks related to the war in Ukraine completed more quickly, whereas those related to climate change taking longer.[39]

We then investigate whether the magnitude of the effects varies with the speed at which the fact-check is produced. We find that the effect of a fact-check is three times larger when the fact-check is published within two days of the story being discussed in the morning meeting (online Appendix Table C.9 Columns (2) and (3)).[40] These findings suggest a clear trade-off between the quality of the fact-check produced and its impact on the circulation of the story, a trade-off we explore further in Section 5.3.

[39] The excluded category includes a mix of very different topics. According to the informed discussions we had with the AFP Factuel journalists, the fact that – on average – it takes longer to produce a fact-check on climate-change related stories than on Ukraine-related stories is most probably due to the type of misinformation in question. For climate change-related stories, it is often necessary for the journalists to consult experts on the subject to describe climatic or meteorological phenomena such as rising sea levels. On the contrary, fake news stories on Ukraine tend to be much more "basic", with no need for the journalist in charge to get in touch with experts. For example, writing a fact-check about a photo reproduced out of context or a fake graffiti is relatively quick for an AFP Factuel journalist because it requires skills that the AFP already has in-house (such as reverse image search or geolocation of an image).

[40] In Appendix Table C.9 Columns (4) and (5), we perform the same exercise, differentiating between stories that were discussed more or less quickly after they appear. Fact-checking is twice as impactful for stories that were considered within four days of the first posting on Facebook than for those considered after four days (though the coefficients are not significantly different from zero).

Heterogeneity depending on the topic Finally, we investigate whether the magnitude of the effects varies depending on the topic of the stories considered for fact-checking; online Appendix Table C.10 reports the results. As highlighted in Table 1, the most common topic of the stories considered for fact-checking during our study period is the Ukraine war and NATO. The Ukraine war started a few months after the start of our partnership with AFP. The magnitude of the effect of fact-checking for the subset of stories on that topic is particularly large, with a 24% drop in the number of engagements following a rating.

One explanation for this result is that, as shown in Appendix Table C.8, fact-checks related to the Ukraine war tend to be produced faster. However, this cannot fully account for the greater effectiveness of Ukraine war fact-checks, since we still observe a stronger effect for these fact-checks even when restricting the sample to those produced within two days (online Appendix Table C.11). Another potential explanation for this heterogeneity is that shifting people's beliefs is considerably more difficult when those beliefs are already well established (e.g. climate change-related beliefs[41]). In contrast, the Ukraine war was still in its early stages, and social media users had not yet formed firmly held views.

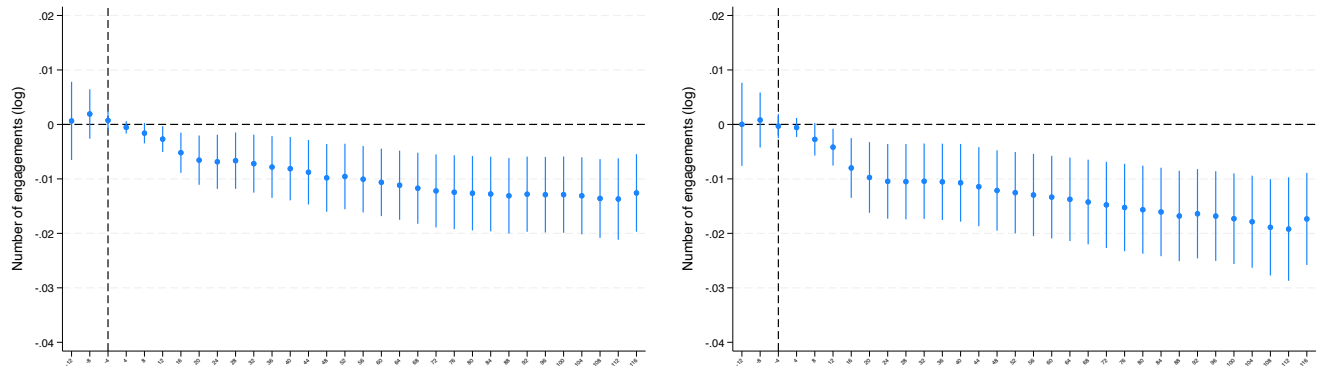
The effects for the other topics are much smaller, with the exception of the stories relevant to the French-speaking offices located outside France, for example in Belgium, Canada and Africa.[42] For stories on topics such as Covid or Climate (i.e mostly climate change), we find no statistically significant effect of fact-checking; note, however, that there is only a very limited number of stories on this topic in our dataset.

Robustness We perform a number of robustness checks on these first results. First, we show that our results also hold when considering each component of engagement separately: shares, likes and comments (online Appendix Table C.12, Columns (2) to (4)). The effects are slightly stronger for comments, but are not statistically different from the effects on shares or likes.

Second, we show that the results are robust to making different restrictions on the sample. In Column (5), we do not use a matching procedure and simply impose the baseline restrictions on the sample. In Columns (6) and (7), we use different covariates for the matching procedure. In Column (6) we use as an additional covariate an indicator variable taking the value of one if the story is more than half a day old and zero otherwise. In Column (7), we add the number of engagements six hours before the discussion date. Note that by adding covariates, we decrease the percentage of the stories that are used in the final analysis. All these additional specifications yield estimates consistent with our baseline results.

[41] On the persistence of climate change attitudes, see e.g. Dewitte (2025).

[42] We classified these stories under the "non-French" label. Those are typically stories centered on and widely spread in former French colonies in Africa; common themes include claims of abuse, post-colonial extraction or discrimination of these countries by the French (for example, a fake story about the petroleum company Total or France prohibiting Niger from signing defense agreement with Russia), or local military or political topics (such as fake claims about protest in Ouagadougou). The few stories involving Belgium, the Netherlands and Canada include a photo mistakenly showing a previous protest in the Netherland.



(a) All types of flags **(b)** Only posts rated "False"

Notes: The figure reports the results of the estimation of equation (4). An observation is a post*time. Standard errors are clustered at the level of the stories. Observations are weighted using weights derived from a propensity score matching procedure using as covariates engagement at consideration date and engagement 12 hours before consideration. The dependent variable is the logarithm of the cumulative number of engagements. Sub-figure 4a includes all the stories considered for fact-checking by the AFP Factuel team. In Sub-figure 4b, only the stories rated "False" are included.

Figure 4: Effect of fact-checking on the number of engagements: Post-level analysis, Event-study estimates

5.1.2 Post-level analysis

We next present the results from the (within-story) post-level analysis (estimation of equations (3) and (4)). This approach directly measures the impact of fact-checking on the interactions users have with the posts. Our final sample contains 240 stories with $4,234$ posts. This represents fewer stories than in the story-level approach, since we only use stories that are fact-checked and for which a portion of posts are not rated (given that – as detailed above – the post-level specification relies, within stories, on the difference in engagement between rated and unrated posts).[43] Appendix Table C.3 provides descriptive statistics on those posts.

The impact of the ratings on the circulation of posts As described in Section 4.1.2 above, to better support our identification strategy we perform propensity score matching between rated and unrated posts.[44] We present the results of the event-study estimates (estimation of equation (4)) in Figure 4. Panel (a) shows the results of the estimation when all the flags are included, and Panel (b) when we only consider the posts rated as false. In both cases, we see that the pre-trends between the treated and the control posts are very similar. We then observe a progressive and very significant drop in the number of engagements with the rated posts compared to the unrated ones. Consistently with the findings of the story-level analysis, the effect is driven by the posts rated "False".

[43] As already noted, we also restrict our sample of posts to those that were present more than 12 hours before the first rating (as in the story-level analysis) in order to be able to estimate the validity of the parallel trends assumption.

[44] Since on average there are only a few posts within each story, we do not restrict the matching to matching within stories in our preferred specification in order to increase the common support between rated and unrated posts and achieve a better balance. However, we show below that our results are robust to restricting the matching to posts published on the same story.

Table 3: Effect of fact-checking on the number of engagements: Post-level analysis, Difference-in-Differences estimates

	All types of flags		Only stories rated False	
	(1)	(2)	(3)	(4)
Post * Rated	-0.011***	-0.011***	-0.014***	-0.014***
	(0.003)	(0.003)	(0.004)	(0.004)
Post FEs	✓	✓	✓	✓
Story * Time FEs	✓	✓	✓	✓
Day of the week		✓		✓
Observations	129,802	129,802	88,680	88,680
Mean DepVar	2.51	2.51	2.52	2.52
Sd DepVar	2.19	2.19	2.18	2.18

Notes: * p<0.10, ** p<0.05, *** p<0.01. The Table provides the results of the estimation of equation (3). An observation is a post*time (standard errors clustered at the story level between parentheses). The dependent variable is the number of engagements. Columns (1) and (2) include all the stories considered for fact-checking by the AFP Factuel team. In Columns (3) and (4), only the stories rated "False" are included in the treated group.

Table 3 reports the Difference-in-Differences estimates. We see that the number of engagements with a rated post drops by 1.1% following the flag, compared to a similar but unrated post. The effect for posts rated false is larger, in the order of a drop of 1.4%. We further discuss the magnitude of those effects below.

Robustness We conduct a series of robustness checks, reported in online Appendix Table C.14, where we focus on the subset of stories rated as false. First, we examine the treatment effects on different components of engagement – comments, shares, and likes. The results are consistent, though attenuated for comments. Second, we test the sensitivity of our results to the matching procedure by limiting matches to posts within the same story (online Appendix Table C.14 Column (5)). This approach narrows the pool of usable stories, as sufficient overlap between rated and control posts is required. To mitigate the loss of sample size, we use only engagement at date 0 as a covariate and show that, if anything, the magnitude of the effect is larger. Finally, we assess robustness to using an alternative treatment definition (Column (6)): whether a post was flagged, based on supplementary data collected by a scraper run after the end of the experiment (in May-June 2023). This measure differs in two important ways from the rated variable used in our preferred specification: first, it does not capture posts deleted prior to scraping, and second it may include flags from fact-checking organizations other than the AFP. Overall, the estimated effects remain broadly consistent, indicating the robustness of our main findings.

Comparison between story-level and post-level analysis Overall, the post-level analysis identifies a very sharp and statistically significant effect of fact-checking on circulation. However, the magnitude of the effect is much smaller than for the story-level analysis. The main reason for this difference is that the set of stories we use in the post-level analysis is different from the general set of stories. Presumably

fact-checkers put more efforts into rating extensively when the story is more likely to circulate.

The two approaches, in any case, capture different dimensions of user behavior. The post-level analysis directly measures the effect of a post being rated on Facebook, whereas the story-level approach captures the broader impact of the fact-check's publication. This includes potential spillover effects on unrated posts linked to the same story. It may also reflect substitution behavior – for example, groups reposting similar content after their original post has been flagged.

5.2 Behavioral responses

Both the story-level and post-level analyses show a sharp decrease in engagement following fact-checking of stories and rating of posts. The effects identified may result from two distinct mechanisms: first, platform-level interventions implemented by Meta (the enforcement channel), and second, behavioral reactions from users (the behavioral channel). The enforcement channel varies depending on the type of rating, as described above: visibility reduction is significantly stronger for posts rated as 'False', and accounts repeatedly sharing such content may even be banned (see e.g., Théro and Vincent, 2022).[45] We find that the observed reductions in engagement are primarily driven by stories or posts rated as "False", rather than those given milder ratings. However, user behavior may also vary by rating type. Posts labeled as "Partly False" or "Missing Context" may cause debate or controversy, potentially stimulating further sharing and commenting. Our engagement-based analysis does not allow us to isolate users' behavioral responses. To address this limitation, we now turn to outcomes that can be more directly attributed to user behavior.

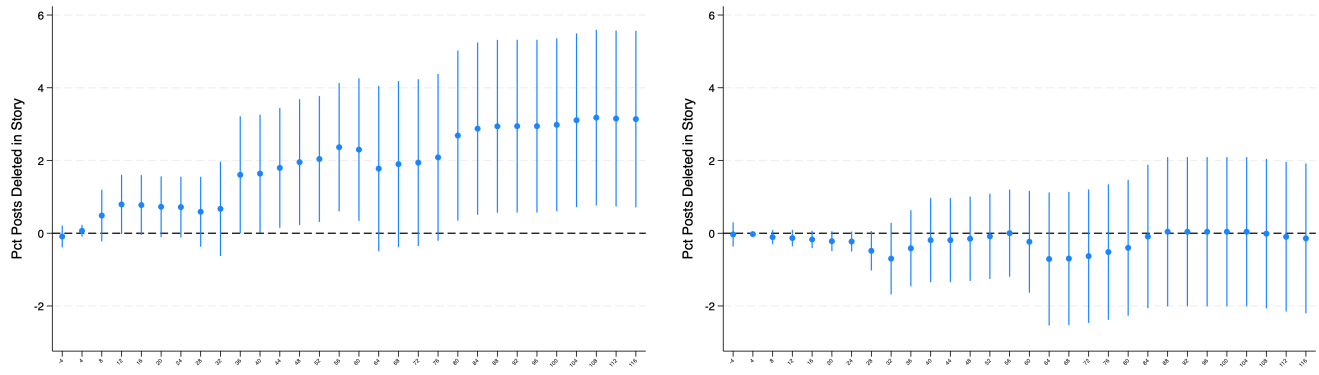
Deletions Deleting a post is a relatively rare outcome and has to be triggered by strong reputation concerns. As reported in Table 1, among non-checked stories in our working sample, only 12% have at least one post deleted at the time of consideration. This figure rises to 24% for fact-checked stories. In this section, we examine formally the extent to which fact-checking affects deletions.

Figure 5 investigates whether the posts related to fact-checked stories are significantly more likely to be deleted than the posts related to similar stories which were ultimately not fact-checked.[46] The outcome variable is the percentage of posts already posted at consideration date that are deleted (estimation of equation (5)). Sub-figure 5a shows that, following the publication of a fact-check related to a story, the number of deleted posts within this story significantly increases. However, as appears on sub-Figure 5b, this effect only holds for the stories rated "False"; we indeed find no statistically significant difference in the tendency to delete posts for the stories rated "Partly False."

Table 4 reports the Difference-in-Differences estimates. Regarding the magnitude of the effect, we show that the share of deleted posts increases by 2.12 percentage points due to fact-checking, and by 2.92

[45] Unfortunately, Facebook does not disclose which accounts are classified as 'repeat offenders', preventing us from directly controlling for this factor. Théro and Vincent (2022) propose a methodology to assess the consequences of the 'repeat offenders' policy using engagement data (see also Vincent et al. 2022). Their analysis, however, focuses exclusively on repeat-offender accounts, whereas our study considers the spread of fact-checked content across all Facebook accounts.

[46] As explained in Section 4, we cannot test for the absence of pre-trends in this analysis since we cannot observe deletions prior to the discussion date.



(a) Only stories rated "False" **(b)** Only stories rated "Partly False"

Notes: The figure reports the results of the estimation of equation (5). An observation is a story*time. Standard errors are clustered at the level of the stories. The dependent variable is the share of deleted posts at date t that were already present at date 0. Sub-figure 5a only includes in the treated group the stories that are rated "False". Sub-figure 5b only includes in the treated group the stories that are rated "Partly False".

Figure 5: Effect of fact-checking on the behavior of users: Story-level analysis, Event-study estimates, Share of deleted posts

percentage points for stories rated false, which represents close to a doubling of the deletion rate. These results suggest that, for some users, being labeled as sharing false information triggers reputational concerns strong enough to prompt them to delete content.

Future behavior Another behavioral dimension we investigate is whether fact-checking can have a disciplining effect on accounts. Specifically, we study whether being fact-checked affects the future behavior of individuals and accounts on other stories. As explained in Section 2, we tracked the daily activity of a number of accounts that had at least one post related to a story discussed by AFP. Exploiting this data, we estimate equation (6). The main variable of interest is $\overline{Y}_{at}$, the logarithm of the cumulative number of posts published by the account and not related to the story discussed by AFP.

Figure 6 shows that accounts whose post appears in a story that is ultimately fact-checked subsequently reduce their posting activity on other stories, relative to accounts whose posts were associated with stories that were considered but not selected for fact-checking by AFP. Prior to the fact-check, there are no pre-trends. The decline is gradual and intensifies after the fifth day, consistent with the time typically required to produce and publish a fact-check. The effect is not very large in magnitude, but reaches 2% at the end of the period.[47]

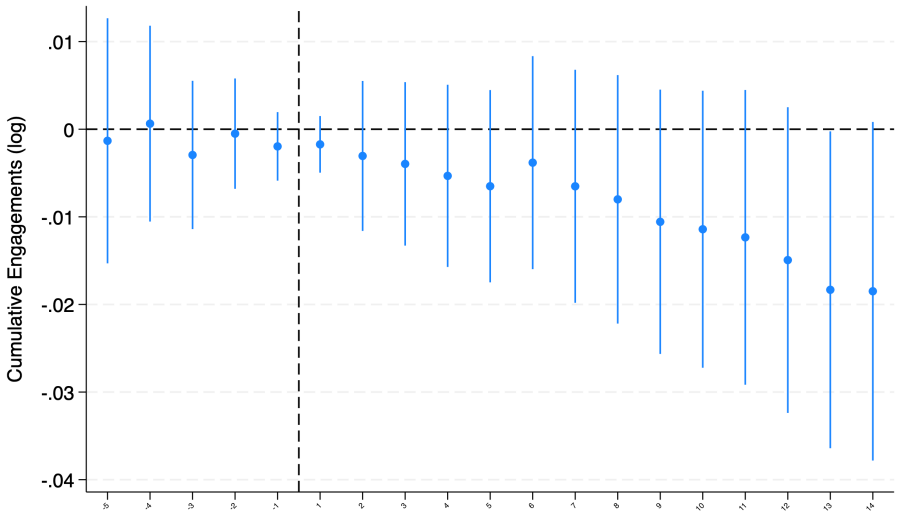
Recidivism The previous result shows that the accounts that are fact-checked become less active. However, it does not shed light on the nature of the content produced, and in particular on whether fact-checking has an effect on the account-holders' behavior with respect to misinformation in the long run.

[47] Table C.15 shows an average reduction of 1% over the 14-day period.

Table 4: Effect of fact-checking on deletions

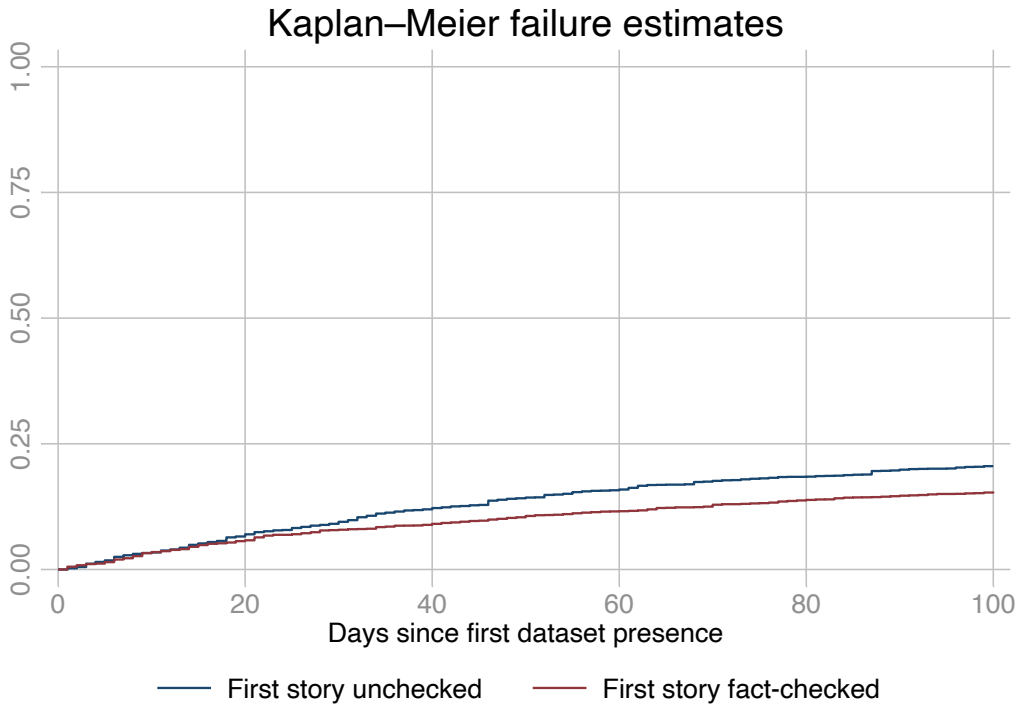
	All types of flags		Only stories rated False	
	(1)	(2)	(3)	(4)
Post * Fact-checked stories	2.13**	2.12**	2.93***	2.92***
	(0.98)	(0.99)	(1.11)	(1.11)
Story FEs	✓	✓	✓	✓
Time FEs	✓	✓	✓	✓
Day of the week		✓	✓	✓
Observations	39,284	39,284	32,208	32,208
Nb of clusters (stories)	644	644	528	528
Mean DepVar	3.38	3.38	3.70	3.70
Sd DepVar	13.97	13.97	15.11	15.11

Notes: * p<0.10, ** p<0.05, *** p<0.01. The Table provides the results of the estimation of equation (5). An observation is a story × time (standard errors clustered at the story level between parentheses). The dependent variable is the share of deleted posts at date t that were already present at date 0. Columns (1) and (2) include all the stories considered for fact-checking by the AFP Factuel team. In Columns (3) and (4), only the stories rated "False" are included in the treated group.



Notes: The figure reports the results of the estimation of equation (6). Observations are weighted using weights derived from a propensity score matching procedure using as covariates the cumulative number of posts at day -5 and day 0 (the day of first story discussion). An observation is an account × day.

Figure 6: Effect of fact-checking on the behavior of the Facebook accounts: Event-study estimates



Notes: The figure reports the cumulative probability of accounts whose first entry in the dataset is fact-checked (with respect to not fact-checked) to be fact-checked again from 0 to 100 days after the first time that it is considered for fact-checking.

Figure 7: Long-term effect of fact-checking on "recidivism": Kaplan-Meier Failure Curve

We now turn to this question, which we tackle by investigating whether an account whose post appeared in a story considered for fact-checking by the AFP Factuel subsequently publishes posts that appear in another story considered for fact-checking. To tackle this question, we estimate a Cox proportional hazard model (equation (7)).

Before turning to the formal estimation of the model, we present some motivating statistics. In Figure 7, we define date 0 as the first date on which the account enters our database (i.e, the first date where a discussion happened that involves some content published by the account) and plot the cumulative probability that the account appears in another story considered for fact-checking. We see that, as time passes since the account first published content related to a story considered for fact-checking, the accounts whose content was fact-checked are less likely to appear again in stories considered for fact-checking, compared to those whose content was not fact-checked.

As the Kaplan-Meier curves suggests, accounts fact-checked once are less likely to be fact-checked again in the future. The Cox model confirms this finding even after controlling for possible confounders. Specifically, Table 5 reports the estimation of equation (7). Across specifications, we find that being fact-checked significantly reduces the hazard of recidivism – by around 20 percent at any given point in time. Importantly, the effect remains robust when controlling for journalist fixed effects, their interactions with time, and day-of-week variation. While the inclusion of time interactions is motivated by the fact that

Table 5: Effect of fact-checking on an account's readmission in the database, Cox Proportional Hazard Model

	Hazard for Recidivism		
	(1)	(2)	(3)
First story checked	-0.31***	-0.21***	-0.21***
	(0.06)	(0.06)	(0.06)
Journalist FEs		✓	✓
Journalist × Time		✓	✓
DOW FEs			✓
DOW × Time			✓
Observations	679,386	448,645	448,645
Nb of accounts	8053	8053	8053

Notes: The Table provides the results of the estimation of equation (7). An observation is an account × day. The coefficients represent the differences in hazard rate of being associated with another story in the sample.

journalist-specific practices may vary over time, tests of the proportional hazard assumption confirm that the main effect of being fact-checked is not time-varying.[48] Taken together, these results provide strong evidence that fact-checking discourages repeat misinformation activity in the long run. Of course, cannot rule out that users intensify their activity and circulation of misinformation on other networks.

5.3 Policy implications

Fact-checking has emerged as one of the most prominent policy tools to combat the spread of disinformation. However, it has faced persistent criticism from certain groups who argue that it infringes upon freedom of expression. This was notably the rhetoric invoked by Mark Zuckerberg when suspending the TFPC program in the US. Others contend that fact-checking consumes substantial resources while delivering limited practical impact. Our findings speak to both of these critiques and offer insights for potential policy reforms.

First, we show that the impact of fact-checking does not primarily result from Facebook's actions to hide or demote content, but rather from users' behavioral responses – challenging the idea that it suppresses free speech. As detailed in Section 5.2, users delete posts flagged as false by fact-checkers and become more cautious in their subsequent sharing behavior. Rather than restricting freedom of expression, fact-checking appears to nudge users toward more responsible expression of opinion and increases their focus on reputational concerns.

Second, our paper provides causal evidence – derived from a real-world setting – on the effectiveness of fact-checking. We show that fact-checking significantly reduces engagement with rated posts and also decreases users' subsequent activity. We present a conservative lower-bound estimate of the return to fact-checking by focusing on its static impact on stories that were actually rated. While we view

[48]The proportional hazard test rejects the hypothesis that the main effect of being fact-checked is time-varying, with a p-value of 0.47.

the dynamic reduction in future user activity as socially beneficial, reflecting greater caution in content selection and a lower likelihood of posting misinformation, it is more difficult to quantify. Table C.13 presents our main result in levels rather than logs, and shows that, on average, a fact-check leads to 274 fewer engagements with the rated story (513 for the stories rated fast).[49] Estimating the average cost of a fact-check is challenging, but we adopt a conservative benchmark of 80 euros.[50] This implies an upper-bound cost of between 0.15 to 0.35 euros per removed engagement. Although the broader social value of preventing a single engagement with false information is hard to quantify, prior research has documented meaningful negative effects of fake news exposure on beliefs and behavior. Moreover, our estimate is highly conservative, as it does not account for longer-term or spillover effects, such as changes in user behavior across other stories or platforms.

Of course it may be argued that fact-checking will not disappear with the end of the TPFC program given that Facebook aims to replace professional fact-checking with crowd-sourced fact-checking through the use of community notes. With the data at our disposal, it is hard to evaluate the relative effectiveness of crowd-sourced vs. professional fact-checking, and there is indeed evidence of the fact that community notes lead to a decrease in the spread of false news (Chuai et al., 2024a; Slaughter et al., 2025). However, it is important to highlight that community notes strongly rely on fact-checking sources (Borenstein et al., 2025), and that the trustworthiness of community notes comes from the context provided in the notes, such as fact-checking explanations (Drolsbach et al., 2024).[51] Hence, the end of the TPFC program will de facto threaten the effectiveness of crowd-sourced fact-checking.

Finally, our results suggest a number of ways in which the policy could be improved to increase the social return to fact-checking. First, the selection of stories to fact-check should be optimized. Our descriptive evidence indicates that fact-checking is most effective when applied to topics where beliefs have not yet fully crystallized – for example, the war in Ukraine during our study period. It also suggests that the detection process through algorithmic methods could be optimized: during our study period, it was only used in 16% of cases as a source to identify fake news. This led fact-checkers to over-rely on the groups or accounts they were already monitoring.

Second, and most importantly, the rating process could be improved. We find that around only half of the posts related to a story identified as false were actually rated – and this is likely a lower bound, given the potential for undetected related posts. This is the result of the incentives provided to fact-checkers, which reward the first rating and furthermore encourage rating individual posts rather than links or photos (which automatically apply ratings across all shared instances). This cautious approach places disproportionate emphasis on avoiding Type I errors (false positives), while underweighting the risk of Type II errors (false negatives).[52]

[49] Table C.13 reproduces Table 3 in levels, but we limit the sample to stories below the 97th percentile based on engagement at date 0. Indeed the distribution of date 0 engagement is very skewed.

[50] On average, it takes one day of work to produce one fact-check and the monthly salary is below 2000 euros net, which would mean a daily salary of approximately 80 euros.

[51] On the importance of links to unbiased external sources in the effectiveness of community-based fact-checking, see also Solovev and Pröllochs (2025) and Kangur et al. (2024).

[52] In the European context, it is worth noting that ratings can be appealed before an independent appeals body established under the Digital Services Act (DSA).

6 Conclusion

This paper provides the first causal estimation using field data on the effect of fact-checking on the spread of misinformation, relying on a unique partnership with the largest fact-checking organization in the world. We show that Facebook posts related to stories that are fact-checked circulate less compared to Facebook posts related to stories that were considered for fact-checking but ultimately not fact-checked. Furthermore, we show that, within fact-checked stories, rated posts receive less engagements than similar unrated ones. We finally show that – due to reputational concerns – users tend to improve their behavior following the publication of a fact-check.

Furthermore, we highlight several low-hanging fruits that could improve the effectiveness of fact-checking. Our findings show that speed matters in two key ways: the time it takes to identify (potential) misinformation, and the time required to produce a fact-check. We argue that while the latter might involve a quality trade-off for fact-checkers, the former only requires better technical support for fact-checkers by Meta to identify relevant misinformation faster. Yet it is possible that the low-quality of the Facebook claim, during the time period of our study, reflects Meta's unwillingness to effectively reduce the speed of misinformation. It has indeed been documented that fake-news stories attract more clicks than real ones, thus leading to additional advertising revenues for social media companies.

We further point out a potentially imperfect incentive structure in the TPFC program. Since journalists are paid for writing fact-checks and not for flagging posts, a large number of posts that share the same misinformation go unrated. This similarly raises the question of a possible misalignment of Meta's incentives to tackle misinformation effectively. Our findings have important implications for policy makers, platforms and fact-checkers.

References

Alesina, A., Miano, A., and Stantcheva, S. (2018). Immigration and redistribution. *CEPR Discussion Paper 13035*.

Allcott, H. and Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2):211–236.

Allcott, H., Gentzkow, M., and Yu, C. (2019). Trends in the Diffusion of Misinformation on Social Media. *Research & Politics*.

Altay, S., De Araujo, E., and Mercier, H. (2020). If this account is true, it is most enormously wonderful: Interestingness-if-true and the sharing of true and false news.

Assenza, T., Cardaci, A., and Huber, S. (2024). Fake news: Susceptibility, awareness and solutions. Technical report, Toulouse School of Economics (TSE).

Barrera, O., Guriev, S., Henry, E., and Zhuravskaya, E. (2020). Facts, alternative facts, and fact checking in times of post-truth politics. *Journal of Public Economics*, 182:104123.

Borenstein, N., Warren, G., Elliott, D., and Augenstein, I. (2025). Can community notes replace professional fact-checkers?

Briole, S., Cagé, J., and Prat, A. (2025). Access to information, news consumption and democratic participation: A nationwide experiment in french high schools.

Cagé, J. (2020). Media Competition, Information Provision and Political Participation: Evidence from French Local Newspapers and Elections, 1944-2014. *Journal of Public Economics*, 185.

Cagé, J. (2016). *Saving the Media*. Harvard University Press.

Cagé, J., Hervé, N., and Viaud, M.-L. (2020). The production of information in an online world. *The Review of Economic Studies*, 87:2126–2164.

Chopra, F., Haaland, I., and Roth, C. (2022). Do people demand fact-checked news? evidence from u.s. democrats. *Journal of Public Economics*, 205:104549.

Chuai, Y., Pilarski, M., Renault, T., Restrepo-Amariles, D., Troussel-Clément, A., Lenzini, G., and Pröllochs, N. (2024a). Community-based fact-checking reduces the spread of misleading posts on social media.

Chuai, Y., Tian, H., Pröllochs, N., and Lenzini, G. (2024b). Did the roll-out of community notes reduce engagement with misinformation on x/twitter? *Proc. ACM Hum.-Comput. Interact.*, 8.

Dewitte, E. (2025). Economic identities and the historical roots of climate change attitudes.

Drolsbach, C. P., Solovev, K., and Pröllochs, N. (2024). Community notes increase trust in fact-checking on social media. *PNAS Nexus*, 3:pgae217.

Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *The Harvard Kennedy School Misinformation Review*.

Gao, Y., Zhang, M. M., and Rui, H. (2024). Can crowdchecking curb misinformation? evidence from community notes. Technical report.

Grigorieff, A., Roth, C., and Ubfal, D. (2016). Does Information Change Attitudes Towards Immigrants? Representative Evidence from Survey Experiments. IZA Discussion Papers 10419, Institute for the Study of Labor (IZA).

Guess, A., Nagler, J., and Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science Advances*, 5.

Guriev, S., Henry, E., Marquis, T., and Zhuravskaya, E. (2023). Curtailing false news, amplifying truth. Technical report, C.E.P.R. Discussion Papers.

Henry, E., Zhuravskaya, E., and Guriev, S. (2022). Checking and sharing alt-facts. *American Economic Journal: Economic Policy*, 14(3):55–86.

Kangur, U., Chakraborty, R., and Sharma, R. (2024). Who checks the checkers? exploring source credibility in twitter's community notes.

Kuziemko, I., Norton, M. I., Saez, E., and Stantcheva, S. (2015). How elastic are preferences for redistribution? evidence from randomized survey experiments. *The American Economic Review*, 105(4):1478–1508.

Larraz, I., Salaverría, R., and Serrano-Puche, J. (2024). Combating repeated lies: The impact of fact-checking on persistent falsehoods by politicians. *Media and Communication*, 12.

Lim, C. (2018). Checking how fact-checkers check. *Research & Politics*, 5:2053168018786848. doi: 10.1177/2053168018786848.

Louis-Sidois, C. (2025). Both judge and party? investigating the political unbiasedness of fact-checkers. *Journal of the European Economic Association*, page jvaf011.

Ma, S., Bergan, D., Ahn, S., Carnahan, D., Gimby, N., McGraw, J., and Virtue, I. (2023). Fact-checking as a deterrent? a conceptual replication of the influence of fact-checking on the sharing of misinformation by political elites. *Human Communication Research*, 49:321–338.

Mattozzi, A., Nocito, S., and Sobbrio, F. (2022). Fact-checking politicians.

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.

Munger, K., Egan, P. J., Nagler, J., Ronen, J., and Tucker, J. (2022). Political knowledge and misinformation in the era of social media: Evidence from the 2015 uk election. *British Journal of Political Science*, 52:107–127.

Nyhan, B., Porter, E., Reifler, J., and Wood, T. (2020). Taking fact-checks literally but not seriously? the effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior*, 42:939?960.

Nyhan, B. and Reifler, J. (2015). The effect of fact-checking on elites: A field experiment on u.s. state legislators. *American Journal of Political Science*, 59:628–640.

Pennycook, G., Bear, A., Collins, E. T., and Rand, D. G. (2020a). 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.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A., Eckles, D., and Rand, D. (2020b). Understanding and reducing the spread of misinformation online.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., and Rand, D. G. (2020c). Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31:770–780. doi: 10.1177/0956797620939054.

Ribeiro, M. H., Zannettou, S., Goga, O., Benevenuto, F., and West, R. (2022). Can online attention signals help fact-checkers fact-check?

Slaughter, I., Peytavin, A., Ugander, J., and Saveski, M. (2025). Community notes moderate engagement with and diffusion of false information online.

Solovev, K. and Pröllochs, N. (2025). References to unbiased sources increase the helpfulness of community fact-checks. *Scientific Reports*, 15:25749.

Swire, B., Berinsky, A., Lewandowsky, and Ecker, U. (2017). Processing political misinformation: comprehending the trump phenomenon. *Royal Society Open Science*, 4(3).

Théro, H. and Vincent, E. M. (2022). Investigating Facebook's interventions against accounts that repeatedly share misinformation. *Information Processing and Management*, 59(2):102804.

Vincent, E., Théro, H., and Shabayek, S. (2022). Measuring the effect of Facebook's downranking interventions against groups and websites that repeatedly share misinformation. *Harvard Kennedy School Misinformation Review*, 3(3).

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false information online. *Science*, 359:1146–1151.

Wintersieck, A. L. (2017). Debating the truth: The impact of fact-checking during electoral debates. *American Politics Research*, 45:304–331. doi: 10.1177/1532673X16686555.

Yaqub, W., Kakhidze, O., Brockman, M. L., Memon, N., and Patil, S. (2020). Effects of credibility indicators on social media news sharing intent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14. Association for Computing Machinery.

Zhou, Y., Hou, J., Gao, Y., and Chen, P.-Y. (2025). How does crowdsourced fact-checking approach tackling misinformation affect audience engagement? evidence from twitter's community notes program. In *Proceedings of the 58th Hawaii International Conference on System Sciences*.

Online Appendix to the paper: Fact-Checking and Misinformation: Evidence from the Market Leader

Julia Cagé[*], Nathan Gallo[†], Moritz Hengel[‡], Emeric Henry[§], Yuchen Huang[¶]

December 5, 2025

Contents

[*]Department of Economics. Email: julia.cage@sciencespo.fr.
[†]Department of Economics. Email: nathan.gallo@sciencespo.fr.
[‡]Department of Economics. Email: moritz.hengel@sciencespo.fr.
[§]Department of Economics. Email: emeric.henry@sciencespo.fr.
[¶]Department of Economics. Email: yuchen.huang@sciencespo.fr.

A Details on the data collection process

A.1 Scrapping of the Facebook accounts

To observe whether the behavior of a Facebook account changes when it is rated by fact-checkers, we retroactively request the information for all posts posted by the account during the observation period. We call this action "tracking", and say that an account is "tracked" if we collect this information.

We started this part of the data collection process in July 2022, and implemented three major waves of collection in August 2022, December 2022 and June 2023 respectively.

First, we extracted the account IDs to be tracked from the post database at the start of the wave. For each wave, we focus on the top $2,000$ accounts defined with respect to the number of posts in the database at the time. The three lists of accounts partially overlap. Note that only the surviving accounts are trackable; thus, if an account is deleted or made private at the time of the tracking, it will not be tracked. At the end of the day, of the $8,054$ accounts that publish at least one post in our working sample, $3,223$ were tracked in at least one of the three waves.

Second, we run a questing script (written in Python). The script retroactively quested Crowdtangle for all the posts that were posted on the tracked accounts between two cutoff dates. The August 2022 wave quested all the posts posted between 2021-12-01 and 2022-07-01; the December 2022 wave all the posts posted between 2021-08-01 and 2022-08-01; and the June 2023 wave all the posts posted between 2022-08-01 and 2023-06-01. Note, however, that the June 2023 wave questing script crashed during the data collection process. While it collected all the posts published between November 2022 and June 2023, it failed to collect the posts published before that date, resulting in a three-month gap in the tracking sample between August 2022 and November 2022.

For each post, we retrieve all the information available on Crowdtangle. Specifically, we are interested in the time at which the post is posted and its engagement statistics. The engagement statistics are reported by Meta in the form of "timesteps."

A.2 Measuring the posting activity of the Facebook account

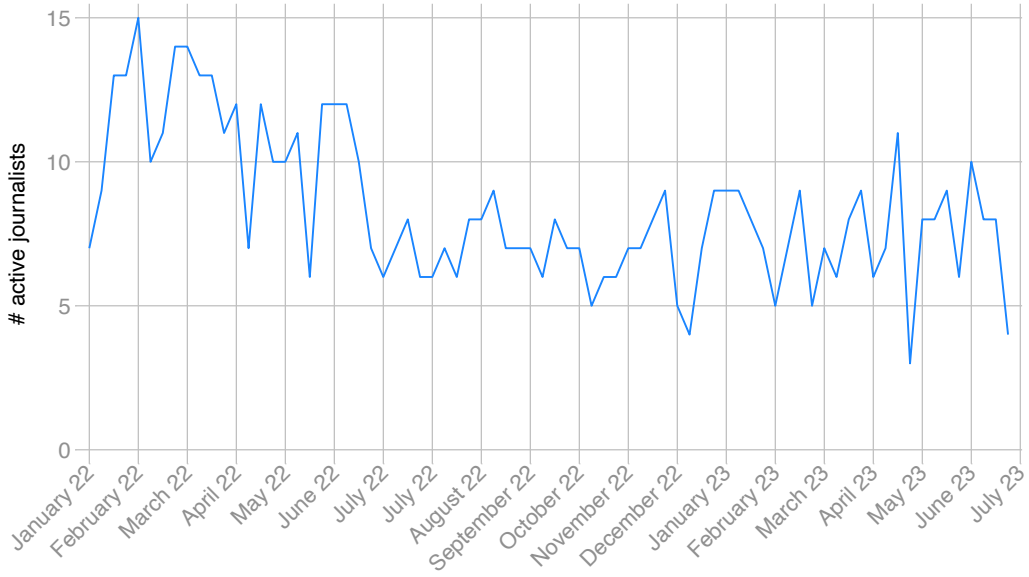
To measure the Facebook accounts activity on a given day, we proceed in three steps.

First, for each post, we track the cumulative engagements that it received in the first 14 days after publishing. To do so, we take engagement at all timesteps that falls within 14 days after posting. We take the mean of those timesteps at a days $\times$ posts level, and use linear interpolation in the few cases when a data point is missing.

Second, for each account $\times$ day, we sum the engagements of the posts that are in the "tracking window" at each given day. That is to say, the account's engagement statistic is the cumulative engagement of the posts published in less than 14 days.

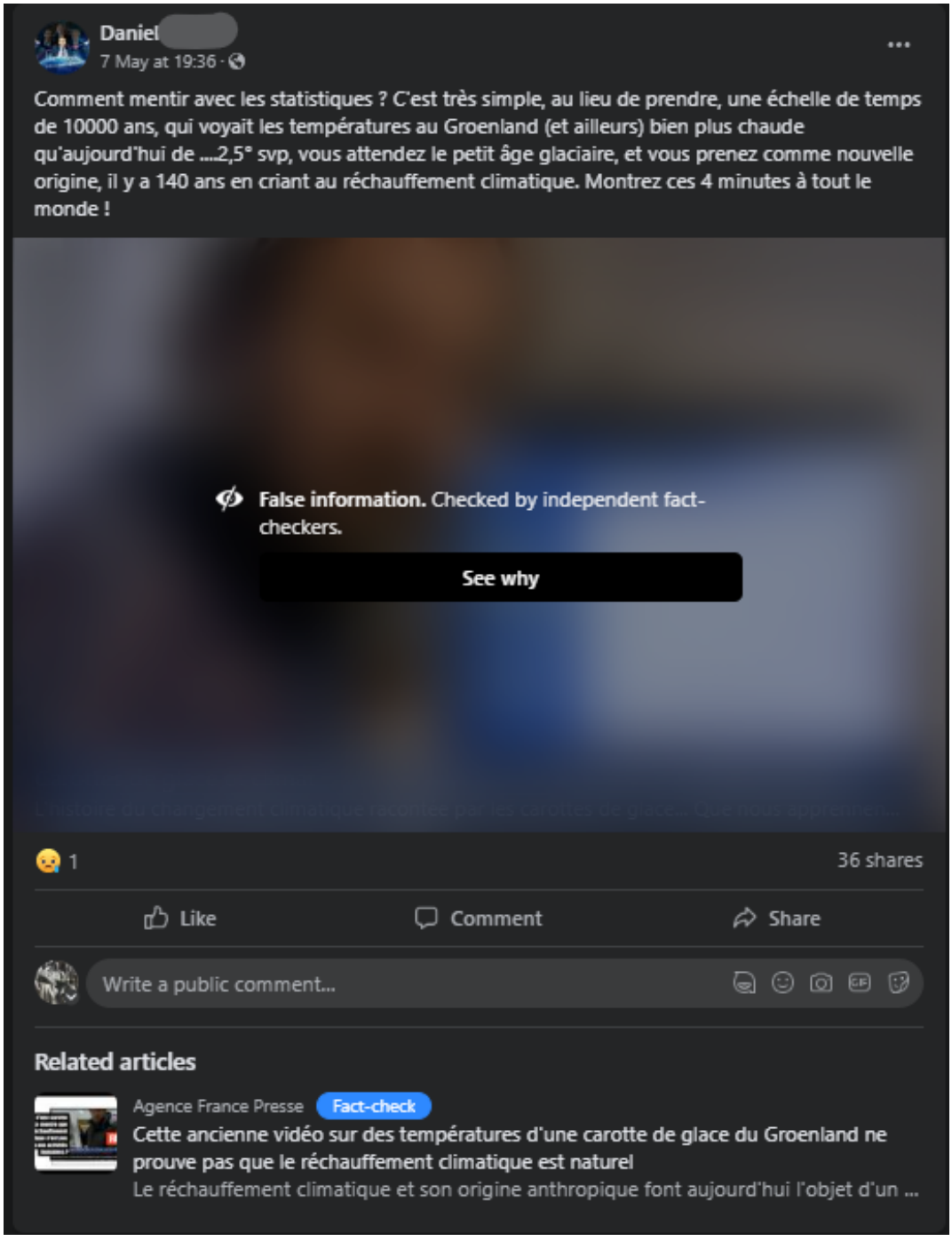
Third, we also document the number of posts that correspond to this engagement statistics at an account $\times$ day level.

B Additional Figures



Notes: The figure plots the number of active fact-checkers (i.e. of journalists working for the AFP Factuel team on a given day) over the study period based on the fact-checks published. This number represents a lower bound, as fact-checkers can work on other tasks than publishing fact-checks (e.g. answering WhatsApp requests or TikTok moderation requests). Large dips usually represent the holiday season around public holidays.

Figure B.1: Evolution of the number of active fact-checkers over time



Notes: The figure illustrates the blurring of the posts rated as "False" by the fact-checkers.

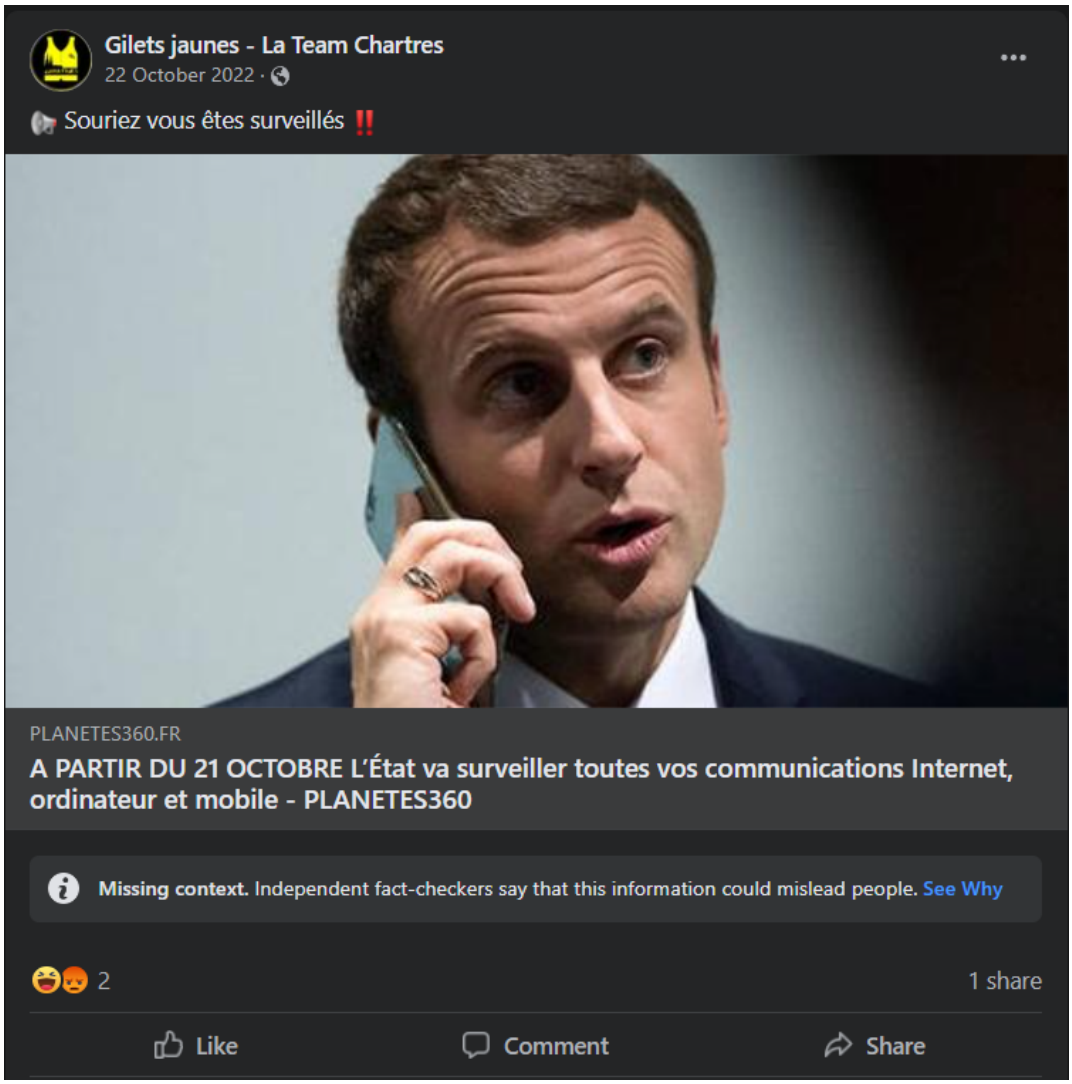
Figure B.2: Illustration: Post rated "False"

Notes: The figure illustrates the warning flag users are exposed to when they decide to see a post rated as "False" by the fact-checkers despite the blurring of the post.

Figure B.3: Illustration: Removing the blurring barrier of a post rated "False"

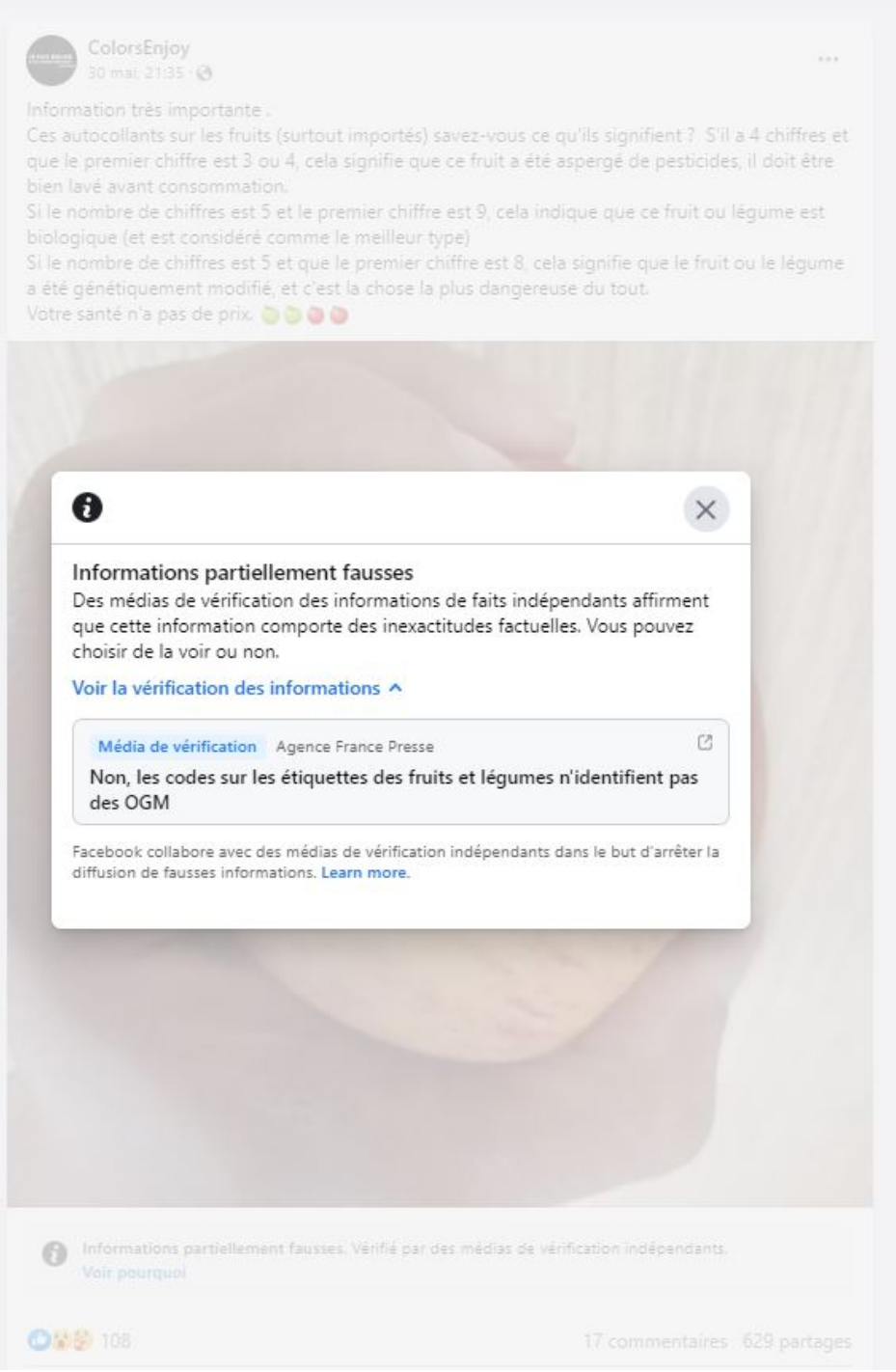
Notes: The figure illustrates the banner that appears on the posts rated as "Partially False" by the fact-checkers.

Figure B.4: Illustration: Post rated "Partially False"



Notes: The figure illustrates the banner that appears on the posts rated as "Missing Context" by the fact-checkers.

Figure B.5: Illustration: Post rated "Missing Context"

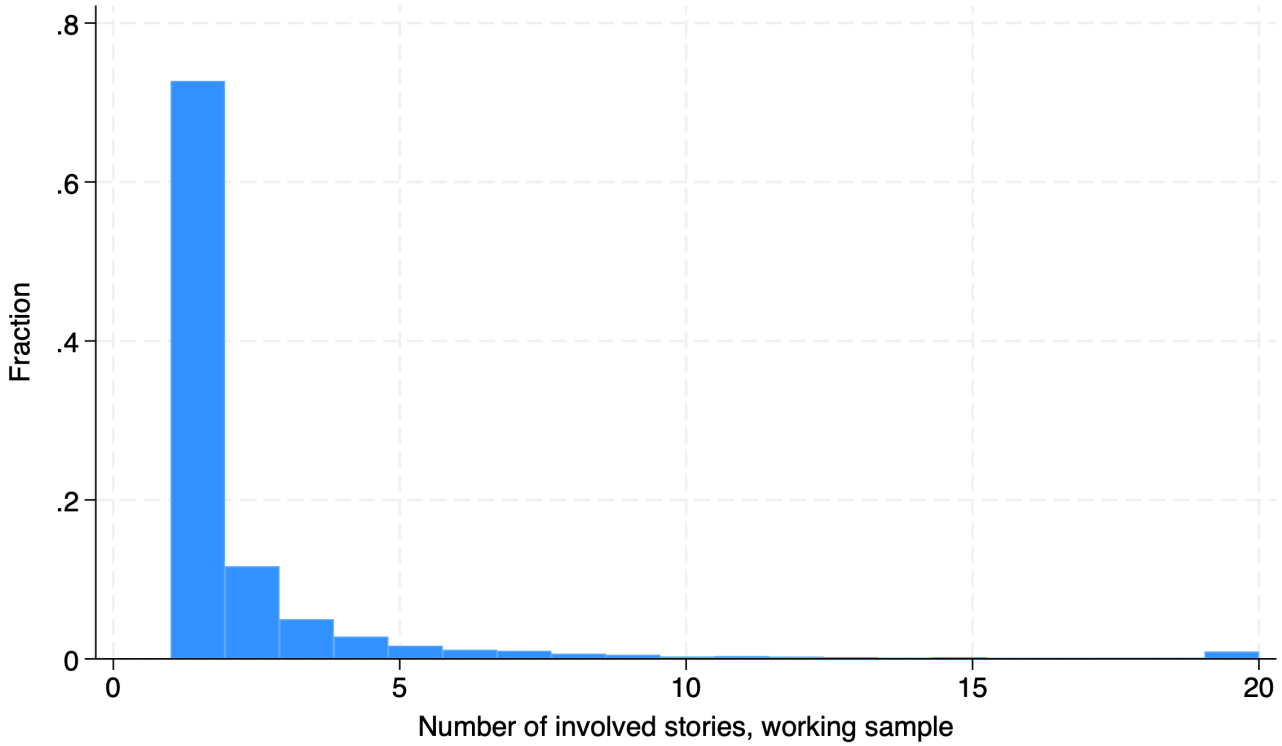


Notes: The figure illustrates the warning flag – "Vous pouvez choisir de la voir ou non" (You can choose whether or not to view it) – users are exposed to when they decide to see a post rated as "Partly False" by the fact-checkers .

Figure B.6: Illustration: Removing the blurring barrier of a post rated "Partly False"

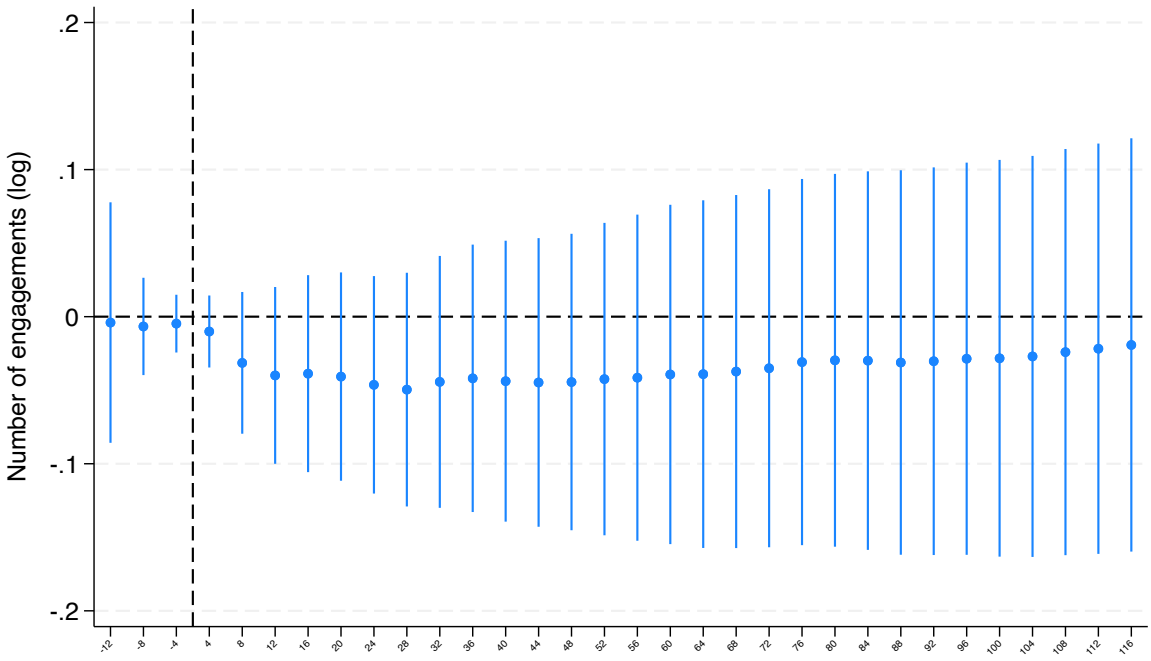
Notes: The Figure illustrates a post that is part of a story rated "Partially False" (see Figure B.4) but was not flagged by the fact-checkers.

Figure B.7: Illustration: Post in a story rated "Partially False" that was not flagged



Notes: The Figure describes the number of stories per account in our working sample. An account is considered as involved in a story if it published a post in the story. The graph is capped at 20 stories (74 accounts are associated with 20 stories or more).

Figure B.8: Distribution of the number of stories considered for fact-checking per Facebook account



Notes: The figure reports the results of the estimation of equation (2). An observation is a story. Observations are weighted using weights derived from a propensity score matching procedure using as covariates the engagement at consideration date and the engagement 12 hours before consideration. The dependent variable is the logarithm of the number of engagements. Only the stories rated "Partly False" are included in the treated group.

Figure B.9: Effect of fact-checking on the number of engagements: Story-level analysis, Event-study estimates, Only stories rated "Partly False"

C Additional Tables

Table C.1: All posts in the working sample: descriptive statistics

	Total (N=19448)	Fact-checked (N=7461)	Not Fact-checked (N=11987)	Difference	P-value
Main content of the post					
Text/Link	0.33	0.40	0.29	0.11	0.00
Video	0.45	0.45	0.45	0.00	0.70
Photo	0.20	0.13	0.24	-0.11	0.00
Other	0.02	0.02	0.02	0.00	0.02
Post features					
Text length (# of characters)	375.36	312.29	414.62	-102.33	0.00
# days btw post and AFP discussion	34.66	11.31	49.10	-37.79	0.00
Account that publishes the post					
Account is a Facebook group	0.78	0.80	0.77	0.03	0.00
Number of subscribers	55457.97	42875.87	63229.71	-2.0e+04	0.01
Type of flag					
Altered Media		0.01			
False		0.53			
Missing Context		0.37			
Partly False		0.08			
Satire		0.01			
Deletion					
Scrapers found deleted	0.31	0.29	0.33	-0.05	0.00

Notes: The Table provides descriptive statistics on the $19,448$ posts related to the 944 stories in our working sample. An observation is a post. Column (1) provides statistics for all the posts, and Column (2) (respectively Column (3)) provides statistics for the posts whose story is fact-checked (respectively unchecked).

Table C.2: Posts related to fact-checked stories: descriptive statistics (working sample)

	Total (N=9070)	Rated (N=5545)	Unrated (N=3525)	Difference	P-value
Main content of the post					
Text/Link	0.29	0.35	0.21	0.14	0.00
Video	0.47	0.51	0.40	0.11	0.00
Photo	0.23	0.14	0.38	-0.23	0.00
Other	0.01	0.00	0.02	-0.01	0.00
Post features					
Text length (# of characters)	262.99	281.17	234.39	46.78	0.00
# days btw post and AFP discussion	4.52	4.21	5.01	-0.80	0.25
Account that publishes the post					
Account is a Facebook group	0.76	0.81	0.70	0.11	0.00
Number of subscribers	54256.71	39763.09	77055.91	-3.7e+04	0.00
Type of flag					
Altered media		0.01			
False		0.54			
Missing context		0.36			
Partly false		0.08			
Satire		0.01			
Deletion					
Scrapers found deleted	0.34	0.29	0.41	-0.11	0.00
Post activity at the time of 1st rating					
Total number of engagements	420.56	356.08	521.98	-165.90	0.08
# of shares	145.81	138.00	158.09	-20.09	0.69
# of comments	46.95	38.61	60.08	-21.47	0.01
# of likes	172.09	137.00	227.29	-90.29	0.03

Notes: The Table provides descriptive statistics on the $9,070$ posts related to the 558 fact-checked stories included in our working sample. An observation is a post. Column (1) provides descriptive statistics for all the posts, and Column (2) (respectively Column (3)) for the posts whose content is rated (respectively unrated).

Table C.3: Posts related to fact-checked stories with internal variation in rating: descriptive statistics (post-level regression sample)

	Total (N=4234)	Rated (N=2902)	Unrated (N=1332)	Difference	P-value
Types of Posts' Main Content					
Text/Link	0.29	0.35	0.17	0.18	0.00
Video	0.42	0.45	0.35	0.10	0.00
Photo	0.28	0.19	0.47	-0.28	0.00
Other	0.01	0.00	0.01	-0.01	0.02
Post features					
Post text length (characters)	268.49	287.44	227.19	60.25	0.00
# days btw post and AFP discussion	6.05	6.08	5.99	0.09	0.95
Information on Publishing Account					
Posted by a group	0.70	0.73	0.62	0.11	0.00
Posting account subscriber count	63448.46	51645.50	89163.33	-3.8e+04	0.00
Flags (if Flagged)					
Altered media		0.00			
False		0.54			
Missing context		0.36			
Partly false		0.08			
Satire		0.01			
Deletion					
Deleted	0.34	0.31	0.40	-0.09	0.00
Post activity at the time of 1st rating					
Total Engagement	585.51	565.14	629.89	-64.76	0.70
Shares	213.19	228.81	179.17	49.63	0.64
Comments	60.62	55.34	72.13	-16.78	0.16
Likes	239.43	217.73	286.70	-68.97	0.28

Notes: The Table provides descriptive statistics on the $4,234$ posts related to the 240 fact-checked stories included in our working sample with internal variations in rating; namely, not all posts within the stories are rated. An observation is a post. Column (1) provides descriptive statistics for all the posts, and Column (2) (respectively Column (3)) or the posts whose content is rated (respectively unrated).

Table C.4: Stories considered for fact-checking: Descriptive statistics on the set of stories used in the regression sample

	Total (N = 589)	Fact-checked (N = 395)	Not Fact-checked (N = 194)	Difference	P-value
Story Origin					
Twitter	0.37	0.36	0.40	-0.04	0.39
Facebook Claim	0.21	0.23	0.18	0.05	0.23
Other Facebook	0.34	0.35	0.32	0.04	0.41
WhatsApp	0.06	0.05	0.07	-0.02	0.40
Media	0.02	0.02	0.02	0.00	0.84
TikTok	0.03	0.03	0.02	0.01	0.39
Other Social Media	0.02	0.02	0.04	-0.02	0.15
Topics					
Covid	0.17	0.14	0.24	-0.10	0.00
Ukraine/NATO	0.18	0.17	0.19	-0.02	0.53
Vaccines	0.14	0.11	0.19	-0.07	0.02
Climate	0.10	0.09	0.10	0.00	0.87
Other	0.27	0.41	0.00	0.41	0.00
Activity at Discussion Date					
N. Active Posts	15.91	16.35	15.03	1.32	0.67
Total Engagement	7275.23	7058.21	7717.11	-658.91	0.73
Shares	2506.19	2463.39	2593.34	-129.96	0.89
Comments	776.93	784.13	762.25	21.88	0.90
Likes	3084.05	2962.50	3331.52	-369.02	0.67
# days btw posting and AFP discussion	47.88	48.18	47.29	0.89	0.96
Flags					
False		0.72			
Altered Media		0.03			
Partially False		0.03			
Missing Context		0.12			
Reasons for Not Fact-checking					
Infeasible			0.73		
Lack of Ressources			0.26		
Deletion at Discussion Date					
Has Deleted Post	0.24	0.28	0.16	0.12	0.00

Notes: The Table provides descriptive statistics on the 589 stories considered for fact-checking that are included in our regression sample. An observation is a story. Column (1) provides descriptive statistics for all the stories, and Column (2) (respectively Column (3)) for the stories that are fact-checked (not fact-checked).

Table C.5: Account related to fact-checked stories: descriptive statistics (working sample)

	Total (N = 8054)	Tracked (N = 3223)	Untracked (N = 4831)	Difference	P-value
Account Information					
Facebook Group	0.69	0.96	0.51	0.45	0.00
Max subscriber count	78404.12	41988.07	1.0e+05	-6.1e+04	0.00
Presence in the Working Sample					
N posts in dataset	2.39	3.33	1.77	1.56	0.00
N stories in dataset	2.11	2.89	1.58	1.31	0.00
Pct posts in dataset rated	23.17	24.72	22.14	2.58	0.00
Pct stories in dataset rated	64.50	56.67	69.72	-13.06	0.00
Pct post in dataset deleted	29.92	27.15	31.78	-4.63	0.00
Pct posts in dataset in seed	57.04	57.30	56.87	0.43	0.67
Activites at First Presence (if Tracked)					
N. new posts per day		51.22			
Cumu. posts from days –20		1091.53			

Notes: The Table provides descriptive statistics on the $8,054$ posts related to the 944 fact-checked stories included in our working sample. An observation is an account. Column (1) provides statistics for all the accounts, and Column (2) (respectively Column (3)) provides statistics for the accounts that are tracked in at least one tracking wave (respectively untracked).

Table C.6: Account related to fact-checked stories: Top Accounts

Name of account	N Posts	N Stories	First entry	Pct from seed	Pct deleted
Pour la demission d'Emmanuel Macron	278	133	22feb2022	59	38
Reinfo Gard collectif extraordinaire	148	92	16feb2022	65	29
Tempete en marche contre les dictatur...	134	90	04mar2022	65	32
Osons l'info	127	64	25apr2022	65	41
NON AU PASS VACCINAL	122	60	21mar2022	65	30
odyssee.com	73	57	21mar2022	63	30
LES RESISTANTS CONTRE LE PASS SANITAIRE	97	53	01mar2022	69	0
Le groupe des non vaccines	63	50	21mar2022	68	21
La liberte commence par nos enfants	71	46	04mar2022	65	46
Les Francais contre Macron	53	45	21mar2022	45	47
La Verite Cachee	50	45	21mar2022	70	12
Collectif ANTI-MASQUES !	56	44	21mar2022	55	23
Le pouvoir du peuple pour le peuple p...	47	43	21mar2022	72	26
LES AMOUREUX D'UNE FRANCE LIBEREE	50	41	21mar2022	58	32
mouvement citoyen ANTI-MACRON	51	40	21mar2022	65	18
S'informer autrement !	47	40	21mar2022	68	15
Stop a la mascarade ! On veut la veri...	63	40	16feb2022	75	37
Oliv oliv 2	54	37	21mar2022	63	46
Contre la vaccination, et la dictatur...	46	37	04mar2022	52	13
Haute Saone liberte egalite fraternit...	50	37	28mar2022	66	48
O M S L ORGANISATION MONDIALE SECURIT...	44	34	21mar2022	68	39
Sentinelle Guadeloupe	43	33	12apr2022	65	44
Gilet Jaune En Direct	44	33	21mar2022	82	39
ANTI-COVID 19 France	39	33	04mar2022	72	33
Collectif des climato-realistes	55	32	30mar2022	49	18
Les maquisards de la republique . (ex...	36	32	21mar2022	64	14
le Rat Porteur	40	31	21mar2022	78	33
groupe de soutien a Alexandra Henrio...	54	31	14apr2022	74	80
Soignants en resistance	35	31	21mar2022	54	34
Stop a la dictature sanitaire (bis)	37	31	04mar2022	46	30

Notes: The Table lists the top 30 accounts in our account dataset by number of stories in the working sample involved.

Table C.7: Differences between stories generated before -12 hours versus after

	Created before -12h (N = 589)	Created after -12h (N = 55)	Difference	P-value
Origins				
Twitter	0.37	0.79	-0.42	0.00
Facebook Claim	0.21	0.09	0.13	0.04
Other Facebook	0.34	0.09	0.25	0.00
WhatsApp	0.06	0.04	0.02	0.66
Translation	0.00	0.00	0.00	.
Media	0.02	0.02	-0.01	0.79
TikTok	0.03	0.00	0.03	0.26
Other Social Media	0.02	0.04	-0.02	0.38
Topics				
Covid	0.17	0.29	-0.12	0.03
Ukraine/NATO	0.18	0.16	0.01	0.81
Vaccines	0.14	0.25	-0.12	0.02
Climate	0.10	0.13	-0.03	0.44
Other	0.27	0.15	0.13	0.04
Activity at Discussion Date				
N. Active Posts	15.91	3.16	12.75	0.01
Total Engagement	7,275.23	292.41	6,982.82	0.02
Shares	2,506.19	59.46	2,446.73	0.08
Comments	776.93	29.97	746.96	0.01
Likes	3,084.05	64.00	3,020.05	0.02

Notes: The Table provides descriptive statistics for the stories in the regression sample (N = 589, exluding translation, unviral, and generated more than $12h$ before the discussion) compared with the stories that satisfy the regression sample's criteria otherwise (no translation, no unviral) but generated less than $12h$ before the discussion. Column (3) reports the difference in means between the two groups, and the Column (4) reports the p-value of the difference.

Table C.8: Determinants of the speed of fact-checking

	(1)	(2)	(3)	(4)
Engagements at consideration date	0.02	0.02	0.02**	0.02*
	(0.01)	(0.01)	(0.01)	(0.01)
Climate		-0.19**	-0.20**	-0.18**
		(0.08)	(0.08)	(0.09)
Covid		-0.02	-0.05	-0.04
		(0.11)	(0.12)	(0.11)
Ukraine/NATO		0.19**	0.19**	0.18*
		(0.09)	(0.09)	(0.09)
Vaccines/Health		-0.11	-0.08	-0.04
		(0.11)	(0.11)	(0.10)
Twitter			-0.19**	-0.27***
			(0.09)	(0.09)
Facebook Claim			-0.35***	-0.34***
			(0.09)	(0.10)
Other Facebook			-0.24***	-0.17*
			(0.08)	(0.09)
Journalist FEs				✓
Observations	329	329	329	329
Mean DepVar	0.47	0.47	0.47	0.47
Sd DepVar	0.50	0.50	0.50	0.50

Notes: * p<0.10, ** p<0.05, *** p<0.01. Robust standard errors are reported in parentheses. An observation is a fact-checked story. The dependent variable is an indicator variable equal to one if the fact-check is produced faster than the median fact-checking time and to zero otherwise. Engagements at consideration date are measured in tens of thousands.

Table C.9: Effect of fact-checking on the number of engagements: Story-level analysis, Difference-in-Differences estimates, Heterogeneity depending on the speed of the fact-checking process

	All stories	Fast Rating (≤2 days)	Slow Rating (>2 days)	Fast Discuss (≤4 days)	Slow Discuss (>4 days)
	(1)	(2)	(3)	(4)	(5)
Post * Fact-checked stories	-0.08	-0.11**	-0.04	-0.07	-0.03
	(0.05)	(0.05)	(0.06)	(0.07)	(0.06)
Story FEs	✓	✓	✓	✓	✓
Time FEs	✓	✓	✓	✓	✓
Day of the week	✓	✓	✓	✓	✓
Observations	18,227	12,056	12,243	10,043	8,184
Nb of clusters (stories)	553	366	371	305	248
Mean DepVar	6.10	6.07	6.04	5.86	6.41
Sd DepVar	2.59	2.61	2.60	2.51	2.65

Notes: The Table provides the results of the estimation of equation (1). An observation is a story. All the regressions include day-of-the-week fixed effects. The dependent variable is the logarithm of the cumulative number of engagements. Column (1) includes all the stories considered for fact-checking by the AFP Factuel team (reproducing Column (2) of Table 3). In Column (2), the treated group only includes the stories that receive a first rating faster than the median fact-checking time (in less than 2 days). In Column (3), the treated group only includes the stories that receive a first rating after the median fact-checking time (in more than 2 days). In Column (4), the treated group only includes the stories for which the time between the story's appearance and the discussion in the editorial meeting is less or equal to the median (less than 4 days). In Column (5), the treated group only includes the stories for which the time between the story's appearance and the discussion in the editorial meeting is higher than the median (more than 4 days).

Table C.10: Effect of fact-checking on the number of engagements: Story-level analysis, Difference-in-Differences estimates, Heterogeneity depending on the topic of the stories

	All stories	Ukraine/NATO	Covid	Climate	Vaccines/Health	Non French	Other
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Post * Fact-checked stories	-0.08	-0.24**	0.01	0.11	-0.01	-0.17	-0.13
	(0.05)	(0.12)	(0.09)	(0.11)	(0.09)	(0.11)	(0.22)
Story FEs	✓	✓	✓	✓	✓	✓	✓
Time FEs	✓	✓	✓	✓	✓	✓	✓
Day-of-the-week FEs	✓	✓	✓	✓	✓	✓	✓
Observations	18,227	3,234	3,267	1,815	3,300	1,419	6,512
Nb of clusters (stories)	553	98	99	55	100	43	198
Mean DepVar	6.10	6.22	5.11	5.29	4.88	7.99	6.62
Sd DepVar	2.59	2.35	2.33	2.76	2.25	1.72	2.44

Notes: * p<0.10, ** p<0.05, *** p<0.01. The Table provides the results of the estimation of equation (1). An observation is a story * time (standard errors clustered at the story level between parentheses). Only the stories considered for fact-checking by the AFP Factuel team are included.The dependent variable is the logarithm of the cumulative number of engagements. All the stories are included in Column (1), and the other columns include the stories related to the topics described in the column names.

Table C.11: Effect of fact-checking on the number of engagements: Difference-in-Differences estimates, Heterogeneity depending on the topic of the stories, only considering stories that are fact-checked within two days

	All stories	Ukraine/NATO	Covid	Climate	Vaccines/Health	Non French	Other
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Post * Fact-checked stories	-0.11**	-0.24**	-0.14	0.06	-0.14	-0.16	-0.13
	(0.05)	(0.12)	(0.10)	(0.21)	(0.10)	(0.11)	(0.21)
Story FEs	✓	✓	✓	✓	✓	✓	✓
Time FEs	✓	✓	✓	✓	✓	✓	✓
Day-of-the-week FEs	✓	✓	✓	✓	✓	✓	✓
Observations	12,056	2,475	2,277	957	2,211	1,122	3,509
Nb of clusters (stories)	366	75	69	29	67	34	107
Mean DepVar	6.07	6.13	5.22	5.66	4.87	7.88	6.50
Sd DepVar	2.61	2.33	2.44	2.87	2.27	1.74	2.49

Notes: * p<0.10, ** p<0.05, *** p<0.01. The Table provides the results of the estimation of equation (1). An observation is a story*time (standard errors clustered at the story level between parentheses). Only the stories considered for fact-checking by the AFP Factuel team are included, and the treated group is reduced to the stories that are fact-checked within two days. The dependent variable is the logarithm of the cumulative number of engagements. All the stories are included in Column (1), and the other columns include the stories related to the topics described in the column names.

Table C.12: Effect of fact-checking on the number of engagements: Difference-in-Differences, Robustness checks

	(1) Engagements	(2) Comments	(3) Shares	(4) Likes	(5) Engagements	(6) Engagements	(7) Engagements
Post * Fact-checked stories	-0.08	-0.10*	-0.08	-0.07	-0.08	-0.06	-0.09*
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Story FEs	✓	✓	✓	✓	✓	✓	✓
Time FEs	✓	✓	✓	✓	✓	✓	✓
Day of the week	✓	✓	✓	✓	✓	✓	✓
Matching	✓	✓	✓	✓			
Matching Alt1						✓	
Matching Alt2							✓
Observations	18,227.00	16,916.00	17,323.00	17,625.00	19,415.00	18,062.00	18,326.00
Mean DepVar	6.10	4.36	5.07	4.98	6.46	6.10	6.15
Sd DepVar	2.59	2.30	2.38	2.62	2.73	2.57	2.68

[plain]Effect by topic

- Possible interpretation: much more difficult to shift beliefs and behaviors for topics where beliefs are already strongly entrenched.

Notes: The Table provides the results of the estimation of equation (1). An observation is a story. All the stories considered for fact-checking by the AFP Factuel team are included. The dependent variable is the logarithm of the number of engagements in Column (1), and of the number of comments, shares, and likes in Columns (2), (3) and (4) respectively. Column (5) reproduces Column (1) but with no matching procedure. In Columns (6) and (7), we use different covariates for the matching procedure. In Column (6), we add as a covariate an indicator variable equal to one if the story is more than half a day old and to zero otherwise. In Column (7), we add the engagements 6 hours before the discussion date as an additional covariate.

Table C.13: Effect of fact-checking on the number of engagements (in level): Story-level analysis, Difference-in-Differences estimates

	All types of flags		Stories rated False		Fast rating	
	(1)	(2)	(3)	(4)	(5)	(6)
Post * Fact-checked stories	-274.24	-270.76	-388.55	-386.59	-513.26	-492.71
	(418.46)	(406.46)	(434.96)	(420.85)	(416.51)	(383.37)
Story FEs	✓	✓	✓	✓	✓	✓
Time FEs	✓	✓	✓	✓	✓	✓
Day of the week		✓		✓		✓
Observations	17,853	17,853	14,784	14,784	11,748	11,748
Nb of clusters (stories)	541	541	448	448	356	356
Mean DepVar	3,219	3,219	3,226	3,226	3,030	3,030
Sd DepVar	6,833	6,833	6,860	6,860	6,404	6,404

Notes: The Table provides the results of the estimation of equation (1). An observation is a story. The dependent variable is the number of engagements. The sample is restricted to stories below the 97th percentile based on engagement at date 0. Columns (1) and (2) include all the stories considered for fact-checking by the AFP Factuel team. In Columns (3) and (4), only the stories rated "False" are included in the treated group. In Columns (5) and (6), only the stories that are fact-checked fast (in less than 2 days) are included.

Table C.14: Effect of fact-checking on the number of engagements: Difference-in-Differences, Post-level analysis, Robustness checks

	(1) Engagements	(2) Comments	(3) Shares	(4) Likes	(5) Engagements	(6) Engagements
Post *Rated	-0.011***	-0.002	-0.015**	-0.007*	-0.017***	
	(0.003)	(0.006)	(0.006)	(0.004)	(0.005)	
Post * Flagged						-0.006**
						(0.003)
Post FEs	✓	✓	✓	✓	✓	✓
Story * Time FEs	✓	✓	✓	✓	✓	✓
Day of the week	✓	✓	✓	✓	✓	✓
Within-story Matching					✓	
Observations	129,802	69,293	99,275	94,500	68,131	121,486
Mean DepVar	2.51	2.06	1.85	2.10	2.37	2.52
Sd DepVar	2.19	1.86	1.80	2.13	2.18	2.21

Notes: The Table provides the results of the estimation of equation (3). An observation is a post*time (standard errors clustered at the story level between parentheses). The dependent variable is the number of engagements. Column (1) reproduces Column (1) of Table 4. Column (2) (respectively Columns (3) and (4)) uses the logarithm of the cumulative number of comments (respectively of shares and of likes). In Column (5), we use a different matching procedure by limiting matches to posts within the same story, using only engagement at date 0 as a covariate. In Column (6), we use as an explanatory variable whether a post was flagged rather than rated.

Table C.15: Effect of fact-checking on accounts activity: Difference-in-Differences, Cumulative Numbers of Posts

	(1)	(2)
Post * Fact-check	-0.01	-0.01
	(0.01)	(0.01)
Account FEs	✓	✓
Time FEs	✓	✓
Day of the week		✓
Observations	51020	51020
Nb of clusters (Account)	2,551	2,551
Mean DepVar	6	6
Sd DepVar	1.63	1.63

Notes: The Table provides the results of the estimation of equation (6). An observation is an account*day(standard errors clustered at the story level between parentheses). The dependent variable is the logarithm of cumulative number of posts from 20 days before the first time the account's story is discussed. The regression is run on the observations from 5 days before the discussion to 14 days after. Column (1) controls for day and account FE, while Column (2) additionally controls for days of the week FE.