GenAI Misinformation, Trust, and News Consumption: Evidence from a Field Experiment*

Filipe Campante[†] Ruben Durante[‡] Felix Hagemeister[§] Ananya Sen[¶]

October 2026

Abstract

We study how AI-generated misinformation affects demand for trustworthy news, using data from a field experiment by *Süddeutsche Zeitung* (SZ), a major German newspaper. Readers were randomly assigned to a treatment that highlighted the difficulty of distinguishing real from AI-generated images. The treatment increased concern over misinformation (+0.3 s.d.) and reduced trust in all news sources (−0.05 to −0.1 s.d.), including SZ itself. Crucially, it also affected post-survey browsing behavior: daily visits to SZ digital content rose by 2.5% in the days following the treatment. In addition, subscriber retention increased by 1.2% over the following five months, corresponding to about a 30% reduction in the attrition rate. These results are consistent with a model in which the relative value of trustworthy news sources rises with the prevalence of misinformation, boosting engagement with these sources even as trust in news content declines.

JEL Codes: D12, L82, L86.

Keywords: Generative AI, Misinformation, News Media, Trust.

---

*We are indebted to many people for helpful feedback, including audiences at multiple conferences and seminars. Special thanks to co-editor Leo Bursztyn and three anonymous referees, as well as Nicolás Ajzenman, Jennifer Allen, Elliott Ash, Manuela Collis, Chuck Eesley, Henry Farrell, Thomas Fujiwara, Sid George, Lorenz Goette, Michael Kummer, Mor Namaan, Brendan Nyhan, Jesse Shapiro, Sonny Tambe, Joachim Voth, David Yanagizawa-Drott, and especially Noam Yuchtman, for detailed comments and suggestions. We thank Alessandro Izzo, Matheus Junqueira, David Schwartsman, Ren Jie Teh, and Nicolás Valle for outstanding research assistance, and Maya Judd for careful proofreading and editorial assistance. All remaining errors are our own. Durante and Sen acknowledge financial support from the European Union's Horizon research and innovation program (Grant 101125953) and from the Andrew Carnegie Fellows Program, respectively. This study was pre-registered in the AEA RCT Registry (ID: AEARCTR-0015359), and determined to be exempt from review by the National University of Singapore IRB (NUS-IRB-2025-598).

[†]Johns Hopkins University & NBER. Email: fcampante@jhu.edu

[‡]National University of Singapore, ICREA-UPF, CESifo, IZA & CEPR. Email: rubendurante@gmail.com

[§]Süddeutsche Zeitung Digitale Medien. Email: felix.hagemeister@sz.de

[¶]Carnegie Mellon University. Email: ananyase@andrew.cmu.edu

1 Introduction

Widespread unease has emerged over the prevalence of online misinformation and its ramifications for politics, business, and society (Allcott and Gentzkow, 2017; Jerit and Zhao, 2020; Pennycook and Rand, 2021; Ahmad et al., 2024).[1] Particularly worrisome is the link between such misinformation and declining trust in news content (Brenan, 2024; Newman et al., 2024), with potentially far-reaching implications for political polarization and democracy (Lazer et al., 2018; Tucker et al., 2018; Acemoglu et al., 2024), and for the economic viability of the news industry.

With recent advances in artificial intelligence (AI), this issue has become even more prominent. Generative AI (GenAI) enables the creation of visual and audio content that is virtually indistinguishable from authentic material ("deepfakes"), heightening concerns over the trust-eroding potential of false or misleading information.[2] From the standpoint of the news media industry, this could pose an existential challenge to outlets whose business model relies on producing costly high-quality news: a collapse in trust could reduce demand to the point where such journalism becomes economically unsustainable.[3]

Yet, when trust grows scarcer, trustworthiness becomes more valuable. This coun-

[1] For a discussion on the meaning and definition of "misinformation," see Vraga and Bode (2020).

[2] See, for instance, Vaccari and Chadwick (2020); Langguth et al. (2021); Veerasamy and Pieterse (2022); Spitale et al. (2023); Endert (2024); Toff and Simon (2025); Simon and Altay (2025). Relatedly, the concept of "AI slop," meant to capture the proliferation of low-quality AI-generated content, has also gained traction (Hoffman, 2024).

[3] This risk is especially grave given the industry's enduring financial difficulties since the rise of the internet in the 2000s (Djourelova et al., 2024; Beattie et al., 2021), compounded by the emergence of news aggregators and social media platforms as preeminent vehicles for the dissemination of news content (Calzada and Gil, 2020; Zhuravskaya et al., 2020).

terpoint creates a potential business opportunity for reputable outlets, as consumers turn to reliable sources for help distinguishing real from synthetic content. Such a dynamic could have important implications for the economic viability of online news and, more broadly, for how misinformation shapes trust in the media ecosystem in democratic societies.

This paper studies the interplay between AI-powered misinformation, trust, and the media ecosystem, taking advantage of a unique field experiment conducted by the German newspaper *Süddeutsche Zeitung* (henceforth SZ). As part of its online marketing operations, SZ regularly surveys online subscribers, digital app users, and website visitors. SZ also explicitly promotes the idea that its journalism helps readers distinguish real from AI-generated content, reflecting the economic and practical significance of our question for media outlets.[4] This experiment allows us to study a sample of individuals who we can presume (and later verify) regard SZ as a highly trustworthy news source, thus providing a suitable testing ground for our hypotheses.

In one of these surveys, SZ implemented a randomly assigned treatment designed to highlight AI's ability to create seemingly authentic material and challenge readers to distinguish real from AI-generated content. Readers in the treatment group were shown three pairs of images – each consisting of one real and one AI-generated – and asked to judge whether either had been created by AI. Those in the control group, in contrast, were shown pairs of real images related to the same set of current affairs, and asked questions unrelated to AI. Both groups (a total of more than 17,000 individuals) were then asked to evaluate the severity of the problem of misinformation, and report

[4]One SZ campaign used the slogan *"Die Wahrheit lässt sich nicht generieren. Nur recherchieren."* ("The truth cannot be generated. Only researched.")

their level of trust in the content on SZ and on several other platforms.[5] For a subset of about 6,000 participants who agreed to have their data tracked and linked to the experiment, we can also observe their actual behavior for several weeks before and after the intervention, namely, the number of daily visits to SZ digital content (website or app), the articles they clicked on, and their subscription status.[6]

To guide our analysis of the experiment, we first develop a simple theoretical framework to study the interplay between misinformation and trustworthiness in news consumption. Modeling AI-powered misinformation as broadly degrading the quality of the information environment, we nevertheless show that the latter can increase demand for a source perceived as sufficiently trustworthy – i.e., a source able to mitigate this degradation. Intuitively, consumer demand depends on the *relative* value of the source, which increases with the prevalence of misinformation, even as the perceived quality of the trustworthy source declines.

With this framework in hand, we turn to the key experimental results. First, regarding survey-based outcomes, we find that the AI exposure treatment increased respondents' reported concern over online misinformation by about one-third of a standard deviation.[7] The AI intervention also reduced trust in the content of all media outlets and platforms, including SZ itself.

Despite that, we find that treated individuals in the tracked sample increased

[5] Unsurprisingly, SZ readers in the experiment reported very high levels of trust in SZ content: in the control group, the average rating was 2.9 on a scale from 0 to 3, with 92% reporting the maximum level and only 0.6% reporting trust levels below 2.

[6] Consent to tracking was obtained before participants started the quiz. Moreover, observables are balanced across treatment and control groups.

[7] The AI quiz was successful in confounding participants: the average respondent answered 0.9 questions correctly, with only about 2.5% answering all three correctly and 36% all three wrong.

their daily visits to SZ digital content immediately after the quiz. Visits rose by about 2.5% relative to the average number of visits for the first 3-5 days, with the effect declining over time but remaining significant for more than two weeks. We also observe a sustained increase in subscriber retention, which after five months was 1.2% higher in the treatment than in the control group, reflecting about a 30% lower attrition rate. This indicates that while the treatment's impact on day-to-day decisions (e.g., online visits) may fade, it remains influential as individuals reassess their commitment as subscribers.[8] Finally, we quantify the economic significance of the results based on simple back-of-the-envelope calculations scaled to the subscriber base. These imply about €456,000 in additional annual subscription revenue, aside from advertising revenue from extra visits in the short run (23,000 visits) and from retained subscribers (over 3 million visits annually).

In sum, the results confirm the key predictions of our framework: inducing greater concern over misinformation via the AI treatment increases consumption of the most trusted source, even while reducing overall trust in news content (the trusted source included).

Additional analysis sheds further light on the potential dynamics at play in the data and on the plausibility of our mechanism. First, for the tracked subsample, we find that individuals who were relatively less interested in political news and who were relatively heavy users of SZ online platforms (both as measured by pre-intervention reading patterns) respond more strongly to the treatment in terms of subsequent

[8] As an additional check, we show that these results do not hinge on the nature of the treatment vs. control quiz, and hold even when we compare the treatment to a non-respondent comparison group of SZ readers who did not take any quiz.

engagement. These findings align with our model's predictions that individuals with less knowledge of the topic at hand and with a higher prior assessment of the source's trustworthiness will exhibit a stronger demand response. Second, we also find that the treatment increases the consumption of articles on hard news topics (politics and economy) by the same subset of individuals, in contrast with a statistically insignificant, though imprecisely estimated, effect on clicks on sports-related content. This indicates that the demand response is concentrated in topics that are arguably more exposed to misinformation concerns raised by our treatment.

The underlying dynamics suggested by our results have potentially broader implications. For the news industry, they point to a possible strategic business response to the challenge posed by AI-generated content. From a broader societal perspective, our findings provide a nuanced counterpoint to concerns that AI (and misinformation more generally) will trigger a downward spiral of distrust in the information environment: increased scarcity may raise the rewards for trustworthiness.[9] Moreover, as AI tools become more proficient, news consumers may find it increasingly difficult to distinguish between real and synthetic content (Kamali et al., 2024), such that our documented effects could represent a lower bound on the impact on trust and demand. Though – as our conceptual framework underscores – this would only hold as long as news outlets' ability to help readers keeps pace with the challenge.

Our model is furthermore suggestive that the same dynamic may apply to the role of expertise in other contexts likewise affected by evolving technologies. For instance,

[9] Global survey evidence from the Reuters Institute Digital News Report (June 2025) provides suggestive corroboration of these findings (Newman et al., 2025). Specifically, when asked where they would turn to check the veracity of suspected false online news, respondents across demographics most often chose the response: "A news source I trust."

large language models (LLMs) can produce convincing expert-sounding content that is difficult for laypeople to evaluate. Our results indicate that, rather than replacing experts, this development could increase demand for trustworthy sources able to evaluate the AI-generated content. This logic extends beyond AI: any technology that favors misinformation – in the sense of making it harder to distinguish authentic from manufactured content – could activate the same dynamics. Likewise, it extends beyond the realm of politics and current affairs, as misinformation and "slop" can affect many different types of content.

Our paper contributes to three strands of literature. First, it relates to prior work on the prevalence and spread of misinformation online (Allcott and Gentzkow, 2017; Pennycook and Rand, 2021; Vosoughi et al., 2018; Hengel et al., 2025). In this context, several studies examine the impact of fact-checking on voter beliefs, demand for news, and subsequent actions (Henry et al., 2022; Guriev et al., 2025; Nyhan et al., 2020; Chopra et al., 2022).[10] More closely related to our paper, two contributions study experimentally the effect of raising individuals' awareness of their limited ability to detect misinformation (Assenza et al., 2024) and improving their perception of their own ability (Harris et al., 2024). These interventions increase willingness to pay for protection against misinformation but have no impact on concern over misinformation or on the consumption of different outlets (e.g., politically-aligned or mainstream). Besides our emphasis on AI-enabled misinformation, our paper differs from these studies in two important ways. First, while these contributions focus on hypothetical misinformation insurance and individuals' stated preferences, our study

[10] See Lazer et al. (2018) and Zhuravskaya et al. (2020) for reviews on these topics.

is based on a field experiment that allows us to observe both survey outcomes and actual online engagement and subscription choices. Second, our conceptual framework enables us to interpret the observed patterns, illustrating how and why demand for a trustworthy outlet can rise even as trust in news content, including that in the trustworthy outlet itself, declines.

Second, our paper intersects with previous research on the role of technology, especially AI, in creating and spreading online news content, including misinformation.[11] Groh et al. (2024) show that GenAI can increase the quantity and quality of deepfakes, making synthetic images, audios, and videos indistinguishable from real ones. Meanwhile, Vaccari and Chadwick (2020) underline the potential of deepfakes to erode trust in news content on social media more broadly. Relatedly, based on online experiments, Toff and Simon (2025) document that news articles generated with AI attract lower trust than those written by humans, and Kreps et al. (2022) shows individuals are largely incapable of distinguishing between AI- and human-generated text. Unlike these studies, our field experiment setting allows us to elicit both survey outcomes and revealed preferences without priming respondents about misinformation. Most importantly, grounded in our theoretical framework, we show that exposure to AI-generated misinformation can reduce consumers' confidence even in their most trusted source, but with the crucial nuance that it may nonetheless also raise demand for this very same source.

Finally, our paper connects to work proposing strategies to curb low-quality information within the broader platform ecosystem. Costello et al. (2024) shows that

[11] See Simon and Altay (2025) for a survey on the topic.

LLMs can be used to durably correct conspiracy theories online. Ahmad et al. (2024) highlight information interventions to prevent companies from inadvertently funding misinformation websites. Allen et al. (2021) study different approaches to scaling fact-checking on social media, and find that crowd-sourced judgments are similar in quality to those of experts. Building on these studies, we emphasize the important role of high-quality, trusted news outlets in providing authentic information, especially as GenAI becomes a more prevalent tool for generating misinformation.[12] In doing that, our evidence reinforces the importance of understanding the role of demand for trusted information, and how it responds to shocks to the supply of misinformation (Simon and Altay, 2025).

The remainder of the paper is organized as follows. Section 2 provides background information on SZ and the German context. Section 3 lays out our conceptual framework for studying the interplay between GenAI, misinformation, and trust and the impact of the latter on news consumption. Section 4 describes the field experiment, and Section 5 presents the results. Section 6 concludes, with a discussion of our findings and their broader implications.

[12] More broadly, our paper also relates to research on raising awareness about misinformation through educational interventions (Badrinathan, 2021), on limiting the impact of fake reviews online (Ananthakrishnan et al., 2020), and on curbing deceptive claims by companies on social media (Fong et al., 2024).

2 Background

2.1 Süddeutsche Zeitung

Süddeutsche Zeitung (SZ) was established in 1945 in Munich (Bavaria). Over the years, SZ has acquired a reputation as one of the most prominent news outlets in Germany. It is the most widely sold broadsheet daily newspaper in the country, with a daily paid circulation of over 260,000 copies as of 2024, as well as 295,000 online subscribers. SZ is considered center-left in its editorial stance and covers topics ranging from politics to business, culture, sports, and entertainment. Most recently, SZ has gained international recognition for playing an important role in breaking the Panama Papers and Paradise Papers stories that exposed tax evasion and financial secrecy on a global scale. The reputation and quality of SZ's journalism are akin to those of the New York Times in the US and the Guardian in the UK.

SZ monetizes its content online through subscriptions and online ads, and therefore, it has a key goal of increasing engagement with its content to ensure its financial viability. There are various subscription tiers for accessing digital content, ranging from 14.99 to 34.99 Euros per month, with a trial subscription priced at 0.99 Euros for the first month. Given the publication's reputation, companies from all industries advertise on SZ's website. Around 75% of SZ's readership is based in Germany, and the largest share of readers is between 40 and 60 years old.

2.2 The German News Industry

The German news industry has a diverse ecosystem of print and digital publications reflecting a broad spectrum of political perspectives. Its landscape mirrors that of other Western democracies in terms of journalistic quality and political slant. Apart from SZ, on the center-left, the weekly Der Spiegel is known for its investigative journalism and comprehensive news coverage. On the center-right, the daily newspapers Frankfurter Allgemeine Zeitung (FAZ) and Die Welt have significant readership, offering conservative perspectives on national and international issues. At the populist end of the spectrum, Bild, published by Axel Springer SE, remains Germany's most widely read tabloid, known for its sensationalist reporting.

As in the United States (Seamans and Zhu, 2014; Djourelova et al., 2024), the rise of the internet and the dominance of digital platforms over the past two decades have impacted readership and revenues of mainstream German news outlets. A decline in print circulation in the 2000s has led news outlets to adopt a variety of digital strategies, such as subscriptions, paywalls, and investing resources in quality investigative stories. The rise of social media and news aggregators such as Google News has led to additional financial uncertainty and regulatory challenges (Calzada and Gil, 2020).

2.3 AI and the News Media

The rise of online misinformation and the potential for AI tools to exacerbate this problem by increasing the ease with which people can manipulate content have com-

plicated the media ecosystem across the world, including in Germany. Indeed, about 42% of German newsreaders (and 58% across the world) have highlighted the inability to tell apart real from fake content online as a major problem (Newman et al., 2025). Moreover, evidence suggests that image-based misinformation spreads more widely than text alone (Garimella and Eckles, 2017). The need for news producers to explicitly label content (text, images, videos, etc.) generated by AI is being put forward as part of regulation (e.g., the EU AI act), precisely because of the potential threat of deepfakes to society.[13]

Beyond regulation, the proliferation of "AI slop" online (Hoffman, 2024) has prompted journalists to highlight their ability to distinguish genuine from machine-manipulated images and videos. For example, the New York Times and CNN have teams assigned to fact-check visual content in particular to clarify for readers which images and videos are real and which are synthetic.[14] To highlight the difficulty in distinguishing real from synthetic images, mainstream news outlets often have such quizzes for their readers. Alongside these efforts, there has been a growing number of explainer articles and guides that educate audiences on "how to" evaluate content critically to defend against AI-manipulated information. The interplay between AI-based misinformation, particularly deepfakes, and the online news ecosystem underscores the broader significance of our experimental setting.

[13]For a broader discussion, see https://cjel.law.columbia.edu/preliminary-reference/2024/deepfake-deep-trouble-the-european-ai-act-and-the-fight-against-ai-generated-misinformation/.

[14]See for example, the New York Times Visual Investigation page here https://www.nytimes.com/spotlight/visual-investigations

3 A Simple Framework

To guide our empirical exercise, we outline a simple conceptual framework. The goal is not to provide a complete theoretical account, but rather to broadly capture the possible implications of misinformation for news consumption and how trustworthiness can affect those implications.

3.1 Basic Environment

A news consumer wants to learn about a variable $X \in \mathbf{R}$, to which they attach a prior probability $X \sim N(0, \sigma_X^2)$. They do so in order to decide on an action $a \in \mathbf{R}$, so as to minimize a loss function $L(a, X) = (a - X)^2$.

In order to learn about X, the consumer can obtain different signals, $s_i = X + \varepsilon_i$, with $\varepsilon_i \sim N(0, \sigma_i^2)$, where we can think of i as a news source or outlet. An outlet i is of higher quality than outlet j if $\sigma_i^2 < \sigma_j^2$.

The value of each signal to the consumer can be determined in a standard signal extraction problem. Under the assumptions of normality and a quadratic loss function, the optimal action that minimizes the expected loss, having observed signal s_i, is given by the posterior mean conditional on s_i. The corresponding loss function is in turn given by the conditional variance; under the normality assumption, this conveniently boils down to a function of the variances. We denote the expected value from observing signal s_i as V_i:

$$V_i = -E[(a - X)^2 | s_i] = -\frac{1}{\phi_X + \phi_i}, \tag{1}$$

where $\phi_i \equiv \frac{1}{\sigma_i^2}$ is the precision of the signal coming from outlet i.

Now let us benchmark all sources by setting σ_0^2 to represent the quality of information generally available to the consumer – for instance, from doing their own research using online sources. We posit that demand for content from source i, denoted as d_i, is increasing in the value that it adds over and above s_0:

$$d_i \equiv V_i - V_0 = \frac{1}{\phi_X + \phi_0} - \frac{1}{\phi_X + \phi_i}. \tag{2}$$

There will only be positive demand for a source that is seen as at least as good as the benchmark ($\sigma_0^2 > \sigma_i^2$), which we will refer to as "high-quality" sources.[15]

3.2 The Impact of GenAI Misinformation

Now consider the advent of a technology that makes all signals worse, by adding a $\Delta > 0$ to their variance.[16] This is meant to capture the role of GenAI content such as "deepfakes", for instance, which would presumably call into question the veracity of information in all platforms: even high-quality sources will be more often confronted with situations where they cannot verify the authenticity of a given piece of content.

Consider the "precision gap" between a high-quality source i and the benchmark, $\frac{1}{\sigma_0^2} - \frac{1}{\sigma_i^2}$, which with the new technology becomes $\frac{1}{\sigma_0^2+\Delta} - \frac{1}{\sigma_i^2+\Delta}$. It is easy to verify

[15] Specifically, the signal extraction problem entails that consumers will be willing to pay for access to the source as long as d_i is greater than the price charged for that access; that includes the possibility that they choose to consume multiple sources. In Appendix D we present a model where outlets optimally set those prices taking demand into account, and show that the insights carry over to that context.

[16] This linearity assumption is for simplicity, but it is not necessary. In Appendix D we present a version of the model where misinformation enters multiplying the signal variance, and show that our key results still hold.

that the precision gap is now smaller; in fact, as Δ grows arbitrarily large, the precision gap becomes arbitrarily small. It follows straightforwardly from (2) that d_i also becomes smaller, and converges towards zero as the size of the misinformation problem increases. We can summarize this as follows:

Result 1. *The introduction of symmetric misinformation (i.e. leading to the same decrease in the quality of all available signals) leads to a reduction in demand for high-quality sources. In the limit, as misinformation grows, the consumer's demand for the high-quality source approaches zero.*

This simple model neatly captures the idea that misinformation might lead to a downward spiral in the news media environment. In particular, it causes a reduction in the demand for the high quality news source, even while it remains recognized as of higher quality. The consumer lowers their assessment of the quality of all sources, but reduces their demand for the higher quality outlet, as the gain relative to the alternative decreases. If it is costly to produce high-quality news, it is easy to see how that could lead to its becoming a non-viable business.

3.3 The Role of Trustworthiness

To bring the role of trustworthiness into the model, let us introduce a modification whereby a trustworthy outlet (denoted s_1), while still suffering from the misinformation problem introduced by the new technology, is seen by the consumer as being able to mitigate it: as misinformation introduces the Δ component to the variance, the trustworthy source can mitigate that by a factor of $\alpha \in (0, 1)$. In other words, we now have the variance for that source be $\sigma_1^2 + \alpha\Delta$ in the misinformation scenario.

The model thus allows us to draw an important distinction between the consumer's trust in the content provided by the outlet, on the one hand, and their perception of the trustworthiness of the outlet in terms of its ability to mitigate the broad issue of misinformation. Trust in content can be captured by the precision of the signal, $\frac{1}{\sigma_1^2+\alpha\Delta}$ as it represents the consumer's assessment of its quality; trustworthiness, in turn, can be captured by $\frac{1}{\alpha}$. In what follows, we will refer to "trust" and "trustworthiness" as these separate concepts; they are obviously related as trust is an increasing function of trustworthiness.

We can rewrite (1) as a function of the variance of the news source:

$$\Lambda(z) = \frac{\sigma_X^2 z}{\sigma_X^2 + z}. \tag{3}$$

It follows that d_1 is given by:

$$d_1 = \Lambda(\sigma_0^2 + \Delta) - \Lambda(\sigma_1^2 + \alpha\Delta) \tag{4}$$

Differentiating that with respect to Δ yields:

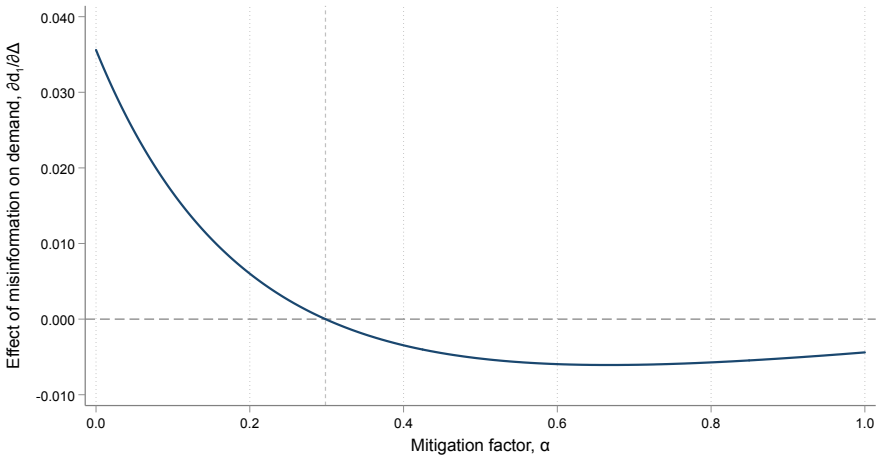
$$\frac{\partial d_1}{\partial \Delta} = \frac{(\sigma_X^2)^2}{(\sigma_X^2 + \sigma_0^2 + \Delta)^2} - \frac{\alpha(\sigma_X^2)^2}{(\sigma_X^2 + \sigma_1^2 + \alpha\Delta)^2}. \tag{5}$$

It is easy to see by inspection that, for any given values of $(\sigma_X^2, \sigma_0^2, \sigma_1^2, \Delta)$, this number will be positive for a sufficiently low value of α. The pattern is depicted in Figure 1 and can be summarized as follows:

Result 2. *A marginal increase in perceived misinformation leads to an increase in*

demand for a sufficiently trustworthy high-quality source (i.e. sufficiently effective in mitigating the harm of misinformation).

Figure 1: Impact of misinformation Δ on demand d_1, as a function of α



Notes: The figure is drawn for parameter values $(\sigma_X^2, \sigma_0^2, \sigma_1^2, \Delta) = (1, 1.3, 1, 3)$.

The intuition is clear: in an environment with widespread misinformation, a sufficiently trustworthy news source – in the sense that the consumer expects it to do a sufficiently good job in mitigating the challenge posed by that misinformation – becomes relatively more valuable in response to an increase in the level of misinformation. This happens even though the increase in misinformation induces a reduction in trust in the content of that source, in the sense of lowering the perceived quality of its signal from the consumer's standpoint. That is because what matters to the consumer's choice is the source's value relative to the alternatives, which are degrading more rapidly.

Note that our framework clarifies the nature of the challenge faced by high-quality news sources over time: if Δ keeps growing without bound, eventually any given value

of α will be insufficient to sustain demand, and d_1 will fall. Yet equation (5) makes it clear that, if α were to go down at least as fast as Δ increases, d_1 could still be sustained. In other words: as the prevalence and quality of misinformation keeps increasing, trustworthy sources have to make sure that their ability to mitigate it, as perceived by their readers, grows at least as fast.

We can also derive additional interesting comparative statics. For instance, it is straightforward to show that $\frac{\partial^2 d_1}{\partial \Delta \partial \sigma_X^2} > 0$ for sufficiently small α.[17] In short, the increase in demand for a sufficiently trustworthy source, as a result of a marginal increase in misinformation, will be greater when the individual has more difficulty interpreting the underlying outcome. In addition, as illustrated by Figure 1, at low levels of α we have $\frac{\partial^2 d_1}{\partial \Delta \partial \alpha} < 0$: intuitively, demand for the outlet will respond more strongly to the marginal increase in misinformation for individuals who see the outlet as more trustworthy.

3.4 Extensions

It is straightforward to extend our model by adding a supply side in which the high-quality outlet endogenously chooses the level of trustworthiness. As we show in Appendix D, under reasonable assumptions, a higher-quality outlet (lower σ_1^2) will choose to invest in the ability to mitigate misinformation (lower α). Intuitively, this complementarity stems from the fact that the increase in demand coming from an improvement in signal quality gets less diluted by the presence of misinformation

[17] Specifically, computing the cross-partial derivative from (5) yields $\frac{2(\sigma_X^2\sigma_0^2+\sigma_X^2\Delta)}{(\sigma_X^2+\sigma_0^2+\Delta)^3} - \frac{2\alpha(\sigma_X^2\sigma_1^2+\alpha\sigma_X^2\Delta)}{(\sigma_X^2+\sigma_1^2+\alpha\Delta)^3}$, which is positive for sufficiently low values of α.

when the reader sees the outlet as better able to mitigate the latter.[18]

We also show in Appendix D that our results survive incorporating price-setting behavior by outlets. Higher-quality outlets are able to charge higher prices and sell higher quantities, which reinforces their incentive to invest in trustworthiness. A marginal increase in misinformation would still benefit outlets of sufficient quality, as their demand increases as per Result 2. Lower-quality outlets, choosing endogenously lower levels of mitigation, would have to cut prices and quantities in response to the decrease in demand stemming from increases in misinformation. This highlights the logic of the impact of mitigation and trustworthiness in the presence of misinformation: trustworthiness increases market power and, for sufficiently trustworthy outlets, an increase in misinformation ends up reinforcing that increase.

3.5 In Sum

Overall, our simple framework underlines the nuanced impact of the expansion of a technology such as GenAI, and the misinformation that it engenders. To the extent that it leads news consumers to perceive a lower quality of information coming from all sources, this technology can be expected to unambiguously reduce trust in the information environment. However, a source that is deemed sufficiently trustworthy, such that the consumer believes it to be effective at mitigating the decrease in overall quality, will see its demand go up: its relative value increases as a result of the overall degradation in the quality of the environment.

[18] That extension also shows that increased misinformation will tend to lead to more investment in trustworthiness, but only as long as (mitigated) misinformation is not the dominant source of noise; or, put differently, as long as mitigation can still keep misinformation in check.

4 Experimental Setting and Data

4.1 Experimental Setting and Design

In line with our theoretical framework, our experimental setting enables us to look at a group of consumers for whom we can plausibly identify a specific trustworthy news source. At a high level, we aim to highlight GenAI's ability to generate synthetic images, thereby drawing attention to its potential to fuel misinformation. In this context, the idea is to induce an exogenous shock to Δ, i.e., the level of concern with misinformation perceived by those consumers. This will enable us to test whether a trusted news source can generate increased demand by making such a concern salient. We implement this idea by highlighting the challenge of distinguishing AI-generated from real visual content using a picture quiz on the SZ website, and evaluating the impact of that awareness on readers' attitudes and engagement with SZ.

Subjects were recruited from two populations: subscribers and other visitors to the SZ platform. For the first group, on March 1, 2025 individuals on the SZ subscriber mailing list received an email inviting them to take a quiz via an enclosed link.[19] For the second group, between February 25 and March 9, 2025, one percent of visitors to the SZ platform were randomly shown a pop-up window that invited them to take the quiz by clicking on a link.[20] The language of both the email and the pop-up ad was neutral, simply inviting readers to take a "survey on the impact of images,"

[19] Because subscribers were reached by email, the vast majority of the respondents took the quiz on March 1 (71%), March 2 (17%), or March 3 (6%), with a dwindling number over the subsequent week. The distribution of non-subscribers was fairly uniform over the period in which the pop-up invitation was displayed.

[20] Note that German federal elections had been held on February 23.

which will include a picture quiz aimed at "examining the comprehensibility and impact of the images."[21] Neither message included any mention of misinformation or technology. Once a user clicked the link, they were randomized in real time to the treatment or control group, as described below.

It is important to note that the experiment was conducted entirely within the SZ platform, relying on the infrastructure routinely used by SZ for fielding surveys and for randomization, and with no involvement or interaction with any other entity (including the research team). In line with SZ standard procedures, participants received no monetary or other compensation for taking the quiz, and responses were not incentivized. In addition, there was no mention of this being part of a study or of there being different experimental conditions, as disclosure of this kind of information is not part of the company's standard operating procedures.

As participants accessed the quiz, and prior to starting it, they were asked for their age and gender. They were then asked for consent to link their quiz performance to their browsing behavior on the SZ website and app, as required by the European Union's stringent privacy regulations (i.e., the General Data Protection Regulation, GDPR). It is important to note that this consent was obtained before participants took the quiz. No prior beliefs were elicited at this stage, because SZ considered this would not be conducive to a natural user experience: while asking for general demographic information occurs occasionally as part of surveys, elicitation of priors is not part of their routine survey or quiz-taking approach.

After the quiz, participants across both treatment and control were asked to re-

[21] The text of the email is reported in Figure B.1 with the machine translated version in Figure B.2 in the Appendix.

port how hard they thought the quiz was and their concern about misinformation, both on a 0 to 6 scale.[22] We then elicited their willingness to pay for an SZ subscription using a sliding scale of 0 to 45 Euros.[23] Finally, we elicited participants' trust in the content available on various news outlets (SZ, Bild, ARD/ZDF) and social media platforms (Instagram, LinkedIn, TikTok, and X/Twitter).[24] Bild is Germany's largest tabloid publication, as previously discussed, while ARD/ZDF is Germany's public TV broadcaster, akin to the BBC in the UK. As the question is framed in terms of trust in content, we take it to be a direct measure of what we defined as "trust" in the context of our conceptual framework.

The choice of what outlets and platforms to include in the trust question was in part driven by the sensitivity of asking such a question for outlets seen by SZ as direct competitors. We were subsequently able to obtain additional information about news media consumption patterns for a subset of our original respondents.[25] Nearly all respondents reported SZ to be an important part of their news diet, and a large majority (about 86%) also reported consuming news from ARD/ZDF. By contrast, only a small fraction (around 3%) reported consuming Bild, while platforms

[22] The questions were phrased as: "How easy were the picture questions for you?" (0: "very easy"; 6: "very difficult"); and "These days, it is easy and deceptively realistic to create and spread disinformation. What is your assessment of this?" (0: "not problematic at all"; 6: "highly problematic").

[23] It is worth noting that it is challenging to interpret this measure of WTP due to potential strategic considerations. Indeed, since SZ was eliciting these valuations, respondents may have had an incentive not to reveal their true valuation, as this information could be used for price discrimination or personalized pricing down the road. As a result, the WTP variable potentially has systematic measurement error. That said, we will later show results for this outcome, since it was included in our pre-registered design.

[24] The question was phrased as follows: "How trustworthy do you rate the content on the following media platforms?" ("not at all"; "not very"; "moderately"; "very"; plus a "no answer" option).

[25] We were able to field a short follow-up survey in February 2026, as we discuss in Section 5.4. This included questions on news consumption, with results reported in Table A.1 in the Appendix.

such as X/Twitter and TikTok play a negligible role as news sources.[26] This overall pattern – which, as we will see, is consistent with the levels of trust recovered in our experiment – suggests that SZ and ARD/ZDF are seen by our respondents as complementary sources of a similar type of news, provided through different channels (daily newspaper vs. television).[27]

Treatment Group: Subjects assigned to the treatment group saw three pairs of pictures in a sequence (shown in Appendix B).[28] Each pair included an authentic picture and one generated with AI. Participants were asked which image they thought was generated with AI, and could choose one of four options: left, right, both, or neither. Following SZ's standard practice, once participants had responded and before they could proceed to the next pair of pictures (or to the follow-up survey), the correct answer was revealed. The first pair included pictures of extreme weather events; the authentic one, on the right, depicts the damage caused by the floods that hit Valencia (Spain) in 2024. The second pair includes images of politicians; the authentic one is a picture of Barack Obama and Donald Trump laughing together during the inauguration of the first Trump presidency in January 2017 in Washington, D.C. Finally, the third pair includes images of protests, with the authentic one depicting a young woman waving a European flag in a pro-EU demonstration held

[26] Respondents also reported consuming weekly hard news outlets with a broad center-left leaning, such as Der Spiegel and Die Zeit (around 30%). Their consumption of center-right hard news, such as FAZ (around 14%), is smaller but non-trivial. This suggests a strong preference for hard news across different formats, tempered by a center-left orientation.

[27] ARD/ZDF also has a digital app with news articles, which is available free of charge and with no advertising, in line with the public funding model behind the channels.

[28] Each respondent saw the pairs and pictures in the same order, for ease of implementation within the SZ platform. That order was randomly assigned.

in Tbilisi (Georgia) in 2024. The non-authentic pictures were generated – entirely or at least substantially – using AI tools such as Midjourney and Magicstudio, based on an authentic seed image and a prompt. These same images were also used in the control group, which we describe next.

Control Group: Subjects assigned to the control group also saw a sequence of three pairs of pictures (Appendix B). In this case, however, all images were authentic. Specifically, each pair included the authentic picture from the treatment quiz, plus the seed image used to create the AI-generated one. The first pair included the picture of the floods in Valencia alongside a picture of floods that hit Bavaria in 2024. The second pair included the Obama-Trump picture plus the picture of two prominent German politicians: Sahra Wagenknecht, leader of the far-left Sahra Wagenknecht Alliance, and Alice Weidel, leader of the far-right Alternative for Germany (AfD); this picture was the seed used to produce the image in the treatment quiz depicting Weidel with Olaf Scholz, member of Germany's Social Democratic Party (SPD) and Chancellor between 2021 and 2025. Finally, the last pair included the image of the protest in Georgia alongside that of a protest in Paris in 2023. The goal of the control quiz was to prompt respondents to engage with visual content in a way that mimicked their daily news consumption without highlighting issues related to AI-assisted manipulation. To this end, for each pair of pictures, respondents were asked in what country each image was taken, choosing one of four options. As was done for the treatment quiz, the correct answer was revealed after subjects had submitted their response.

Design Choices: The aim of keeping pictures similar across treatment and control was to hold constant across groups any thoughts or emotions evoked by the images, and thus tease out the effect of highlighting how AI can make it difficult to tell apart real and synthetic images. Moreover, having a quiz in the control group also accounts for any entertainment value coming from actively doing a quiz itself. This approach is akin to an active control experimental design (Haaland et al., 2023) and ensures that the act of taking a quiz or being exposed to certain images does not lead in itself to the effects we capture. As a robustness exercise, we draw on samples of subscribers who received the email invitation to take the quiz but chose not to, as a non-respondent comparison group, when measuring the impact of the AI treatment on online behavior.

Our unique collaboration with SZ to implement such a field experiment also raised practical considerations that are not present in lab or online experimental settings. For instance, we were constrained in terms of having a third experimental condition simply because of engineering and software considerations, as the SZ A/B testing tool only supports two testing conditions.[29] Moreover, alternative options also posed major practical challenges: while we did not want to prompt readers about AI in the control group, we could not use a version using AI-generated images without explicitly disclosing their existence in some form. This is because Google indexes SZ webpages, and there are terms and conditions on AI-generated content that could otherwise be violated.[30]

[29] While potentially surmountable in the medium term, these considerations could hamper the timely execution of the experiment, with limited potential upside given a relatively clean existing treatment and control setup.

[30] As an additional example, we considered a different experimental condition using older AI tools

From our design perspective, it is also important to note that we did not mention the term "misinformation" in either quiz, so as to keep the language as neutral as possible. Along with the neutral and simple language used in solicitation emails and ads, this minimizes issues related to experimenter demand effects. In addition, the experiment was conducted online, allowing individuals to take it on their own devices as part of their normal browsing experience. As documented in the literature (De Quidt et al., 2018), conducting the study online, without the physical presence of a researcher, minimizes the potential for experimenter demand effects.[31]

Quiz Performance: In terms of actual and perceived quiz performance, the AI quiz was not easy, which demonstrates the current state of GenAI technology in its ability to create seemingly authentic images. On average, there were about 29% correct responses in the treatment group.[32] The level of difficulty was also reflected in their assessment of perceived difficulty, which was 4.7 out of 6. In contrast, the current events control quiz was less difficult, with an average correct rate of 89% and 72% getting all three answers right. This performance is in line with other

(e.g., Deepdream from 2015 or initial versions of Dall-e from 2019) generating less realistic images. Input from SZ suggested that such a condition would not seem natural to their readers, undermining the ability to claim that the quiz was aimed at improving the readers' experience on the website. In essence, this would effectively amount to deception, raising legitimate concerns for our partner.

[31] Our experimental setup also allows us to observe both survey outcomes and revealed preferences, which typically requires explicit participant recruitment into the experiment (in the field, online surveys, or lab settings). In our setting, there is no mention of researchers being involved, which limits issues of experimenter demand effects as well.

[32] The performance for each question was as follows: 20% correct for the climate pair, 36% for the politicians pair, and 30% for the protest pair. Performance was better than random overall, against a benchmark of 25%, though it is possible that "neither" may have been in practice a less appealing option than the other three for guessing at random. Relative to other contexts where individuals attempted to distinguish AI-generated images from real ones, individuals in our treatment group performed better (see Miller et al. (2023) and https://journalism.columbia.edu/news/cjr-new-ai-campaign as additional examples.

quizzes on general political information involving distinguishing real from fake news stories, such as in Angelucci and Prat (2024). The perceived difficulty was in line with performance at 1.52.[33] While the quiz difficulty differed across treatment and control, so as to demonstrate the challenge posed by dealing with AI-powered content, the amount of time taken by the two groups was not too different, highlighting the significant effort put in by both groups. The median individual took 4:06 minutes to complete the quiz. In the treatment, the median time was 4:20, while in the control group it was 3:54.

4.2 Data and Empirical Framework

The bulk of our analysis focuses on two connected samples of about 17,000 and 6,000 SZ readers. For the first sample, which we refer to as the "survey sample", we gathered data through the survey that hosted the quiz. This means we have data on individuals who took the quiz and answered at least one endline survey question.[34] For these individuals, we have information on self-reported demographics (age and gender) along with stated preference questions at endline. The treatment and control

[33]The real images in the quiz were chosen to mimic, in part, the readers' standard daily information consumption. Some images from the main stories on the homepage (at 9 am local time each day) during the sample period are in Appendix C, suggesting that the control quiz was more challenging than their daily information environment. The day-to-day information on the website includes images, titles, and text to provide context, which is absent from our quiz, making the task more challenging than their daily experience, even for individuals in the control group. In Appendix Table A.14, focusing solely on the control group, we show that results for our key outcomes did not vary depending on whether individuals reportedly found the quiz more or less difficult. Moreover, Appendix Table A.17 shows individuals are largely similar in observable characteristics except for marginal differences in the probability of being a subscriber.

[34]Following SZ's protocol, respondents were not forced to answer any survey question. A proportions test, based on our focal analysis sample, comparing treatment and control for individuals who answered the first question of the survey pre-intervention (on gender) and the final post-intervention question (on WTP) cannot reject the null of no differential responses (p=0.66).

groups are balanced on a number of observables, as seen in Panel A of Table A.2. About 45% of the quiz takers are female and 94% are subscribers, balanced across treatment and control groups. Treatment and control groups are also balanced in terms of age and of the time they took the quiz (most individuals took the quiz in early March). Importantly, the proportion of individuals who consented for their survey responses to be merged with actual browsing data is very similar across the two groups, around 34%. These checks provide validity to our random assignment.

The second sample is focused on about 6,000 SZ subscribers who consented to tracking (which we refer to as the "tracked sample"), and for whom we have additional browsing and subscription status data over time, along with the quiz information. For these individuals, apart from the observables mentioned earlier, we also observe past visits to the SZ website, political interest (based on pre-treatment readership data), location, and subscription type. As shown in Panel B of Table A.2, both the survey-elicited characteristics and those based on actual behavior are balanced between treatment and control groups. Crucially, the number of daily visits to SZ in the three weeks prior to the experiments is also balanced.

In terms of the analysis, we use two approaches based on the type of data. For survey-based outcomes, we estimate the following regression at the individual i:

$$Y_i = \alpha + \beta Treated_i + \epsilon_i \tag{6}$$

where $Treated_i$ indicates whether the individual i took the AI-awareness quiz, and β captures its causal effect on the outcome of interest. Because of the successful randomization, the control variables should not affect the estimate of β; we therefore

estimate the regression without controls as the baseline specification.

The analysis focusing on browsing behavior, for the sample with such information, uses a difference-in-differences (DiD) specification with individual-day level data:

$$y_{it} = \alpha_i + \tau_t + \beta PostQuiz_{it} \times Treated_i + \gamma PostQuiz_{it} + \epsilon_{it} \tag{7}$$

where $PostQuiz_{it}$ is the event variable which takes a value of 1 if date t is following the date individual i took the quiz and 0 otherwise. α_i and τ_t capture fixed effects for individual and date, respectively.[35] The coefficient of interest β captures the average impact of an individual taking the AI quiz relative to the current affairs quiz.[36] We can separately identify the coefficient γ on $PostQuiz_{it}$, as there is a modicum of variation in the dates when individuals took the quiz.[37] We use a three-week pre-period before the quiz and varied the post-period using different time windows.

[35] The individual fixed effects allow us to control for fixed individual characteristics, which will be especially helpful for comparisons with non-respondent comparison groups not exposed to the quiz, as we will discuss in our robustness section. The date fixed effects capture the fact that dates differ in the presence of newsworthy events that would influence engagement with the SZ platforms.

[36] Since we have a binary treatment and our tracked sample took the quiz within a very short time period (see footnote 19), our specification closely approximates the canonical differences-in-differences case, where the TWFE estimator is robust (de Chaisemartin and D'Haultfoeuille, 2025).

[37] In practice, again as per footnote 19, there is a strong correlation between that variable and the date fixed effects. We choose the more conservative (less efficient) approach of including $PostQuiz_{it}$ in the specification; we will also show the baseline results with the more parsimonious specification excluding that variable in Section 5.

5 Results

5.1 Survey Outcomes

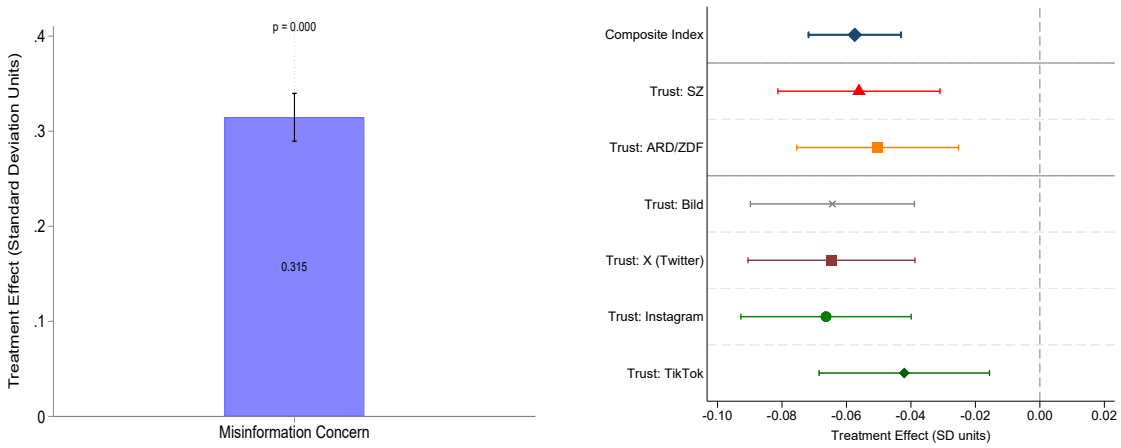
We begin with the average treatment effect (ATE) for the survey sample. The first outcome we analyze is concern with misinformation (measured on a scale of 0 to 6). Panel (a) of Figure 2 shows that it is significantly higher in the treatment group than in the control. The increase is also economically meaningful – 0.31 of a standard deviation – particularly given that misinformation was never explicitly mentioned before this question. This result also serves as a manipulation check: exposure to the AI quiz did in fact raise concern about misinformation, even among a population predisposed to worry about the issue, such as SZ's readers.[38]

Having established our "first stage" in inducing increased concern with misinformation, as captured by the Δ parameter in our model, we then look at reported trust in media outlets and platforms (henceforth platforms). This is a key outcome in the context of our framework, as a shock to Δ induces a decreased trust in content, as captured by $\frac{1}{\sigma^2+\alpha\Delta}$.

The first thing to note is that the different platforms command very different levels of trust in content, consistent with different perceptions of quality (σ^2) and trustworthiness (α). Focusing on the control group, we see that, unsurprisingly, SZ readers and subscribers have high trust in SZ: an average score of 2.91 on our 0-3 scale. This confirms our prior that the experimental sample has an identifiable trustworthy

[38] In the control group, 62% of respondents gave a score of 6, and an additional 30% answered 5. The numbers are 77% and 18%, respectively, for the treatment group.

Figure 2: Average Treatment Effect: Survey Outcomes



(a) Concern with Misinformation (b) Trust in Media Platforms

Notes: Figure (a) captures the treatment effect on concern with misinformation across the treatment and control groups. The dependent variable is on a scale of 0 to 6. The vertical lines represent 90 percent confidence intervals. Figure (b) plots the average treatment effects capturing the impact of the AI intervention on the treatment group relative to the control for trust in SZ, X, Instagram, TikTok, Bild, ARD/ZDF, and a composite index. The horizontal lines represent 90 percent confidence intervals. The treatment effect in Figures (a) and (b) is captured in terms of standard deviation units.

news source. Respondents also report high trust in the public broadcaster ARD/ZDF (2.80), lower trust in LinkedIn (1.32), lower still for Bild (0.82) and Instagram (0.76), and the lowest levels for TikTok (0.26) and X/Twitter (0.25).

In line with Result 1 of the model, Figure 2 (b) shows a negative impact of the AI Awareness treatment on trust in all media platforms, including a composite index, equivalent to approximately 0.05-0.1 standard deviations. While this is consistent with prior findings that exposure to AI-generated content reduces trust in social media (Vaccari and Chadwick, 2020), we document that this ATE applies to low- and high-trust platforms alike – including, notably, SZ itself.[39] The decline in trust is

[39] We find a null ATE of the AI intervention on the reported willingness to pay (WTP) for a subscription to SZ (on a scale of 0 to 45 Euros) (Figure A.2 in the Appendix). As discussed previously, it is difficult to interpret this stated WTP measure since, in our context, there could be strategic

economically meaningful, comparable to treatment effects found for other outcomes such as attitudes towards political correctness (Braghieri, 2024) and the impact of social media usage on mental health outcomes (Braghieri et al., 2022).

The underlying estimates to Figure 2 are reported in Table A.3, in the Appendix: given the different baseline levels of trust, we see a substantially smaller relative decline in trust for SZ and ARD/ZDF than for other, less-trusted sources, in line with the prediction of our model.[40] Interpreted through the lens of the model, the pattern of relative decline in trust across different platforms is inconsistent with what would be produced by symmetric misinformation, as we show in Appendix E. Instead, the data suggest that different outlets are indeed perceived by our subjects as having different α_i: more trustworthy outlets offer greater mitigation ability.[41] In fact, from our estimates we can back out relative α_i parameters: the implied α_i for SZ is about 5 percent of that for X/Twitter (under the multiplicative specification) and as low as 0.5 percent under the additive one, with ARD/ZDF close to SZ and the other platforms an order of magnitude higher.

5.2 Revealed Preference Outcomes: Browsing

A key advantage of our setting is the ability to track the actual behavior of experimental subjects in their engagement with a trusted online news source, in this case SZ. We can do so for our tracked sample of SZ subscribers who agreed to have

incentives to misreport.

[40] Treatment effects are robust to controlling for observables as shown in Table A.4.

[41] Specifically, in the absence of different levels of mitigation ($\alpha_i = 1$), the proportional decline in trust would be largest for the most trusted outlets, in the baseline model with additive misinformation ($\partial \log t_i / \partial \Delta = -t_i$), or identical across outlets, in the model with non-linear misinformation. Both versions are soundly rejected at conventional levels.

their information matched with the survey: following their online engagement with SZ after taking the quiz allows us to peer into their revealed preferences regarding demand for the platform's content.

As noted in Section 4.2, because we have pre-period data for this sample, our preferred approach is to use a DiD strategy, which improves precision. In addition, although attrition in the survey was relatively small, we use an intent-to-treat specification with individuals assigned to the treatment and control groups, as long as they started taking the quiz embedded in the survey link.

5.2.1 Daily Visits: Baseline

We start with a key outcome: the number of daily visits to the SZ website. This is one of the main metrics tracked by SZ, as it aims to increase overall engagement with its content to generate additional advertising revenue and subscriptions. Online visits are an important indicator of the financial health of news outlets and are widely used in academic research (e.g., Cagé et al., 2020; Peukert et al., 2024).

For an initial, model-free look, we consider the t-test for the difference in mean daily visits between the two groups, within the five-day post-intervention period. We find that the means for the two groups (4.83 for the treated and 4.69 for the control) are significantly different (p-value: 0.032). The balance checks in Table A.2, together with a statistically and economically significant (2.9%) estimate, suggest a meaningful impact of the intervention on engagement.[42]

[42] Table A.5 shows the survey estimates when looking only at the tracked sample. The estimates across the eight survey measures are qualitatively similar across the samples; the tracked sample estimates for trust in ARD/ZDF and X are statistically different from the full sample, but of the same sign.

We then turn to the main DiD analysis, using the data available before and after the survey, which allows us to explore the timing in greater depth. We exclude from the analysis the day on which respondents answered the quiz, to avoid the possibility of a mechanical effect on website visits driven by the survey itself. We take the three weeks prior as the pre-treatment period and consider a range of post-treatment windows.

In column (1) of Table 1, we see that for the AI treatment group in the three days after taking the quiz, the number of visits increased by about 2.5% relative to the control group. This effect is detectable for about two weeks post-treatment, declining in magnitude as we increase the post-treatment window to five (column 2), 10 (column 3), and 14 days (column 4). These magnitudes are economically meaningful from SZ's perspective, a point to which we return below when discussing the results for the subscription outcomes.

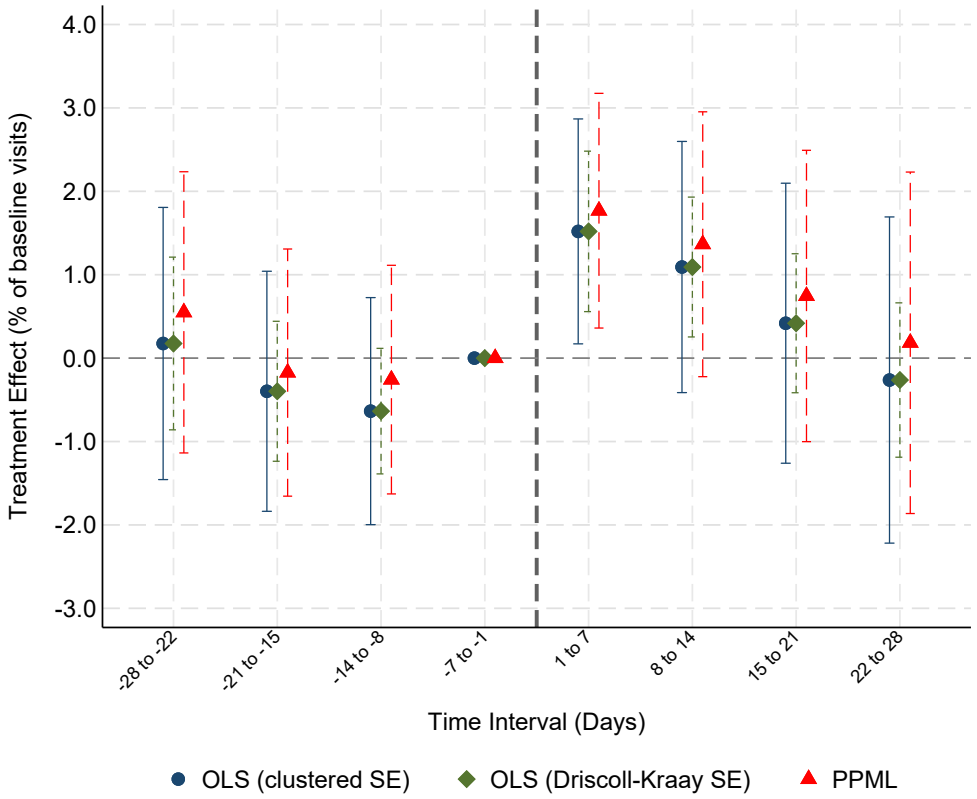
Table 1: Post-Intervention Engagement: Differences-in-Differences Analysis

Dep. variable	(1)	(2)	(3)	(4)
	Number of daily visits			
Timespan	(-21;+3)	(-21;+5)	(-21;+10)	(-21;+14)
AI Treatment × Post	0.114**	0.085**	0.074**	0.067*
	(0.052)	(0.042)	(0.035)	(0.034)
Sample mean	4.690	4.653	4.615	4.614
	(5.534)	(5.493)	(5.457)	(5.463)
Individual FE	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes
Observations	143,208	155,142	184,977	208,845
Adjusted-R^2	0.768	0.768	0.766	0.765
F statistic	6.756	6.224	6.800	6.868

Notes. Standard errors are clustered at the individual level. Significance at the 10% level is represented by *, at the 5% by **, and at the 1% by ***. The dependent variable is the number of daily visits by the SZ reader. The time window for analysis varies from 21 days prior to the intervention to 3 days post in column (1), 5 days post in column (2), 10 days post in column (3), and 14 days post in column (4). We control for Post, which equals 1 if the day is after the date the individual took the quiz. AI Treatment is equal to 1 if individual i was in the treated group and zero otherwise.

To further analyze the timing of the effect, we conduct an event-study-type analysis reported in Figure 3. To minimize noise from daily fluctuations, we estimate interactions between the treatment dummy and indicators for each week before and after the week in which the individual took the quiz, with the week before that date as the reference point. As shown in Figure 3, there is no evidence of pre-trends prior to the experimental intervention. Next, in the post-period, the effect is strongest in the first week, diminishes in the second week, and then largely goes back to the pre-intervention baseline.

Figure 3: Visits to SZ Online: Event Study Estimates



Notes: Figure 3 depicts estimates obtained from linear (OLS) and non-linear (Poisson) models, and also using Driscoll-Kraay standard errors. The vertical lines represent 90 percent confidence intervals. AI treatment is equal to one if an individual was in the AI quiz group and zero otherwise. The treatment effect is measured in percent of baseline visits to ensure comparability of estimates across specifications.

5.2.2 Daily Visits: Heterogeneity

We then conduct heterogeneity analyses to shed light on the mechanism proposed by our conceptual framework. To that end, we classify readers (in the tracked sample) based on their pre-experiment behavior on SZ digital platforms. First, using data

on reading activity, we code dummy variables for whether a reader engaged with particular news topics prior to the treatment; we look at the three weeks before the experiment to reduce noise from short-term reading patterns.[43] The quiz embedded in the survey was focused on political issues (climate, political leaders, and protests); for that reason, we are particularly interested in identifying readers with a specific affinity for political topics. We thus split the sample according to whether the individuals are classified as having a politics affinity (57% of the sample) or not (43%). Second, we compare relatively heavy and light users of SZ, classifying them by contrasting individuals in the top and bottom quartiles of the number of visits in the three weeks before the experiment (145 and 20 visits, respectively).

We can conceptualize these two dimensions of heterogeneity through the lens of the model. First, we can think of individuals without political affinity as having a higher σ_X^2, i.e., a more diffuse prior. More broadly, an intervention could provide novel information to previously uninformed individuals, or serve as a nudge or reminder to those already knowledgeable about the topics mentioned (Nelson, 1974). Political affinity serves as a marker of prior knowledge, helping distinguish between informative and persuasive channels. Second, pre-treatment engagement levels can serve as a proxy for the degree of trustworthiness (i.e., $\frac{1}{\alpha}$) attached to SZ.

Table 2 shows the results of these heterogeneity exercises.[44] First, for politics

[43] The comparison between heavy and light users was pre-registered, but the affinity dimension was not, as it was not clear before the experiment that we would have access to this type of information. Specifically, SZ computes reading affinity variables based on the Piano machine-learning tool (more details here: https://docs.piano.io/user-interest-segments/).

[44] We also look at another (pre-registered) heterogeneity dimension: performance in the quiz. Specifically, to capture differences across groups, we define a dummy variable $Hard$ equal to 1 if the individual is above the median of their group for assessing the quiz as difficult, and 0 otherwise. As shown in Appendix Table A.6, those who found the quiz relatively hard and those who performed

affinity, we see a distinct pattern: individuals with relatively low prior affinity respond more strongly to the treatment, with an increased number of visits (column 1). The effect size is about 3.4% for those who do not have an affinity for reading politics, while the rest display a statistically insignificant effect, less than one-half the magnitude of the former group (column 2). This is consistent with the prediction from our conceptual framework, where relatively uninformed individuals (with a more diffused prior on related topics) would be expected to display a stronger demand response to increased misinformation, and suggests an informational rather than a persuasive mechanism. Next, focusing on the model's trustworthiness prediction, we find that it is the heavy readers who respond more strongly to the treatment by increasing their visits (columns 3 and 4), again in line with the model's predicted comparative statics for lower α (i.e., higher trustworthiness, $\frac{1}{\alpha}$).[45]

worse exhibit significantly larger treatment effects across both misinformation concern and SZ website visits. We should use caution in interpreting these estimates as causal, since quiz performance assessment was not randomly assigned; in particular, individuals who found the quiz (relatively) hard in the control group could be systematically different from those in the treatment group.

[45] Interestingly, Appendix Figure A.3 shows that the first stage of misinformation concern is equally strong among both heavy and light readers, suggesting that they are not systematically different despite differences in their prior engagement with SZ. The first stage is stronger for individuals with low politics affinity; note that this is not a prediction from the model, since it has no predictions about the strength of the shock to misinformation perceptions.

Table 2: Visits Heterogeneity by Politics Affinity and Reading Intensity

Dep. variable	(1)	(2)	(3)	(4)
	Number of daily visits			
Subgroup	Low Politics	High Politics	Light Reader	Heavy Reader
AI Treatment × Post	0.1385**	0.0522	0.0003	0.2584**
	(0.0618)	(0.0603)	(0.0365)	(0.1264)
Sample mean	4.030	5.514	0.347	11.480
	(5.103)	(5.760)	(0.863)	(5.988)
Individual FE	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes
Observations	63,778	84,838	38,584	38,532
Adjusted-R^2	0.7738	0.7627	0.1512	0.5760
F statistic	2.914	5.441	0.075	2.459

Notes. Standard errors clustered at the individual level. Significance: * 10%, ** 5%, *** 1%. The dependent variable is the number of daily visits by the SZ reader. The time window for analysis varies from 21 days prior to the intervention to 5 days post. AI Treatment equals 1 if individual i was in the treated group and 0 otherwise. We control for Post, which equals 1 if the day is after the date the individual took the quiz. Columns (1) and (2) are based on low and high politics affinity individuals while in columns (3) and (4) is based on engagement, both using pre-experiment data.

5.2.3 Article Types

Another window into the mechanism is provided by a different set of browsing outcomes: the types of articles individuals clicked before and after our intervention. Using the sections ("ressorts") into which SZ content is classified, we draw a contrast between what we may call "hard news" (politics and the economy) versus sports. This is meant to capture differences in news topics that may be more prone to misinformation concerns. Sports is a canonical example of content in which the typical news piece naturally has greater verifiability and is thus less subject to manipulation or

slant (Gentzkow and Shapiro, 2006).[46] On the other hand, heightened concerns about misinformation could lead individuals, especially those with low political knowledge, to seek out more political and economic news. This analysis comes with the caveat that clicks on specific article categories are a sparser outcome and hence, can have more imprecise estimates.

The results in Table 3 below highlight a coherent pattern. In column (1), we see that for the full sample, engagement with politics increased after the treatment, though the estimate is slightly noisy (p=0.11). Given our model's prediction and the fact that we are examining political and economic news, we analyze heterogeneity using our pre-experiment politics affinity measure. Column (2) highlights a clear pattern, in line with our model and informational mechanism: those with low politics affinity increase their engagement with political news after being treated with the AI quiz. In contrast, column (3) shows that those with high politics affinity prior to the experiment kept their behavior unchanged across treatment and control. Similarly, we can see that individuals with higher levels of prior engagement with SZ increased their consumption of hard news articles (column 4), while those with lower levels of prior engagement (a proxy for trust) didn't see a change (column 5). In Table A.7, in the Appendix, we analyze engagement with sports articles to find statistically insignificant, though imprecise, effects overall and across all sub-groups (columns 1-6) suggesting that the response of the effect is concentrated in hard news. With

[46] That noted, a priori, the impact of our intervention on hard news relative to soft news (such as sports) is ambiguous. For example, it could well be that individuals induced to visit SZ more because of increased concern about misinformation would, once on the platform, simply scroll more on the homepage, browsing through different articles or even click on other content that piques their interest (say, an article on the football club FC Bayern).

the caveat of limited power and imprecise estimates, the results do suggest that individuals who respond more are those whom our conceptual framework would predict to be most sensitive: individuals who would feel the more pressing need for help with detecting misinformation (as captured by a low politics affinity), and those who attach greater perceived trustworthiness to SZ (as proxied by their pre-existing levels of engagement).

Table 3: Post-Intervention Content Engagement: Hard News

	(1)	(2)	(3)	(4)	(5)
Subgroup	Full	Low Politics	High Politics	Heavy Reader	Light Reader
AI Treatment × Post	0.3140	0.6442**	0.1055	1.0049*	-0.0499
	(0.1981)	(0.2729)	(0.2982)	(0.5879)	(0.1309)
Sample mean	10.251	8.116	12.638	22.785	0.738
	(20.286)	(18.400)	(21.929)	(29.296)	(4.075)
Individual FE	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes
Observations	155,142	63,778	84,838	38,532	38,584
Adjusted-R^2	0.5685	0.5791	0.5503	0.5473	0.1899
F statistic	3.594	2.786	4.630	2.605	0.450

Notes. Standard errors clustered at the individual level. Significance: * 10%, ** 5%, *** 1%. The dependent variable is the number of daily clicks on politics and economy articles by an individual reader. The time window for analysis varies from 21 days prior to the intervention to 5 days post. AI Treatment equals 1 if individual i was in the treated group and 0 otherwise. We control for Post, which equals 1 if the day is after the date the individual took the quiz. Column (1) contains the full sample. Columns (2) and (3) are based on low and high politics affinity individuals, while columns (4) and (5) are based on engagement, both using pre-experiment data.

We also analyze whether our treatment led to changes in engagement with AI-related articles in the days following our intervention, to ensure that it did not merely nudge readers to browse them. Using a large set of keywords to identify AI-related

articles on SZ, we find no change in the consumption of AI-related articles between treatment and control (Panel A of Table A.8 in the Appendix); we also find no pattern in the daily number of AI-related articles published on the SZ digital platform before and after the intervention (Panel B). This result alleviates the concern that our results may be driven by omitted factors related to the coverage of specific topics.

5.2.4 Alternative Mechanisms and Robustness

To highlight the robustness of our results and rule out the possibility that the increased demand may reflect alternative mechanisms, we conduct a series of additional analyses.

First, for website engagement as the outcome variable, we can construct a non-respondent comparison group. We drew a random sample of 3,000 individuals who received the invitation email but did not take the quiz, out of whom 2,344 visited SZ during the sample window. This group provides a distinct counterfactual benchmark, with individuals who were not exposed to any quiz; while the choice to not take the quiz is endogenous, note that the individual fixed effects can account for unobserved characteristics affecting that selection.[47] As can be seen in column (2) of Appendix Table A.9, there is a statistically and economically significant effect in the first five days after the experiment, suggesting our treatment effect is not necessarily attributable to the nature of the control quiz.[48]

[47] We set the treatment date to be the 1st of March since it is the day that subscribers received the email and most took the quiz.

[48] We also conduct several additional exercises to test the robustness of our results to different choices of specifications and variables. In Appendix Table A.9, we find that results are robust to a non-linear Poisson model, as in the event study, to account for the count nature of our dependent variable (column 3); to using total daily time spent by a user on SZ online (columns 4 and 5), to using a

Next, we draw on additional browsing data from SZ to rule out short-term attention or entertainment effects resulting from our AI intervention. In particular, we analyze the propensity for individuals in our sample to click on other survey links on SZ in the post-intervention period, to test whether the additional engagement simply reflected a more captivating experience with the AI quiz relative to the control quiz. We find that there is no differential engagement with surveys on SZ across treatment and control (Table A.9, columns 8-9). In addition, our baseline results are robust to excluding from the sample any individuals or visits that engaged with surveys and quizzes on the SZ platforms after our experiment (Table A.11). These results confirm that our findings are not driven by individuals seeking to engage with other surveys and quizzes online after our intervention.

As a final step, we also analyze whether these individuals were more likely to seek more entertaining content. While we do not observe browsing activity off-platform, we do observe incoming traffic to the SZ platform from social media websites. With the caveat of traffic from social media links being sparse, we find that the overall effect is driven by organic visits, with no change in social media-mediated visits to the SZ platforms (Table A.12). These results also provide suggestive evidence that the treatment increased engagement with SZ itself, but not with other platforms.

14-day pre-period window (column 6), and to a more saturated specification with time-since-quiz fixed effects in addition to date fixed effects (column 7). Finally, following the discussion in Section 4.2, in light of the high correlation between $PostQuiz_{it}$ and the date fixed effects (as per footnote 19), we show in Appendix Table A.10 that results are robust to excluding that dummy variable.

5.3 Revealed Preference Outcomes: Subscription Retention

Continuing with revealed preference outcomes, we turn to individual subscription behavior and analyze, for individuals in the tracked sample, whether they discontinue or maintain their subscription to SZ.[49] Note that there is no variation in subscription status before the treatment, since all tracked individuals are subscribers and variation at the individual-day level is limited: the vast majority of subscribers retain their status on any given day, and dropped subscriptions are an absorbing state in the context of our analysis.

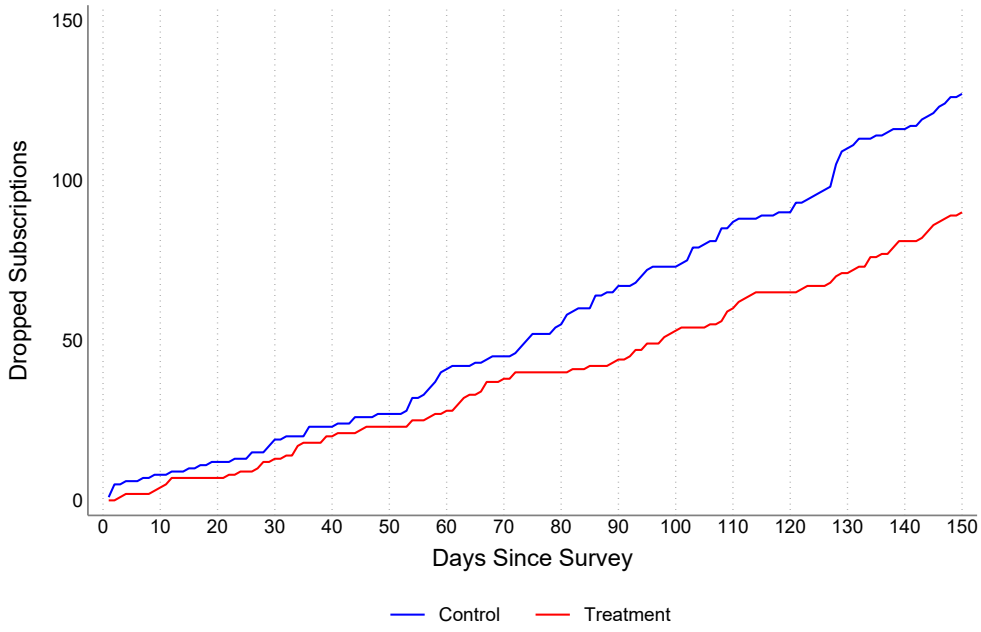
5.3.1 Baseline

We begin with a descriptive look at the data before turning to formal statistical analysis. Figure 4 shows the cumulative number of dropped subscriptions, and the pattern is striking: an immediate gap emerges, with more dropped subscriptions in the control group, and the gap persists and widens over time.

This translates into the retention pattern in Table 4, where the dependent variable in each column is the individual subscription retention two through five months after the intervention. The treatment group displays a consistently higher probability of staying on as SZ subscribers. The magnitudes are instructive: as over time the probability of remaining a subscriber naturally declines, the treatment effect correspondingly increases. This reflects a stable effect in terms of attrition rate, with the treatment group consistently displaying a rate around 30% lower than the control

[49] Almost all subscriptions are on auto-renewal. While there are many different kinds of subscription packages, they are largely either on a monthly or yearly basis; yearly subscriptions may be structured with either monthly or yearly payments.

Figure 4: Post-Intervention Dropped Subscriptions



Notes: This figure plots the cumulative number of dropped subscriptions to SZ for the first five months post-intervention, by group.

group. After five months, the effect accumulates to the point where the probability of remaining a subscriber is about 1.2% higher for the treatment group.[50]

[50]These magnitudes are comparable to the three most recent email experiments run by SZ that measured retention three months post-intervention. Our estimated treatment effect of 0.74 percentage points falls well within the range across those experiments, and is comparable to the average of the treatment effects estimated in them (0.55 p.p.).

Table 4: Post-intervention Subscription Status

Dep. variable	(1)	(2)	(3)	(4)
	Subscription Status			
Months Post-Intervention	2 Months	3 Months	4 Months	5 Months
AI Treatment	0.0042	0.0074**	0.0080**	0.0119**
	(0.0027)	(0.0034)	(0.0040)	(0.0047)
Sample mean	0.989	0.982	0.975	0.965
	(0.105)	(0.133)	(0.157)	(0.184)
Observations	6,158	6,158	6,158	6,158
R-squared	0.0004	0.0008	0.0007	0.0010
F-statistic	2.436	4.783	4.057	6.423

Notes. Robust standard errors in parantheses. Significance at the 10% level is represented by *, at the 5% by **, and at the 1% by ***. The dependent variable is whether the SZ reader is still subscribed to SZ on that day. The post intervention time window for analysis varies from 60 days (2 months) post in column (1), 90 days (3 months) post in column (2), 120 days (4 months) post in column (3), and 150 days (5 months) post in column (4). AI Treatment is equal to 1 if individual i was in the treated group and zero otherwise.

The effect size is economically consequential for SZ, especially given the light-touch, low-cost nature of the intervention. To put the magnitude of the effects in context, it is important to note that about 80% of experiments in digital markets yield null results (Kohavi et al., 2013), including interventions aimed at fighting misinformation (Betzer et al., 2025; Aslett et al., 2022). Among more recent interventions, the magnitude is comparable to what has been found in response to an explicit advertising approach designed to combat misinformation (Lin et al., 2024).[51]

A conservative back-of-the-envelope calculation suggests that, if scaled across the outlet, this mild intervention would increase annual revenues from subscriptions by about 456,000 euros (base annual price × increased retention × subscriber base =

[51] It is pertinent to note that we mitigate issues related to the Hawthorne effect since we had both a treatment and control group take a quiz, and they were unaware of being part of an experiment.

129 euros $\times$ 0.012 $\times$ 295000). The outlet would also benefit from increased advertising revenue due to more visits by existing subscribers in the short-run. To get a lower bound on total daily visits, we use data from the alternate control group to estimate that SZ gets at least 900,000 daily visits (3.1 $\times$ 295000) from (less-engaged) subscribers. This implies that our intervention, administered from time to time, could increase visits by about 23,000 (0.025 $\times$ 900,000). Additionally, there would be an increase in visits by new or retained subscribers annually by over 3.9 million (3500 $\times$ 365 $\times$ 3.1).

5.3.2 Heterogeneity and Robustness

Next, we conduct the same heterogeneity analysis that we pursued for the browsing outcomes, again in the spirit of providing additional evidence for our proposed mechanism. This analysis comes with the caveat that as we split the sample we run into potential issues of sparsity in the outcome variable and hence noisier estimates. We focus on subsamples based on prior politics affinity and pre-treatment levels of engagement with SZ. Table 5 shows the results for the probability of remaining a subscriber, focusing on the horizon of five months after the intervention.

The results dovetail with what we found for daily visits and hard news articles. First, the increase in retention was driven by individuals who had relatively low prior politics affinity (column 1) relative to those who had higher levels of engagement with politics prior to the intervention (column 2). Next, we also see that individuals who had higher prior engagement with SZ overall (top quartile) are more likely to retain their subscription than those with lower levels (bottom quartile) (columns 3 and 4).

Overall, these heterogeneity results are in line with the visits heterogeneity analyzed before, and in line with the predictions of our theoretical model.

Finally, to ensure the robustness of our results for subscriptions, we adapt the idea of a non-respondent comparison group based on SZ subscribers who were invited to take the quiz but did not eventually take it. We do so by focusing on 479 highly engaged non-respondents who had more than 234 visits to SZ between 8-28 February 2025, thereby aligning them with the top 10% of the treatment group. Despite selecting highly engaged individuals, Table A.13 in the Appendix shows that this non-respondent comparison group is less likely to remain subscribers in the subsequent period than individuals who took the AI quiz.

Table 5: Subscription Retention: By Politics Affinity and Reading Intensity

Dep. variable	(1)	(2)	(3)	(4)
	Subscription status (5 Months)			
Subgroup	Low Politics	High Politics	Heavy Reader	Light Reader
AI Treatment	0.0174**	0.0086	0.0145*	0.0044
	(0.0075)	(0.0063)	(0.0085)	(0.0111)
Sample mean	0.964	0.967	0.972	0.952
	(0.186)	(0.178)	(0.164)	(0.213)
Observations	2,433	3,234	1,477	1,472
R-squared	0.0022	0.0006	0.0019	0.0001
F statistic	5.409	1.877	2.882	0.157

Notes. Robust standard errors in parentheses. Significance at the 10% level is represented by *, at the 5% by **, and at the 1% by ***. The dependent variable is whether the SZ reader is still subscribed to SZ 5 months (150 days) after the intervention. Columns (1) and (2) are based on low and high politics affinity individuals while in columns (3) and (4) is based on engagement, both using pre-experiment data..

5.4 Long(er)-Run Outcomes

Our baseline analysis reveals an interesting pattern: although the treatment's impact on daily visit decisions may fade, it remains strong when individuals face the choice to discontinue or keep their subscription. As they think this through, exposure to the AI treatment may provide an additional nudge to keep paying for SZ access.

To further explore the persistence of effects on stated attitudes, we examine data from a follow-up survey conducted approximately one year after the original experiment. Specifically, in February 2026, SZ agreed to email the individuals in our sample again, asking them to respond to a quick two-minute survey. The survey contained the same questions on the concern with misinformation and on trust in the content of different platforms as in the original survey.[52] A total of 1,710 individuals responded; a subset of 1,278 were part of the tracked sample, enabling us to assign them to their treatment status (655 control, 623 treatment). Selection into response being evidently non-random, we should be careful to interpret any estimates causally; that said, this unique opportunity to engage with some of our experimental subjects one year after the original intervention sheds light on the persistence patterns.[53]

Consistent with the treatment effect dissipating over time, we find no difference between treatment and control groups in their survey response related to the concern with misinformation, one year after the experiment (Table A.16 in the Appendix).

[52] In addition, for broader contextualization of their news consumption diet, the survey included questions about consumption of other news outlets and platforms, as well as SZ, as we discussed in Section 4.1.

[53] Regarding selection, Appendix Table A.15 reports balance for follow-up respondents by treatment status. Response rates are similar across arms and pre-treatment characteristics are mostly balanced (joint tests do not reject balance for the survey (p=0.25) and tracked characteristics (p = 0.83)).

By the same token, we find that differences between treatment and control in trust numbers for both high- and low-trust outlets are statistically insignificant and economically small.

In parallel, to complement the analysis, we also managed to gather additional data on subscription status over time. The difference between treatment and control does persist for the cumulative number of dropped subscriptions in the first five months, as shown in Figure A.4 in the Appendix. Looking at the monthly cumulative effects, however, we see that the treatment effect stabilizes after five months; this indicates that after that horizon, there is no longer a difference between treatment and control in the likelihood of dropping their subscription.

Overall, these results suggest that our treatment effects across both stated and revealed preferences dissipated in the long run. This is in line with intuition, given our relatively light-touch intervention.

6 Discussion and Conclusion

Our results indicate that increased concern over misinformation, due to difficulty distinguishing between real and AI-generated content, has important and nuanced effects on trust in news media and news consumption. Though we find a negative average effect on trust in news content, including in the source considered trustworthy, we also uncover an important counterpoint in news consumer behavior. Namely, this unease leads to increased engagement with the trustworthy source in an economically significant manner, as evidenced by visits to the website, consumption of hard news

articles, and subscription retention. The subtle nature of our intervention heightens the significance of these results: the quiz was not a sustained campaign, and made no explicit mention of misinformation.

These findings confirm the intuition outlined in our conceptual framework. Deterioration in the information environment, engendered by the emergence of technologies like GenAI, leads to reduced trust in the information environment as a whole. However, an outlet perceived as sufficiently trustworthy may still witness increased demand, even as trust in its content also suffers. That is, its relative value increases in the eyes of readers, who deem it trustworthy enough to mitigate the effects of misinformation technology.

The average treatment effects are further substantiated by observed patterns of heterogeneity, which align with the predictions of our model. The fact that the increase in engagement is driven by individuals who were relatively less interested in political news and by heavy users of SZ digital content underscores the logic highlighted by our framework.

More broadly, SZ readers clearly care enough about the news to regularly engage with what they consider to be high-quality news content—arguably not unlike consumers of other similar outlets across the globe.[54] Indeed, our key results are corroborated by global survey evidence in the Reuters Institute Digital News Report released in June 2025 (Newman et al., 2025). In the report, based on online surveys conducted in Jan-Feb 2025, a question asked which source they would turn to if they felt a piece of news online was false. The top choice for survey takers, across

[54]This is underscored by the fact that respondents in our follow-up survey reported significant levels of consumption of sources such as Die Zeit, Der Spiegel, or FAZ.

political leanings and education levels, was "a news source I trust". This survey data, from a well-known third party, provides further support for the internal and external validity of our core results.

Less clear is how low-trust or less engaged readers would react to the same intervention. Additionally, we do not observe off-platform behavior on other social media and news websites. Both questions beg further research, which could be tackled using experiments, for example, by deploying a browser extension that could track broader online browsing behavior in the field. Along similar lines, while our focus here is on the misinformation challenge posed by GenAI in terms of distinguishing synthetic from real content, it would be interesting to assess other systemic or societal forces that might also generate a perception of difficulty in that distinction, and whether they could have similar effects. While our conceptual framework suggests this would be the case, it remains an open empirical question.

Our findings have potentially important business and societal implications. From the standpoint of the news industry, they offer a possible business strategy in response to the challenge posed by the advent of AI-generated content. As long as the outlet or platform can build trust with a segment of their audience, this challenge presents an opportunity to increase revenue: as trust becomes scarcer, trustworthiness becomes more valuable, and readers might be willing to pay more for it. Of course, this begs the larger question of how one builds trustworthiness in the first place. From a societal perspective, our results indicate that AI-powered misinformation may not lead to a downward spiral in trust in the information environment. The logic that cheap and high-quality fake content will have a competitive advantage and displace

real content, to the point that trust in all content collapses, may be counterbalanced by increased scarcity, creating greater potential rewards for trustworthiness.

Our conceptual framework, however, highlights important issues going forward. Trust in the content of an outlet depends essentially on the balance between the overall prevalence of misinformation and the outlet's perceived ability to mitigate its effects. For a given level of misinformation, sufficiently trustworthy outlets will benefit from an increase in misinformation; however, as the latter continues to grow, the trustworthiness threshold continues to rise. The lesson is thus that media outlets must ensure that their ability to help readers distinguish between real and AI content evolves at least as quickly as the difficulty of this task. As AI technology continues to evolve at a blistering pace, the challenge will only become greater, and the dynamics we highlight ever more important.

References

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius, "A Model of Online Misinformation," *The Review of Economic Studies*, 2024, *91* (6), 3117–3150.

Ahmad, Wajeeha, Ananya Sen, Charles Eesley, and Erik Brynjolfsson, "Companies Inadvertently Fund Online Misinformation Despite Consumer Backlash," *Nature*, 2024, *630* (8015), 123–131.

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

Allen, Jennifer, Antonio A Arechar, Gordon Pennycook, and David G Rand, "Scaling Up Fact-Checking Using the Wisdom of Crowds," *Science Advances*, 2021, *7* (36), eabf4393.

Ananthakrishnan, Uttara M, Beibei Li, and Michael D Smith, "A Tangled Web: Should Online Review Portals Display Fraudulent Reviews?," *Information Systems Research*, 2020, *31* (3), 950–971.

Angelucci, Charles and Andrea Prat, "Is Journalistic Truth Dead? Measuring How Informed Voters Are about Political News," *American Economic Review*, 2024, *114* (4), 887–925.

Aslett, Kevin, Andrew M Guess, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker, "News Credibility Labels Have Limited Average Effects

on News Diet Quality and Fail to Reduce Misperceptions," *Science Advances*, 2022, *8* (18), eabl3844.

Assenza, Tiziana, Alberto Cardaci, and Stefanie J. Huber, "Fake News: Susceptibility, Awareness and Solutions.," 2024. Toulouse School of Economics Working Paper.

Badrinathan, Sumitra, "Educative Interventions to Combat Misinformation: Evidence from a Field Experiment in India," *American Political Science Review*, 2021, *115* (4), 1325–1341.

Beattie, Graham, Ruben Durante, Brian Knight, and Ananya Sen, "Advertising Spending and Media Bias: Evidence from News Coverage of Car Safety Recalls," *Management Science*, 2021, *67* (2), 698–719.

Betzer, C., M. Booth, B. Cappio, A. Cook, M. Gochee, B. Grayzel, L. Jacoby, S. Majumder, M. Manda, J. Qian, M. Ransden, M. Rubens, M. Sardesai, E. Sullivan, H. Tekriwal, R. Waaland, and B. Nyhan, "State Media Tagging Does Not Affect Perceived Tweet Accuracy: Evidence from a U.S. Twitter Experiment in 2022," *Harvard Kennedy School (HKS) Misinformation Review*, 2025.

Braghieri, Luca, "Political Correctness, Social Image, and Information Transmission," *American Economic Review*, 2024, *114* (12), 3877–3904.

_ , Ro'ee Levy, and Alexey Makarin, "Social Media and Mental Health," *American Economic Review*, 2022, *112* (11), 3660–3693.

Brenan, Megan, "Americans' Trust in Media Remains at Trend Low," *Gallup News*, 14 Oct 2024. (Available at https://news.gallup.com/poll/651977/americans-trust-media-remains-trend-low.aspx, accessed on 27 Feb 2025).

Cagé, Julia, Nicolas Hervé, and Marie-Luce Viaud, "The Production of Information in an Online World," *The Review of Economic Studies*, 2020, *87* (5), 2126–2164.

Calzada, Joan and Ricard Gil, "What Do News Aggregators Do? Evidence from Google News in Spain and Germany," *Marketing Science*, 2020, *39* (1), 134–167.

Chopra, Felix, Ingar Haaland, and Christopher Roth, "Do People Demand Fact-Checked News? Evidence from US Democrats," *Journal of Public Economics*, 2022, *205*, 104549.

Costello, Thomas H, Gordon Pennycook, and David G Rand, "Durably Reducing Conspiracy Beliefs Through Dialogues with AI," *Science*, 2024, *385* (6714), eadq1814.

de Chaisemartin, Clément and Xavier D'Haultfoeuille, *Credible Answers to Hard Questions: Differences-in-Differences for Natural Experiments*, Princeton University Press, 2025.

Djourelova, Milena, Ruben Durante, and Gregory J Martin, "The Impact of Online Competition on Local Newspapers: Evidence from the Introduction of Craigslist," *The Review of Economic Studies*, 05 2024, *92* (3), 1738–1772.

Endert, Julius, "Generative AI is the Ultimate Disinformation Amplifier," *DW Akademie*, 17 Mar 2024. (Available at https://akademie.dw.com/en/generative-ai-is-the-ultimate-disinformation-amplifier/a-68593890, accessed on 27 Feb 2025).

Fong, Jessica, Tong Guo, and Anita Rao, "Debunking Misinformation About Consumer Products: Effects on Beliefs and Purchase Behavior," *Journal of Marketing Research*, 2024, *61* (4), 659–681.

Garimella, Kiran and Dean Eckles, "Image-Based Misinformation on WhatsApp," in "Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)" 2017.

Gentzkow, Matthew and Jesse M. Shapiro, "Media Bias and Reputation," *Journal of Political Economy*, 2006, *114* (2), 280–316.

Groh, Matthew, Aruna Sankaranarayanan, Nikhil Singh, Dong Young Kim, Andrew Lippman, and Rosalind Picard, "Human Detection of Political Speech Deepfakes Across Transcripts, Audio, and Video," *Nature Communications*, 2024, *15* (1), 7629.

Guriev, Sergei, Emeric Henry, Théo Marquis, and Ekaterina Zhuravskaya, "Curtailing False News, Amplifying Truth," 2025. Unpublished.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart, "Designing Information Provision Experiments," *Journal of Economic Literature*, 2023, *61* (1), 3–40.

Harris, Elizabeth A., Stephanie L. Demora, and Dolores Albarracín, "The Consequences of Misinformation Concern on Media Consumption.," *Harvard Kennedy School Misinformation Review*, 2024.

Hengel, Moritz, Julia Cagé, Emeric Henry, and Nathan Gallo, "Fact-Checking and Misinformation: Evidence from the Market Leader," *Working Paper*, 2025.

Henry, Emeric, Ekaterina Zhuravskaya, and Sergei Guriev, "Checking and Sharing Alt-Facts," *American Economic Journal: Economic Policy*, 2022, *14* (3), 55–86.

Hoffman, Benjamin, "First Came "Spam." Now, With AI, We've Got "Slop."," *The New York Times*, 2024, *11*.

Jerit, Jennifer and Yangzi Zhao, "Political Misinformation," *Annual Review of Political Science*, 2020, *23*, 77–94.

Kamali, Negar, Karyn Nakamura, Angelos Chatzimparmpas, Jessica Hullman, and Matthew Groh, "How to Distinguish AI-Generated Images from Authentic Photographs," *arXiv preprint arXiv:2406.08651*, 2024.

Kohavi, Ron, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann, "Online Controlled Experiments at Large Scale," in "Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining" 2013, pp. 1168–1176.

Kreps, Sarah, R. Miles McCain, and Miles Brundage, "All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation," *Journal of Experimental Political Science*, 2022, *9* (1), 104–117.

Langguth, Johannes, Konstantin Pogorelov, Stefan Brenner, Petra Filkuková, and Daniel Thilo Schroeder, "Don't Trust Your Eyes: Image Manipulation in the Age of DeepFakes," *Frontiers in Communication*, 2021, *6* (632317).

Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain, "The Science of Fake News," *Science*, 2018, *359* (6380), 1094–1096.

Lin, Hause, Haritz Garro, Nils Wernerfelt, Jesse Shore, Adam Hughes, Daniel Deisenroth, Nathaniel Barr, Adam Berinsky, Dean Eckles, Gordon Pennycook et al., "Reducing Misinformation Sharing at Scale Using Digital Accuracy Prompt Ads," *PsyArXiv*, 2024.

Miller, Elizabeth J, Ben A Steward, Zak Witkower, Clare AM Sutherland, Eva G Krumhuber, and Amy Dawel, "AI Hyperrealism: Why AI Faces Are Perceived as More Real than Human Ones," *Psychological Science*, 2023, *34* (12), 1390–1403.

Nelson, Phillip, "Advertising as Information," *Journal of Political Economy*, 1974, *82* (4), 729–754.

Newman, Nic, A Ross Arguedas, Craig T Robertson, Rasmus Kleis Nielsen, and Richard Fletcher, *Reuters Institute Digital News Report 2025*, Reuters Institute for the Study of Journalism, 2025.

_ , Richard Fletcher, Craig T Robertson, A Ross Arguedas, and Rasmus Kleis Nielsen, *Reuters Institute Digital News Report 2024*, Reuters Institute for the Study of Journalism, 2024.

Nyhan, Brendan, Ethan Porter, Jason Reifler, and Thomas J Wood, "Taking Fact-Checks Literally but Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability," *Political Behavior*, 2020, *42*, 939–960.

Pennycook, Gordon and David G Rand, "The Psychology of Fake News," *Trends in Cognitive Science*, 2021, *25* (5), P388–402.

Peukert, Christian, Ananya Sen, and Jörg Claussen, "The Editor and the Algorithm: Recommendation Technology in Online News," *Management Science*, 2024, *70* (9), 5816–5831.

Quidt, Jonathan De, Johannes Haushofer, and Christopher Roth, "Measuring and Bounding Experimenter Demand," *American Economic Review*, 2018, *108* (11), 3266–3302.

Seamans, Robert and Feng Zhu, "Responses to Entry in Multi-Sided Markets: The Impact of Craigslist on Local Newspapers," *Management Science*, 2014, *60* (2), 476–493.

Simon, Felix M. and Sacha Altay, "Don't Panic (Yet): Assessing the Evidence and Discourse Around Generative AI and Elections," Technical Report, Knight First Amendment Institute at Columbia University 2025.

Spitale, Giovanni, Nikola Biller-Andorno, and Federico Germani, "AI Model GPT-3 (Dis)informs us Better than Humans," *Science Advances*, 2023, *9* (26).

Toff, Benjamin and Felix M. Simon, "Or They Could Just Not Use It?: The Dilemma of AI Disclosure for Audience Trust in News," *The International Journal of Press/Politics*, 2025, *0* (0).

Tucker, Joshua A., Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan, "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature.," 2018. Hewlett Foundation.

Vaccari, Cristian and Andrew Chadwick, "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News," *Social Media + Society*, 2020, *6* (1).

Veerasamy, Namosha and Heloise Pieterse, "Rising Above Misinformation and Deepfakes," in "Proceedings of the 17th International Conference on Cyber Warfare and Security (ICCWS 2022)" 2022, pp. 340–348.

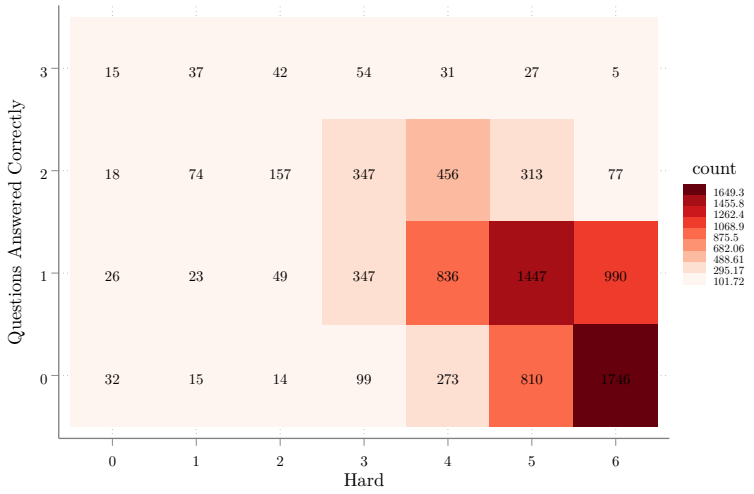
Vosoughi, Soroush, Deb Roy, and Sinan Aral, "The Spread of True and False News Online," *Science*, 2018, *359* (6380), 1146–1151.

Vraga, Emily K. and Leticia Bode, "Defining Misinformation and Understanding Its Bounded Nature: Using Expertise and Evidence for Describing Misinformation," *Political Communication*, 2020, *37* (1), 136–144.

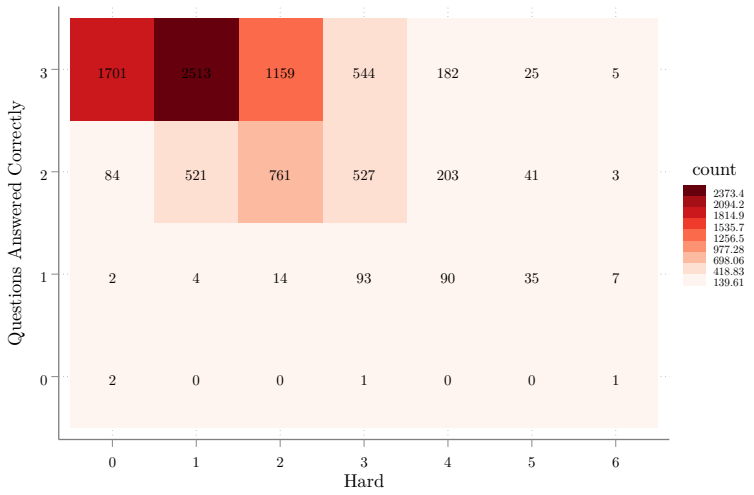
Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov, "Political Effects of the Internet and Social Media," *Annual Review of Economics*, 2020, *12*, 415–438.

A Appendix: Additional Results

Figure A.1: Joint distribution of perceived difficulty and performance



(a) Treatment



(b) Control

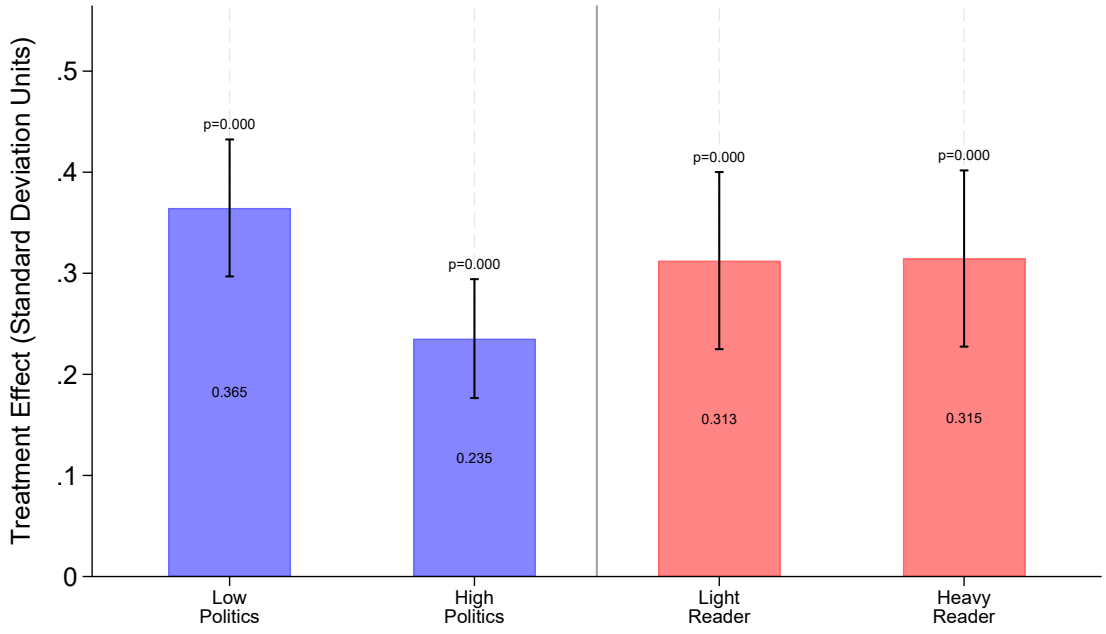
Notes: The figures capture the perceived and actual performance for the Treatment (figure (a)) and Control group (figure (b)). Perceived performance is measured on a scale of 0-6 of how hard an individual found the quiz while the total number of correct responses could range from 0 to 3.

Figure A.2: Treatment Effect on Stated Willingness to Pay



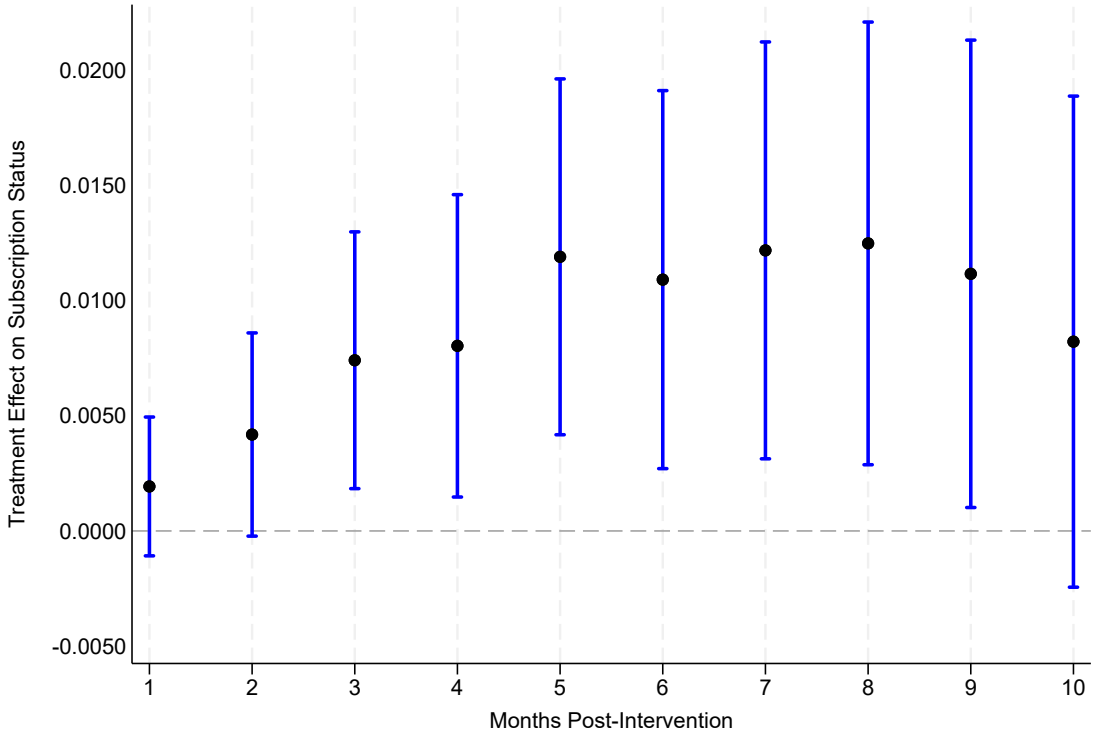
Notes:The figure shows the treatment effect on willingness to pay, captured in standard deviation units, for an SZ subscription across the treatment and control groups. The dependent variable is measured in Euros per month. The vertical lines represent 90 percent confidence intervals.

Figure A.3: Treatment Effect on Misinformation Concern: Heterogeneity by Politics Affinity and Reading Intensity



Notes: The figure captures the treatment effect on concern with misinformation across the treatment and control groups with heterogeneity by politics affinity and reading intensity. The dependent variable is on a scale of 0 to 6. The treatment effect in the figure is captured in terms of standard deviations. The vertical lines represent 90 percent confidence intervals

Figure A.4: Treatment Effect on Subscription Status by Months Post-Intervention



Notes: The dependent variable is whether the SZ reader is still subscribed at the end of the month. The figure shows subscription status at approximately 1-10 months after the intervention. The estimates capture the treatment effect, with AI Treatment equal to 1 if individual i was in the treated group and 0 otherwise. The vertical lines represent 90 percent confidence intervals.

Table A.1: News Media Consumption

Media Source	**Respondents**	**% of Sample**
Süddeutsche Zeitung	1,661	97.1
ARD/ZDF	1,475	86.3
Die Zeit	520	30.4
Der Spiegel	503	29.4
FAZ	239	14.0
TAZ	225	13.2
NZZ	143	8.4
Handelsblatt	134	7.8
Instagram	121	7.1
YouTube	109	6.4
Bild	54	3.2
X (Twitter)	20	1.2
TikTok	6	0.4

Note: The survey question in the follow-up survey from February 2026 was: "I get news about current events from the following media." An individual could opt for as many outlets from the above list as they wanted.

Table A.2: Balance Checks

		Control			Treatment			Difference	
	Total Obs	Obs	Mean	s.d.	Obs	Mean	s.d.	T - C	p-value
				Survey sample					
Female	17,199	8,674	0.46	0.50	8,525	0.45	0.50	-0.006	0.413
SZ subscriber	17,199	8,674	0.94	0.24	8,525	0.94	0.23	0.003	0.340
Tracked	17,199	8,674	0.35	0.48	8,525	0.34	0.47	-0.006	0.388
Early March	17,199	8,674	0.84	0.37	8,525	0.83	0.37	-0.007	0.229
Old (60-80)	17,199	8,674	0.49	0.50	8,525	0.48	0.50	-0.011	0.150
Young (<50)	17,199	8,674	0.21	0.41	8,525	0.22	0.42	0.009	0.169
				Tracked sample					
Female	5,915	3,010	0.40	0.49	2,905	0.40	0.49	0.001	0.929
Early March	5,915	3,010	0.88	0.33	2,905	0.87	0.33	-0.004	0.609
Old (60-80)	5,915	3,010	0.52	0.50	2,905	0.53	0.50	0.008	0.537
Young (<50)	5,915	3,010	0.19	0.39	2,905	0.19	0.40	0.001	0.913
Pre-intervention Daily Visits	5,967	2,988	96.31	98.77	2,979	97.51	102.08	1.196	0.646
High Pre-intervention Affinity									
Politics	5,718	2,871	0.57	0.50	2,847	0.57	0.49	0.008	0.527
Sports	5,718	2,871	0.27	0.44	2,847	0.27	0.44	0.003	0.824
Music	5,718	2,871	0.18	0.38	2,847	0.18	0.38	0.001	0.909
Crime	5,718	2,871	0.39	0.49	2,847	0.40	0.49	0.012	0.351
Location									
Germany	5,585	2,860	0.80	0.40	2,725	0.79	0.41	-0.015	0.157
Bayern	5,585	2,860	0.15	0.36	2,725	0.14	0.35	-0.009	0.330
Nordrhein-Westfalen	5,585	2,860	0.13	0.34	2,725	0.13	0.34	-0.002	0.846
Schleswig-Holstein	5,585	2,860	0.11	0.31	2,725	0.10	0.30	-0.004	0.635
Rheinland-Pfalz	5,585	2,860	0.09	0.28	2,725	0.08	0.27	-0.006	0.420
Standard Monthly Subscription	5,915	3,010	0.36	0.48	2,905	0.36	0.48	-0.006	0.605

Notes. Significance at the 10% level is represented by *, at the 5% by **, and at the 1% by ***. This table shows the balance along several observable dimensions between users in the treatment condition and those in the control condition. The first column provides the total observations across treatment and control. *p-value* is obtained based on a two-sided t-test on the equality of means across treatment and control.

Table A.3: ATE on Full Sample

Dep. variable	(1) Misinfo Concern	(2) WTP	(3) Trust SZ	(4) Trust X	(5) Trust IG	(6) Trust TikTok	(7) Trust Bild	(8) Trust ARDZDF	(9) Index
AI Treatment	0.223***	-0.085	-0.019***	-0.031***	-0.047***	-0.020***	-0.049***	-0.023***	-0.0374***
	(0.011)	(0.184)	(0.005)	(0.008)	(0.011)	(0.008)	(0.012)	(0.007)	(0.006)
Sample mean	5.605	21.002	2.898	0.236	0.740	0.247	0.794	2.789	1.357
	(0.729)	(11.057)	(0.334)	(0.482)	(0.703)	(0.472)	(0.755)	(0.458)	(0.397)
Observations	17,284	14,461	17,113	16,096	15,498	15,496	16,718	17,153	17,294
Adjusted-R^2	0.023	-0.000	0.001	0.001	0.001	0.000	0.001	0.001	0.002
F-statistic	416.980	0.214	13.470	16.846	17.066	6.857	17.326	10.865	38.32

Notes. Robust standard errors in parentheses. Significance at the 10% level is represented by *, at the 5% by **, and at the 1% by ***. The dependent variable is misinformation concern in column (1), WTP for SZ in column (2), Trust in SZ in column (3), Trust in X in column (4), Trust in Instagram (IG) in column (5), Trust in Tiktok in column (6), Trust in Bild in column (7) Trust in ARDZDF in column (8), and a composite index in column (9). AI Treatment is equal to 1 if individual i was in the treated group and zero otherwise.

Table A.4: Survey Robustness: Controls

Dep. variable	(1) Misinfo Concern	(2) Trust SZ	(3) WTP
AI Treatment	0.224***	-0.020***	-0.089
	(0.011)	(0.005)	(0.177)
Sample mean	5.605	2.898	21.002
	(0.729)	(0.334)	(11.057)
Controls	Yes	Yes	Yes
Observations	17,284	17,113	14,461
R-squared	0.040	0.032	0.078
F-statistic	87.205	36.155	189.204

Notes. Robust standard errors in parentheses. Significance at the 10% level is represented by *, at the 5% by **, and at the 1% by ***. The dependent variable is misinformation concern in column (1), Trust in SZ in column (2), and WTP for SZ in column (3). AI Treatment is equal to 1 if individual i was in the treated group and zero otherwise. Controls include whether the individual is female, a subscriber, has consented to tracking, took the quiz in early March, is under 50 years old, or is between 60 and 80 years old.

Table A.5: ATE on Tracked Sample

Dep. variable	(1) Misinfo Concern	(2) WTP	(3) Trust SZ	(4) Trust X	(5) Trust IG	(6) Trust TikTok	(7) Trust Bild	(8) Trust ARDZDF	(9) Index
AI Treatment	0.197***	0.198	-0.019***	-0.004	-0.034*	-0.012	-0.045**	-0.042***	-0.026***
	(0.017)	(0.311)	(0.007)	(0.012)	(0.019)	(0.013)	(0.020)	(0.011)	(0.010)
Sample mean	5.624	22.749	2.938	0.221	0.774	0.252	0.778	2.823	1.372
	(0.682)	(10.977)	(0.256)	(0.462)	(0.705)	(0.468)	(0.755)	(0.410)	(0.387)
Observations	5,923	4,980	5,888	5,527	5,329	5,316	5,736	5,902	5,936
Adjusted-R^2	0.021	-0.000	0.001	-0.000	0.000	-0.000	0.001	0.002	0.001
F-statistic	127.328	0.406	8.286	0.081	3.049	0.901	4.995	15.357	6.81

Notes. Robust standard errors in parentheses. Significance at the 10% level is represented by *, at the 5% by **, and at the 1% by ***. The dependent variable is misinformation concern in column (1), WTP for SZ in column (2), Trust in SZ in column (3), Trust in X in column (4), Trust in Instagram (IG) in column (5), Trust in Tiktok in column (6), Trust in Bild in column (7) Trust in ARDZDF in column (8), and a composite index in column (9). AI Treatment is equal to 1 if individual i was in the treated group and zero otherwise. The sample is based on those individuals who opt into tracking, allowing us to link survey and revealed preference measures.

Table A.6: Heterogeneous Effects: Quiz Difficulty

	Misinfo Concern		Visits (-21,+5)		Subscription (5 months)	
	(1) Hard	(2) Not Hard	(3) Hard	(4) Not Hard	(5) Hard	(6) Not Hard
AI Treatment	0.3070***	0.0846***			0.0161**	0.0072
	(0.0143)	(0.0182)			(0.0069)	(0.0073)
AI Treatment × Post			0.1213**	0.0221		
			(0.0604)	(0.0626)		
Sample mean	5.668	5.533	4.717	4.665	0.964	0.964
	(0.648)	(0.803)	(5.485)	(5.563)	(0.185)	(0.187)
Individual FE	No	No	Yes	Yes	No	No
Date FE	No	No	Yes	Yes	No	No
Observations	9,318	7,859	81,770	66,898	3,258	2,655
R-squared	0.0541	0.0026	0.7729	0.7836	0.0018	0.0003

Notes. Significance at the 10% level represented by * , at the 5% level by ** , and at the 1% level by ***. Hard = 1 if the individual provides an above median difficulty rating relative to others in the group they are in. Columns (1), (2), (5), and (6) have robust standard errors in parentheses, while columns (3) and (4) have standard errors clustered at the individual level. The dependent variable in columns (1) and (2) is the concern for misinformation, daily visits to SZ (with a window $[-21, 5]$ days) in columns (3) and (4), and subscription retention after 5 months in columns (5) and (6) for an individual SZ reader. The coefficient of interest is AI Treatment × Post in columns (3) and (4) while it is AI Treatment in columns (1), (2), (5), and (6).

Table A.7: Post-Intervention Content Engagement: Sports articles (daily)

	(1)	(2)	(3)	(4)	(5)
Subgroup	Full	Low Politics	High Politics	Heavy Reader	Light Reader
AI Treatment × Post	-0.0275	-0.0769	0.0054	0.0021	-0.0674
	(0.0572)	(0.0858)	(0.0820)	(0.1775)	(0.0418)
Sample mean	1.206	0.971	1.476	2.791	0.052
	(6.007)	(5.707)	(6.425)	(9.070)	(1.160)
Individual FE	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes
Observations	155,142	63,778	84,838	38,532	38,584
Adjusted-R^2	0.4618	0.4817	0.4471	0.4849	0.1998
F statistic	1.337	0.672	3.280	0.502	1.305

Notes. Standard errors clustered at the individual level. Significance: * 10%, ** 5%, *** 1%. The dependent variable is the number of daily clicks on sports articles by an individual reader. The time window for analysis varies from 21 days prior to the intervention to 5 days post. AI Treatment is equal to 1 if individual i was in the treated group and zero otherwise. We control for Post, which equals 1 if the day is after the date the individual took the quiz. Column (1) contains the full sample. Columns (2) and (3) are based on low and high politics affinity individuals while in columns (4) and (5) is based on engagement, both using pre-experiment data.

Table A.8: AI Article Demand and Supply

Panel A: AI Article Demand (reader clicks)

Dep. variable	AI articles clicked (daily)	
	(1)	(2)
Window	$[-21, +5]$	$[-21, +21]$
AI Treatment × Post	-0.0043	0.0007
	(0.0087)	(0.0070)
Sample mean	0.087	0.071
	(0.904)	(0.806)
Individual FE	Yes	Yes
Date FE	Yes	Yes
Observations	155,142	250,360
Adjusted-R^2	0.0837	0.0662
F statistic	1.706	0.861

Panel B: AI Article Supply (articles published)

Dep. variable	AI articles published (daily)		
	(1)	(2)	(3)
Window	Feb 10–Mar 9	Jan 1–Apr 9	Jan 1–Jun 9
Post	-0.1667	-0.1667	-0.1667
	(0.3165)	(0.3048)	(0.3043)
Sample mean	1.036	0.929	0.919
	(1.138)	(1.272)	(1.176)
Week FE	Yes	Yes	Yes
Observations	28	98	160
Adjusted-R^2	0.5774	0.3350	0.2754
F statistic	0.277	0.299	0.300

Notes. Statistical significance at 10% denoted by *, at 5% by **, and 1% by ***. Panel A: Standard errors clustered at the individual level. The dependent variable is the number of AI articles clicked by an individual in a day. AI treatment is equal to 1 if the individual was in the treatment group and zero otherwise. Panel B: Robust standard errors in parentheses. The dependent variable is the number of AI articles published daily on SZ digital. Post = 1 from Feb 25, 2025.

Table A.9: Post-Intervention Engagement: Robustness

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dep. variable	Number of daily visits			Time Spent (Seconds)		Number of daily visits		Other Survey Engagement	
Timespan	(-21;5)	(-21;5)	(-21;5)	(-21;5)	(-21;5)	(-14;5)	(-21;5)	(-21;5)	(-21;5)
Method	OLS	OLS	Poisson	OLS	OLS	OLS	OLS	OLS	LPM
Sample	Baseline	T + No Quiz	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline
AI Treatment × Post	0.0847**	0.1125***	0.0176**	227.7855*	178.2037*	0.0838**	0.0845**	0.0001	0.0000
	(0.0418)	(0.0432)	(0.0088)	(126.1053)	(91.8784)	(0.0421)	(0.0418)	(0.0005)	(0.0002)
AI Treatment			0.0121						
			(0.0268)						
Sample mean	4.653	4.017	0.001	5,699.834	4,119.070	4.699	4.653	0.0007	0.0002
	(5.493)	(5.169)	(0.031)	(9,927.587)	(8,816.244)	(5.530)	(5.493)	(0.0549)	(0.0134)
Individual FE	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time Since FE	No	No	No	No	No	No	Yes	No	No
Observations	155,142	142,470	155,142	112,104	155,142	113,373	155,142	155,090	155,142
R-squared	0.777	0.779	.	0.468	0.504	0.780	0.777	0.039	0.039

Notes. Standard errors are clustered at the individual level. Significance at the 10% level is represented by *, at the 5% by **, and at the 1% by ***. The dependent variable is the number of daily visits by the SZ reader in columns (1)-(3), (6)-(7) and time spent in columns (4)-(5). The analysis period is 21 days before the intervention to 5 days post-intervention in all columns except where we use a 14-day pre-intervention window. AI Treatment is equal to 1 if individual i was in the treated group and zero otherwise. Post is equal to 1 if the day is after the date the individual took the quiz. Column (1) replicates the baseline result in Table 1, Column (2) uses a non-respondent comparison group while column (3) uses a Poisson model. Columns (4) and (5) uses time spent as an alternative dependent variable (with the caveats about censoring and other measurement issues). Column (7) uses time since quiz fixed effects in addition. Columns (8) and (9) look at engagement with other surveys by individuals in our sample on the news website with the number of clicks on survey links in Column (8) and the probability of any click in Column (9) as the dependent variable.

Table A.10: Differences-in-Differences Robustness

Dep. variable	(1)	(2)	(3)	(4)
	Number of daily visits			
Timespan	(-21;+3)	(-21;+5)	(-21;+10)	(-21;+14)
AI Treatment × Post	0.141***	0.104**	0.086**	0.077**
	(0.050)	(0.041)	(0.035)	(0.034)
Sample mean	4.690	4.653	4.615	4.614
	(5.534)	(5.493)	(5.457)	(5.463)
Individual FE	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes
Observations	143,208	155,142	184,977	208,845
Adjusted-R^2	0.768	0.768	0.766	0.765
F statistic	8.102	6.515	6.148	5.232

Notes. Standard errors are clustered at the individual level. Significance at the 10% level is represented by *, at the 5% by **, and at the 1% by ***. The dependent variable is the number of daily visits by the SZ reader. The time window for analysis varies from 21 days prior to the intervention to 3 days post in column (1), 5 days post in column (2), 10 days post in column (3), and 14 days post in column (4). We do not control for the Post variable. AI Treatment is equal to 1 if individual i was in the treated group and zero otherwise.

Table A.11: Dropping Visits and Individuals Engaging with Other Surveys

Dep. variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Number of daily visits							
Timespan	(-21;+3)	(-21;+5)	(-21;+10)	(-21;+14)	(-21;+3)	(-21;+5)	(-21;+10)	(-21;+14)
AI Treatment × Post	0.113**	0.0848**	0.0745**	0.0673*	0.116**	0.0860**	0.0742**	0.0664*
	(0.0517)	(0.0418)	(0.0354)	(0.0344)	(0.0517)	(0.0418)	(0.0354)	(0.0344)
Sample mean	4.690	4.653	4.615	4.614	4.690	4.653	4.615	4.614
	(5.534)	(5.493)	(5.457)	(5.463)	(5.534)	(5.493)	(5.457)	(5.463)
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	143,133	155,062	184,887	208,745	142,992	154,908	184,698	208,530
R-squared	0.778	0.777	0.774	0.772	0.778	0.777	0.774	0.772

Notes. Standard errors are clustered at the individual level. Significance at the 10% level is represented by *, at the 5% by **, and at the 1% by ***. The dependent variable is the number of daily visits by the SZ reader. The time window for analysis varies from 21 days prior to the intervention to 3 days post in column (1 and 5), 5 days post in column (2 and 6), 10 days post in column (3 and 7), and 14 days post in column (4 and 8). AI Treatment is equal to 1 if individual i was in the treated group and zero otherwise. Post is equal to 1 if the day is after the date the individual took the quiz. The sample in columns (1)-(4) drops any visit that engages with a survey or a quiz on the SZ website, while in columns (5)-(8), we exclude individuals who clicked on any survey or quiz on the SZ website in the post-experimental period.

Table A.12: Post-Intervention Engagement: Traffic Sources

	(1)	(2)	(3)
Dep. variable	Number of daily visits		
Source	Overall	Organic	Social
AI Treatment × Post	0.0847**	0.0847**	-0.0001
	(0.0418)	(0.0418)	(0.0003)
Sample mean	4.653	4.653	0.001
	(5.493)	(5.493)	(0.031)
Individual FE	Yes	Yes	Yes
Date FE	Yes	Yes	Yes
Observations	155,142	155,116	155,116
R-squared	0.768	0.768	0.037
F-statistic	6.224	6.274	0.331

Notes. Standard errors are clustered at the individual level. Significance at the 10% level is represented by *, at the 5% by **, and at the 1% by ***. The dependent variable is the number of daily visits by the SZ reader by source. Column (1) has total traffic, column (2) has organic traffic (non-social media sources), and column (3) has social media traffic to the website and app. AI Treatment is equal to 1 if individual i was in the treated group and zero otherwise. Post is equal to 1 if the day is after the date the individual took the quiz.

Table A.13: Subscription Status: Non-Respondents vs. Treatment

Dep. variable	(1)	(2)	(3)	(4)
	Subscription Status			
Months Post-Intervention	2 Months	3 Months	4 Months	5 Months
AI Treatment (vs. Non-respondent group)	0.0132*	0.0206**	0.0270**	0.0237**
	(0.0079)	(0.0095)	(0.0111)	(0.0120)
Constant	0.9810***	0.9689***	0.9556***	0.9513***
	(0.0072)	(0.0091)	(0.0109)	(0.0115)
Controls	Yes	Yes	Yes	Yes
Observations	3,284	3,284	3,284	3,284
R-squared	0.0086	0.0094	0.0070	0.0061

Notes. Robust standard errors in parentheses. Significance at the 10% level is represented by *, at the 5% by **, and at the 1% by ***. The dependent variable is whether the SZ reader is still subscribed to SZ on that day. The post-intervention time window for analysis varies from 2 months post in column (1), 3 months post in column (2), 4 months post in column (3), and 5 months post in column (4). AI Treatment is equal to 1 if individual i was in the AI Treatment group and zero if the individual was in the non-respondent comparison group. Controls include reading affinity across politics, crime, music, and sports in the month prior to the experiment.

Table A.14: Control Group Placebo

VARIABLES	(1) Visit Count	(2) Subscription Status	(3) Misinfo Concern	(4) Trust SZ
Hard × Post	-0.0510			
	(0.0588)			
Hard		-0.0063	-0.0148	-0.0065
		(0.0074)	(0.0173)	(0.0071)
Constant	4.5822***	0.9612***	5.5008***	2.9109***
	(0.0256)	(0.0047)	(0.0119)	(0.0047)
Individual FE	Yes	No	No	No
Date FE	Yes	No	No	No
Observations	75,608	3,005	8,654	8,582
R-squared	0.7741	0.0002	0.0001	0.0001

Notes. Robust standard errors in parentheses. Significance at the 10% level is represented by *, at the 5% by **, and at the 1% by ***. The dependent variable is the number of daily visits to SZ in column (1), whether the SZ reader is still subscribed to SZ on that day in column (2), misinformation concern in column (3), and Trust in SZ in column (4). Post is equal to 1 if the day is after the date the individual took the quiz. Column (1) includes individual and date fixed effects; standard errors clustered at the individual level.

Table A.15: Balance of Follow-Up Survey Respondents

		Control			Treatment			Difference	
	Total Obs	Obs	Mean	s.d.	Obs	Mean	s.d.	T - C	p-value
Panel A. Response to the follow-up survey (tracked sample)									
Responded to follow-up survey	5,915	3,010	0.212	0.409	2,905	0.201	0.401	-0.010	0.330
Panel B. Pre-treatment characteristics of follow-up respondents									
Female	1,278	655	0.34	0.47	623	0.30	0.46	-0.034	0.191
Tracked	1,278	655	0.99	0.08	623	0.98	0.13	-0.012*	0.058
Early March	1,278	655	0.90	0.30	623	0.87	0.33	-0.026	0.142
Old (60-80)	1,278	655	0.64	0.48	623	0.65	0.48	0.012	0.654
Young (<50)	1,278	655	0.10	0.30	623	0.11	0.32	0.012	0.503
Pre-intervention Daily Visits	1,217	632	100.08	103.33	585	99.91	100.13	-0.169	0.977
High Pre-intervention Affinity									
Politics	1,170	617	0.61	0.49	553	0.61	0.49	-0.003	0.910
Sports	1,170	617	0.30	0.46	553	0.29	0.46	-0.010	0.705
Music	1,170	617	0.20	0.40	553	0.17	0.38	-0.027	0.230
Crime	1,170	617	0.43	0.49	553	0.41	0.49	-0.021	0.463
Location (non-missing region)									
Germany	1,190	626	0.79	0.41	564	0.84	0.37	0.046**	0.042
Bayern	1,190	626	0.15	0.35	564	0.14	0.35	-0.002	0.932
Nordrhein-Westfalen	1,190	626	0.14	0.35	564	0.14	0.35	-0.000	0.994
Schleswig-Holstein	1,190	626	0.09	0.29	564	0.10	0.30	0.003	0.849
Rheinland-Pfalz	1,190	626	0.08	0.28	564	0.08	0.27	-0.005	0.760
Subscription type									
Monthly subscription	1,263	651	0.30	0.46	612	0.28	0.45	-0.020	0.431
Yearly subscription	1,263	651	0.09	0.29	612	0.08	0.27	-0.009	0.572
Other subscription	1,263	651	0.61	0.49	612	0.64	0.48	0.029	0.287
Joint test, survey characteristics (all respondents): F = 1.35									0.249
Joint test, all characteristics (tracked respondents): F = 0.61									0.825

Notes. Significance at the 10% level is represented by *, at the 5% by **, and at the 1% by ***. Panel A uses the tracked sample of Table A.2 and compares the share of individuals who responded to the February 2026 follow-up survey and could be linked to their treatment status. Panel B uses these linked follow-up respondents, the sample of Table A.16 (variables defined as in Table A.2); tracked-sample characteristics are for respondents in the tracked sample. All respondents were SZ subscribers at the time of the original survey. The first column gives the total observations across treatment and control. The difference and its p-value come from a regression of each variable on the treatment indicator with robust standard errors. Joint tests regress the treatment indicator on the listed characteristics (robust standard errors) and report the F-statistic and p-value for the hypothesis that all coefficients are zero.

Table A.16: Follow-Up Survey Treatment Effects: Misinformation Concern and Trust

Dep. variable	(1) Misinfo Concern	(2) Trust SZ	(3) Trust X	(4) Trust IG	(5) Trust TikTok	(6) Trust Bild	(7) Trust ARDZDF
AI Treatment	-0.016	0.005	-0.018	-0.011	-0.023	-0.007	-0.009
	(0.046)	(0.005)	(0.026)	(0.035)	(0.026)	(0.036)	(0.016)
Sample mean	5.441	2.997	0.220	0.532	0.210	0.538	2.976
	(0.816)	(0.087)	(0.421)	(0.513)	(0.414)	(0.568)	(0.244)
Observations	1,261	1,190	1,065	866	1,030	972	995
Adjusted-R^2	-0.001	-0.000	-0.000	-0.001	-0.000	-0.001	-0.001
F-statistic	0.115	1.000	0.487	0.093	0.820	0.038	0.346

Notes. Robust standard errors in parentheses. Significance at the 10% level is represented by *, at the 5% by **, and at the 1% by ***. The dependent variable is misinformation concern in column (1), Trust in SZ in column (2), Trust in X in column (3), Trust in Instagram (IG) in column (4), Trust in TikTok in column (5), Trust in Bild in column (6), Trust in ARDZDF in column (7). AI Treatment is equal to 1 if individual i was in the treated group and zero otherwise.

Table A.17: Balance checks on test difficulty: Hard subsample

		Control			Treatment			Difference	
	Total Obs	Obs	Mean	s.d.	Obs	Mean	s.d.	T - C	p-value
Female	9,262	3,776	0.44	0.50	5,486	0.45	0.50	0.010	0.348
SZ subscriber	9,262	3,776	0.93	0.25	5,486	0.95	0.21	0.018	0.000
Tracked	9,262	3,776	0.35	0.48	5,486	0.35	0.48	-0.002	0.877
Early March	9,262	3,776	0.84	0.36	5,486	0.83	0.37	-0.011	0.150
Old (60-80)	9,262	3,776	0.47	0.50	5,486	0.48	0.50	0.009	0.412
Young (<50)	9,262	3,776	0.24	0.43	5,486	0.23	0.42	-0.013	0.146

Notes. This table shows the balance along several observable dimensions between users in the treatment and control conditions for the subsample of individuals who find the quiz hard. The first column provides the total observations across treatment and control for this subsample. *p-value* is obtained based on a two-sided t-test on the equality of means across treatment and control.

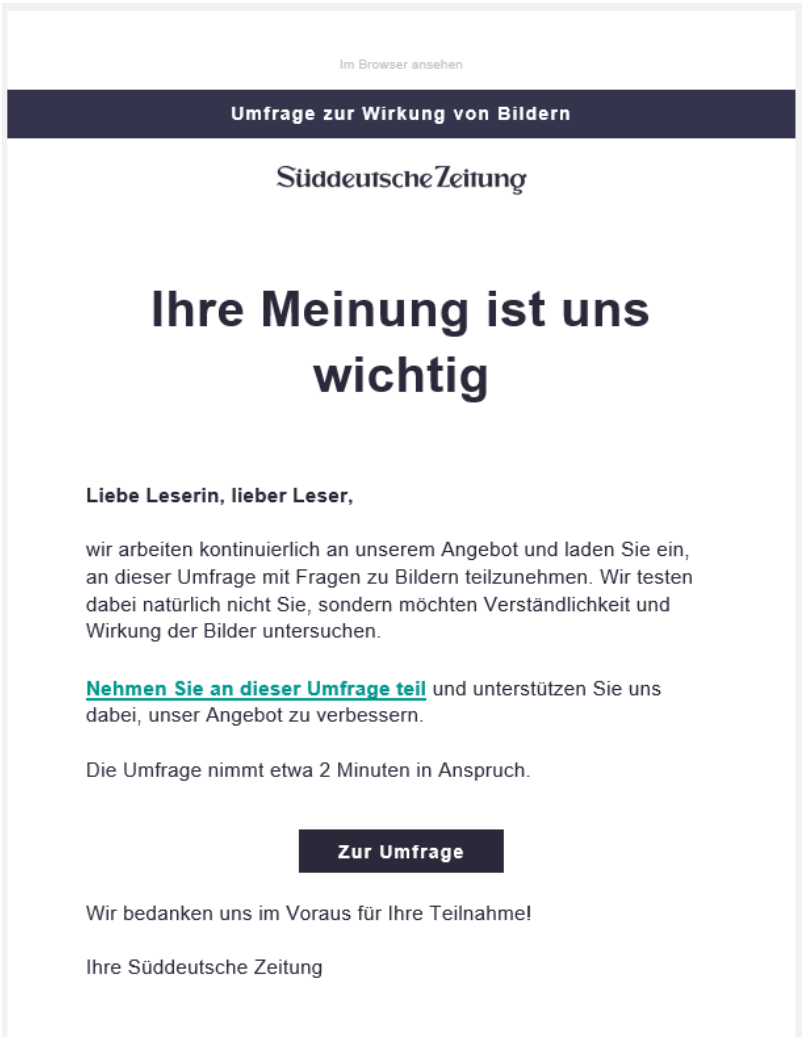
B Intervention Details

The AI images in B.3 and B.5 were created using Midjourney version 6.1 with an aspect ratio of 3:2. The prompt used for the AI image in B.3 (left image) was "A realistic photo from a scene during a flooding in a street of a small town in Bavaria.

In the center of the photo are three rescue men walking through knee-high water with a red rescue boat. In the background you can see a tree and some buildings, but the street is covered in (calm) standing water. The men look a bit exhausted. One older person in a dark jacket sits in the rescue boat. Photo is taken at midday". The prompt used for the AI image in B.5 (left image) was "A single protester on the streets of Georgia waving a European Union flag in front of a bunch of policemen in dark outfits". The AI image in B.4 (right image) was created by modifying B.7, using freely available AI tools in Canva. No prompt was used to generate this image.

We report all outcome variables specified in the pre-registration in the paper. For the regression analyses of survey outcomes, we didn't pre-specify whether we would use control variables because it was unclear what data SZ would make available to us. Hence, we provide regression results both with and without control variables.

Figure B.1: Quiz Email Invitation



Im Browser ansehen

Umfrage zur Wirkung von Bildern

Süddeutsche Zeitung

Ihre Meinung ist uns wichtig

Liebe Leserin, lieber Leser,

wir arbeiten kontinuierlich an unserem Angebot und laden Sie ein, an dieser Umfrage mit Fragen zu Bildern teilzunehmen. Wir testen dabei natürlich nicht Sie, sondern möchten Verständlichkeit und Wirkung der Bilder untersuchen.

Nehmen Sie an dieser Umfrage teil und unterstützen Sie uns dabei, unser Angebot zu verbessern.

Die Umfrage nimmt etwa 2 Minuten in Anspruch.

Zur Umfrage

Wir bedanken uns im Voraus für Ihre Teilnahme!

Ihre Süddeutsche Zeitung

Figure B.2: Quiz Email Invitation (Machine Translated)

Survey on the effect of images

Süddeutsche Zeitung

Your opinion is important to us

Dear reader (female), dear reader (male),

we are continuously working on our offering and invite you to take part in this survey with questions about images. Of course, we are not testing you, but rather want to examine comprehensibility and effect of the images.

Take part in this survey and support us in improving our offering.

The survey takes about 2 minutes.

To the survey

We thank you in advance for your participation!

Your Süddeutsche Zeitung

Figure B.3: Treatment Picture Question 1

Bilder-Quiz



Welches Bild wurde mit KI generiert?

- ○ Links
- ○ Rechts
- ○ Beide
- ○ Keines

Figure B.4: Treatment Picture Question 2

Bilder-Quiz





Welches Bild wurde mit Hilfe von KI generiert?

- ○ Links
- ○ Rechts
- ○ Beide
- ○ Keines

Figure B.5: Treatment Picture Question 3

Bilder-Quiz





Welches Bild wurde mit KI generiert?

- ○ Links
- ○ Rechts
- ○ Beide
- ○ Keines

Figure B.6: Control Picture Question 1

Bilder-Quiz



In welchen Ländern sind diese Aufnahmen von 2024 entstanden?

- ○ Frankreich und Belgien
- ○ Deutschland und Belgien
- ○ Deutschland und Spanien
- ○ Belgien und Spanien

Figure B.7: Control Picture Question 2

Bilder-Quiz



Die abgebildeten Personen sind Politiker und Politikerinnen welcher Länder?

- ○ Kanada und Deutschland
- ○ Kanada und Frankreich
- ○ Amerika und Italien
- ○ Amerika und Deutschland

Figure B.8: Control Picture Question 3

Bilder-Quiz




In welchen Ländern sind diese Bilder entstanden?

- ○ Frankreich und Georgien
- ○ Frankreich und Deutschland
- ○ Deutschland und Spanien
- ○ Italien und Finnland

C Examples of Daily Information Environment

Figure C.1: Homepage Images and Headlines at 9 am



Image and Headline 25th Feb 2025



Image and Headline 26th Feb 2025

Figure C.2: Homepage Images and Headlines at 9 am



Image and Headline 27th Feb 2025



Image and Headline 28th Feb 2025

Figure C.3: Homepage Images and Headlines at 9 am



Image and Headline 1st March 2025



Image and Headline 2nd March 2025

Figure C.4: Homepage Images and Headlines at 9 am

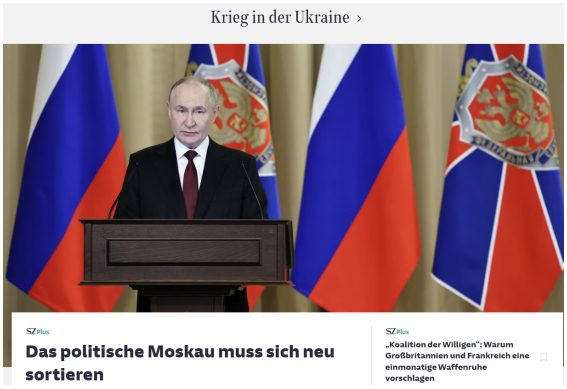


Image and Headline 2nd March 2025

D Model Extensions

Here we consider extensions of the model in Section 3.

D.1 Non-linear misinformation

Let us posit an alternative formulation for misinformation: instead of adding Δ to the signal variance σ_i^2, we now have a multiplicative impact, $(1+\Delta)\sigma_i^2$. As before, this is then modified to incorporate the mitigation factor $\alpha \in (0,1)$: $(1+\alpha\Delta)\sigma_i^2$.

This entails that, instead of (4), demand is now given by the expression:

$$d_1 = \Lambda(\sigma_0^2(1+\Delta)) - \Lambda(\sigma_1^2(1+\alpha\Delta)) \tag{D.1}$$

Differentiating that with respect to Δ now yields:

$$\frac{\partial d_1}{\partial \Delta} = (\sigma_X^2)^2 \left[\frac{\sigma_0^2}{(\sigma_X^2 + (1+\Delta)\sigma_0^2)^2} - \frac{\alpha\sigma_1^2}{(\sigma_X^2 + (1+\alpha\Delta)\sigma_1^2)^2} \right]. \tag{D.2}$$

It is again clear from inspection that, for any given values of $(\sigma_X^2, \sigma_0^2, \sigma_1^2, \Delta)$, this number will be positive for a sufficiently low value of α, confirming our Result 2.

As for Result 1, we have to consider the case where $\alpha = 1$. We can rewrite (D.2) as $\frac{\partial d_1}{\partial \Delta} = (\sigma_X^2)^2 \left[g(\sigma_0^2) - g(\sigma_1^2) \right]$, where $g(a) = \frac{a}{\left(\sigma_X^2 + (1+\Delta)a\right)^2}$. Given that $\sigma_0^2 > \sigma_1^2$, it follows that $\frac{\partial d_1}{\partial \Delta}$ will be greater (resp. less) than zero if $g(\cdot)$ is an increasing (resp. decreasing) function. Computing $g'(\cdot)$, we can see that $g'(a) \lesseqgtr 0$ when $a \gtreqless \frac{\sigma_X^2}{1+\Delta}$. It follows that, when the degree of misinformation Δ is sufficiently large, we will have $g'(\cdot) < 0$, and $\frac{\partial d_1}{\partial \Delta} < 0$. This confirms that Result 1 holds for sufficiently high levels

of misinformation: as misinformation increases, the advantage of the high-quality outlet gets diluted, and demand for it eventually declines as a result.

D.2 Endogenous trustworthiness

We now allow an outlet s_1, with a given level of signal quality σ_1^2, to invest in its trustworthiness. Specifically, we assume that it can choose the level of α, paying a cost $C(\alpha)$, with $C'(\cdot) < 0$ and $C''(\cdot) > 0$: more trustworthiness (lower α) is more expensive to achieve, and convexly so.

Assuming initially for simplicity that the outlet is a price-taker, we can ignore the price (setting it equal to 1) and have the outlet choose the level of α to maximize demand net of costs. Using (4), this yields the following FOC:

$$-C'(\alpha) = \frac{(\sigma_X^2)^2 \Delta}{(\sigma_X^2 + \sigma_1^2 + \alpha\Delta)^2}. \tag{D.3}$$

It is easy to see that, as long as $C(\cdot)$ is sufficiently convex, we have $\frac{\partial \alpha}{\partial \sigma_1^2} > 0$: higher σ_1^2 decreases the marginal benefit of an improvement in trustworthiness, on the right-hand side, which requires the left-hand side to decrease; this in turn requires a higher α because of the convexity of $C(\cdot)$.[55] In words: higher-quality outlets (lower σ_1^2) will choose to invest more in trustworthiness (lower α). The intuition is simple: lower-quality signals elicit less demand, which reduces the marginal return to mitigating disinformation. (Note also that endogenizing σ_1^2 while holding α to be exogenous yields the same complementarity property: the RHS of the corresponding FOC would

[55] More specifically, using the implicit function theorem, we have $\frac{\partial \alpha}{\partial \sigma_1^2} > 0 \iff C''(\cdot) > \frac{2\Delta^2 (\sigma_X^2)^2}{(\sigma_X^2 + \sigma_1^2 + \alpha\Delta)^3}$. This holds for sufficiently large σ_X^2 and/or Δ, as long as $C''(\cdot)$ is bounded away from zero.

be the same as in (D.3), up to a factor of Δ.)

The impact of a marginal increase in misinformation is more nuanced: under the same convexity conditions, we will have $\frac{\partial \alpha}{\partial \Delta} < 0$ as long as $\sigma_X^2 + \sigma_1^2 > \alpha\Delta$. In other words, a marginal increase in misinformation will increase investment in trustworthiness (lower α), as long as the role of (mitigated) misinformation is not the dominant source of noise.[56] Interestingly, in an environment where there is greater inherent noise (higher σ_X^2), this will be more likely to be the case: increased misinformation is more likely to trigger investment in mitigation for topics where discerning the true outcome is more challenging to begin with.

D.3 Price-setting

Our framework has so far assumed that outlets are price-takers. We now extend the model to explicitly consider the implications of outlets optimally adjusting prices in response to demand shocks.

To maintain simplicity, let us posit a linear demand function, such that the quantity q_i demanded of outlet i, as a function of the price p_i that it sets, is given by $q_i = max\{0, d_i - p_i\}$, where d_i comes from the signal extraction problem that yields (4). It is easy to verify that $\frac{\partial d_i}{\partial \alpha_i} < 0$: more trustworthiness (lower α) leads to higher demand. Similarly, $\frac{\partial d_i}{\partial \sigma_i^2} < 0$: higher quality outlets will have more demand from consumers.

Let us start from our baseline case, where trustworthiness (α) is taken as given.

[56] This nuance is absent when we consider σ_1^2 as the endogenous variable: increased misinformation unambiguously decreases investment in the quality of the signal. This is because misinformation unambiguously dilutes the marginal benefit of a higher-quality signal, given that investment in the latter does not mitigate the effect of misinformation itself.

Outlet i's profit maximization problem is as follows:

$$max_{\{p_i\}}\{p_i(d_i - p_i)\} \tag{D.4}$$

Taking the FOC with respect to p_i yields the optimal price and corresponding quantity: $p_i^* = \frac{d_i}{2}$, $q_i^* = \frac{d_i}{2}$. It follows immediately that higher-quality and more trustworthy outlets will be able to charge a higher price and sell more, because of the higher demand for their product; with the simple linear demand function, an increase in consumer demand will be split in half between price and quantity.

From this it follows that our comparative statics with respect to misinformation, as captured in Result 2, translates naturally to the context with price-setting: sufficiently trustworthy sources would benefit from a marginal increase in misinformation. That increase would lead to higher demand, as per Result 2, and allow them to increase prices and unambiguously increase profits. Outlets that are not sufficiently trustworthy would face lower demand and optimally choose to charge lower prices because of that.

Now let us endogenize α, as in the previous subsection. The profit maximization problem is now:

$$max_{\{p_i,\alpha_i\}}\{p_i(d_i(\alpha_i) - p_i) - C(\alpha_i)\} \tag{D.5}$$

From this we obtain the FOC with respect to α, which modifies (D.3) to:

$$-C'(\alpha_i) = \frac{d_i(\alpha_i)(\sigma_X^2)^2\Delta}{2(\sigma_X^2 + \sigma_i^2 + \alpha_i\Delta)^2}. \tag{D.6}$$

Note that the complementarity between α_i and σ_i^2 is made stronger in this case: higher-quality outlets will have a stronger incentive to invest in trustworthiness. Because they are able to charge higher prices to begin with, the marginal benefit of increased trustworthiness is higher for them.

In sum, higher-quality outlets would invest more in trustworthiness, because they would be able to obtain a greater benefit, given their ability to charge higher prices. Lower-quality outlets would be less trustworthy and charge lower prices.[57]

E Relative Declines in Trust through the Lens of the Model

This appendix elaborates on the discussion in Section 5.1, on the interpretation of the treatment effects on trust as declines relative to baseline levels of trust, through the lens of our conceptual framework.

We start by computing the relative declines in trust for each outlet implied by our estimates. Specifically, in the model, trust in the content of outlet i is the precision of its signal, $t_i = 1/(\sigma_i^2 + \alpha_i \Delta)$, which we thus map onto our survey measure. Let β_i denote the ATE on trust in outlet i and $\bar{\tau}_i^C$ its mean in the control group. The

[57] While in this setting we have left out strategic interactions between different outlets' pricing decisions, it is easy to see that the intuition extends to when they are present. For instance, we could specify a two-outlet discrete-choice model with logit demand, with the consumers' deterministic utility component described by (4) and the outside option given by s_0, and where each outlet chooses its price to maximize profit, taking the rival's price as given in Bertrand-Nash competition. It is easy to check that Result 2 would still hold: sufficiently trustworthy outlets would increase their market share and price in response to a marginal increase in misinformation. Similarly, the complementarity between quality and investment in trustworthiness would also hold, given similar convexity conditions as before.

relative decline is $r_i = \beta_i/\bar{\tau}_i^C$, and its log analogue is $\ell_i = \log(1 + r_i)$. Table E.1 reports the results.

Table E.1: Relative Declines in Trust

	ATE	Control mean	Relative decline (%)		
Outlet	β_i	$\bar{\tau}_i^C$	r_i	s.e.	95% CI
SZ	-0.019	2.908	-0.65	0.18	[-0.99, -0.30]
ARD/ZDF	-0.023	2.800	-0.82	0.25	[-1.31, -0.34]
Bild	-0.049	0.818	-5.94	1.38	[-8.65, -3.23]
Instagram	-0.047	0.763	-6.10	1.43	[-8.91, -3.30]
TikTok	-0.020	0.257	-7.72	2.83	[-13.26, -2.18]
X/Twitter	-0.031	0.252	-12.37	2.80	[-17.85, -6.88]

Notes. ATEs and robust standard errors are from the regressions of Table A.3. Control means are the means of each trust outcome in the control group (estimation sample of Table A.3). The relative decline is the ATE divided by the control mean. Standard errors are computed by the delta method from the robust standard errors of the ATEs and the sampling variance of the control means, $\mathrm{Var}(\bar{\tau}_i^C) = s_i^2/N_i^C$, using $\mathrm{Cov}(\beta_i, \bar{\tau}_i^C) = -\mathrm{Var}(\bar{\tau}_i^C)$.

Note that in our model with symmetric misinformation, every outlet's signal variance increases by the same amount: $t_i = 1/(\sigma_i^2 + \Delta)$. Then

$$\frac{\partial \log t_i}{\partial \Delta} = -\frac{1}{\sigma_i^2 + \Delta} = -t_i, \tag{E.1}$$

so the proportional decline in trust is increasing in the level of trust and should be largest for SZ and ARD/ZDF.[58]

We can thus test the hypothesis of symmetric misinformation, taking X/Twitter as our reference point. Specifically, we test each null hypothesis jointly with a Wald test of the five restrictions that equate each outlet's statistic with that of X/Twitter.[59]

[58] Under the non-linear specification of Appendix D.1, $t_i = 1/\big((1+\Delta)\sigma_i^2\big)$ and $\frac{\partial \log t_i}{\partial \Delta} = -\frac{1}{1+\Delta}$, which does not depend on σ_i^2: proportional declines ℓ_i should be equal across outlets.

[59] Because the trust questions are answered by the same respondents, the estimates for different

Table E.2 shows that both the additive and non-linear versions of the model with symmetric misinformation are rejected at the 0.1 percent level. The last panel compares the average log decline for the two high-trust outlets (SZ and ARD/ZDF) with that for the four low-trust platforms: the difference is 7.7 log points ($z = 4.70$).

Table E.2: Joint Tests of Models with Symmetric Misinformation

Multiplicative: H_0: ℓ_i equal across outlets	
$\chi^2(5)$	30.0
p-value	< 0.001
Additive: H_0: $\beta_i/(\bar{\tau}_i^C)^2$ equal across outlets	
$\chi^2(5)$	41.4
p-value	< 0.001
High- vs. low-trust outlets (log points)	
Difference	7.68
z-statistic	4.70
p-value	< 0.001

Notes. Wald tests of the five restrictions equating each outlet's statistic with that of X/Twitter. $\ell_i = \log(1 + \beta_i/\bar{\tau}_i^C)$. The last panel compares the mean of ℓ_i for SZ and ARD/ZDF with its mean for Bild, Instagram, TikTok and X/Twitter. Covariances of the ATEs and control means are from a stacked regression of the six trust outcomes on outlet indicators and their interactions with treatment, with standard errors clustered by respondent. The p-values for the multiplicative and additive tests are 1.5×10^{-5} and 7.8×10^{-8}. $N = 98{,}074$ respondent–outlet observations from 17,294 respondents.

Allowing α_i to differ across outlets reconciles the model with these patterns.

outlets are correlated. We estimate this covariance directly by stacking the six trust outcomes, one observation per respondent and outlet with a valid answer, and regressing trust on outlet indicators and their interactions with treatment, clustering standard errors by respondent. The outlet intercepts are the control means and the interactions are the ATEs; the clustered covariance matrix of these twelve parameters then yields the covariance of every statistic by the delta method. The estimated correlations have a clear block structure: about 0.55 between the estimates for SZ and ARD/ZDF, 0.3–0.4 among Instagram, TikTok and X/Twitter, and close to zero between the two groups.

Under the multiplicative specification, $t_i = 1/\big((1+\alpha_i\Delta)\sigma_i^2\big)$ and

$$\frac{\partial \log t_i}{\partial \Delta} = -\frac{\alpha_i}{1+\alpha_i\Delta} \approx -\alpha_i \quad \text{for small } \alpha_i\Delta, \tag{E.2}$$

so that $\ell_i \approx -\alpha_i\, d\Delta$. Under the additive specification of Section 3.3, $\partial \log t_i/\partial\Delta = -\alpha_i t_i$, so that $\beta_i/(\bar\tau_i^C)^2 \approx -\alpha_i\, d\Delta$. In both cases the treatment-induced change $d\Delta$ is common to all outlets; ratios of the observed statistics therefore identify the ratios α_i/α_j:

$$\left.\frac{\alpha_i}{\alpha}\right|_{\text{mult.}} = \frac{\ell_i}{\ell}, \qquad \left.\frac{\alpha_i}{\alpha}\right|_{\text{add.}} = \frac{\beta_i/(\bar\tau_i^C)^2}{\beta/(\bar\tau^C)^2}. \tag{E.3}$$

Only ratios are identified, because the units of Δ are arbitrary. We normalize by X/Twitter, which we take to play the role of source s_0 in the model: that source is unmitigated, with $\alpha = 1$.[60]

Table E.3 presents the results. Under the multiplicative specification, SZ's implied α is about 5 percent of X/Twitter's; equivalently, readers perceive SZ as mitigating about 95 percent of the misinformation shock faced by X/Twitter (95 percent interval: 89–98 percent). ARD/ZDF is close to SZ. The additive specification yields the same ordering with more extreme magnitudes. The intervals for Bild and Instagram lie below one but well above those for SZ and ARD/ZDF, and the interval for TikTok includes one.

[60] Because these are ratios of noisy estimates, with denominators that are themselves imprecise, their sampling distributions are skewed. We therefore report percentile intervals from 200,000 Monte Carlo draws from the joint normal distribution of $(\beta_i, \bar\tau_i^C)_i$, usingthe estimated covariance matrix

Table E.3: Calibrated Trustworthiness Relative to X/Twitter (α_i/α)

	Multiplicative (App. D.1)		Additive (Sec. 3.3)	
Outlet	Estimate	95% interval	Estimate	95% interval
SZ	0.049	[0.021, 0.108]	0.0045	[0.0020, 0.0094]
ARD/ZDF	0.063	[0.023, 0.140]	0.0060	[0.0023, 0.0127]
Bild	0.464	[0.239, 0.903]	0.148	[0.080, 0.272]
Instagram	0.477	[0.256, 0.897]	0.163	[0.091, 0.289]
TikTok	0.609	[0.177, 1.188]	0.612	[0.187, 1.120]
X/Twitter	1	–	1	–

Notes. Point estimates are plug-in values based on the ATEs and control-group means from the stacked regression underlying Table E.2. Intervals are the 2.5th and 97.5th percentiles of 200,000 Monte Carlo draws (seed 1, Stata/Mata) from the joint normal distribution of the ATEs and control means, using the covariance matrix from that regression, with standard errors clustered by respondent.