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

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

October 2025

**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 (-0.1 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 1.1% over the following five months, corresponding to about a one-third drop in 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 Nicolás Ajzenman, Elliott Ash, Manuela Collis, Chuck Eesley, Henry Farrell, Thomas Fujiwara, Sid George, Lorenz Goette, Mor Namaan, Brendan Nyhan, Jesse Shapiro, Joachim Voth, David Yanagizawa-Drott, and Noam Yuchtman for detailed comments and suggestions, and to Alessandro Izzo, Matheus Junqueira, David Schwartsman, Ren Jie Teh, and Nicolás Valle for outstanding research assistance. We are grateful to 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).

†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., 2024b), 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 ("deepfakes") that is virtually indistinguishable from authentic material, 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 becomes relatively scarce, trustworthiness becomes more valu-

[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). Relatedly, the concept of "AI slop," meant to capture the proliferation of low-quality AI-generated content, has also started to gain 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).

able. This counterpoint 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)

[4]One SZ campaign used the slogan *"Die Wahrheit lässt sich nicht generieren. Nur recherchieren."* ("The truth cannot be generated. Only researched.") *The New York Times* is another outlet that emphasizes verification in its marketing strategy, creating a visual investigations page that authenticates content and directs readers to subscribe.

were then asked to evaluate the severity of the problem of misinformation, report their level of trust in SZ content and in a number of other outlets and platforms, and state the monthly amount they would be willing to pay for an SZ digital subscription.[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 after the intervention; namely, the number of daily visits to SZ digital content (website or app) 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

[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 0.40% 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 2% answering all three correctly and 36% all three wrong.

media outlets and platforms, including SZ itself. Yet, in spite of this reduced trust, reported willingness to pay (WTP) for an SZ subscription did not fall.

For the revealed preference outcomes in the tracked sample, we find that treated individuals increased 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 1.1% higher in the treatment than in the control group, reflecting about a one-third lower attrition rate. This indicates that while the impact of the treatment on day-to-day decisions (i.e., online visits) may fade, it remains influential as individuals reassess their commitment as subscribers. Importantly, these results do not hinge on the nature of the treatment vs. control quiz, and hold even when we compare the treatment group to a "pure control" group of SZ readers who did not take any quiz. 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 €400,000 in additional annual subscription revenue, plus any advertising revenue that would come 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).[8]

[8]Global survey evidence from the Reuters Institute Digital News Report (June 2025) provides suggestive corroboration of these findings (Newman et al., 2024a)). Specifically, when asked where they

Heterogeneity patterns shed further light on the dynamics at play in the data and on possible mechanisms. Respondents who found the quiz to be relatively hard reported larger increases in misinformation concern and daily visits, as well as higher WTP and no significant decline in trust in SZ content. In addition, for the tracked subsample, we find that individuals who were relatively less interested in political news – as measured by their reading patterns in the three weeks before the experiment – respond more strongly to the treatment in terms of subsequent engagement. We also find suggestive evidence that the engagement response is stronger for heavier users of the website (i.e., above average). These 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.

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. 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 – only as long as news outlets' ability to help readers keeps pace with this growing challenge.

would turn to check the veracity of suspected false online news, respondents across demographics most often chose the response: "A news source I trust."

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, large language models (LLMs) can produce convincing expert-sounding content that is difficult for laymen 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).[9] 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

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

from these studies in two important ways. First, while these contributions focus on hypothetical misinformation insurance and individuals' stated preferences, our study 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. 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, Longoni et al. (2022) and Toff and Simon (2025) document that news articles generated with AI attract lower trust than those written by humans. 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

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.[10]

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.

# 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

[10] 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).

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 provides coverage of topics across politics, 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), 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 complicated 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., 2024a). Moreover, evidence suggests that image-based misinformation spreads more widely than text alone (Garimella and Eckles, 2017). The need for content 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.[11]

Beyond regulation, the proliferation of "AI slop" online (Hoffman, 2024) has

[11] 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/.

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.[12] 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.

# 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 $L(a, X) = (a - X)^2$.

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

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 $\sigma_0^2$ 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 $s_0$:

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

We posit for simplicity that 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.

## 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. 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.

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 that the precision gap is now smaller; in fact, as $\Delta$ 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 $\Delta$ 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:

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

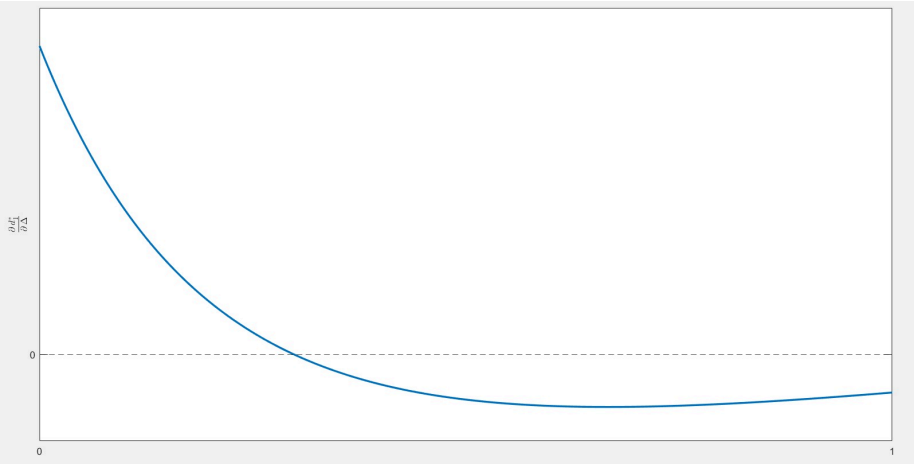
Differentiating that with respect to $\Delta$ yields:

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

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 $\alpha$. 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 $\Delta$ on demand $d_1$, as a function of $\alpha$




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 reduces trust in the 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.[13]

Note that our framework clarifies the nature of the challenge faced by high-quality news sources over time: if $\Delta$ keeps growing without bound, eventually any given value of $\alpha$ will be insufficient to sustain demand, and $d_1$ will fall. Yet equation (5) makes it clear that, if $\alpha$ were to go down at least as fast as $\Delta$ 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 $\alpha$.[14] 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 $\alpha$ 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

[13] It is easy to extend our model by adding a supply side in which the high-quality outlet endogenously chooses the quality of its signal. As we show in Appendix D, under reasonable assumptions, a more trustworthy outlet (lower $\alpha$) will choose to invest in a higher-quality signal (lower $\sigma_1^2$). 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 when the reader sees the outlet as better able to mitigate the latter.

[14] 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 $\alpha$.

as more trustworthy.

Overall, our simple framework underlines the nuanced impact of the expansion of a technology such as GenAI, and the misinformation that arises after. 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.

# 4 Experimental Setting and Data

## 4.1 Experimental Setting and Design

In line with our theoretical framework, our experimental setting enables us to focus on 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. 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.

The experimental subjects were recruited from two populations: website visitors

and subscribers. For the former, between February 25 and March 9, 2025, one percent of visitors were randomly shown a pop-up that encouraged them to take a picture quiz, which could be accessed by clicking the link presented to them.[15] For the latter, individuals on the SZ subscriber mailing list received an email on March 1, with a link to the quiz.[16] The language of both the email and the pop-up ad was neutral, and simply asked respondents to take a "survey on the impact of images," which will include a picture quiz, with the goal of "examining the comprehensibility and impact of the images."[17] There was no mention of anything related to misinformation or technology. Once a user clicked on the link, they were randomized in real time into the treatment or control group, which we describe below.

It is important to note that the experiment was entirely conducted within the SZ platform, using the infrastructure routinely used by SZ for fielding the survey and for randomization, and with no interface or interaction with any other entity (including ourselves, as the research team). In line with their standard procedures, there was no monetary compensation for taking the quiz, and the responses were not incentivized. In addition, there was no mention of this being part of a study or to there being different experimental conditions; as this is not something that is disclosed as part of the company's regular operations.

As an individual accesses the quiz, and prior to starting it, they are asked for

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

[16] Because subscribers were reached by email, the vast majority of the respondents took the quiz on March 1 (71%), 2 (17%) or 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 pushed.

[17] The email can be seen in Figure B.1 with the machine translated version in Figure B.2 in the Appendix.

their age and gender. Due to stringent privacy regulations in the European Union (i.e., the General Data Protection Regulation, GDPR), we then asked for consent allowing us to link their quiz performance to their browsing behavior on the website and on the SZ App. It is important to note that this is done prior to them taking the quiz. At this stage, we do not elicit any prior beliefs, again because SZ did not think it would be a natural user experience: while asking for general demographic information happens occasionally as part of surveys, elicitation of priors is not part of their routine survey or quiz-taking approach.

After the quiz, we ask individuals across both treatment and control about how hard they thought the quiz was and their concern for misinformation, both on a 0 to 6 scale.[18] We then elicit their willingness to pay for an SZ subscription using a sliding scale of 0 to 45 Euros. Finally, we elicit their trust in the content available on various media (SZ, Bild, ARD/ZDF) and social media outlets (Instagram, LinkedIn, TikTok, and X/Twitter).[19] 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 directly measuring what we have termed "trust" in the context of our conceptual framework.

**Treatment Group:** A user assigned to the treatment group sees three pairs of pictures in a sequence. For each pair, there is one image that is AI-generated, and

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

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

participants were asked which of the two images they thought were generated using AI. They are presented with four options (left, right, both, neither); following the practices suggested by SZ, we provided the correct answer after each question.

As seen in Appendix B, the first pair is related to climate events.[20] The one on the right is real, based on floods in Valencia (Spain) in 2024. The second pair has Barack Obama and Donald Trump laughing together as the authentic image taken during Jimmy Carter's funeral in January 2025 in Washington, D.C. Finally, in the third pair, the real image is that of the young woman waving an EU flag in a protest in Tbilisi (Georgia), in 2024. In each pair, one image was generated completely (or at least significantly) using an AI tool (Midjourney, Magicstudio). For this process, we provided some reference image for the GenAI tool to begin with as part of the prompt. We utilized these images in the control group, which we describe next.

**Control Group:** If a user is assigned to the control group, then, similar to the treatment group, they see a sequence of three pairs of pictures, but here all images are authentic and not manipulated by AI tools. The idea of the control group is to provide a counterfactual mimicking daily news consumption without highlighting issues related to AI-assisted manipulation. Hence, we have a quiz that focuses on testing their current affairs knowledge: for each pair of images, we ask which geographical location (country) the subjects of those images are from. As with the treatment, they are presented with four options and are told the correct answer upon submitting their chosen option.

For each pair, we use the corresponding real image presented in the treatment

[20] Each individual saw the pairs and pictures in the same order, for ease of implementation.

group. As for the second picture, we use the real image that was provided as part of the base prompt for the AI tool to mimic. In Appendix B, in the first set of images we have the same picture of the Valencia floods and in addition, a picture of the floods in the Bavaria region of Germany, of which Munich is the capital. For the second pair, the additional real image is of two prominent German politicians – Sahra Wagenknecht and Alice Weidel.[21] We used AI tools to generate a different picture for the treatment group where Weidel is seen interacting with Olaf Scholz, who served as German Chancellor from 2021 to 2025, from the Social Democratic Party of Germany (SPD). Finally, in the last set, we have an image from a protest in Paris.

**Design Choices:** The aim of keeping pictures similar across treatment and control is 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. Additionally, although we did not want to prompt readers about AI in the control group, we could not use 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

[21] Wagenknecht had been a leader of the left-wing party The Left, who in 2024 formed her own party (the Sahra Wagenknecht Alliance). Weidel was the leader of the far-right Alternative for Germany (AfD) at the time of the survey.

otherwise be violated.

As a robustness exercise, we draw on samples of subscribers who received the email invitation to take the quiz but chose not to, as an alternative "pure" control group, when measuring the impact of the AI treatment on online behavior.

Our unique collaboration with SZ to implement such a field experiment also comes with practical considerations that are not present in lab or online experimental settings. For instance, we considered a different experimental condition using older AI tools (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.

Moreover, we were constrained in terms of having a third condition for the experiment simply because of engineering and software considerations, as the SZ A/B testing tool only supports two testing conditions.[22]

From our design perspective, it is also important to note that we did not mention the term "misinformation" in either of the quizzes, 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

[22] 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.

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.[23]

**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.[24] The level of difficulty was also reflected in their assessment of perceived difficulty, which was 4.7 out of 6. Overall, individuals in the treatment group performed better than guessing at random. Although this may seem like a challenging quiz, it effectively highlights the technology's capabilities. 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) as an example).[25] In contrast, the current events control quiz was less difficult, with an average correct rate of 88% and 72% getting all three answers right. This performance is in line with other 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.

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

[23] 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.

[24] 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.

[25] See https://journalism.columbia.edu/news/cjr-new-ai-campaign as an additional example.

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. While the control quiz was less hard than the treatment, the images were chosen to mimic, in part, the readers' standard daily information consumption. Some images from the main stories on the homepage (at 9am 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. Moreover, 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.

## 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.[26] For these individuals, we have information on self-reported demographics (age and gender) along with stated preference questions at endline. The treatment and control groups are balanced on a number of observables, as seen in Panel A of Table 1. About 45% of the quiz takers are female and 94% are subscribers,balanced across treatment

[26] A proportions test to look at differences in 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 attrition (p=0.73).

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), and location.

Table 1: 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+) | 17,199 | 8,674 | 0.49 | 0.50 | 8,525 | 0.48 | 0.50 | -0.011 | 0.150 |
| Young (<40) | 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+) | 5,915 | 3,010 | 0.52 | 0.50 | 2,905 | 0.53 | 0.50 | 0.008 | 0.537 |
| Young (<40) | 5,915 | 3,010 | 0.19 | 0.39 | 2,905 | 0.19 | 0.40 | 0.001 | 0.913 |
| Pre-intervention Daily Visits (21 days) | 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 |

*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.

As shown in Panel B of Table 1, 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 $i$:

$$Y_i = \gamma + \beta T_i + \epsilon_i \tag{6}$$

where $\beta$ captures the causal effect of the AI-awareness quiz on the outcome of interest. Because of the successful randomization, controls ($Z_i$) should not affect the estimate of $\beta$. We therefore estimate the regression without control variables as the baseline specification.

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

$$y_{it} = \gamma_i + \tau_t + \beta(PostTreatment_{it} \times Treated_i) + \epsilon_{it} \tag{7}$$

where $PostTreatment_{it}$ is a dummy equal to one for the days after individual $i$ took the quiz and 0 otherwise. $Treated_i$ indicates whether individual $i$ took the AI-focused quiz. $\gamma_i$ and $\tau_t$ are individual and date fixed effects, respectively. The coefficient of interest $\beta$ captures the average impact of taking the AI quiz relative to the current

affairs quiz.[27] We consider three weeks as the pre-quiz period and different windows for the post-quiz period.

# 5 Results

## 5.1 Survey Outcomes

### 5.1.1 Baseline

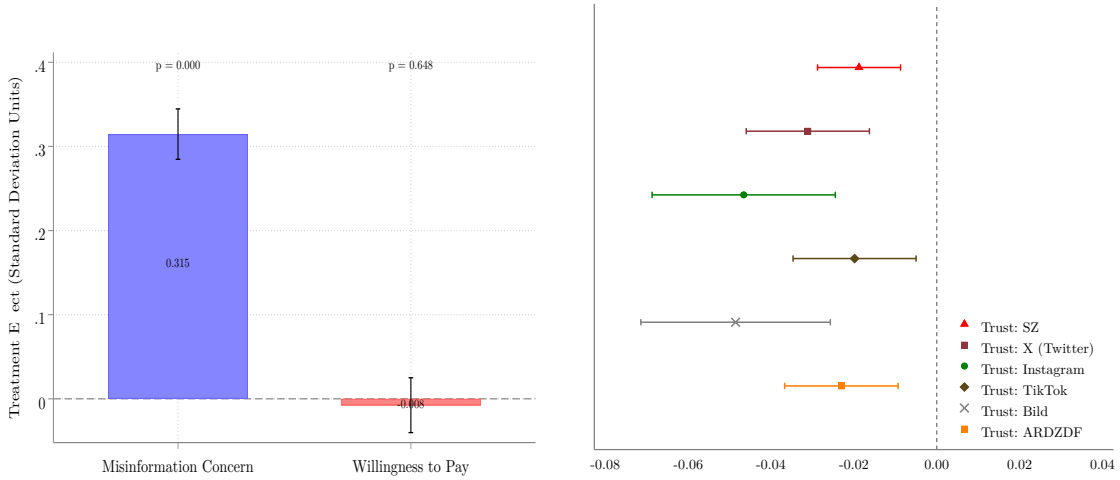
We begin with the average treatment effect (ATE) for the survey sample. The first outcome we analyze is concern with misinformation (measured on a scale of 0 to 6). Panel (a) of Figure 2 shows that it is significantly higher in the treatment group than in the control. The increase is also economically 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.[28]

Having established our "first stage" in inducing increased concern with misinformation, as captured by the $\Delta$ parameter in our model, we then look at reported trust in media outlets and platforms (henceforth platforms). The first thing to note is that the different platforms command very different levels of trust. Focusing on

[27] Note that, since we have a binary treatment and our tracked sample took the quiz within a very short time period (see footnote 16), our specification boils down to the canonical differences-in-differences case, where the TWFE estimator is robust (de Chaisemartin and D'Haultfoeuille, 2025).

[28] 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




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

Notes: Figure (a) captures the treatment effect on concern with misinformation and WTP for SZ across the treatment and control groups. The dependent variable is on a scale of 0 to 6 for concern with misinformation while it 0 to 45 Euros for the WTP question. The treatment effect in the figure is captured in terms of standard deviations for ease of interpretation. The vertical lines represent 95 percent confidence intervals. Figure (b) plots the average treatment effects capturing impact of the AI intervention on the treatment group relative to the control for trust in SZ, X, Instagram, Tiktok, Bild, and ARD/ZDF. The treatment effect in the figure is captured in terms of standard deviations for ease of interpretation. The horizontal lines represent 95 percent confidence intervals.

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 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, equivalent to approximately 0.1 standard deviations. While this is in line 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-trust and high-trust platforms alike – including, notably, SZ itself. The decline in trust is 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).

Interestingly, as shown in Figure 2 (a), the AI intervention has no effect on the willingness to pay (WTP) for a SZ subscription: despite the decline in trust, individuals do not report being less willing to pay to access SZ. [29]

## 5.2 Revealed Preference Outcomes

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 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.

[29] The underlying estimates to Figure 2 are reported in Table A.1, in the Appendix. Table A.2 confirms that the results are qualitatively similar when looking only at the tracked sample. Treatment effects are also robust to controlling for observables as shown in columns (1)-(3) of Table A.3.

### 5.2.1 Browsing Outcomes: 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 at the 5% level (p-value 0.032). The balance checks in Table 1, together with a statistically and economically significant (2.77%) estimate, suggest a meaningful impact of the intervention on engagement.

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 the day on which respondents answered the quiz from the analysis, 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 2, 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 we return to below when discussing the results for the subscription outcomes.

Table 2: Post-Intervention Engagement: Differences-in-Differences Analysis

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Dep. variable | 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) |
| Post | 0.192** | 0.202** | 0.208** | 0.215*** |
| | (0.089) | (0.087) | (0.083) | (0.082) |
| 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 |
| Obs | 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). 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.

We further analyze how the treatment effect decays over time in Table A.4 of the Appendix, across four mutually exclusive post-period windows. Column (1) replicates the result for the five days post-intervention as in column (2) of Table 2, with the coefficient of 0.085 statistically significant at the 5% level. In days 6-10 post-treatment, this coefficient decreases by about 25%, but remains sizable at 0.062, although only marginally insignificant at conventional levels (p-value: 0.13).

The point estimate then decreases to 0.05 (days 11-15) and 0.032 (days 16-20), with the effects being statistically insignificant.

### 5.2.2 Browsing Outcomes: Alternative Mechanisms and Robustness

Our conceptual framework attributes the patterns we have found in the data to a mechanism based on an increased concern with misinformation on the part of news consumers. Our results from the survey outcomes, documenting precisely such an increased concern, provide prima facie validation of that mechanism, as encapsulated in Result 2 of the model. To rule out the possibility that the increased demand may reflect some alternative mechanisms, and to highlight the robustness of our results, we conduct a series of additional analyses.

First, for website engagement as the outcome variable, we can construct an alternative control group. Recall that, in the baseline results, the control group comprises individuals who took the control picture quiz, for whom we have both survey and browsing outcomes. As an alternative, we draw a random sample of 3,000 individuals who received the invitation email but chose not to take the quiz, and thus were not part of our experiment. While this group may differ in some dimensions from the treatment group due to potential selection issues, it provides a distinct counterfactual benchmark, a "pure" control group with individuals who were not exposed to any quiz.

We set the treatment date to be the 1st of March since it is the day that the email invitation to the survey was sent to subscribers, and when most of those who chose to take it did so. As can be seen in column (2) of Appendix Table A.5, there

is a statistically and economically significant effect in the first five days after the experiment. The effect size is nearly 50% larger than the baseline five-day post-period result, at about 2.8%. Such a check thus establishes that the effect we document is not attributable to the nature of the control quiz.

Next, we draw on additional browsing data from SZ to rule out short-term attention or entertainment effects resulting from our AI intervention, as opposed to the mechanism related to misinformation concerns that our model highlights. 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.

In columns (8) and (9) of Table A.5, we find that there is no differential engagement with surveys on SZ across treatment and control. In addition, our baseline results are robust to excluding from the sample any engagement with surveys and quizzes on the SZ platforms after our experiment. Table A.6 shows the results when we exclude any visit that included engagement with a survey (columns 1-4) or any individual who engaged with any survey after the experiment (columns 5-8). 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 next 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. This allows us to break down overall traffic into two categories: traffic from social media

referrals and organic visits.

With the caveat of traffic from social media links being sparse, we find that the overall effect in column (1) of Table A.7 is driven by organic visits (column 2). There is no change in social media-mediated visits to the SZ website and app (column 3), with the sign of the effect being negative. These results also provide suggestive evidence that the treatment increased engagement with SZ itself, but not with other platforms.

We then analyze the news diet of the individuals in the experiment to test whether they returned to SZ in anticipation of specific content. While our experiment was a standalone exercise with no coordination with the editorial team, the readers might have anticipated certain types of content in the treatment relative to the control group. We draw on additional fine-grained browsing data to analyze how consumption of different types of news articles across categories (politics, sports, entertainment, etc.) changed in the aftermath of the experiment. As seen in columns (1)-(4) of Table A.8, there was no change in news consumption diet (topics such as politics, sports, music, and crime) right after the intervention for the treatment group relative to those in the control. These results strengthen our confidence that the effect we document is not driven by alternative short-term entertainment mechanisms triggered by the intervention.

We also conduct several additional exercises to test the robustness of our results. First, Appendix Table A.5 shows in column (3) results for a non-linear Poisson model, to account for the count nature of our dependent variable. The coefficient on $Treatment \times Post$ is positive and statistically significant at the 5% level, highlighting

a result qualitatively similar to our OLS estimates. As an additional outcome, we analyze the total daily time spent by a user on SZ online, though this variable, even when captured by advanced analytics, comes with significant caveats.[30] That noted, we find that the treatment increases the total time spent on SZ digital content, both conditional on visiting the website or app (column 4) and unconditionally (column 5), by about 4%. In column (6), we see that the results are robust to using a 14-day rather than a 21-days pre-period window. Finally, when we use a more saturated specification with time-since-quiz fixed effects in addition to date fixed effects, we find similar results (column 7), which is unsurprising since 87% of the tracked sample were treated within two days. These results demonstrate the robustness of our baseline results along various dimensions.

### 5.2.3 Subscription Outcomes

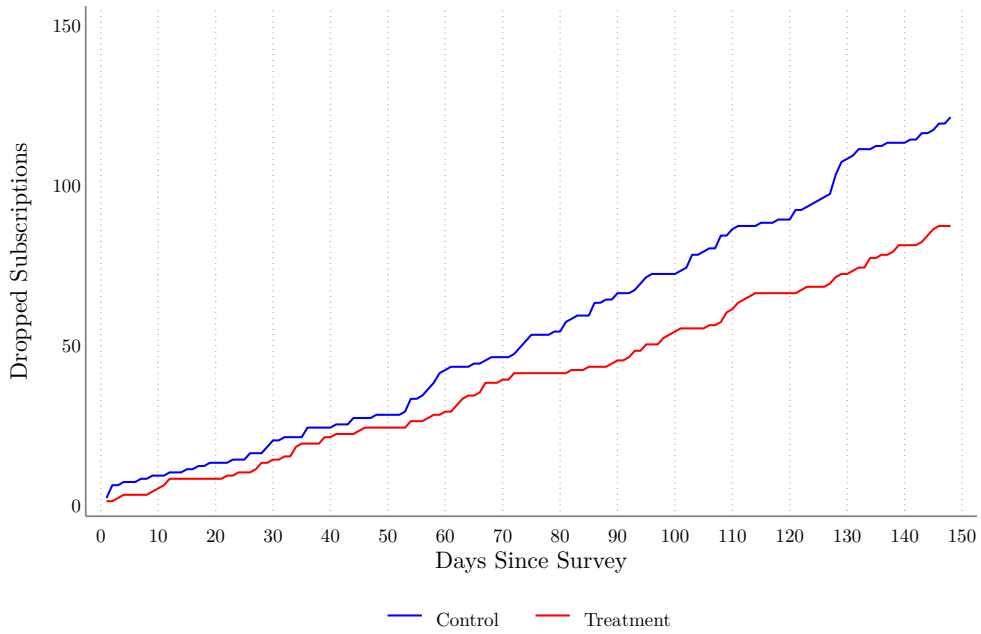
Continuing with revealed preference outcomes, we turn to individual subscription behavior and analyze, for individuals in the tracked sample, whether they discontinue or maintain their subscription to SZ.[31] Note that there is no variation in subscription status before the treatment, since all tracked 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. We begin with a descriptive look at the data before turning

[30] These include censoring (such as the need for multiple page views to estimate duration), the inability to track active attention, ad blockers, and challenges in detecting tab-switching behavior, all of which are further complicated by privacy regulations.

[31] 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.

to formal statistical analysis.

Figure 3: 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.

We begin with the cumulative number of dropped subscriptions, which is illustrated in Figure 3. The results are 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 3, where the dependent variable in each column is the individual subscription status two through five months after the intervention. We see that 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 1/3 lower than the control group. After five months, the effect accumulates to the point where the probability of remaining a subscriber is about 1.1% higher for the treatment group.

Table 3: 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 | 0.0068** | 0.0074* | 0.0109** |
| | (0.0027) | (0.0034) | (0.0040) | (0.0046) |
| Sample mean | 0.989 | 0.982 | 0.975 | 0.967 |
| | (0.105) | (0.132) | (0.156) | (0.180) |
| Observations | 6,158 | 6,158 | 6,158 | 6,158 |
| R-squared | 0.0004 | 0.0007 | 0.0006 | 0.0009 |
| F-statistic | 2.436 | 4.053 | 3.472 | 5.699 |

*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 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 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).[32]

[32]It is pertinent to note that we mitigate issues related to the Hawthorne effect since we had both a

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 380,000 euros (base annual price $\times$ increased retention $\times$ subscriber base = 129 euros $\times$ 0.01 $\times$ 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 pure 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 about 3.3 million (2900 $\times$ 365 $\times$ 3.1).

Finally, to ensure the robustness of our results for subscriptions, we adapt the idea of the "pure control" group from SZ subscribers who were invited to take the quiz but did not respond. Unlike for the analysis of browsing outcomes, in which we could include individual fixed effects in the DiD specification to account for potential selection, in this case, we do so by focusing on 500 highly-engaged non-respondents with at least 234 visits to SZ in February 2025, aligning them with the top 10% of the treatment group. We thus compare treated individuals to subscribers with the highest expected loyalty, a comparison that should, if anything, bias the results against finding a treatment effect on retention.

Yet Table A.9 shows that the pure control subgroup is in fact less likely to remain subscribers over the subsequent period, compared to the individuals who took the

treatment and control group take a quiz, and they were unaware of being part of an experiment.

AI quiz.[33] As with the browsing outcomes, this underscores that the subscription results are not driven by the specific characteristics of the control quiz in comparison to the treatment.

Our baseline analysis reveals an interesting pattern: although the impact of the treatment on daily visit decisions may fade, it remains strong when individuals face the choice of whether 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.[34]

Overall, our results demonstrate that highlighting the difficulty of distinguishing real from synthetic images increased concern about misinformation. In line with our theoretical framework, this increased concern led to an increase in engagement with SZ, both in terms of daily visits in the short term and subscription outcomes over the following months. Our key results are corroborated by global survey evidence in the Reuters Institute Digital News Report released in June 2025 (See Newman et al. (2024a)). In the report, based on online surveys conducted in Jan-Feb 2025, a question asked what source they would go to if they felt a piece of news online was false. The top choice for survey takers, across political leanings and education levels, was "a news source I trust". This survey data, from a well-known third party,

[33] Moreover, we find no significant difference between this "pure control" and the control group that took the current affairs quiz.

[34] It is also interesting to consider the result in relation to the null effect in terms of the survey-reported WTP, as described above. The effect we detect on subscriptions highlights the different nature of the choice at hand – individuals already subscribing to the platform do not face, as a matter of course, the actual choice of which price to pay – and as such the value added by our experimental context. In the absence of the revealed preference outcomes, such a change in behavior – evidently important from the platform's business perspective – may have gone unnoticed. It is also the case, as we will see in the next subsection, that certain subgroups within the population that are more strongly responsive to the treatment do report an increased WTP.

provides further support for the internal and external validity of our core results.

## 5.3 Heterogeneity

### 5.3.1 Quiz Performance

We begin with a key (pre-registered) heterogeneity dimension: performance in the quiz. In line with the literature that has emphasized individuals' perception of their ability to identify misinformation (Assenza et al., 2024; Harris et al., 2024), we focus on self-assessed relative performance. Specifically, to capture differences across groups, we define a dummy variable $Hard$ equal to one if an individual reported finding the quiz harder than the median individual in their group (treatment or control).[35] In column (1) of Table 4, we see that those who found the quiz relatively hard display a significantly larger treatment effect than those who found it relatively easy, when it comes to concern with misinformation, indicating that the treatment left a stronger impression on individuals who found it more challenging.[36]

This was matched by heterogeneity in the other survey outcomes. Interestingly, that group did not reduce their trust in SZ, as the effect is precisely estimated zero (column 2). Since SZ trust levels are already close to the maximum in the control group, and a positive effect would thus have been essentially impossible,

[35] Median values are 1 for the control group, and 5 for the treatment group. To keep relative balance, in light of the discrete nature of the measures, we set $Hard = 1$ for individuals scoring 2 and above in the control group (about 44%) and 5 and above in the treatment group (about 65%). Setting the thresholds to 1 and above and 4 and above, respectively, would have yielded 79% and 84%.

[36] We also define an objective measure of performance, $WeakPerformance$, taking the value of 1 if the individual had fewer correct answers than the median individual in their group. The two performance measures are substantially positively related (correlation 0.44). Figure A.1 displays the joint distribution for the number of correct answers and the self-assessed difficulty, separately for treatment and control. Results are qualitatively similar, as shown in Appendix Table A.11.

Table 4: Survey Outcomes: Heterogeneity

| Dep. variable | (1) Misinfo Concern | (2) Trust in SZ | (3) WTP | (4) Misinfo Concern | (5) Trust in SZ | (6) WTP | (7) Misinfo Concern | (8) Trust in SZ | (9) WTP |
|---|---|---|---|---|---|---|---|---|---|
| Treatment | 0.085*** | -0.045*** | -1.148*** | 0.248*** | -0.008 | 0.266 | 0.219*** | -0.024** | 0.747 |
| | (0.018) | (0.008) | (0.284) | (0.028) | (0.010) | (0.498) | (0.025) | (0.010) | (0.481) |
| AI Treatment × Hard | 0.222*** | 0.043*** | 1.881*** | | | | | | |
| | (0.023) | (0.011) | (0.378) | | | | | | |
| AI Treatment × PoliticsAffinity | | | | -0.093** | -0.015 | -0.086 | | | |
| | | | | (0.037) | (0.014) | (0.650) | | | |
| AI Treatment × HeavyReader | | | | | | | -0.047 | 0.013 | -0.927 |
| | | | | | | | (0.035) | (0.014) | (0.633) |
| Sample mean | 5.606 | 2.899 | 20.996 | 5.625 | 2.939 | 22.887 | 5.626 | 2.939 | 22.867 |
| | (0.726) | (0.334) | (11.050) | (0.678) | (0.254) | (10.873) | (0.677) | (0.254) | (10.919) |
| Observations | 17,177 | 17,006 | 14,374 | 5,493 | 5,460 | 4,613 | 5,728 | 5,693 | 4,811 |
| R-squared | 0.033 | 0.002 | 0.002 | 0.022 | 0.001 | -0.001 | 0.021 | 0.001 | 0.001 |
| F-statistic | 256.178 | 10.312 | 8.349 | 40.977 | 2.563 | 0.181 | 41.411 | 2.608 | 1.649 |

*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 columns (1), (4), and (7) while it is Trust in SZ in columns (2), (5), and (8), and is the WTP for SZ in columns (3), (6), and (9). AI Treatment is equal to 1 if individual $i$ was in the treated group and zero otherwise.

this underscores the resilience of trust in the outlets seen ex ante as trustworthy, when the challenge of distinguishing between real and AI-generated content is seen as especially acute. In contrast, trust in the platforms deemed as low-trust ex ante did decrease, consistent with the idea of an overall decline in trust in the broader information environment. Moreover, the same group increased their reported WTP for an SZ subscription by about 3.9%, an effect significant at conventional levels (column 3). Turning to the revealed preference outcomes, we see in Table 5 that the increase in daily visits that we detected in the ATE is in fact driven by SZ readers who found the quiz relatively hard, as captured by $Hard = 1$ (column 1), and not those who did not, $Hard = 0$ (column 2). In other words, this result is reassuringly in line with the patterns in the survey results.

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.[37] Overall, though, the suggestive pattern is consistent with the idea that awareness of the challenge of discerning AI-generated from real content induces increased demand for relatively more trusted news sources, while reducing trust in the information environment more broadly.

[37] To provide evidence that the groups might be similar, in Table A.10, we test the balance on observables to find no statistical or economic differences across different observables (gender, age, etc.). There is a higher proportion of subscribers in the treatment group who find the quiz hard, but economically it is small, with the difference being 1.9%. Moreover, our results are qualitatively similar if we control for subscriber status and other controls in our regressions, as seen in columns (4)-(6) of Table A.3 in the Appendix.

Table 5: Post-Intervention Engagement: Heterogeneity

| Dep. variable | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Number of daily visits | | | | | |
| Sample | Hard | Not Hard | High Politics Affinity | Low Politics Affinity | Heavy Reader | Light Reader |
| AI Treatment × Post | 0.121** | 0.022 | 0.060 | 0.139** | 0.148* | 0.029 |
| | (0.060) | (0.063) | (0.060) | (0.062) | (0.076) | (0.033) |
| Post | 0.181 | 0.180 | 0.338*** | 0.050 | 0.329** | 0.049 |
| | (0.120) | (0.134) | (0.123) | (0.136) | (0.165) | (0.077) |
| Sample mean | 4.693 | 4.644 | 5.497 | 4.029 | 8.128 | 1.153 |
| | (5.470) | (5.554) | (5.747) | (5.108) | (5.747) | (1.696) |
| Individual FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Date FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 81,770 | 66,898 | 85,783 | 64,507 | 79,002 | 77,868 |
| R-squared | 0.764 | 0.775 | 0.752 | 0.765 | 0.648 | 0.319 |
| F-statistic | 4.463 | 1.079 | 5.365 | 2.892 | 5.164 | 0.716 |

*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 analysis period is 21 days before the intervention to 5 days post-intervention. 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.

### 5.3.2 Politics Affinity

Next, we examine heterogeneity based on pre-experiment behavior on SZ digital platforms. This information, available only for the tracked sample, allows us to shed further light on the mechanisms through which the documented effects may operate.

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.[38] 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

[38]This dimension was not pre-registered, 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/).

to whether the individuals are classified as having a politics affinity (54% of the sample) or not (46%).

The politics affinity variable is helpful in two ways. First, through the lens of our conceptual framework, we can think of individuals without political affinity as having a higher $\sigma_X^2$, i.e. a more diffuse prior. Second, similar to the effects, an intervention could provide novel information to previously uninformed individuals, or alternatively, it could serve as a nudge or reminder to individuals who were already knowledgeable about the topics mentioned (Nelson, 1974). Political affinity serves as a marker of prior knowledge, helping to distinguish between an informative and a persuasive channel.

The results for the survey outcomes, in columns 4-6 of Table 4, show that the effect of the treatment on misinformation concern (our focal manipulation check) is larger for the individuals who had a relatively low level of prior affinity for political topics. Interestingly, these individuals did not significantly reduce their trust in SZ either, while those with high political affinity displayed a more negative coefficient in that regard – even if short of conventional statistical significance thresholds.

Turning again to the engagement outcomes, the results in columns 3-4 of Table 5 are consistent with the pattern for the survey outcomes: the increase in visits as a result of the treatment is driven by individuals who do not have a significant interest in politics. 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, of less than one-half of the magnitude compared to the former group.

Overall, this suggests that our intervention highlighted the issue of AI-generated

synthetic content significantly more to those who are not usually as informed about politics, and would thus have more diffuse priors. This is consistent with the prediction from our conceptual framework, where these individuals would be expected to display a stronger demand response to increased misinformation, and suggests an informational rather than a persuasive mechanism.

### 5.3.3 Prior SZ Usage and Other Demographics

An additional (pre-registered) comparison is between relatively heavy and light users of SZ, which we classify by splitting the sample based on the median number of visits in the two weeks before the experiment (44 visits).[39]

Regarding the number of visits to SZ digital platforms, columns 5-6 of Table 5 indicate that heavy readers respond more strongly to the treatment. This is in line with the idea that it is readers who attached a higher level of trustworthiness to the outlet prior to the intervention, as indirectly revealed by their prior engagement, who respond to increased awareness of the AI challenge regarding misinformation by increasing that engagement. This is again consistent with the prediction from our framework. Interestingly, although the patterns from the survey outcomes are noisy (columns 7-9, Table 4), the signs of the coefficients are also in line with the predictions of the model: individuals who attach greater trustworthiness to the outlet naturally have a smaller drop in their trust in the outlet's content when faced with a shock to their perception of the spread of misinformation.[40]

[39] These results are robust to using the entire pre-period window data of three weeks.

[40] Specifically, individuals who attach a smaller $\alpha$ (greater trustworthiness) to the outlet will have a smaller increase in their perceived assessment of the variance of the signal, $\sigma_X^2 + \alpha\Delta$, in response to an increase in $\Delta$. It is natural to expect that they would report a smaller increase in the concern

In Table A.12 in the Appendix, we examine additional heterogeneity (not pre-registered) by age and gender. For age, we find no significant differences between younger and older readers across the three main survey outcomes. For gender, we find that female readers report greater concern about misinformation than male readers, but that the treatment effect is somewhat weaker for women. We find no significant gender differences in the treatment effect on trust in SZ or willingness to pay for a SZ subscription.

# 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 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 technology like GenAI leads to reduced trust in the information environment as a whole. However, an outlet perceived as sufficiently trustworthy may still witness increased

with misinformation in general, as it is arguably affected by their perception of the outlet itself.

demand, even as trust in its content also suffers. That is, its relative value goes up in the eyes of readers, who deem it trustworthy enough to mitigate the effects of the misinformation technology.

The average treatment effects are further substantiated by observed heterogeneity patterns. First, we see that the effects are driven by those individuals who have (or admit to) a stronger perception of the task's difficulty. While task difficulty was not randomly assigned, this descriptive fact is consistent with our hypothesized causal pathway. Second, the heterogeneity patterns align with the predictions of our framework: individuals with low political affinity (higher $\sigma_X^2$ in the model) are more sensitive to increased misinformation in their demand. This – along with the suggestive evidence of a differential response from heavy users of SZ digital content (arguably attaching a smaller $\alpha$ to the outlet, in terms of the model) – 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 (e.g., *The New York Times* or *The Wall Street Journal* in the US, *The Financial Times* or *The Guardian* in the UK, or *Frankfurter Allgemeine Zeitung* in Germany). Indeed, as discussed, Newman et al. (2024a) provide prima facie support for the external validity of our study. 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. 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 does, however, highlight certain issues that should be considered going forward. Recall that, in our model, trust in the content of a trustworthy outlet depends essentially on the balance between trustworthiness ($\frac{1}{\alpha}$) and the overall prevalence of misinformation ($\Delta$). Our Result 2 establishes that, for a given level of misinformation, there will be a sufficiently high level of trustworthiness such that the outlet can benefit from an increase in the overall prevalence of misinformation; however, as the latter continues to grow, the threshold for trustworthiness

keeps rising. In other words, from a dynamic perspective, it is not enough to maintain a given level of trustworthiness. 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.

**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.

**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.

**Longoni, Chiara, Andrey Fradkin, Luca Cian, and Gordon Pennycook**, "News from Generative Artificial Intelligence is Believed Less," in "Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency" 2022, pp. 97–106.

**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, 2024.

**_ , 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.

**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.

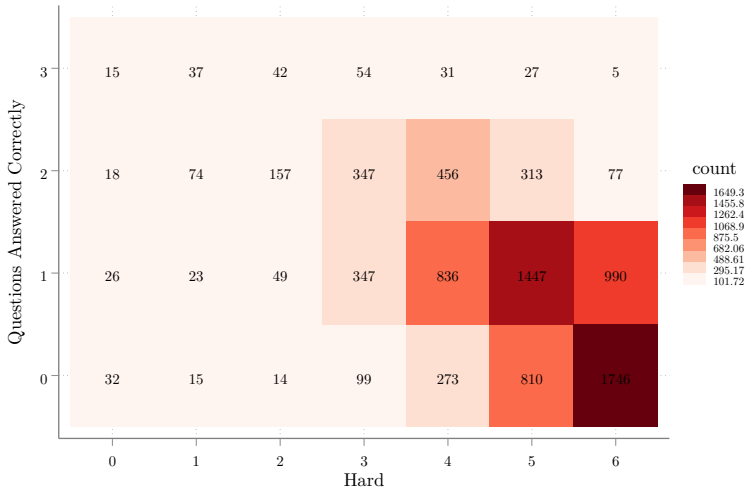
**Vosoughi, Soroush, Deb Roy, and Sinan Aral**, "The Spread of True and False News Online," *Science*, 2018, *359* (6380), 1146–1151.

**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.

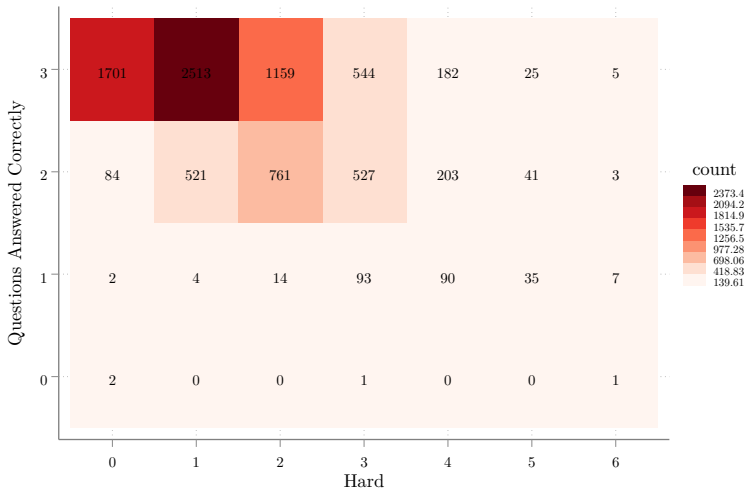
**Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov**, "Political Effects of the Internet and Social Media," *Annual Review of Economics*, 2020, *12*, 415–438.

# A Appendix: Additional Results

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




(a) Control




(b) Treatment

Notes: The Figures captures the perceived and actual performance for the Control (figure (a)) and Treatment 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.

Table A.1: 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 |
|---|---|---|---|---|---|---|---|---|
| AI Treatment | 0.223*** | -0.085 | -0.019*** | -0.031*** | -0.047*** | -0.020*** | -0.049*** | -0.023*** |
| | (0.011) | (0.184) | (0.005) | (0.008) | (0.011) | (0.008) | (0.012) | (0.007) |
| Sample mean | 5.605 | 21.002 | 2.898 | 0.236 | 0.740 | 0.247 | 0.794 | 2.789 |
| | (0.729) | (11.057) | (0.334) | (0.482) | (0.703) | (0.472) | (0.755) | (0.458) |
| Observations | 17,284 | 14,461 | 17,113 | 16,096 | 15,498 | 15,496 | 16,718 | 17,153 |
| Adjusted-$R^2$ | 0.023 | -0.000 | 0.001 | 0.001 | 0.001 | 0.000 | 0.001 | 0.001 |
| F-statistic | 416.980 | 0.214 | 13.470 | 16.846 | 17.066 | 6.857 | 17.326 | 10.865 |

*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). AI Treatment is equal to 1 if individual $i$ was in the treated group and zero otherwise.

Table A.2: 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 |
|---|---|---|---|---|---|---|---|---|
| AI Treatment | 0.197*** | 0.198 | -0.019*** | -0.004 | -0.034* | -0.012 | -0.045** | -0.042*** |
| | (0.017) | (0.311) | (0.007) | (0.012) | (0.019) | (0.013) | (0.020) | (0.011) |
| Sample mean | 5.624 | 22.749 | 2.938 | 0.221 | 0.774 | 0.252 | 0.778 | 2.823 |
| | (0.682) | (10.977) | (0.256) | (0.462) | (0.705) | (0.468) | (0.755) | (0.410) |
| Observations | 5,923 | 4,980 | 5,888 | 5,527 | 5,329 | 5,316 | 5,736 | 5,902 |
| Adjusted-$R^2$ | 0.021 | -0.000 | 0.001 | -0.000 | 0.000 | -0.000 | 0.001 | 0.002 |
| F-statistic | 127.328 | 0.406 | 8.286 | 0.081 | 3.049 | 0.901 | 4.995 | 15.357 |

*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). 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.3: Survey Robustness: Controls

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Dep. variable | Misinfo Concern | Trust SZ | WTP | Misinfo Concern | Trust SZ | WTP |
| AI Treatment | 0.224*** | -0.020*** | -0.089 | 0.093*** | -0.041*** | -0.858*** |
| | (0.011) | (0.005) | (0.177) | (0.018) | (0.008) | (0.271) |
| Hard | | | | -0.008 | -0.006 | -0.537** |
| | | | | (0.017) | (0.007) | (0.247) |
| AI Treatment × Hard | | | | 0.208*** | 0.035*** | 1.338*** |
| | | | | (0.023) | (0.011) | (0.363) |
| Sample mean | 5.605 | 2.898 | 21.002 | 5.606 | 2.899 | 20.996 |
| | (0.729) | (0.334) | (11.057) | (0.726) | (0.334) | (11.050) |
| Controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 17,284 | 17,113 | 14,461 | 17,177 | 17,006 | 14,374 |
| R-squared | 0.040 | 0.032 | 0.078 | 0.048 | 0.032 | 0.078 |
| F-statistic | 87.205 | 36.155 | 189.204 | 103.428 | 28.402 | 147.045 |

*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 columns (1) and (4), Trust in SZ in columns (2) and (5), and WTP for SZ in columns (3) and (6). 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 40 years old, or is over 60 years old.

Table A.4: Post-Intervention Engagement: Split Window

| Dep. variable | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Number of daily visits | | | |
| Post Window | (1;5) | (6;10) | (11;15) | (16;20) |
| AI Treatment × Post | 0.0847** | 0.0627 | 0.0509 | 0.0326 |
| | (0.0418) | (0.0424) | (0.0462) | (0.0475) |
| Post | 0.2023** | -0.2632 | | |
| | (0.0873) | (0.4014) | | |
| Sample mean | 4.653 | 4.586 | 4.621 | 4.558 |
| | (5.493) | (5.437) | (5.479) | (5.404) |
| Individual FE | Yes | Yes | Yes | Yes |
| Date FE | Yes | Yes | Yes | Yes |
| Observations | 155,142 | 155,142 | 155,142 | 155,003 |
| R-squared | 0.768 | 0.764 | 0.763 | 0.761 |
| F-statistic | 6.224 | 1.259 | 1.213 | 0.470 |

*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 has a pre-period of 21 days from -21 to -1. The post period in the analysis varies from 1-5 days post in column (1), 6-10 days post in column (2), 11-15 days post in column (3), and 16-20 days post in column (4). AI Treatment is equal to 1 if individual $i$ was in the treated group and zero otherwise. Post is equal to 1 in the time window after the quiz in the analysis after the individual took the quiz.

Table A.5: 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) |
| Post | 0.2023** | 0.0475 | 0.2075*** | 206.7250 | 190.3241 | 0.1915** | | 0.0004 | 0.0002 |
| | (0.0873) | (0.1054) | (0.0528) | (264.6938) | (176.9965) | (0.0886) | | (0.0005) | (0.0001) |
| 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 2, Column (2) uses an alternative control 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.6: Dropping Visits and Individuals Engaging with Other Surveys

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| Dep. variable | 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) |
| Post | 0.195** | 0.204** | 0.210** | 0.219*** | 0.189** | 0.200** | 0.206** | 0.214*** |
| | (0.0888) | (0.0875) | (0.0834) | (0.0822) | (0.0887) | (0.0873) | (0.0833) | (0.0821) |
| 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.7: 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) |
| Post | 0.2023** | 0.2035** | -0.0009 |
| | (0.0873) | (0.0873) | (0.0014) |
| 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.8: Post-Intervention Engagement: Reading Topics

| Dep. variable | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Reading Topics | | | |
| | (Politics) | (Sports) | (Music) | (Crime) |
| Timespan | (-21;+5) | (-21;+5) | (-21;+5) | (-21;+5) |
| AI Treatment × Post | 0.000232 | -0.00227 | -0.000640 | 0.000659 |
| | (0.00423) | (0.00332) | (0.00330) | (0.00370) |
| Post | 0.0212** | 0.00327 | 0.00920 | 0.0144* |
| | (0.00870) | (0.00615) | (0.00682) | (0.00744) |
| Constant | 0.157*** | 0.0751*** | 0.0451*** | 0.111*** |
| | (0.00163) | (0.00114) | (0.00127) | (0.00139) |
| Individual FE | Yes | Yes | Yes | Yes |
| Date FE | Yes | Yes | Yes | Yes |
| Observations | 155,116 | 155,116 | 155,116 | 155,116 |
| R-squared | 0.416 | 0.487 | 0.439 | 0.434 |

*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 topic (affinity) read by the individual at the daily level. 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. Post is equal to 1 if the day is after the date the individual took the quiz.

Table A.9: Post-intervention Subscription Status: Pure Control vs. Treatment

| Dep. variable | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Subscription Status | | | |
| Months Post-Intervention | 2 Months | 3 Months | 4 Months | 5 Months |
| AI Treatment (vs. Pure Control) | 0.0132* | 0.0206** | 0.0270** | 0.0249** |
| | (0.00785) | (0.00946) | (0.0111) | (0.0120) |
| Constant | 0.980*** | 0.968*** | 0.955*** | 0.950*** |
| | (0.00259) | (0.00318) | (0.00393) | (0.00460) |
| Controls | Yes | Yes | Yes | Yes |
| Observations | 3,284 | 3,284 | 3,284 | 3,284 |
| R-squared | 0.009 | 0.009 | 0.007 | 0.007 |

*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 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 pure control group. Controls include reading affinity across politics, crime, music, and sports in the month prior to the experiment.

Table A.10: 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+) | 9,262 | 3,776 | 0.47 | 0.50 | 5,486 | 0.48 | 0.50 | 0.009 | 0.412 |
| Young (<40) | 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.

Table A.11: Heterogeneous Effects: Quiz Actual Performance

| Dep. variable | (1) Misinfo Concern | (2) Trust SZ | (3) WTP |
|---|---|---|---|
| AI Treatment × Weak performance | 0.095*** | 0.024* | 1.197*** |
| | (0.027) | (0.013) | (0.438) |
| AI Treatment | 0.165*** | -0.031*** | -0.480 |
| | (0.019) | (0.010) | (0.334) |
| Weak performance | -0.035* | -0.015* | -0.986*** |
| | (0.020) | (0.008) | (0.286) |
| Sample mean | 5.606 | 2.900 | 21.000 |
| | (0.726) | (0.331) | (11.039) |
| Observations | 16,856 | 16,700 | 14,136 |
| R-squared | 0.024 | 0.001 | 0.001 |
| F-statistic | 141.364 | 6.711 | 4.117 |

*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.

Table A.12: Heterogeneous Effects: Age and Gender

| Dep. variable | (1) Misinfo Concern | (2) Trust SZ | (3) WTP | (4) Misinfo Concern | (5) Trust SZ | (6) WTP |
|---|---|---|---|---|---|---|
| AI Treatment | 0.235*** | -0.015** | -0.140 | 0.248*** | -0.021*** | 0.059 |
| | (0.015) | (0.007) | (0.232) | (0.016) | (0.007) | (0.254) |
| AI Treatment × Old (60+) | -0.023 | -0.007 | 0.199 | | | |
| | (0.022) | (0.010) | (0.371) | | | |
| Old (60+) | 0.035** | -0.011 | 2.265*** | | | |
| | (0.017) | (0.007) | (0.257) | | | |
| AI Treatment × Female | | | | -0.053** | 0.006 | -0.320 |
| | | | | (0.022) | (0.010) | (0.368) |
| Female | | | | 0.162*** | 0.008 | -0.004 |
| | | | | (0.017) | (0.007) | (0.256) |
| Sample mean | 5.605 | 2.898 | 21.002 | 5.605 | 2.898 | 21.002 |
| | (0.729) | (0.334) | (11.057) | (0.729) | (0.334) | (11.057) |
| Observations | 17,284 | 17,113 | 14,461 | 17,284 | 17,113 | 14,461 |
| R-squared | 0.024 | 0.001 | 0.011 | 0.032 | 0.001 | -0.000 |
| F-statistic | 140.181 | 6.908 | 54.319 | 191.251 | 6.127 | 0.586 |

*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 columns (1) and (4), Trust in SZ in columns (2) and (5), and WTP for SZ in columns (3) and (6). AI Treatment is equal to 1 if individual $i$ was in the treated group and zero otherwise.

# B Intervention Details

Figure B.1: Quiz Email Invitation

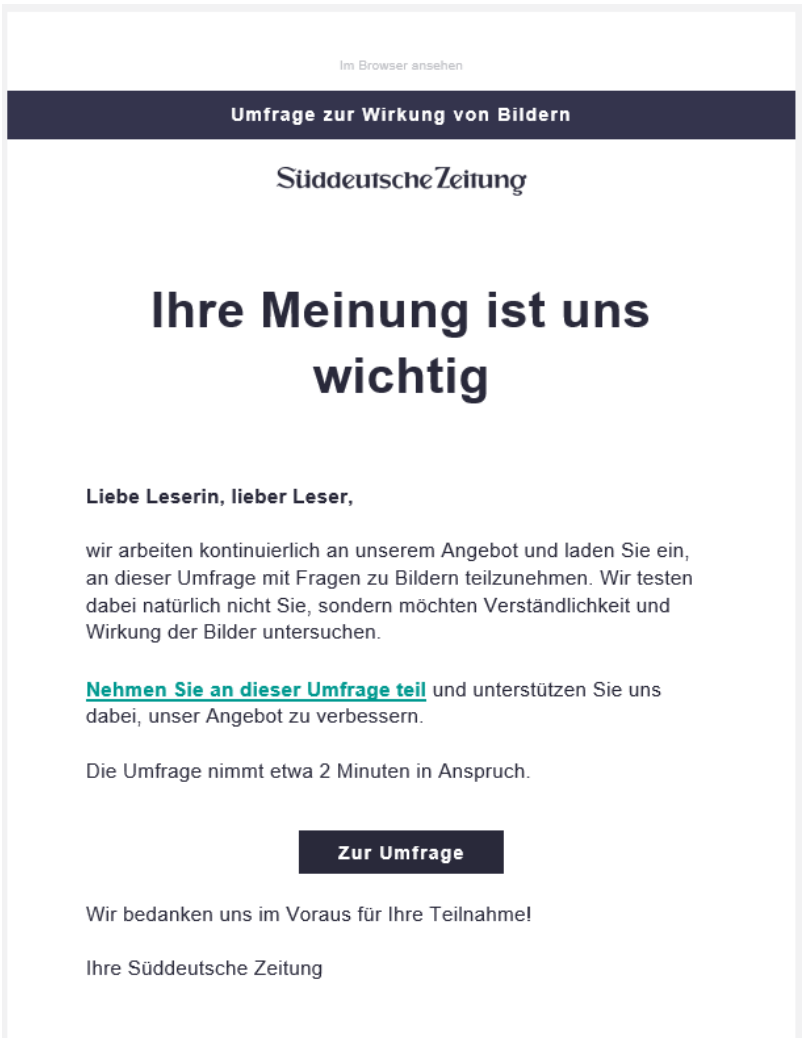


Figure B.2: Quiz Email Invitation (Machine Translated)

**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

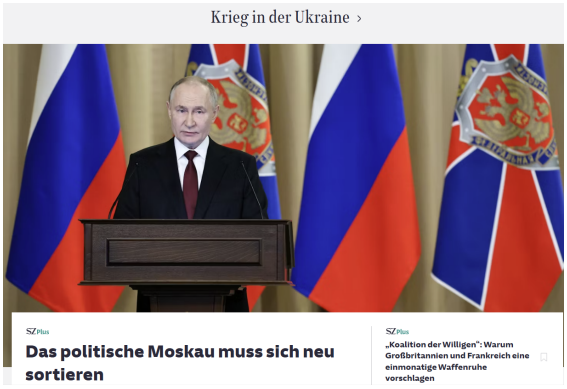



Image and Headline 2nd March 2025

# D Model Extension

Here we consider a simple extension of the model in Section 3, in which we allow the trustworthy outlet $s_1$ to invest in improving the quality of its signal. Specifically, we assume that it can choose the level of $\sigma_1^2$, paying a cost $C(\sigma^2)$, with $C'(\cdot) < 0$ and $C''(\cdot) > 0$: a higher-quality signal (lower $\sigma^2$) is more expensive to achieve, and convexly so.

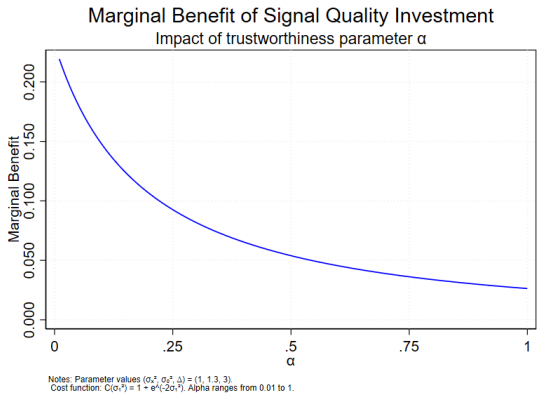
Assuming for simplicity that the outlet is a price-taker, we can ignore the price (setting it equal to 1) and have the outlet choose the level of $\sigma_1^2$ to maximize demand net of costs. Using (4), this yields the following FOC:

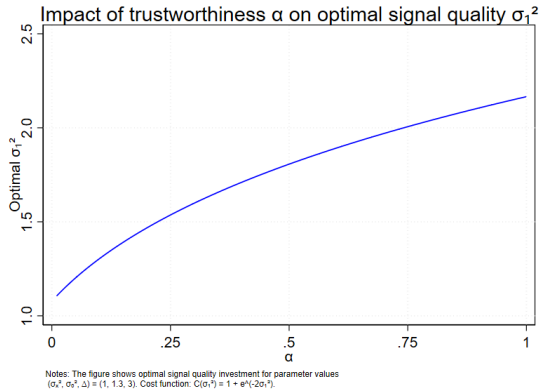
$$-C'(\sigma_1^2) = \frac{(\sigma_X^2)^2}{(\sigma_X^2 + \sigma_1^2 + \alpha\Delta)^2}. \tag{D.1}$$

It is easy to see that, as long as $C(\cdot)$ is sufficiently convex, we have $\frac{\partial \sigma_1^2}{\partial \alpha} > 0$: higher $\alpha$ decreases the marginal benefit of an improvement in the quality of the signal, on the right-hand side, which requires the marginal cost on the left-hand side to decrease; this in turn requires a higher $\sigma_1^2$ because of the convexity of $C(\cdot)$.[41] This can be seen in the following figures, depicting the FOC in (D.1) and the resulting optimal choice of $\sigma_1^2$, as functions of $\alpha$.

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

Figure D.1




In words: more trustworthy outlets (lower $\alpha$) will choose to invest more in the quality of their signal (lower $\sigma_1^2$). The intuition is simple: the benefit of an improved signal, in terms of higher demand, gets less diluted for the more trustworthy outlet, as it can mitigate the impact of increased disinformation. This induces additional investment in signal quality.