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

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

July 2025

Abstract

We study how AI-generated misinformation affects demand for trustworthy news, using data from a field experiment by a major German outlet, Süddeutsche Zeitung (SZ). Readers were randomly assigned to a treatment highlighting the challenge of distinguishing real from AI-generated images. The treatment raised concern with misinformation (0.3 s.d.) and reduced trust in news (0.1 s.d.), including SZ. Importantly, it affected post-survey browsing behavior: daily visits to SZ digital content rose by 2.5% in the immediate aftermath of the treatment. Moreover, we find that subscriber retention increased by 1.1% after five months, corresponding to about a one-third drop in attrition rate. Results are consistent with a model where the relative value of trustworthy news sources increases with the prevalence of misinformation, which may thus boost engagement with those sources even while lowering trust in news content.

JEL Codes: D12, L82, L86.
Keywords: Generative AI, Misinformation, News Media, Trust.

---

*We are indebted to many people for helpful feedback, including seminar participants at the Economics of AI Colloque (Collège de France), NUS, University of Rome Tor Vergata, Marketing Science Conference, VQMS. Special thanks to Nicolás Ajzenmán, Manuela Collis, Chuck Eesley, Henry Farrell, Thomas Fujiwara, Sid George, Lorenz Goette, Mor Namaan, Brendan Nyhan, Jesse Shapiro, and Noam Yuchtman for detailed comments and suggestions, and to Alessandro Izzo, David Schwartsman, Ren Jie Teh, and Nicolás Valle for outstanding research 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: ruben-durante@gmail.com

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

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

1 Introduction

There has been widespread concern over the prevalence of online misinformation and its implications for politics, business, and society (Allcott and Gentzkow, 2017; Jerit and Zhao, 2020; Ahmad et al., 2024).[1] In particular, this has been seen as fostering a new environment of markedly diminished trust in news content (Brenan, 2024; Newman et al., 2024), with potentially significant implications for political polarization and democracy (Lazer et al., 2018; Tucker et al., 2018; Acemoglu et al., 2024), as well as for the economic viability of the news industry.

This issue has gained further urgency with recent developments in artificial intelligence (AI). Generative AI (GenAI) has enabled the creation of visual and audio content ("deepfakes") that is virtually indistinguishable from authentic material, and has thus heightened concerns over the trust-eroding potential of false or misleading information.[2] From the standpoint of the news media industry, this could represent an existential challenge for outlets whose business model relies on producing costly high-quality news: a collapse in trust could lead to reduced demand, eventually to the point that high-quality news becomes economically unsustainable.[3]

Yet there is an important counterpoint: when trust becomes relatively scarce, trustworthiness becomes relatively more valuable. As such, a potential business

[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); DWAkademie (2024); Toff and Simon (2025).

[3] This comes about as the media industry has struggled financially since the rise of the internet in the 2000s (Djourelova et al., 2024; Beattie et al., 2021), followed 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).

opportunity emerges for an outlet with a reputation for trustworthiness within a segment of consumers, as the latter seek help to distinguish between real and synthetic content. This could, in turn, have important implications for the viability of online news from a business perspective and, more broadly, for how misinformation shapes trust in the media ecosystem in democratic societies.

This paper aims to study this interplay between AI-powered misinformation, trust, and the media ecosystem, taking advantage of a unique field experiment conducted by a major newspaper in Germany, Süddeutsche Zeitung (henceforth SZ). As part of their online marketing operations, SZ conducts regular surveys of online subscribers, digital app users, and website visitors. We leverage this to study a sample for which we can presume (and then verify) that SZ is seen as a highly trustworthy news source, providing a viable testing ground for our set of hypotheses.

Specifically, in one of those surveys, SZ implemented a randomly assigned treatment with the key aim of making salient the ability of AI to generate seemingly authentic content for readers, and the resulting challenge of distinguishing between real and AI-generated content. Readers assigned to the treatment group were shown three pairs of images, each with one real image and one created using AI tools, and were asked to evaluate whether either of them was, in fact, AI-generated. Those in the control group, in contrast, were shown pairs of real images related to the same set of current affairs, and asked questions unrelated to AI. Both groups (a total of more than 17,000 individuals) were then asked to evaluate how severe the problem of misinformation is and to report their level of trust in the content available on SZ and on a number of other media outlets and platforms, as well as a monthly amount they

would be willing to pay for an SZ digital subscription.[4] Most notably, for a subset of about 6,000 participants who agreed to have their data tracked and linked to the experiment, we are able to also track their actual behavior over several weeks after the experiment, namely the number of daily visits to SZ digital content (website or online app) and their subscription status.[5]

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 a technology that degrades the quality of the information environment across the board, we show that such a technology can nevertheless increase demand for a source that is deemed sufficiently trustworthy, in the sense of being able to mitigate that degradation. Intuitively, consumer demand is driven by the *relative* value of the source, which in that case increases with the prevalence of misinformation, even though the perception of the quality of the source itself decreases.

With this framework in hand, we turn to the key experimental results. First, regarding survey-based outcomes, we find that the AI exposure treatment substantially increased respondents' reported concern with online misinformation by about one-third of a standard deviation.[6] The AI intervention had a negative impact on trust in the content of all media platforms and outlets, including SZ itself. We also

[4]Unsurprisingly, SZ readers in the experiment have very high levels of trust in SZ content, rating it as 2.9/3 on average in the control group. 92% of that group report the maximum trust level (on a 0-3 scale), and a mere 0.40% report trust levels lower than 2.

[5]We elicited their consent to tracking prior to them starting the quiz. Moreover, there is balance in observables across the treatment and control groups.

[6]The AI quiz was successful in creating confusion regarding the images: the average respondent got 0.9 correct answers, with only 2% answering all three correctly and 36% getting all three wrong.

find that, in spite of the reduced trust, there is no decrease in individual reported willingness to pay (WTP) for an SZ subscription.

As 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. We document an increase of about 2.5% over the average number of visits for the first 3-5 days, with the effect declining over time, but still significant for a window of over two weeks. Furthermore, we observe a sustained increase in subscriber retention, which, after five months, was 1.1% higher in the treatment group (with the caveat of sparsity in the outcome variable). This reflects a consistently lower attrition rate over the period, by approximately 1/3, providing suggestive evidence that, although the impact of the treatment may diminish in day-to-day decisions, such as online visits, the treatment remains influential as individuals consider the value of continuing their commitment as subscribers.

In sum, the results confirm the key predictions of our framework: inducing increased concern with misinformation, via the AI treatment, increases consumption of the trustworthy source, even while reducing overall trust in news content (including for that source).

Heterogeneity patterns shed further light on the dynamics at play in the data. We first consider the (pre-registered) dimension of quiz performance, focusing on the respondents' own assessment of how hard they had found the quiz. For individuals who found the quiz relatively hard, there was a substantially larger increase in misinformation concern (0.4 s.d. vs 0.15 s.d.) and a strong increase in daily visits (over 4% in the first 3-5 days). Moreover, those individuals do report an increase

in WTP (3.5%), and no significant decrease in trust in SZ content relative to the control group, even as trust in other sources judged as untrustworthy decreased.

Additional heterogeneity dimensions, available for the tracked subsample, also let us gauge the possible mechanisms at play, through the lens of our framework. In particular, we find that it is individuals who were relatively less interested in political news (as measured by their reading patterns in the three weeks before the experiment) who respond more strongly to the treatment in terms of their subsequent engagement. This is consistent with the predictions of our model, in which individuals with less knowledge about the topic at hand would have a stronger demand response. In addition, we find suggestive evidence that the engagement response is stronger for the group of individuals who were heavier (i.e., above average) users of the website, again consistent with our framework, if we take those individuals as having a higher pre-existing assessment of the source's trustworthiness.

The underlying dynamics suggested by our results have potentially broader implications. For the news industry, they suggest a possible business strategy in response to the challenge posed by the advent of AI-generated content. From a broader societal perspective, our findings provide a nuanced counterpoint to the concern over AI (and misinformation more broadly) leading to a downward spiral in trust in the information environment: increased scarcity creates increased potential rewards for trustworthiness. Moreover, it is plausible that as AI-powered tools become more proficient over time, news consumers may find it even more challenging to distinguish between real and synthetic content (Kamali et al., 2024). In that case, the effects we document might provide a lower bound on trust and demand effects, but as our

conceptual framework highlights, only as long as an outlet's ability to help readers keeps up with the pace of the increase in the challenge they face.

Even more broadly, our model highlights that the forces at play could apply to the role of expertise in other contexts, in the face of evolving technologies. For instance, large language models (LLMs) can at this point produce convincingly expert-sounding content that is hard for a consumer to evaluate outside of their own domains of expertise. The logic of our model and our empirical results suggest that, instead of replacing experts, this could increase demand for trustworthy sources that could help adjudicate the value of the AI-generated content.

Our paper contributes to three strands of the literature. The first one focuses on the prevalence (Allcott and Gentzkow, 2017) and spread (Vosoughi et al., 2018; Hengel et al., 2025) of misinformation online. Several studies have looked at the role of fact-checking misinformation on voter beliefs, the demand for news, and subsequent actions (Henry et al., 2022; Guriev et al., 2023; Nyhan et al., 2020; Chopra et al., 2022).[7] In a similar vein, related to our paper, Assenza et al. (2024) and Harris et al. (2024) study experimentally the impact of raising awareness of individuals' inability to detect misinformation and their perception of their own ability, respectively. They find a positive impact on their willingness to pay for protection against it, but no impact of perceived ability on misinformation concerns or consumption of different (politically-aligned or mainstream) sources. Besides our emphasis on the role of AI-enabled misinformation, we depart from their focus on a hypothetical misinformation insurance, and on individuals' stated preferences (such as media con-

[7]See Lazer et al. (2018) and Zhuravskaya et al. (2020) for a review on these topics.

sumption), by studying a field experiment where we observe both survey outcomes and actual digital reading and subscription behavior. In addition, our conceptual framework helps us make sense of the patterns in the data, illustrating how and why demand for a trustworthy outlet can increase even while trust in content declines in general, including the trustworthy outlet itself.

Next, our work is related to the role of technology, in particular AI, in the creation and diffusion of online news content, including misinformation. Longoni et al. (2022) and Toff and Simon (2025) use online experiments to demonstrate that the use of AI to generate news articles leads to lower trust in news content (relative to articles written by humans). Groh et al. (2024) show how GenAI can be used to increase the quantity and quality of deepfakes, making synthetic content indistinguishable from real images, audio, and video. Consequently, Vaccari and Chadwick (2020) highlight the potential of deepfakes to reduce trust in news content circulating on social media in general. Relative to these studies, we utilize our setting in the field to elicit survey outcomes and revealed preferences without priming them about misinformation. Most importantly, backed by our theoretical framework, we are able to show that exposure to AI-generated misinformation can reduce an individual's trust even in their most trustworthy outlet, but with the crucial nuance that the latter can nevertheless benefit from increased demand as a result of that exposure.

Finally, this paper is also linked to a strand of the literature that proposes several strategies to combat low-quality information within the broader platform ecosystem. Costello et al. (2024) demonstrate that LLMs can be utilized to durably correct conspiracy theories online. Ahmad et al. (2024) highlight information interventions that

will limit companies from inadvertently funding misinformation websites. Allen et al. (2021) analyze different ways of scaling up fact-checking on social media, showing that the wisdom of the crowd is similar in quality to that of experts in this regard. Building upon these studies, we demonstrate the role of high-quality, trusted news outlets in providing authentic information, especially as GenAI becomes a more prevalent tool in generating misinformation.[8]

The remainder of the paper is organized as follows: Section 2 presents the background information on SZ and the German context. Section 3 lays out the conceptual framework to study the interplay between misinformation and trust in affecting news consumption. Section 4 describes the field experiment. Section 5 presents the results. Section 6 concludes, with a discussion of the results 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 a daily paid circulation upward of 260,000 copies as of 2024, as well as 295,000 online

[8]More broadly, our paper is also related to studies that aim to increase awareness about misinformation through educational interventions (Badrinathan, 2021), limit the impact of fake reviews online (Ananthakrishnan et al., 2020), and that of deceptive claims made by companies on social media (Fong et al., 2024). In this paper, we focus on AI-powered misinformation and how making readers aware of the technological capabilities through a simple exercise often utilized by news outlets, without explicit reference to misinformation, can have significant implications for AI-misinformation and the media ecosystem.

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 is 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 across all industry sectors advertise on SZ's website. The SZ readership is mainly in Germany (over 75%), and the largest proportion of readers are between 40-60 years old.

2.2 The German News Industry

The German news industry has a diverse ecosystem of print and digital publications reflecting a wide spectrum of political perspectives. Its landscape mirrors other Western democracies regarding the quality of journalism 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.

Much like 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 within the industry, along with regulatory challenges (Calzada and Gil, 2020).

The rise of online misinformation and the potential for AI tools to exacerbate this problem has further complicated the German news ecosystem. Indeed, about 42% of German newsreaders have highlighted the inability to tell apart real from fake content online as a major problem (Newman et al., 2024). While AI tools are being used for the creation of news content, there is a tension because of the misuse of such tools to create synthetic content which is virtually indistinguishable from real content online. These challenges are not unique to Germany; around the world, news organizations and governments are grappling with the need to balance technological innovation with the preservation of information integrity (Ahmad et al., 2024).

3 A Simple Framework

Before getting into our experimental setting, it is worth sketching 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. This will help guide our empirical exploration.

3.1 Basic Environment

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

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

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

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

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

Now let us benchmark all sources by setting σ_0^2 to represent the quality of in-

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

Result 1. *The introduction of symmetric misinformation (i.e. leading to the same decrease in the quality of all available signals) leads to a reduction in demand for*

high-quality sources. In the limit, as misinformation grows, the consumer's demand for the high-quality source approaches zero.

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

3.3 The Role of Trustworthiness

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

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

"trustworthiness" as these separate concepts; they are obviously related as trust is an increasing function of trustworthiness.

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

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

It follows that d_1 is given by:

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

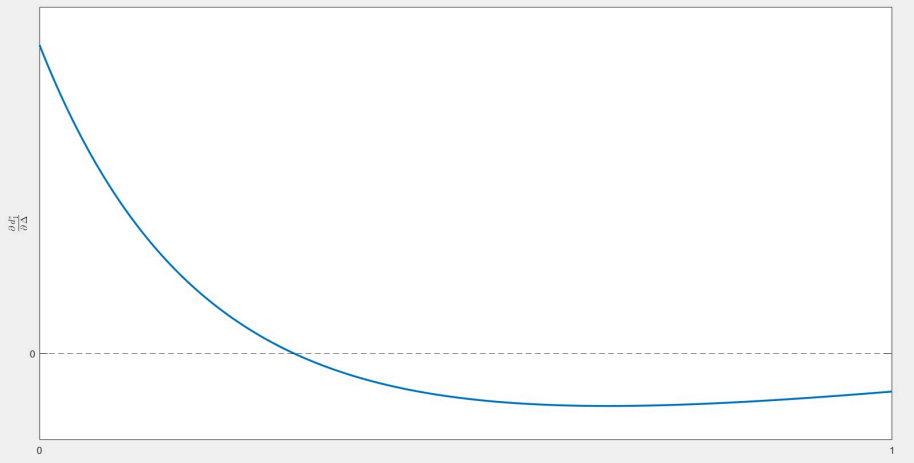
Differentiating that with respect to Δ yields:

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

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

Result 2. *A marginal increase in perceived misinformation leads to an increase in demand for a sufficiently trustworthy high-quality source (i.e. sufficiently effective in mitigating the harm of misinformation).*

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



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

The intuition is clear: in an environment with widespread misinformation, a sufficiently trustworthy news source – in the sense that the consumer expects it to do a sufficiently good job in mitigating the challenge posed by that misinformation – becomes relatively more valuable in response to an increase in the level of misinformation. This happens even though the increase in misinformation 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.

Note that our framework clarifies the nature of the challenge faced by high-quality news sources over time: if Δ keeps growing without bound, eventually any given value of α will be insufficient to sustain demand, and d_1 will fall. Yet equation (5) makes it clear that, if α were to go down at least as fast as Δ increases, d_1 could still be sustained. In other words: as the prevalence and quality of misinformation keeps

increasing, trustworthy sources have to make sure that their ability to mitigate it, as perceived by their readers, grows at least as fast.

We can also derive additional interesting comparative statics. For instance, it is straightforward to show that $\frac{\partial^2 d_1}{\partial \Delta \partial \sigma_X^2} > 0$ for sufficiently small α.[9] In short, the increase in demand for a sufficiently trustworthy source, as a result of a marginal increase in misinformation, will be greater when the individual has more difficulty interpreting the underlying outcome. In addition, as illustrated by Figure 1, at low levels of α we have $\frac{\partial^2 d_1}{\partial \Delta \partial \alpha} < 0$: intuitively, demand for the outlet will respond more strongly to the marginal increase in misinformation for individuals who see the outlet as more trustworthy.

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 it's demand go up: its relative value increases as a result of the overall degradation in the quality of the environment.

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

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 the capabilities of GenAI as a technology in its ability to create synthetic images, thereby drawing attention to the potential for AI to create 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 visual content from a real one using a picture quiz on the SZ website, and evaluating the impact of that awareness on readers' attitudes and engagement with SZ.[10]

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 where randomly shown a pop-up that encouraged them to take a picture quiz, which could be accessed by clicking the link presented to them.[11] For the latter, individuals on the SZ subscriber mailing list received an email on March 1, with a link to the quiz.[12] The language of both the email and the pop-up ad was

[10] In our analysis, we examine whether the AI quiz experience operates through an informative or persuasive mechanism. This helps contextualize the results relative to studies about warning labels (Williams-Ceci et al., 2024) or inoculation interventions (Roozenbeek et al., 2022) aimed to tackle online misinformation.

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

[12] 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.

neutral, and simply asked them 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."[13] 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. The infrastructure used for fielding the survey and subsequent randomization is used routinely by SZ. To mimic the user experience of other surveys fielded by SZ, there was no monetary compensation for taking the quiz and the responses were not incentivized.

As an individual accesses the quiz, they were asked for their age and gender prior to starting the quiz. Due to stringent privacy regulations in the European Union (eg. the General Data Protection Regulation, GDPR), we then ask for consent which allows 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, 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 would not mimic their routine survey or quiz-taking behavior. 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.[14] We then elicit their willingness to pay for an SZ subscription with a sliding scale of 0 to 45 Euros. Finally, we elicit

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

[14]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").

their trust in the content available on various media (SZ, Bild, ARD/ZDF) and social media outlets (Instagram, LinkedIn, TikTok, and X/Twitter).[15] 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 participants are asked which of the two images they think are 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.[16] 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 provide 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.

[15] 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).

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

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 create a counterfactual that mimics their daily news consumption but does not highlight any 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 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.[17] 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.

[17] 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.

Design Choices: The aim of keeping pictures similar across treatment and control is to hold constant across groups the role of 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. Such an approach is akin to an active control experimental design (Haaland et al., 2023) to ensure that the act of taking a quiz or being exposed to certain images does not lead in itself to the effects we capture. As an additional robustness exercise, we draw a random sample of subscribers who received the email invitation to take the quiz but chose not to, as an alternative control group when measuring the impact of the AI treatment on online behavior.

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 being 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. Such an online setting, without the physical presence of a researcher, minimizes the potential for experimenter demand effects, as evidenced in the literature (De Quidt et al., 2018).[18]

Quiz Performance: In terms of actual and perceived quiz performance, the AI quiz proved to be hard, in line with the goal of the intervention. The average respondent got under 0.9 correct answers and a mere 2% got the three answers right, while

[18] 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).

a full 36% got zero.[19]. This was reflected in their assessment of perceived difficulty, which was 4.7 out of 6. In contrast, the current events control quiz was less hard, with an average correct rate of 2.7 and 72% getting the three answers right. The perceived difficulty was in line with performance at 1.52. While the quiz difficulty differed, the amount of time taken by the two groups was not too different, highlighting the significant effort put in by both groups. The median individual took 4:06 minutes to complete the quiz. The numbers were 4:20 for the treatment group and 3:54 for the control group. 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 harder 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 will focus on two connected samples of about 17,000 and 6,000 SZ readers. For the first sample, which we call the "survey sample", we gathered data through the survey that hosted the quiz. This implies that we have data on individuals who took the quiz and answered at least one endline survey ques-

[19]The performance for each individual question was as follows: 20% correct for the climate pair, 36% for the politicians pair, and 30% for the protest pair.

tion during the two-week window.[20] For these individuals, we have information on self-reported demographics (age and gender) along with endline stated preference questions. 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, with these groups balanced across treatment and control. Treatment and control are also balanced in terms of age of quiz takers and when they took the quiz (the bulk of individuals taking 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, at 34%. These checks provide validity to our random assignment.

The second sample is focused on around 6,000 SZ subscribers who consented to tracking (which we call 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 have information on past visits to the SZ website, whether they are interested in politics (based on pre-treatment readership data), and their location.

[20] 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).

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	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 can be seen in Panel B of Table 1, we see a balance in observables across both the survey-elicited characteristics and those based on actual behavior. Importantly, we see balance on the number of daily visits in the three-week window prior to the experimental intervention.

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

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

where β captures the causal effect of the AI-awareness quiz on the outcome of interest.

Because of the successful randomization, we do not expect any controls (Z_i) to affect the estimate of β. Therefore, we estimate the regression without control variables as the baseline results.

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

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

where $PostTreatment_{it}$ is the event variable which takes a value of 1 if date t is following the date individual i took the quiz and 0 otherwise. $Treated_i$ indicates whether the individual i took the AI-focused quiz. γ_i and τ_t capture fixed effects for individual and date, respectively. The coefficient of interest β captures the average impact of an individual taking the AI quiz relative to the current affairs quiz.[21] We use a three-week pre-period before the quiz and varied the post-period using different time windows.

[21] Note that, since we have a binary treatment and our tracked sample took the quiz within a very short time period (see footnote 12), our specification boils down to the canonical differences-in-differences case, where the TWFE estimator is robust (de Chaisemartin and D'Haultfoeuille, 2025).

5 Results

5.1 Survey Outcomes

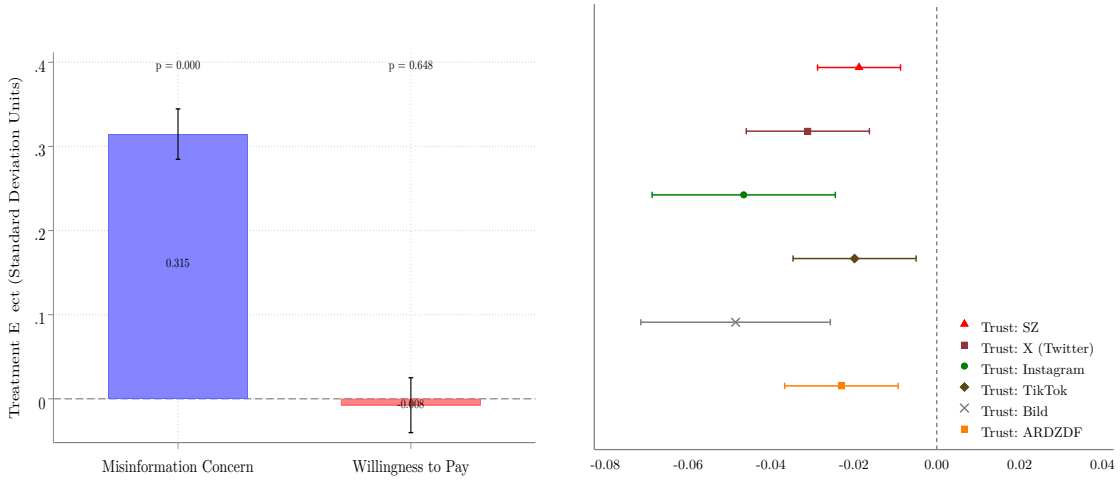
5.1.1 Baseline

We start with the average treatment effect (ATE) on the outcomes for the survey sample. The first outcome we analyze is concern with misinformation (measured on a scale of 0 to 6). In panel (a) of Figure 2, we see that it is significantly higher in the treatment group relative to the control. The magnitude of the increase is also economically meaningful (0.31 of a standard deviation), especially since there was no explicit mention of misinformation prior to this question. This result can also be viewed as a manipulation check: exposure to the AI quiz indeed induced greater concern over the issue of misinformation, even in a population already predisposed to high levels of concern with the topic – as are, unsurprisingly, readers of an outlet such as SZ.[22]

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 when it comes to their content. Focusing on the control group, we can see that, unsurprisingly, our sample of SZ readers and subscribers holds SZ in high trust: an average score of 2.91 on our 0-3 scale. This confirms our prior that our experimental sample has an identifiable trustworthy news source. They also report high trust in the public broadcaster ARD/ZDF (2.80), but have less trust in

[22] 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.

LinkedIn content (1.32), even less so for Bild (0.82) and Instagram (0.76), and report the lowest levels of trust in content on TikTok (0.26) and X/Twitter (0.25).

In line with Result 1 of the model, we observe in Figure 2 (b) a negative impact of the AI Awareness treatment on trust in the content of 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. Interestingly, we find a null ATE of the AI intervention on the willingness to pay (WTP) for a subscription to SZ (on a scale of 0 to 45 Euros) (Figure 2 (a)): in spite of the decline in trust, individuals do

not report a lower WTP.[23]

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 an online news source they hold in high trust – in this case, SZ itself. We can do that with 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 mentioned in Section 4.2, since we have pre-period data for this sample, our preferred approach uses a DiD strategy, which improves precision. In addition, despite attrition being relatively small in the survey, as discussed above, we use an intent-to-treat specification with individuals assigned to the treatment and control groups, as long as they started taking the quiz embedded in the survey link.

5.2.1 Browsing Outcomes: Baseline

We start with a key outcome variable of interest: the number of daily visits to the SZ website. This is one of the key metrics tracked by SZ, as it aims to increase overall engagement with its content to generate additional advertising revenue and subscriptions. Online visits are a key indicator of the financial health of news outlets

[23] The underlying estimates to Figure 2 can be seen in Table A.1, in the Appendix. Table A.2 confirms that all the survey outcome findings are qualitatively similar when we focus on the tracked sample only. Moreover, the treatment effects are robust to controlling for observables as seen in columns (1)-(3) of Table A.3.

and are widely used in academic research (for example Cagé et al. (2020); Peukert et al. (2024)).

For an initial, model-free look at the evidence, we consider the t-test for the difference in mean daily visits across the two groups, within a 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 statistically significantly different at the 5% level (p-value 0.032). The balance checks in Table 1, along with statistically and economically significant (2.77%) estimates, suggest a meaningful impact of the intervention on engagement.

We then move on to the main DiD analysis, deploying the data available before and after the survey, and allowing 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 before that date as the pre-treatment period, and consider a range of alternative time windows after the treatment.

Table 2 shows the baseline results for the DiD specification. In column (1), we see that for the AI treatment group in the three days after taking the quiz, the number of visits increased by about 2.5% relative to the control group. This effect is detectable for about two weeks post-treatment, declining in magnitude as we increase the post-treatment window to five (column 2), 10 (column 3), and 14 days (column 4). This effect size is nontrivial, especially given the relative lightness of the intervention, and contrasted with the fact that about 80% of the experiments in digital markets lead to null results (Kohavi et al., 2013). Moreover, the magnitude is comparable to what has been found in response to an explicit advertising approach designed to combat

misinformation (Lin et al., 2024).[24]

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 the rate at which our 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

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

sizable at 0.062, although marginally statistically insignificant at conventional levels with a p-value of 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.[25]

Overall, the AI Awareness treatment led the tracked SZ subscribers to meaningfully increase their engagement with the newspaper in the days after exposure. This revealed preference result, along with the prior result on the increased concern with misinformation, further validates Result 2 of the model. In particular, by making AI-generated synthetic content more salient, there was an increased concern for misinformation, which created additional demand for the trustworthy news outlet (SZ). This increased engagement occurred in spite of the reduction in trust we documented in our survey results, again in line with Result 2.

5.2.2 Browsing Outcomes: Robustness

To ensure the robustness of our key results, we carry out a few checks. First, for website engagement as the outcome variable, we can construct an alternative control group. In the baseline results, the control group has individuals who took the control picture quiz, for whom we have both survey and browsing outcomes. For the alternative control group, we draw a random sample of 3,000 individuals who received the email to take the quiz but chose not to do so, and hence were not part

[25] While we do not observe browsing activity off-platform, we do observe incoming traffic to the SZ platform from social media websites. This data enables us to break down overall traffic into two categories: traffic originating from social media and organic visits (those not coming from social media links). With the caveat of traffic from social media links being sparse, we find that the overall effect in column (1) of Table A.5 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. This provides some suggestive evidence of the increase in engagement with the focal website, but not other sources.

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. 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.6, there is a statistically and economically significant effect in the first five days after the experiment. The effect size is almost 50% higher than the baseline (using a five-day post-period window) results at about 2.8%. Such a check thus establishes the robustness of our treatment on our key revealed preference measure of interest.

Second, as an additional check, Appendix Table A.6 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 then analyze the total daily time spent by a user on SZ online, though this variable as captured even by advanced analytics tools comes with significant caveats.[26] 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 also find that our estimates are robust when we use an alternative pre-period window of 14 days instead of our baseline 21 days. Next, when we use a more saturated specification with time-since-quiz fixed effects in addition to date fixed

[26] 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.

effects, we find similar results (column 7), which is unsurprising given 87% of the tracked sample were treated within two days. Finally, we also undertake a check to provide evidence against short-term attention effects. 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 see whether the additional engagement was simply a result of an engaging experience with the AI quiz. In columns (8) and (9) we find that there is no differential engagement with surveys on SZ across treatment and control suggesting that our results are not a result of individuals simply wanting to engage with other surveys and quizzes online after our intervention.[27]

Overall, these results demonstrate the robustness of our baseline results along various dimensions.

5.2.3 Subscription Outcomes

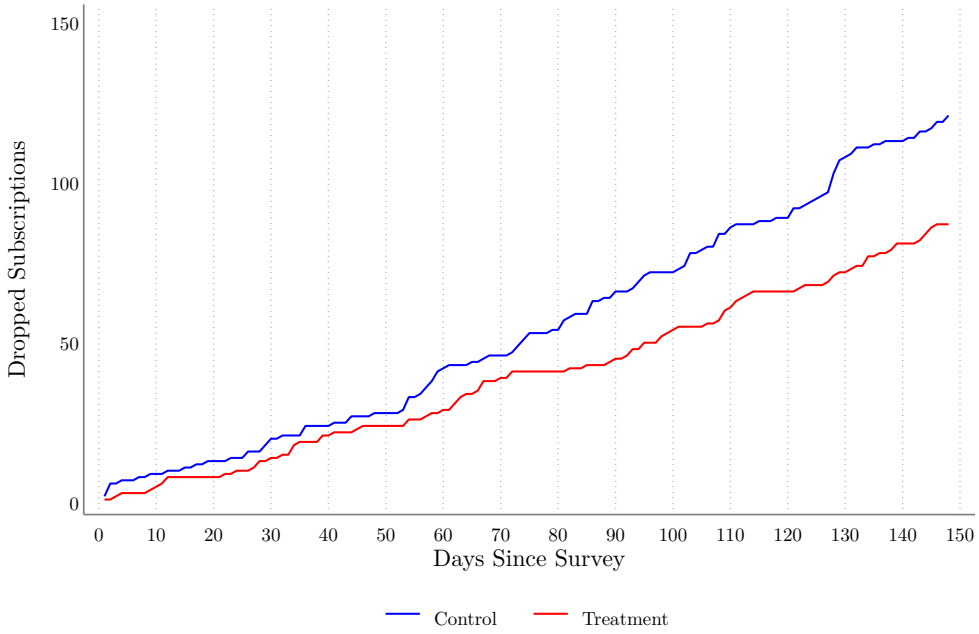
Continuing within the context of our revealed preference outcome analysis, we turn our attention to individual subscription behavior: for the tracked sample, we can analyze their decision to drop their subscription, versus allowing it to continue.[28] Note that there is no variation in subscription status before the treatment, since all tracked individuals are subscribers, and that the 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. This

[27] Our baseline results are robust to dropping those individuals in our sample who engaged with surveys and quizzes on the SZ website and app in the post-treatment period.

[28] 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 monthly or yearly payments.

means that the numbers in question are relatively sparse, even over several months. For this reason, we will focus on a descriptive analysis of the evolution of subscription retention.

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: there is an immediate gap, with a higher number of dropped subscriptions in the control group; that gap persists and in fact grows wider 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.

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 magnitudes are instructive: over time, the probability of remaining a subscriber naturally goes down, and the treatment effect goes up accordingly. This reflects a stable effect in terms of the attrition rate, with the treatment group consistently displaying a rate around 1/3 lower than that of 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.

This suggests an interesting pattern in terms of the revealed preference outcomes: the impact of the treatment may recede in choices such as whether to visit the platform on a given day, yet it seems to retain its strength at the point where individuals face the decision to discontinue their subscription. As they think this through, exposure to the AI treatment may provide an additional nudge to keep

paying for SZ access.[29]

5.3 Heterogeneity

We now turn to heterogeneity patterns, which allow us to further unpack the ATEs we documented in the previous subsection.

5.3.1 Quiz Performance

We start with a (pre-registered) heterogeneity dimension: quiz performance. 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 $Hard$ that equals one if the individual reported finding the quiz harder than the median individual in their group (treatment or control).[30] In column (1) of Table 4, we see that those who found the quiz relatively hard display a treatment effect that is significantly larger than that for those who found it relatively less hard, when it comes to concern with misinformation, indicating that the

[29] 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.

[30] 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%.

treatment left a stronger impression on individuals who found it more challenging.[31]

This was matched by heterogeneity in the other survey outcomes. Interestingly, that group of individuals did not reduce their trust in SZ: there is a precisely estimated zero effect (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, 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 actually increased their reported WTP for a subscription to SZ, by about 3.9%, with the effect being statistically 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.

[31] 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.8.

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.

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.

We should take caution in interpreting these estimates as causal, given that quiz performance assessment was not randomly assigned – in particular, the subgroup who found the quiz (relatively) hard in the control group could potentially be systematically different from those in the treatment group. To provide evidence that the groups might be similar, in Table A.7, 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.

Overall, the suggestive pattern is consistent with the idea that awareness of the challenge of discerning AI-generated from real content induces increased attachment

to trustworthy news sources, while reducing trust in the information environment more broadly.

5.3.2 Politics Affinity

Next, we look at heterogeneity from pre-experiment behavior on SZ digital platforms, which is available for the tracked sample. These allow us to pursue further inquiries into the mechanisms that could be driving our main results.

We start with data on reading affinity dummy variables that capture whether the reader engaged with a given set of topics in the three weeks before the experiment, to dampen any noise arising in short-term reading patterns.[32] The quiz embedded in the survey was focused on political issues (climate, political leaders, and protests); for that reason, we are particularly interested in identifying readers with a specific affinity for political topics. We thus split the sample according to whether the individuals are classified as having a politics affinity (54% of the sample) or not (46%).

The politics affinity variable is helpful in two ways. First, through the lens of our conceptual framework, it can be interpreted as capturing the strength of the priors regarding political information: we can think of individuals without political affinity as having a higher σ_X^2, i.e. a more diffuse prior. Second, a broader question in terms of mechanism is whether the effects of our AI intervention work through information or persuasion. Similar to advertising effects, an intervention could provide novel information to previously uninformed individuals, or alternatively, it could provide a

[32] 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/).

nudge or reminder to individuals who were already knowledgeable about the topics mentioned (Nelson, 1974). Political affinity works as a marker of prior knowledge, and thus helps tease out an informative versus 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 those 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 versus light users of SZ, which we classify by splitting the sample based on the median number of visits in the two-week period prior to the experiment (44 visits).[33]

When it comes to visits to SZ digital platforms, columns 5-6 of Table 5 display a suggestive heterogeneity pattern, indicating 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, while 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.[34]

We also look at additional heterogeneity (not pre-registered) along age and gender, in Table A.9 in the Appendix. We find limited heterogeneity by age with no statistically significant differences across the three main survey outcomes: misinformation concern, WTP, and trust in SZ. For gender, we find that female readers in

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

[34] Specifically, individuals who attach a smaller α (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 Δ. It is natural to expect that they would report a smaller increase in the concern with misinformation in general, as it is arguably affected by their perception of the outlet itself.

the treatment group show a heightened concern for misinformation but lower than that of male readers. We see a higher level of trust in SZ from female readers relative to male readers, but it is slightly noisy and not statistically significant. We also do not find any difference across genders of the treatment on WTP for SZ.

6 Discussion and Conclusion

Our results indicate that increased awareness of the challenge posed by having to distinguish between real and AI-generated content has important and nuanced effects on trust in the news media and on news consumption. We find a negative average effect on trust in news content, including that of a source they find trustworthy, but also uncover an important counterpoint in the behavior of news consumers: it also leads them to increase their engagement with the trustworthy source, as evidenced by visits to the website, and also to an increased willingness to pay for its content, as evidenced by the decreased subscription attrition.

The magnitudes we document are also nontrivial: a 2.5% increase in the number of daily visits, even if it dissipates over time, or a decrease of approximately 30% in subscription attrition (albeit on a small baseline attrition rate), which persists for over five months, are substantial. The subtle nature of our intervention heightens the significance of these results: our quiz is not a sustained campaign, and it contains no explicit mention of misinformation in the intervention itself.

These findings confirm the intuition outlined in our conceptual framework. The deterioration in the information environment engendered by the emergence of a tech-

nology like GenAI leads to reduced trust in the information environment as a whole. However, an outlet perceived as sufficiently trustworthy may nevertheless witness increased demand as a result, even as trust in its content also suffers, as 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 the heterogeneity patterns we uncover. First, we see that the effects are driven by those individuals who have (or are willing to admit to) a stronger perception of the task's difficulty. While task difficulty was not randomly assigned, the descriptive fact that these individuals do not report the same decrease in trust in the trustworthy source as others, and report an increased willingness to pay for it, is consistent with the causal pathway we hypothesize.

Second, the heterogeneity patterns align with the predictions of our framework. The fact that the increase in engagement is also driven by individuals who were relatively less interested in political news, and that those individuals do respond more strongly in the survey, is consistent with individuals with less prior knowledge (higher σ_X^2) being more sensitive to increased misinformation in their demand. It also highlights an interesting possible informational mechanism, indicating that there is additional margin for activation, even within a population of readers who are already engaged with the news source they perceive as trustworthy, to the point of being paying subscribers. Along with the suggestive evidence of a differential response from heavy users of SZ digital content (arguably attaching a smaller α to the outlet, in terms of the model), this underscores the logic highlighted by our model.

When it comes to the broader applicability of our results, it is clear that the SZ readers in our sample are not a random selection; these are individuals who care about the news enough to regularly engage with what they consider high-quality news content. Yet, we see no reason to believe that they would be systematically different from engaged consumers of other similar outlets across the globe – say, The New York Times or The Wall Street Journal in the US, The Financial Times or The Guardian in the UK, or FAZ in Germany, for that matter. That noted, it is unclear how low-trust or less engaged readers would react to the same intervention. Additionally, although we have actual data on engagement behavior and willingness to pay for SZ, we do not observe off-platform behavior on other social media and news websites. As we evaluate the role of GenAI on misinformation within the platform ecosystem, including the intervention in our field experiment, these are fruitful avenues for future research.

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 an outlet or platform can build trust with a segment of the audience, that challenge presents an opportunity to increase revenue: as trust becomes scarcer, trustworthiness becomes more valuable, and readers might be willing to pay more for access to it. Of course, this leaves open for future research the question of how one can build trustworthiness in the first place. From a broader societal perspective, our findings highlight that the availability of AI-powered misinformation need not necessarily 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 to displace real content, to the point that trust in all content collapses, can be counterbalanced by the fact that increased scarcity creates increased potential rewards for trustworthiness.

Yet our conceptual framework also highlights the challenges involved 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 (Δ). Our Result 2 establishes that, for a given level of misinformation, there will be a sufficiently high level of trustworthiness such that an outlet can benefit from an increase in the overall prevalence of misinformation; however, as the latter continues to grow, the threshold for trustworthiness continues to rise. 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 task becomes more challenging. As AI technology continues to evolve at an increasingly blistering speed, the challenge is bound to become even greater, and the dynamics we highlight ever more important.

References

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius, "A Model of Online Misinformation," *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.

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.

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.

DWAkademie, "Generative AI is the ultimate disinformation amplifier," 26 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.

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," *Amplifying Truth (October 29, 2023)*, 2023.

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.

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.

Nelson, Phillip, "Advertising as information," *Journal of political economy*, 1974, *82* (4), 729–754.

Newman, Nic, 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.

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.

Roozenbeek, Jon, Sander Van Der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky, "Psychological inoculation improves resilience against misinformation on social media," *Science advances*, 2022, *8* (34), eabo6254.

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, *(forthcoming)*.

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.

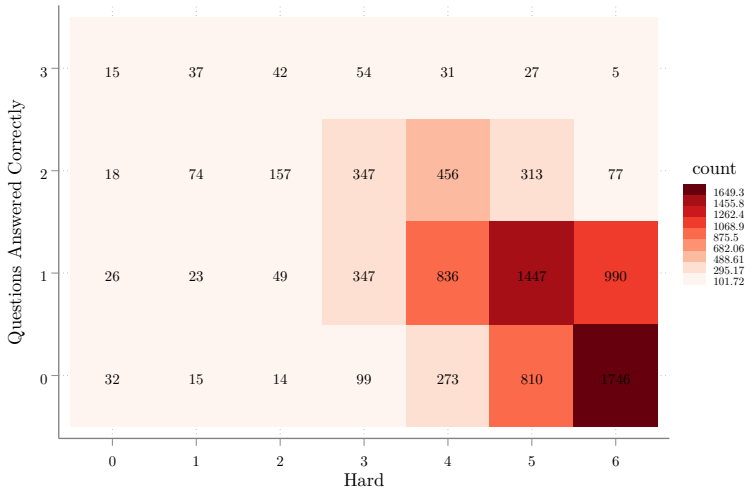
Vraga, Emily K. and Leticia Bode, "Defining Misinformation and Understanding its Bounded Nature: Using Expertise and Evidence for Describing Misinformation," *Political Communication*, 2020, *37* (1), 136–144.

Williams-Ceci, Sterling, Michael W Macy, and Mor Naaman, "Misinformation does not reduce trust in accurate search results, but warning banners may backfire," *Scientific Reports*, 2024, *14* (1), 10977.

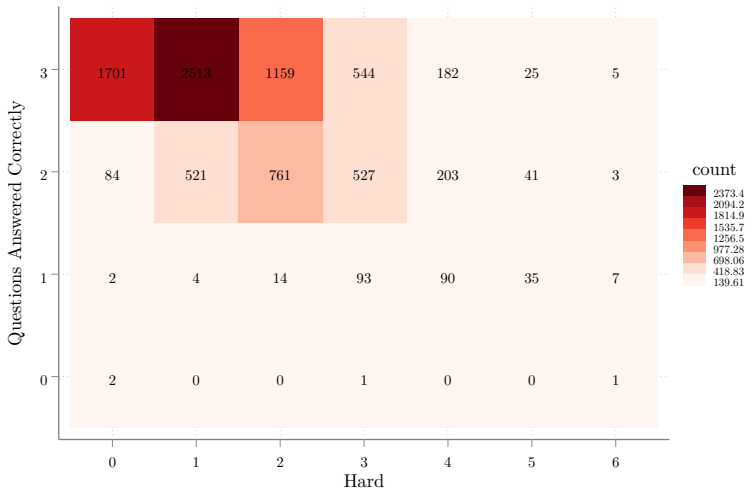
Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov, "Political Effects of the Internet and Social Media," *Annual Review of Economics*, 2020, *12*, 415–438.

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: 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.6: Post-Intervention Engagement: Robustness

Dep. variable	(1) Number of daily visits	(2) Number of daily visits	(3) Number of daily visits	(4) Time Spent (Seconds)	(5) Time Spent (Seconds)	(6) Number of daily visits	(7) Number of daily visits	(8) Other Survey Engagement	(9) 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.7: 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.8: 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.9: 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

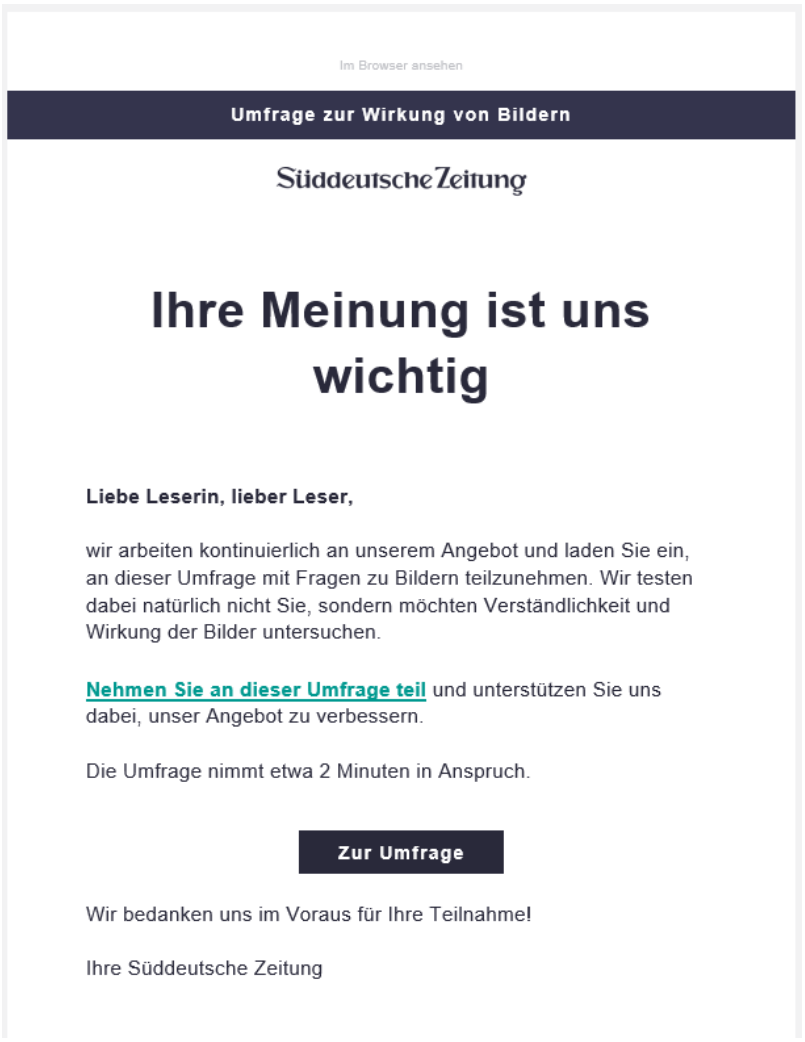


Figure B.2: Quiz Email Invitation (Machine Translated)

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

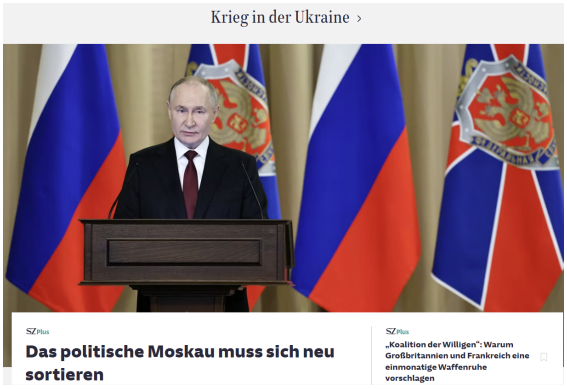


Image and Headline 2nd March 2025