Engagement or Knowledge Retention: Exploring Trade-offs in Promoting Discussion at News Websites

BRIAN MCINNIS, University of California, San Diego, USA
LEAH AJMANI, University of Minnesota, USA
STEVEN P. DOW, University of California, San Diego, USA

How does presenting comments in a news article affect the ways that readers engage with and retain information about news? This paper presents results from a controlled experiment investigating effects related to different strategies for promoting discussion at news websites (N=336 participants). The strategies include highlighting specific comments about a data visualization, providing prompts with the comments, and annotating prompts on the visualization. By comparison to a simple list of comments (baseline), our analysis found that annotations contributed to higher levels of participant engagement in the discussion, yet lower levels of knowledge retention related to the article. These findings raise new considerations about whether and how to integrate discussion content into news and points toward future content moderation systems that assist in representing and eliciting discussion at news websites.

CCS Concepts: • **Human-centered computing → Empirical studies in interaction design**.

Additional Key Words and Phrases: Online discussion, news outlet, engagement, knowledge retention

ACM Reference Format:
Brian McInnis, Leah Ajmani, and Steven P. Dow. 2022. Engagement or Knowledge Retention: Exploring Trade-offs in Promoting Discussion at News Websites. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 303 (November 2022), 38 pages. https://doi.org/10.1145/3555194

1 INTRODUCTION

Many news outlets have tried to foster online discussion around news articles, but often only a few comments add relevant questions, personal stories, and other valuable information to the discussion [21, 36, 59]. Content moderation processes can increase the salience of the valuable content in a discussion system, by reducing the noise caused by off-topic and toxic comments [15, 21] and curating discussion threads to prioritize comments that exhibit valuable qualities (e.g., relevance [20, 58], thoughtfulness [84], or divergent perspectives [49, 56]). People who read and post comments find value in a well moderated discussion; however, most people who read news online avoid the comment sections altogether [23, 39, 80]. Another (and complementary) approach to highlight valuable online discussion is to use visual and interactive features in the news website user interface to orient people toward relevant comments as they engage with an article.

This later approach involves modifying how online discussion is elicited by and represented in the user interface design of a news website [30]. Modifying how online discussion systems elicit participation can affect discussion norms. For instance, requiring people to restate points in another users' comment before posting a reply can promote social norms of active listening [46].

Authors' addresses: Brian McInnis, bmcinnis@ucsd.edu, University of California, San Diego, USA; Leah Ajmani, ajman004@umn.edu, University of Minnesota, USA; Steven P. Dow, spdow@ucsd.edu, University of California, San Diego, USA.

2573-0142/2022/11-ART303 $15.00
https://doi.org/10.1145/3555194

Similarly, modifying how online discussion content is represented can shape how people access existing information and find opportunities to contribute. News website discussions are commonly represented as a chronologically ordered list of comments [59], but there are many other ways to represent discussion, including argument maps [3, 32, 37], pro-con lists [45], thread summaries [62, 96], as well as socially constructed annotations [7, 90, 92, 93]. Grimmelmann [30] refers to such decisions about how to elicit and represent online discussion in the user interface as *Content Organization*, which is "the verb of moderation that most takes advantage of the informational capabilities of computers, [by] remixing authors' contributions to give readers a more satisfying experience" [30, pg. 58].

Increasing the salience of online discussion content in the design of a news website may increase participation in the discussion, but doing so may also introduce risks for the news outlet [21]. Exposure to toxic comments can polarize opinions about news articles [4] and diminish the trust that people have in news reporting [17, 69]. Many news outlets mitigate these risks by designing visual distance between articles and online discussion in the user interface, such as by hiding the comments by default, positioning the discussion interface at the end of an article, and turning the commenting features off entirely [21, 23, 39].

This paper presents results from an online experiment to investigate how various user interface design strategies that increase the salience of reader comments might affect what people remember about a news article and how they engage with the online discussion. By "salience of discussion content" we mean the degree to which design elements separate and integrate content from the discussion into the user experience of reading news articles online. Specifically, we investigated the following research questions:

> **RQ 1**: How does adjusting the salience of discussion content in an online news article affect participant engagement with the online discussion?
>
> **RQ 2**: If discussion content is salient in an online news article, what factors affect the amount of information that people remember about the article?

Inspiration for the design strategies in this study include existing commenting systems at news websites (e.g., *The New York Times* [20, 55]) as well as discussion systems in other contexts, such as deliberation platforms (e.g., *RegulationRoom* [24, 71]) and data visualization systems (e.g., *ManyEyes* [90], *CommentSpace* [92]). Specifically, we developed the following conditions for the study: ("All") listing all comments chronologically in a single pane alongside a news article, ("Curated") adding a second tab to highlight specific comments, ("Prompts") presenting discussion prompts along with subsets of the highlighted comments, and ("Annotations") using socially-constructed annotations that connect each discussion prompt to relevant sections of the news article.

As a motivating case, the study involved a scenario where moderators want to use design elements to highlight comments related to data visualizations in a news article. Existing computer supported cooperative work (CSCW) research has found that comments about data visualizations can offer journalists useful feedback and help to center an online discussion around information presented in reporting, such as how data collection and visualization choices play into different narratives of an issue [36, 43, 59]. The experiment involved three stages (N=336 participants): (1) pre-survey to evaluate existing knowledge about a specific topic, (2) a news website experience that involved reading an article about the topic, and (3) post-survey to assess the amount of information participants remembered from the article. During the news website experience (Stage 2), each participant was randomly assigned to join as a newcomer to an existing online discussion about one of two news articles and one of four design conditions (i.e., All, Curated, Prompts, Annotations), in a 2x4 factorial study design.

Our analysis presents a design trade-off between participant engagement with a news website (e.g., time spent, comments generated, positively recommending an article) and the amount of information that participants are able to retain from a news article. Specifically, we find that by comparison to the All condition, participants assigned to the Annotation condition spent more time interacting with the discussion interface and were more likely to post new comments that responded to the annotated content. However, we also find that participants in the Annotation condition recalled fewer information items presented in the articles than participants in the All condition, even after controlling for factors related to participant background (e.g., prior knowledge about the topic, trust in the reporting), as well as the item-level difficulty associated with each article. Additionally, the findings highlight how some types of discussion content, such as comments about data visualizations, may be more memorable than other content irrespective of the discussion interface design.

The findings illustrate the various considerations for integrating discussion content into a news article, including where and how to represent discussion content. Participant characteristics should weigh into these decisions, such as existing knowledge, ability, and feelings of trust in a news source. Our findings also offer implications for content moderation systems [30], and specifically at news websites [55]. While our analysis identified a stark trade-off between the Annotation and All (baseline) conditions, we discuss how adding conversational prompts to a thread of comments may promote engagement without negatively affecting knowledge retention.

2 RELATED WORK

More than eight in ten adults in the United States of America consume news digitally (e.g., browsers, smartphones, tablets) [77]. There are many ways that journalists can use digital technologies to increase the salience of specific content in a news article [16]. Presenting news online enables journalists to create interactive experiences around geographic [29] and statistical information [35, 36, 59]. For example, Sultanum et al. [87] present a tool called *VizFlow*, which journalists can use to compose articles that integrate dynamic data visualizations that update as a user scrolls up-and-down through the content (called "scrollytelling"). Digital technologies can also collect information about the technology users [1, 16]. For instance, *PersaLog* [1] is a digital publishing protocol that journalists can use to insert rule-based content templates into an article that update based on known and inferred information about the reader (e.g., location, demographics), thereby transforming an article in order to make specific content personally appealing.

It is less clear whether it would be practical to highlight specific sections of a news article by referencing comments from online discussions about the article. A primary concern is that online discussions tend to elicit many comments that include off-topic and toxic content (e.g., profanity, trolling) [15, 21], which can threaten the credibility of the reporting [17, 69] and drive other news readers away from the comment section [80]. Another concern is that comments can affect how news readers interpret topics in the news. Specifically, people are more persuaded by comments that present disagreements in a civil tone than with toxic language [14]. Anderson et al. [4] found that people are more likely to polarize towards their initial perceptions of an issue when they are exposed to comments that contain toxic content (by comparison to civil content). Due to these and other considerations, many news outlets have opted to shut down the commenting sections on their websites altogether [23, 39].

In addition to these concerns, it is less clear how content from an online discussion might be represented within the body of a news article. For example, Wood et al. [93] present a system called *Newsr* that invites users to annotate, revise, and even redact parts of an article by using a graffiti function, rather than text-based commenting. While graffiti in Newsr can draw reader attention to specific sections of an article, in a field trial of the system, users reported that the graffiti

was difficult to read and often hard to understand [93]. In this study, we investigate how several common ways of integrating text-based comments into news articles (e.g., curating, prompting, annotating) affect participant engagement with the discussion as well as what people recall about an article. Specifically, this paper raises new considerations about whether and how to integrate online discussion content into news articles.

2.1 System design strategies to promote engagement with an online discussion

The first research question asks: *How does adjusting the salience of discussion content in an online news article affect participant engagement with the online discussion?* This study involved system design strategies that can help people to discover content in an online discussion [20, 24, 65], generate responses to existing comments [82, 97], and orient the online conversation toward relevant sections of an article [7, 12, 90, 92]. Each strategy reflects a slight shift in the degree to which content from the online discussion is integrated with content presented in a news article. The following sections present related research about each design strategy.

2.1.1 Curate the discussion to help people find valuable content. For interactive online discussion to occur there must be an initial critical mass of interest; however, as the volume of discussion expands, individual participants reach a cognitive limit in their ability to process new information in the discussion [41]. Content curation is a common strategy to address this limit by making choices about which comments to present, when, and in what order [56]. Many news website commenting sections are curated in chronological order by default, but some discussion systems use comment-level ratings (e.g., likes, up/down voting) to curate the discussion (called "distributed moderation") [48, 49]. While commenting systems offer a range of tools for people to sort, filter, and search the discussion on their own, it is worth noting that most people do not change the content curation settings away from the defaults [48].

Moderators can play a key role by identifying valuable content to showcase in a curated discussion [55]. For example, the moderators of *RegulationRoom* [71], an online deliberation platform, were trained to recommend comments that they felt would promote desired social norms in the discussion [24]. Without automated [20] or semi-automated support [65], it can take human moderators a substantial amount of time and effort to search a discussion for comments worth recommending [55]. *CommentIQ* [65] helps to alleviate these costs, by providing moderators with automated tools to intelligently filter discussion comments based on preset criteria, such as article relevance and whether the comment contains a personal experience.

After identifying comments to highlight, news websites can apply user interface design elements to represent discussion curation, such as by listing comments in priority order, adding badges to prioritize specific comments, and creating a designated space to showcase prioritized comments. How do different strategies for representing the curation in an online discussion contribute to participant engagement with the discussion?

2.1.2 Add conversational prompts to foster interactivity. Previous research about participant engagement at news websites has found that many comments do not elicit a response, which reflects a shallow level of interactivity among commenters [6, 15, 33, 81]. Similar to how a small group discussion facilitator nudges conversation forward by occasionally offering questions [60, 66], adding conversational prompts into news articles is a common way to promote interactivity at news websites [82, 97]. Stroud et al. [82] suggest that asking people to respond to conversational prompts in an article encourages them to reflect on the reporting from their own perspective, which in turn can motivate people to write a response.

Ziegele et al. [97] offer some evidence of this. In a content analysis of 619 news articles and 11,218 related user comments on nine nation-wide Facebook news pages, Ziegele et al. [97] found that

when reporters include conversational prompts in an article there are multiple potential benefits to the online discussion. Articles that include conversational prompts are associated with higher levels of inclusive and rational commenting in the discussion, as well as lower levels of toxicity, than articles that do not include prompts [97]. How does providing conversational prompts about comments in an online discussion affect participant engagement with the discussion?

2.1.3 Use annotations to link parts of an article to relevant online discussion. Many online comments about news articles are off-topic [15, 33, 59, 74, 81]. Adding annotations may be a viable strategy to promote online conversation that builds on specific parts of a news article [93]. Balestra et al. [7] define socially-constructed annotation systems as including three primary components: i.e., a resource, users, and metadata created by users as they identify parts of the resource to annotate with keywords, descriptive labels, categories, ratings, comments, and notes, etc. Socially-constructed annotations enable a "doubly-linked" discussion, in that users are able to follow links from a resource to relevant conversations in the discussion or they can follow links from the discussion to observations in the resource [90, 92].

Existing research has explored how annotation-based systems can facilitate interactive online discussion around budget documents (e.g., *FactFul* [44]), data visualizations (e.g., *ManyEyes* [90], *CommentSpace* [92]), and news articles (e.g., *Newsr* [93]). In a field study involving college students, Brush et al. [12] found that socially-constructed annotations can help anchor an online discussion to specific topics in course material. Additionally, annotation-based systems yield a higher volume of participant comments as well as replies to comments, by comparison to traditional discussion systems [12]. While socially-constructed annotations can help anchor conversation around specific topics, Brush et al. [12] also note fewer comments in the annotation condition about broad and general topics related to the materials. How might annotation-based strategies for representing discussion content in a news article affect participant engagement with the discussion?

2.2 Factors that contribute to the knowledge retention of news article content

User interface design strategies to promote participant engagement with an online discussion may have an unintended impact on other desired outcomes for people reading news online (e.g., feelings of trust in reporting [17, 69], accurately recalling information about news [4]). In this study, we investigate how strategies to promote participant engagement in an online discussion may also affect what people remember about a news article. Specifically, the second research question asks: *If discussion content is salient in an online news article, what factors affect the amount of information that people remember about the article?*

For several decades cognitive science research has investigated alternative ways to structure news articles for learn-ability, by studying how formating decisions in publishing play into *knowledge retention* [5, 18, 19, 22, 51, 89]. Knowledge retention has been operationalized by assessing the amount of information that people remember correctly after reading an article, often by employing open-response questions (e.g., "list all of the facts that you remember") and closed-response questions about specific facts from an article. Previous literature has identified several factors that play into the amount of information people retain from reading news, which we group into two broad categories: personal characteristics and information processing behaviors.

2.2.1 Personal characteristics play into how much people retain from news articles. News websites are complex information spaces which can be challenging for people to access. *Information foraging theory* is a cognitive model describing how aspects of a complex information space interact with personal characteristics of people immersed in the space, such as their prior knowledge and capabilities [68]. News websites present information with text, visual, and interactive components, which Pirolli and Card [68] describe as the various "patches" of information where people gather

knowledge throughout a space. People make choices about the amount of effort they are willing to spend gathering knowledge from a specific patch versus looking for new patches. These choices are informed by information scarcity within the space, as well as their personal appetite for information and skill searching for it (by tracking information "scents" through the complex space).

People also have different motivations for searching an information space. For example, when reading about a controversial topic people tend to (unwittingly) prefer information that supports their own stance on a topic, potentially ignoring information about alternative stances (referred to as *confirmation bias* [63]). To account for this cognitive bias, existing studies about knowledge retention that involve controversial topics have included metrics related to participant opinions and knowledge about the topic (e.g., abortion [14], genetically modified foods [50], fracking and synthetic biology [83], nanotechnology [4], protest movements [38]). Additionally, people have more and less trust in various news outlets [9, 10, 75], which may play into whether they prioritize information reported by a specific news source [5, 17, 69].

However, elements of an information spaces design can also play into the trust that people have for both the provided information and information provider(s) [68]. Balestra et al. [7] experimentally varied whether socially-constructed annotations were included in a consent form to participate in genomics research. While the rate of consent to participate did not vary by condition, Balestra et al. [7] found that participants assigned to the annotation condition reported a heightened sense of confidence in their decision, but felt less confident in their understanding of the consent material and less trusting of the research organization requesting their consent, than participants in the control condition without annotations. How (if at all) do participant characteristics play into the information people remember from reading a news article online?

2.2.2 Information processing behaviors can affect knowledge retention. Information foraging theory suggests a framework for how people search complex information spaces [68], but particular search behaviors within a space can also affect information recall. For example, many people scan an article for relevant information, by quickly reading headlines [51]; prior research about knowledge retention has found that this information processing behavior is negatively associated with remembering facts in an article [26, 89]. Thorndyke [89] investigates how information density in print news articles contribute to recall. Participants exposed to a full text article recalled fewer facts about the article than participants exposed to a condensed version, which excluded extraneous information (e.g., tangential background, commentary), but participants also took more time to read the condensed article. Thorndyke [89] suggests that extraneous information in the full-text version increased the likelihood that participants "selectively scanned" the article for facts, but in doing so, skipped over relevant information.

Different from a print article, at a news website there may be interactive components that can capture a readers' attention [19, 22, 26, 64, 88]. Eveland and Cortese [26] found that participants retained more information when they incremented through a health news website using "next" and "back" buttons (called linear navigation), by comparison to a nonlinear navigation system that included hyperlinks within the text and a sidebar site menu. Eveland and Cortese [26] also found a negative relationship between knowledge retention and (self-reported) selective scanning of the material; however, the selective scanning behavior was not related to the website design manipulations (i.e., linear versus nonlinear navigation). Journalists can promote information recall by paying attention to information density in an article [89] and system designers can promote recall by paying attention to how people navigate information online [26], but some user behaviors, such as selective scanning, can counteract these efforts [26, 89].

A limitation of prior research about information processing behaviors, like selective scanning, is that the metrics have been based on self-reported data [26, 89]. Digital news publishing protocols,

such as *IdylL* [16] and *PersaLog* [1], have streamlined the process of collecting information about online news reading behaviors, by accessing system logged user activity data. Grinberg [31] demonstrates how a few easy-to-calculate user behavior statistics—including dwell time, maximal depth reached by a user, and the amount of interaction with the page through any form of input—can be used to investigate information density in news articles. Through an analysis of 7.7 million page views and 66,821 news articles from seven popular publishers, Grinberg [31] found that articles presenting more information in the first few paragraphs are associated with longer reading times than articles with comparably less information early on. How (if at all) do user behaviors at a news website correspond with what people remember about an article?

3 METHOD

This paper presents results from an online controlled experiment designed to investigate how various user interface strategies for representing relevant information from a news website discussion might affect the ways that (**RQ 1**) people engage with the discussion and (**RQ 2**) how much they remember from news articles. As a motivating case, the study involved a scenario where a team of content moderators want to use design elements in a news website user interface to highlight a set of comments that are related to data visualizations in a news article. Prior CSCW research about online discussion related to data visualizations in news articles has emphasized the potential value that these comments can generate for journalists and toward an informed discussion [36, 43, 59].

3.1 Materials

3.1.1 News website system design. Many news outlets have opted to shut down their commenting systems, but the New York Times (NYT) has continued to scale up content moderation resources to facilitate larger online audiences [55]. The quality of discussion at the NYT website is highly regarded by regular commenters [54]. The NYT commenting system was custom built and is regularly updated [25, 61, 85, 86]. The NYT discussion is actively moderated by a team of people who remove problematic content (e.g., toxic, off-topic, misleading) and investigate high quality comments that are worth elevating in the discussion [55]. For these reasons the NYT website is a viable context for our motivating case, which is why we chose to model the baseline design decisions after the NYT commenting system.

The NYT commenting system includes several features that are similar and different from other platforms. Similar to other platforms the NYT discussion interface is not visible on page load, but can be accessed through buttons at the top and bottom of each article. Unlike other platforms the NYT presents comments in a panel to the right of an article, in a two-column design, which is similar to how online deliberation platforms present discussion alongside policy materials to promote informed commenting (e.g., *RegulationRoom* [71]). Additionally, NYT community moderators use a badge to signify high quality comments (called "NYT Picks"), which are also listed in a separate tab within the commenting column. The most active commenters at the NYT have referred to the coveted NYT Pick badge as akin to "the gold medal of the commenting olympics" [54]

We applied some aspects of the NYT interface design to all conditions in the study: (1) comments are presented alongside the articles, in a two-column design, (2) highlighted comments are presented in a separate tab within the commenting column. Additionally, we chose to make the comments visible on page load, rather than hide them until a user decides to access the discussion. We felt that hiding the comments would give an advantage to the Annotation condition, because discussion content in this condition is linked to relevant sections of the news article. Since the research questions are about what people contribute and remember when exposed to different representations of online discussion, we also felt that hiding the discussion by default would limit our analysis to the subset of people who opened the discussion panel.

3.1.2 News article selection and discussion content. Study participants were presented with one of two articles on a simulated news website. The articles were selected from prior NYT reporting, so that we could incorporate actual user comments posted to a commenting system resembling the study interface (see section 3.1.1). We selected articles about climate science, which is considered a controversial topic [11, 42, 52], so that we could evaluate whether participant perspectives on the topic factor into what people remember from the articles.

Modifications to the articles involved reducing the number of words, changing the data units for consistency and removing extraneous graphics within the article text. The location of the data visualizations in each article reflects the relative position of the data visualizations within the original reporting (see Appendix sections A.1 & A.2).

> **Article 1.** Based on "Greenhouse Gas Emissions Accelerate Like a *Speeding Freight Train* in 2018," written by Kendra Pierre-Louis (New York Times, Dec. 5, 2018).
>
> **Article 2.** Based on "As Coal Fades in the U.S., Natural Gas Becomes the Climate Battleground," written by Brad Plumer (New York Times, Jun. 26, 2019).

It remains possible that participants may have seen the articles before participating in the study, but we felt that specific details about the articles may have faded in memory given the 1-2 year time lag between when participants were recruited and the articles were originally published.

3.1.3 Independent variable: Different interface designs for presenting discussion. The simulated news website included authentic comments associated with the original articles (i.e., [67, 70]). In prior research, the comments were hand-coded for topical relevance to the article and relevance to the data visualization within the article [59]. We selected 24 comments from each article to include in the study interface: i.e., 20 topically relevant and 4 that relate to the data visualizations. Each comment was assigned a random timestamp and presented in chronological order.

Each design condition presented the 24 comments in a distinct way. The salience of discussion content was adjusted incrementally, in that each condition builds from the last by adding emphasis—in the news reading experience—to the online discussion (Figures 1 & 2).

All (Baseline). In the baseline condition, all 24 comments were listed in chronological order in a single column next to the article. This design strategy is similar to how online discussion is represented in deliberation platforms [24, 71] and commenting at the NYT website [55].

Curated. The design conditions (i.e., Curated, Prompts, Annotations) were presented in a two-tabbed interface that included an "All" tab containing the 24 comments in a chronological list and a "Highlights" tab containing the design intervention. Specifically, the "Highlights" tab included the 4 moderator selected comments related to the data visualizations, similar to how "NYT Picks" are represented at the NYT website [54].

Prompts. This design condition built on the Curated condition by grouping the comments in the "Highlights" tab into 2-3 relevant conversational prompts [57, 82, 97], such as "Should carbon emissions be displayed differently on the graph?" with comments that argue both for and against displaying emissions in per capita numbers rather than totals. In addition to article specific prompts, we also included a general prompt, "What additional graphs might help to improve the article?" in each condition. The prompts were displayed in an accordion style menu where the initial question could be expanded into a deeper description of the prompt with a "See Comments" button that would display the relevant comments when clicked.

Annotations. This design condition builds on the Prompts condition by incorporating annotations (e.g., [7, 12, 92, 93]) that link content in the "Highlights" tab to conversational prompts that are displayed next to visualizations in the news article (Figures 2 and 3). The Annotations condition is the only system design where engaging with a prompt will automatically scroll the

user interface to the visualization in the article and relevant comments in the discussion. This interaction was inspired by the "doubly-linked" discussion navigation possible in collaborative visualization systems [44, 90, 92].

To maintain some parity in the initial experience of the commenting interface across the baseline and treatment conditions, in all cases the interface defaulted open to the "All" tab.

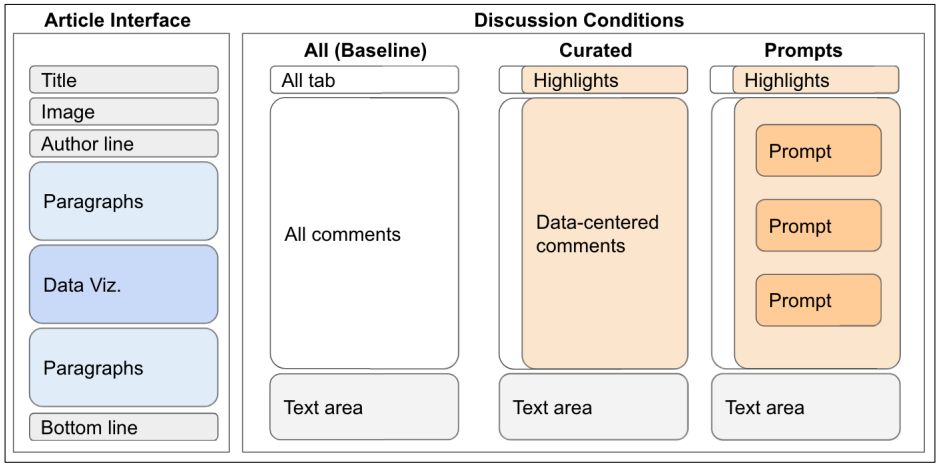

Fig. 1. Layouts for the All, Curated, and Prompts conditions. The article interface and "All" tab were the same across all 3 conditions. The "All" tab included all 24 comments. The "Highlights" tab showcased four comments that relate to observations in the data visualizations. In the Curated condition, the four comments were presented in a chronologically ordered list within the "Highlights" tab. In the Prompts condition, conversational prompts were added to organize the highlighted comments.

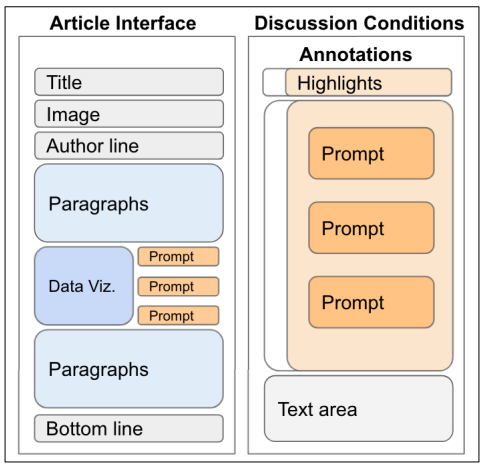

Fig. 2. Layout for the Annotations condition. The Annotations condition is the only design that changes the article interface by adding conversational prompts next to the data visualization. Similar to the Prompts condition, the conversational prompts also appear in the "Highlights" tab.

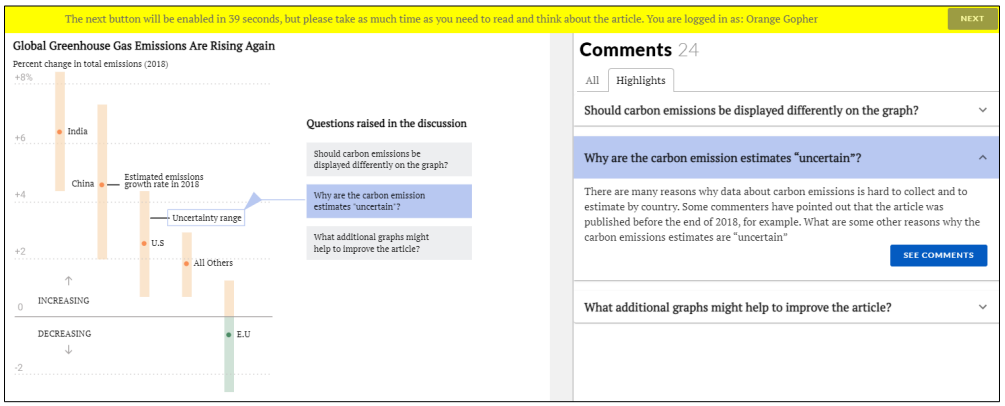

Fig. 3. The *Annotation* condition included several conversational prompts alongside the data visualization. Two interface features are triggered when a prompt is clicked: (1) labels are added to identify parts of the data visualization that relate to the prompt, and (2) the discussion interface changes to present information and comments related to the prompt. Similar to how socially-constructed annotations "doubly-link" resources to the discussion [92], participants can also access relevant parts of the data visualization by clicking open a conversational prompt in the discussion interface.

3.2 Recruitment and Procedure

Participants were recruited online through a university study participant recruitment portal, called SONA. Prior to signing up, participants were informed that they would receive 1 academic credit for completing the task and that the task was estimated to take 30-40 minutes to complete. We received IRB-approved informed consent from all participants.

Prior to entering the news website stage, all participants were required to complete a pre-survey which included questions related to their personal interest in various news topics, news consumption patterns, and familiarity with the specific topics included in the study, e.g., greenhouse gas emissions. After completing the pre-survey, participants received the following instruction:

> "Please read and think about the news article, take as much time as you need. After 1 minute a button will appear in the upper right corner of the user interface, which will direct you to the final stage of the study."

Participants were also prompted to "create an avatar" by selecting a color and animal (e.g., *@blue-Monkey*), which is similar to how many news websites require new users to create a pseudonym for the online discussion [28].

Upon entering the news website stage the participants were randomly assigned to article and design conditions, in a 2x4 factorial study. The news website stage included an article, discussion interface, and a banner at the top of the page with the following instructions alongside a Next button that was initially *disabled*:

> "The next button will be enabled in [60 and lower] seconds, but please take as much time as you need to read and think about the article."

The Next button was *enabled* automatically after the first minute within the experience, permitting participants to progress to the post-survey whenever they felt inclined.[1]

[1]To be clear, we opted for a minimum required period of participation (1-minute), rather than an upper bound, such as 10-minutes, because we were concerned about artificially inducing participation. Reported in the findings, average dwell time in the news website stage was 7.6 minutes (SD 8.3 minutes).

During the post-survey participants were asked knowledge retention questions about the news article they were exposed to during the news website experience, and then they read and responded to questions about a text-only version of the other article. For example, participants presented Article 1 in the news website experience also read a text-only version of Article 2 as part of the post-survey. As information density [89] and website navigational systems [26] can affect knowledge retention, this within-subjects step was taken to account for participant-level ability to retain information from news articles, in general, separate of the study website and conditions we created (referred to as "Retention ability" in the Metrics section 3.3.2). This step in the procedures also allowed us to evaluate differences in article-level difficulty and item-level difficulty (see Tables 8 & 9 in the Appendix).

In total, 414 people attempted the study; however, only 354 completed the pre-survey and post-survey (85%). During our preliminary analysis of the data we retroactively removed 18 participants who had large gaps in their session activity, resulting in 336 participants included in the analysis (81%). Table 1 presents the participant assignment by article and discussion system design conditions. In the Findings, we also present direct quotes from participant contributed comments. To preserve anonymity, we assigned each participant to a unique identifier: P1-P336.

Conditions	Article 1	Article 2	*Total*
All	40	42	82
Curated	40	42	82
Prompts	48	43	91
Annotations	39	42	81
Total	167	169	**336**

Table 1. Participants were randomly assigned in a 2 (Article) x 4 (Design Condition) experiment.

3.3 Metrics

The study measured participant engagement with an online news article discussion (**RQ 1**) and what participants remember about the article (**RQ 2**). The metrics are based on responses to the pre- and post-surveys as well as participant contributions and system logged behaviors during the news website experience. To account for factors unrelated to the user interface design, we also include metrics related to participant characteristics (e.g., prior knowledge, ability, trust in news reporting) [63, 68] and information processing behaviors (e.g., selective scanning) [26, 31, 89].

3.3.1 Participant Engagement. The analysis included measures of participant engagement that were automatically generated by the system and hand-coded by the research team. In order to build a qualitative understanding of participant engagement with the news website experience, we tracked participant scrolling activity, clicking activity, and comment posts through Google Analytics. The user activity data was transformed into "Activity Plots" for each participant, which were used to study how participants engaged with the experience (Figure 7).

Similar to existing research about information density in news articles based on system logged data [31], we calculated four user activity metrics per participant session:

- *Session dwell time.* From article load to when the "Next" button was clicked.
- *Session active time.* Summation of periods when the participant actively engaged with the interface (e.g., scrolling, clicking, hovering, keystroke).
- *Highlights dwell time.* Total time a participant had the "Highlights" tab open.
- *Highlights active time.* Summation of periods when the "Highlights" tab was engaged.

In addition to measures based on user activity, an implicit measure of participant engagement is whether people are willing to "recommend" or "share" an article.

- *Recommends the article (Agree, Neutral, Disagree)*: Participants indicated whether they agree, disagree, or feel neutral about the statement, "I think that other people should read this article to learn about climate change."

3.3.2 Knowledge Retention. Our operationalization of knowledge retention involved open-response and closed-response questions related to the news articles. Participants were asked in the post-survey several open-response questions, "What topics do you remember in the comments about the news article?" with the option to respond "*I don't remember*" to deter participants from guessing. The hand-coding of the responses was to identify participants who remembered anything about the comments included in the "Highlights" tab within the discussion. The following are excerpts from two of the highlighted comments:

- "[...] this graph doesn't include nuclear (along with a bunch of other energy sources) [...]"
- "[...] per capita numbers are used to deceive and manipulate people to ignore the important facts. Mother nature does not care one whit about per capita emissions. [...]"

Additionally, many participants chose to add a comment to the discussion (42.5%, N=143). Two researchers hand-coded each to evaluate whether the comments contained information related to topics expressed in the highlighted comments. Due to the relatively small collection of participant generated comments, coding disagreements were resolved by consensus among the researchers.

- *Responded to highlighted topics (Binary)*: Comment contains information related to topics in the comments highlighted in the user interface.

The research team also developed twelve closed-response questions to evaluate participant's ability to retain information from each news article (twenty-four total questions included in Appendix sections A.1 & A.2). Each question included four assessment options as well as a standard fifth option "*I don't remember*" as a way to reduce the incentive to guess. As the interventions were designed to highlight comments in the discussion that engage with the data visualization, we also tried to balance the number of response questions based on information in the paragraphs above and below the data visualization included in each news article.

Performance on the closed-response questions is reported in terms of the total number of items answered correctly as well as with a standardized score that accounts for question difficulty. To account for question difficulty we applied methods from Item Response Theory (IRT), which includes a collection of statistical models that attempt to establish a link between participant responses on an assessment and an underlying ability being measured by the assessment [72, 78, 95]. For the purpose of this study we considered a Rasch Model of participant ability, which assumes each item has a constant discrimination factor of 1, but may vary in difficulty [94].

All of our quantitative analyses were conducted with the R software environment for statistical computing and graphics. We used the R 'ltm' Package "Latent Trait Models under IRT" (version 1.1-1) to fit the one parameter IRT models (a.k.a., *Rasch* model) [73]. The R 'ltm' package includes a suite of functions which we used to fit and test a model, evaluate item difficulty for each closed-response item, as well as to estimate participant ability scores. The models for Article 1 and Article 2 are based on 169 and 167 responses respectively, which is above the minimum acceptable sample size for a Rasch Model.[2]

[2]The minimum sample size for a Rasch model analysis is subject to some debate. Foundational work by Wright and Stone [94] presents a Rasch model analysis based on 35 subjects and an 18 item test; however, Linacre [53] estimated that ~61 responses is a more conservative minimum sample size.

In the study procedures, each participant was asked to read two articles; the first was incorporated into the news website stage and the second was presented text-only and as part of the post-survey. We fit a Rasch Model for each article based on responses from the post-survey to estimate a participant's general ability to retain news article content, separate of the study interface design conditions (called "Retention ability" and treated as an independent variable). Then we applied each model to evaluate participant ability to retain information from the simulation (called "Article retention" and treated as a dependent measure in the analysis).

- *Retention ability (logit scale score)*: Recall based on a text-only version of the news articles during the post-survey. Included as an independent variable in the analysis.
- *Article retention (logit scale score)*: Recall based on the articles presented in the news website experience. Included in the analysis as a dependent variable (Table 7).

The item-level statistics for the Retention ability models are presented in the Appendices A.1 Table 8 & A.2 Table 9. Listed below are sample questions that reflect low and high levels of difficulty.

- **Low difficulty** (difficulty = -0.44): According to the article, approximately how much did extreme disasters cost the United States in 2018?
  - *Answer*: $300 billion.
- **High difficulty** (difficulty = 1.51): What reason does the article offer to explain China's recent rise in greenhouse gas emissions?
  - *Answer*: More coal-based manufacturing.

At the end of the study, we also asked participants to report if they had decided to take notes during the study and, if so, whether they had changed their note-taking behavior when presented with the text-only article in the post-survey. While we did not ask (or suggest) that participants take notes during the study, we thought that some may choose to do so independently. This data was hand-coded to consider the extent to which our measure of retention ability was inflated due to familiarity with the assessment format.

3.3.3 Personal Characteristics. Several factors beyond user interface decisions may affect what people remember about a news article. Personal characteristics, such as existing knowledge and ability, can affect how people search for information [68]. As the study context involved reading and commenting about online news articles, prior to entering the news website experience we also asked participants about their news consumption habits and experience participating in online discussions about news [76, 77].

- *News consumption habits*: Participants were asked several questions related to the amount of time they spend consuming news, by media (e.g., print journalism, television), and whether they have ever participated in online discussions about news.

Personal beliefs about a topic can play into whether people pay attention to or skim over some information [63]. As reporting about climate science has been politicized and can be polarizing [11, 42, 52], we asked participants a series of questions about their personal interests and prior knowledge about specific topics covered by the selected news articles.

- *Interest and beliefs about climate science*: Participants were asked several questions about their personal interest in climate science research and feelings about climate change policy. The questions were based on statements in the selected news articles (i.e., greenhouse gas emissions [67], shifting away from coal energy [70]).
- *Self-rated knowledge (Low, Moderate, High)*: Self-rated familiarity with specific climate science topics, "How knowledgeable are you about global climate change?"
- *Assessed knowledge (Binary)*: Participants were asked one ranking question related to each of the article conditions (two questions total). Example: "Based on your own intuition, please

drag the following *country* tiles to reorder them from largest (#1) to smallest (#8) in total carbon emissions per year." If participants correctly ranked the first item in each ranking question then the variable was coded TRUE, otherwise marked FALSE.

Additionally, personal trust in the credibility of a news outlet can factor into whether and how people value information in a news article [9, 10, 75]. As the study context involved system design decisions and news articles based on the NYT, we asked participants to rate their trust in various news outlets, including NYT, The Guardian, Breitbart News, BuzzFeed, etc. However, in the analysis we chose to only include participant ratings of trustworthiness related to the NYT.

- *Trust in NYT Reporting (Agree, Neutral, Disagree)*: Participants were asked, to what extent do you agree with the statement, "This news website is a trustworthy source of information."

3.3.4 Information Processing Behaviors. In addition to personal characteristics, information processing behaviors may also affect what people remember from reading a news article online. Rather than rely on a self-reported measure of information processing behaviors, such as selective scanning [26, 89], our analysis involved a qualitative coding of system logged activity data presented in the Activity Plots for each participant (Figure 4).

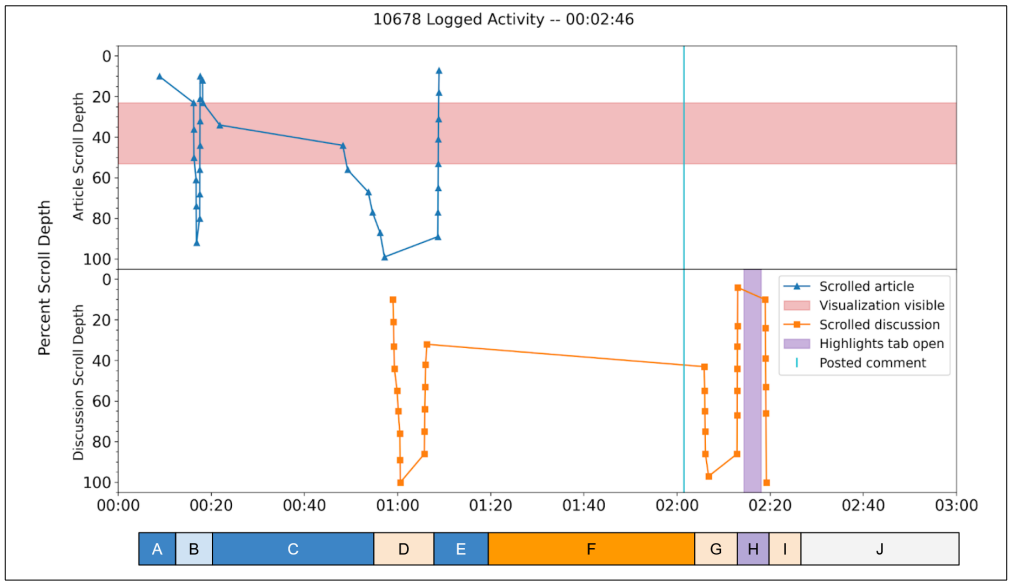

Fig. 4. An activity plot based on system logged data. Each plot was hand-coded for participant engagement features. As an example, the following refer to time periods A-J on the x-axis: (A) read article above the visualization, (B) performed a selective scan of the article, (C) read the data visualization (D) performed a selective scan of the comments, (E) read below the visualization, (F) posted comment, (G) performed a selective scan of the comments, (H) highlights tab opened, (I) performed a selective scan of the comments, and (J) exits session.

To determine whether a behavior was present or not, we defined several thresholds for what constitutes scrolling "quickly" or spending a "considerable time" reading. Specifically, we estimated an approximate reading time of 3 minutes for each article (~750 words) based on the average adult reading speed (~200-250 words per minute) and then defined approximate time (t) thresholds for the following hand-coded measures.

- *Selective scan of the news article (Binary)*: Quickly scrolled (t < 30 seconds) through the entirety of the news article at the beginning of their session.
- *Selective scan of the All tab (Binary)*: Quickly (t < 30 seconds) scrolled through the entirety of the comments in the All tab before slower engagement with the commenting interface.
- *Read above the visualization (Binary)*: Considerable time (t > 30 seconds) scrolling through the news article before the visualization comes into view.
- *Read the visualization (Binary)*: Considerable time (t > 30 seconds) scrolling or paused while the visualization is in view.
- *Read below the visualization (Binary)*: Considerable time (t > 30 seconds) scrolling to the end of the news article after the visualization goes out of view.

Two researchers trained on the Activity Plot coding scheme with a sample of 40 user activity plots and then iteratively tested on samples of 20 until a Cohen's Kappa score above 0.70 was reached for each code. See Table 2 for inter-rater reliability scores.

	Obs.	*IRR*
Selective Scanning		
Article	93	0.87
Comments	106	0.80
Reading		
Above	312	0.77
Visualization	265	1.0
Below	194	0.73
Total participants	336	

Table 2. The number of observations and the Cohen's Kappa score for inter-rater reliability (IRR) by code.

3.4 Statistical Analysis

Multiple linear regression analysis was used to evaluate the effect of the baseline and treatment design conditions on article knowledge retention, while accounting for several factors unrelated to the user interface design, including differences among the articles (e.g., Greenhouse gas emissions). Factors related to personal characteristics were included, such as prior knowledge, ability, and trust in the reporting [63, 68]. Additionally, factors related to information processing behaviors were also included, such as behaviors identified through the activity plot analysis (e.g., selective scanning, reading the visualization) [26, 31, 89]. Table 6 presents descriptive statistics for each factor and Table 7 presents the regression analysis.

As the dependent measure is on a logit scale from $(-\infty, +\infty)$ the coefficients are interpreted as the change in predicted *Article retention* expected from a one-unit increase in an independent variable, holding others constant. The linear regression model significance was evaluated with R^2 and F-test statistic. The multiple R^2 considers the variance among the fixed-effects, whereas the adjusted R^2 accounts for the number of fixed-effects in the model. The F-test statistic is also included, demonstrating the overall Fit of the model to the data.

4 FINDINGS

In total, 336 participants are included in the analysis. The majority of participants identified as Female (70%), speak English as a primary language (75%), and are 21 years of age or younger (83%). These sample characteristics are typical of studies that recruit from university student populations in the United States of America [34].

As the study context involved reading and commenting about online news articles, we asked participants about their news consumption habits and experience participating in online discussions about news. More than half of the participants report that they spend more than 1 hour per day consuming news (58%). During a typical week, participants might consume news through a variety of media. A few participants read print journalism (4%) and listen to radio/podcasts (14%), but the most common news sources include news websites (42%), online video (74%), and social media (85%). By comparison to Pew Research Center studies of news consumption patterns among U.S. adults [76, 77], participants in our sample make higher use of social media and lower use of news websites to get news. While most participants regularly consume news, few report participating in online discussions about news. Specifically, two in five participants report having posted a comment to social media about a news article (40%), though just one in five have ever posted a comment at a news website (20%).

We also asked participants about their familiarity with the specific news topics selected for the study: i.e., global greenhouse gas emissions [67], shifting trends in U.S. domestic energy generation [70]. Most participants believe that there is solid evidence that the average temperature on Earth has been getting warmer due to greenhouse gas emissions (95%), global warming requires immediate government action (92%), and rich countries have a moral obligation to reduce greenhouse gas emissions (86%). In contrast, there are considerable divisions among the general public about evidence surrounding climate science [11, 52].

The following sections present how the interface conditions contributed to different levels of participant engagement with the online discussion (**RQ 1**) and retention of information from the news articles (**RQ 2**). The findings illustrate a design trade-off in that some discussion interfaces that result in relatively high levels of participant engagement with the discussion, specifically the Annotations condition, can also have a detrimental impact on what information participants are able to recall from the news article. Additionally, the findings highlight how some types of comments are more memorable than others irrespective of the discussion interface design. Prior research demonstrates that toxic news is more memorable than average news stories [18]; our findings indicate that comments referencing the data visualizations in a news article may also be more memorable than other topics in the discussion.

4.1 RQ1. How does adjusting the salience of discussion content in an online news article affect participant engagement with the online discussion?

In this section, we describe engagement with the discussion in terms of how participant's spent their time within the news website stage, interacted with the discussion interface, and whether they would be likely to recommend the news article. The time that people spend engaging with an online system can be an important predictor of their future involvement with the platform. Pursuant to the study protocol the participants were permitted to exit the news website stage after 1 minute, but most spent more time using the system: Average dwell of 7.6 minutes (SD 8.3 minutes). Our analysis found no significant difference in the total dwell time or total active time spent in the news website stage (Table 3), but we found that the study conditions contributed to different levels of involvement with the discussion interface (Table 4).

Specifically, participants in the Annotation condition spent significantly more time engaged with the "Highlights" tab, by comparison to participants assigned to the other design conditions (i.e., Curated, Prompts). Table 4 presents summary statistics related to the time spent actively engaged with the "Highlights" tab. Our measure of active time captures when participants directly interact with the news website interface, such as by scrolling or hovering over comments. By contrast, dwell time reflects the total amount of time that the "Highlights" tab was clicked open and visible to the participant. In the Limitations we discuss how dwell and active time capture an upper and lower

bound of interaction, yet without eye-tracking we have no way of knowing whether participants may have glanced back to the comments as they read the news article.

	News website session duration			
	Active time (*minutes*)		**Dwell time** (*minutes*)	
Conditions	Mean (SD)	Max	Mean (SD)	Max
All (Baseline)	4.43 (3.37)	21.40	7.40 (4.82)	26.83
Curated	4.34 (4.34)	28.37	7.30 (8.18)	57.10
Prompts	3.59 (3.06)	13.09	7.14 (6.06)	30.87
Annotations	4.26 (3.86)	17.32	8.82 (12.55)	101.67
	$F(3, 332) = 0.95$, $p = 0.417$		$F(3, 332) = 0.71$, $p = 0.546$	

Table 3. The design conditions did not significantly affect participant *dwell time* within the news website stage of the study or the total amount of time participants spent actively engaging with the user interface (called *active time*). Dwell time is the duration between the moment a participant enters the news website stage and when they click the "next" button to exit the stage. Active time is the summation of each period when the participant showed some cursor or keystroke activity.

	Discussion "Highlights" tab session duration			
	Active time (*seconds*)		**Dwell time** (*seconds*)	
Conditions	Mean (SD)	Max	Mean (SD)	Max
All (Baseline)	NA	NA	NA	NA
Curated	0.11 (0.98)	8.87	26.25 (72.48)	361.63
Prompts	1.69 (6.44)	59.30	41.33 (75.71)	304.07
Annotations	6.51 (14.19)	63.08	114.64 (181.08)	775.49
	$F(2, 81) = 11.43$, $p < 0.001$		$F(2, 81) = 4.16$, $p = 0.018$	

Table 4. Participants in the Annotation condition spent significantly more time interacting with the Highlights tab (*active time* in seconds) than participants assigned to other design conditions (Curated, Prompts). Additionally, participants in the Annotation condition also left the tab open longer as they engaged with other parts of the simulated news article experience (*dwell time*). As the "All (Baseline)" did not include a "Highlights" tab, the active and dwell time within the tab are null.

Another common measure used to gauge participant involvement at a news website is whether they add to the online discussion. Across all of the conditions, many participants chose to add a comment to the discussion (42.5%, N=143). The conditions were intentionally designed to promote specific topics related to each data visualization. Specifically, the Highlights tab in each condition showcased four of the twenty-four comments from the existing discussion (all comments were listed chronologically in the All tab). *How many participant comments respond to the seeded topics?* Our qualitative analysis identified 17 comments that respond to the highlighted topics (11.48%). Sixteen of the comments were generated by participants in either the Curated, Prompts, or Annotations; however, most were contributed by participants in the Annotation condition (N=9), based on a Fisher's Exact Test for count data ($p = 0.031$).

For participants in the Prompt and Annotation conditions the Highlights tab included several prompts specific to each data visualization as well as a general prompt that appeared in both of the news articles that read: "What additional graphs might help to improve the article?" Several participants responded to this prompt, for instance, six participants recommended modifications to

a graph depicting the share of total electricity generation in the U.S. by energy source. Most of the comments were about the lack of additional energy sources in the graph, which merely presents trends in coal, natural gas, wind and solar. Participants wondered about other sources, such as nuclear energy, but also about whether presenting each energy source categorically is an accurate representation of the data as, according to the article, many community power-grids rely on a mix of energy sources to provide consistent electricity.

- (P45) "Although renewable energy is seen as the right choice, it requires other sources of energy [...] It may be more efficient to do research on hybrid energy sources and graph different hybrid combinations rather than individual sources."
- (P20) "It would be more helpful if the graph were more comprehensive and included other sources of power generation. Overall, finding the right mix of energy sources (wind, solar, nuclear, etc.) across the country is key to tackling the issue of climate change."
- (P105) "Renewable and clean Energy like wind or solar must become far more common than it already is, replacing coal and gas as the main producers of energy for our country, with gas acting only as a backup in case these sources become unavailable to due weather."

Rather than depict individual energy sources in a line graph, the authors might have responded to these suggestions by presenting the data as a series of percent stacked bar charts or chord diagrams to depict how strategies for combining sources may have shifted over time.

The article about *Greenhouse gas emissions* included a visualization of the percent change in total emissions by country (e.g., India, China, USA). The prompt receiving the highest response from participants asked, "Should carbon emissions be displayed differently on the graph?" and highlighted two existing comments that discuss how total emissions versus per-capita emissions shift the relative ranking of the countries to either present the USA or China as most responsible for climate change. Several participants responded directly to the highlighted comments, posting comments in support and opposition to changing the units:

- (P8) "@tomatoSquirrel I agree that looking at total emissions helps us understand which countries are most responsible [...]"
- (P31) "@tomatoSquirrel I agree [...], but the reality is that in order to help the problem of global warming (which effects everyone), it is worth considering what can be done on an individual level and how countries can aim to reduce their emissions to be the same per capita as other countries that are doing better in that regard."

Participants also proposed improvements to the visualization. A participant discussed how only presenting the most recent greenhouse gas emissions data is limited in that, "Only showing the greenhouse gas emissions for one year makes it difficult to put the data in perspective, since the article is discussing change over time and not just emissions that year" (P80). This example of constructive feedback can help data journalists to improve the clarity of their reporting [36].

Although only a few of the participants posted comments about the topics seeded in the Highlights tab, our analysis of the post-survey responses indicate that more people recalled the topics than commented about them. In the post-survey participants were asked in an open-response question, "*What topics do you remember in the comments?*" which we hand-coded to identify responses related to the highlighted topics. Our analysis identified 52 participants (15%) who recalled relevant information about the data visualization based on the comments highlighted in the discussion. However, we found no significant difference by condition, which indicates that even in the All condition—without additional scaffolding to promote the discussion topics—the comments about the data visualizations were memorable [χ^2 (3, 336) = 0.074, p = 0.994].

While time spent and interaction with an online discussion are useful measures of participant engagement at a news website, an important question for news outlets is whether readers are likely

to recommend an article (Table 5). Significantly more participants exposed to the design conditions agreed (and fewer disagreed) with the statement "*I think that other people should read this article to learn about climate change*," by comparison to participants in the baseline [χ^2 (2, 336) = 10.152, p = 0.006]. Whether agreeing with this statement may manifest into a behavioral response—e.g., a *like*, *share*, or feelings of trust—are open questions for future research.

	Other people should read this article		
Conditions	Agree	Neutral	Disagree
All	57 (69.5%)	14 (17.0%)	11 (13.4%)
Curated	70 (85.3%)	8 (9.7%)	4 (4.8%)
Prompts	71 (78.0%)	17 (18.6%)	3 (3.2%)
Annotations	65 (80.2%)	13 (16.0%)	3 (3.7%)
Total	263 (78.2%)	52 (15.4%)	21 (6.2%)

Table 5. Descriptive statistics related to how participants responded to the statement (i.e., Agree, Neutral, Disagree): "*I think that other people should read this article to learn about climate change.*"

In response to RQ1, we found that participants in the design conditions by comparison to the baseline, spent more time in the discussion interface, generated more comments about the data visualizations, and were more likely to recommend the article to others. These observations were especially pronounced among the participants assigned to the Annotations condition. In short, increasing the salience of discussion content within a news article can translate into various forms of engagement with the online discussion.

4.2 RQ2. If discussion content is salient in an online news article, what factors affect the amount of information that people remember about the article?

A primary purpose for a news outlet is to effectively communicate insights about current events [89, 91]. For that reason, it is important for system designers to consider how interactive components related to the online presentation of news articles may affect information recall. In this section, we present results from an analysis of what participants recalled from the article presented in the news website stage of the study based on a closed-response assessment.

Consistent with existing literature comparing knowledge retention across low levels (e.g., print journalism) and high levels of interactivity (e.g., online news articles with hyperlinks) [22, 88], participants who read just the text of an article as part of the post-survey recalled more information items (M=5.38, SD=2.84) than participants who read the same article as part of the news website stage (M=4.52, SD=2.42) [F(1, 670) = 17.58, $p < 0.001$]. Our analysis found that the total number of items recalled correctly varied by condition [F(3, 332) = 3.70, p = 0.012]. Specifically, a Tukey based post-hoc comparison of the conditions indicates a significant difference in the number of items recalled between the Annotation condition (M=3.83, SD=2.21) and the All condition (M=5.01, SD=2.31) [diff=-1.18, p-adjusted=0.009] as well as a marginally significant difference between the Annotation and Prompt conditions (M=4.75, SD=2.67) [diff=-0.92, p-adjusted=0.059].

As the items reflect information that was presented throughout each article, we investigated whether information presented earlier in the article (i.e., above the data visualization) was more or less memorable than information presented later in the article (i.e., below the data visualization). Our analysis found a significant difference in recall by position in the article, such that more of the information *above* the data visualization (M=2.40, SD=1.50) was recalled than information *below* (M=2.11, SD=1.50) [F(1, 670) = 5.92, p = 0.015]. Figure 5 presents the mean number of items recalled

(with standard-error bars) by condition as well as by the item position in the article relative to the data visualization.

A Tukey based post-hoc comparison of each condition revealed that participants assigned to the Annotation condition recalled significantly fewer items below the data visualization (M=1.54, SD=1.03) than participants in the All condition (M=2.32, SD=1.49) [diff=-0.78, p-adjusted = 0.017] as well as the Prompt condition (M=2.45, SD=1.66) [diff=-0.91, p-adjusted = 0.001]; however, participant recall of information above the visualization in the article (M=2.28, SD=1.60) was not significantly different from that of participants assigned to the All (M=2.68, SD=1.40) [diff=-0.39, p-adjusted = 0.679] or Prompt conditions (M=2.29, SD=1.58) [diff=0.01, p-adjusted = 1.000].

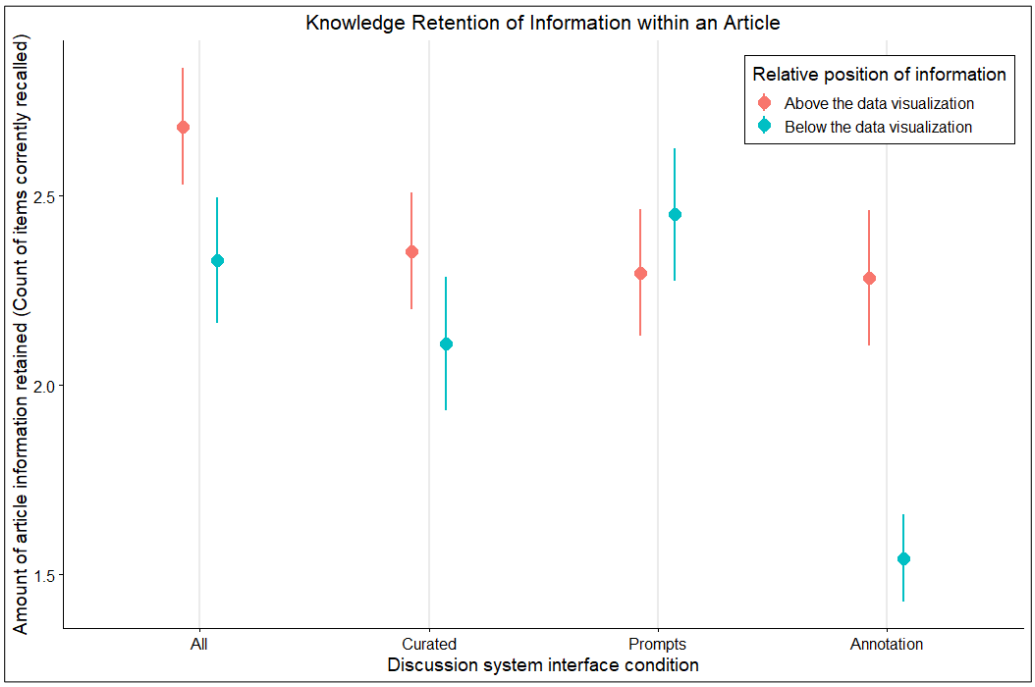

Fig. 5. Presents the amount of article information retained (average count of items recalled and standard error), by condition and relative position of the information to the data visualizations.

Beyond the relative position of information within the news article, many other factors might play into these findings, such as item-level difficulty as well as factors related to participant characteristics and information processing behaviors. To account for item-level difficulty, we constructed a Rasch Model [73] of participant ability to recall information from each article, presented in Tables 8 & 9 in the Appendix, and then applied the model to generate factor scores based on participant recall of information presented during the news website stage of the study. The article retention factor score is normally distributed (on a logit scale) with a mean of -0.23 (SD = 0.76).

We applied a multiple linear regression analysis to investigate possible relationships between the article recall factor scores and the system design conditions as well as other factors related to the news article topic, participant characteristics, and information processing behaviors. Descriptive statistics related to participant background and logged activity with the system are presented in Table 6. Results from the multiple linear regression analysis are presented in Table 7. The reported R^2 (0.412) and adjusted R^2 (0.379) indicate that the model explains ~37-41% of the variability in the

Trust in NYT Reporting	*Count (%)*
Agree	70 (20.8%)
Neutral	246 (73.2%)
Disagree	20 (5.9%)
Self-rated knowledge of climate science	*Count (%)*
High	183 (54.4%)
Moderate	78 (23.2%)
Low	75 (22.3%)
Assessed knowledge related to the article	*Count (%)*
Correct	143 (42.5%)
Incorrect	193 (57.4%)
General ability based on post-survey	*Mean (SD)*
Retention ability	0.01 (0.79)
News website logged activity	*Count (%)*
Scanned article	93 (27.0%)
Read [Above]	312 (92.8%)
Read [Visualization]	265 (78.8%)
Read [Below]	194 (57.7%)
Scanned discussion	106 (31.5%)
Posted comment(s)	143 (42.5%)
News website active time	*Mean (SD)*
Article (seconds)	146.6 (155.2)
Discussion (seconds)	101.71 (139.0)

Table 6. Descriptive statistics related to participant background (e.g., trust in NYT reporting), prior knowledge (e.g., self-rated, assessed), and retention ability based on the article recall assessment in the post-survey stage. Additionally, the table presents summary statistics related to the logged activity and time spent interacting with the news website system.

article recall factor scores. Additionally, the significant F-statistic indicates that the explanatory variables meaningfully add to the model fit with the data.

The model Intercept is significant and negative, which reflects the lower average level of recall from the news website stage by comparison to the text-only versions of each article presented in the post-survey. Additionally, the small, yet significant Intercept indicates that there is some additional variability in the data that is not captured by the predictors. The *News article condition* is included as a control for differences among the news article content and recall assessments. Our analysis found that the participants assigned to read Article 2 (U.S. emissions [70]) recalled more information about the article, by comparison to participants assigned to read Article 1 (Greenhouse gas emissions [67]). These differences are important to account for in analysis, but it is less clear what factors related to the article content may have caused this difference.

After accounting for item-level difficulty and differences among the articles, our analysis found a significant and negative general effect of exposure to the Annotation condition by comparison to the All condition (Est. = -0.23, $p = 0.019$). While our findings related to RQ1 by in large demonstrate that the Annotation condition resulted in higher levels of participant engagement (e.g., active time spent in the discussion, posted comments, willingness to recommend the article), the heightened levels of engagement may also contribute to lower levels of article recall. However, factors related to

Predictors	*Est.*	*Conf. Int.*	*p*
(Intercept)	-0.39	-0.74 – -0.04	**0.030** *
Article condition (compared to Article 1)			
Article 2: U.S. Energy	0.48	0.34 – 0.61	**0.000** ***
Discussion condition (compared to "All")			
Highlight	-0.13	-0.31 – 0.06	0.177
Prompt	-0.02	-0.20 – 0.17	0.840
Annotation	-0.23	-0.42 – -0.04	**0.019** *
Trust in reporting (compared to "Neutral")			
Agree	0.21	0.05 – 0.38	**0.012** *
Disagree	0.27	-0.03 – 0.58	0.080 ·
Self-rated knowledge (compared to "Moderate")			
High	-0.13	-0.30 – 0.03	0.105
Low	-0.11	-0.28 – 0.05	0.187
Assessed prior knowledge and ability			
Article knowledge [Correct]	0.19	0.05 – 0.33	**0.006** **
Retention ability (SD)	0.27	0.20 – 0.34	**0.000** ***
Article interface logged activity			
Scanned article	0.10	-0.05 – 0.25	0.202
Read [Above]	-0.43	-0.70 – -0.15	**0.003** **
Read [Visualization]	0.20	0.01 – 0.38	**0.039** *
Read [Below]	0.20	0.05 – 0.34	**0.008** **
Active time (SD)	0.06	-0.02 – 0.13	0.127
Discussion interface logged activity			
Scanned discussion	-0.14	-0.29 – 0.01	0.059 ·
Posted comment(s)	-0.08	-0.25 – 0.08	0.322
Active time (SD)	0.07	-0.02 – 0.15	0.114
Model performance			
Observations	336		
Multiple R^2 / Adjusted R^2	0.412 / 0.379		
F-statistic	12.35 ($p < 0.001$)		

Table 7. Presents a linear regression predicting the article knowledge retention. Fixed-effects include the news article condition, discussion system design condition, participant trust in the news reporting, prior knowledge and ability to retain information from news articles, as well as participant activity related to the news article and discussion interfaces. The multiple R^2 considers the variance among the fixed-effects, whereas the adjusted R^2 accounts for the number of fixed-effects in the model. The F-test statistic is also included, demonstrating the overall Fit of the model to the data.
Significance codes: p-value <0.001 '***', 0.001 '**', 0.01 '*', 0.05 '·', 0.1 ' '

participant background as well as engagement with the news website system appear to counteract the negative effect on article recall associated with the Annotation condition.

Prior knowledge about climate science affects article recall. Most participants in the study have a favorable view of climate science (95% believe that there is solid evidence of global warming) and perceive their prior knowledge of climate science as high (54.4%). Our analysis found that participants with actual (assessed) prior knowledge related to the article exhibited a higher on

average recall score (Est. = 0.19, $p = 0.006$); however, we found no significant effect of participants' self-rating of their own knowledge.

Additionally, some people are better at retaining information generally than other people. Our analysis found that a one standard deviation increase in retention ability—based on the post-survey assessment—relates to a higher on average article recall score (Est. = 0.27, $p < 0.001$). We asked participants to report any change in their reading behavior between the first article presented during the news website stage and the second article they read as part of the post-survey. About half of the participants (52.6%, N=177) reported that they paid more attention to the second article, by reading more carefully and taking notes; however, we found no significant difference in reported change in reading behavior by condition [χ^2 (3, 336) = 3.512, $p = 0.319$].

While most participants in the study have a favorable view and prior knowledge of climate science, their feelings of trust in New York Times reporting were mixed. Our analysis found that participants with an opinion, either *disagreeing* or *agreeing* with the statement "The New York Times website is a trustworthy source of information," had a higher on average recall of information in the news article than participants rating the statement *neutral*. This phenomena may in part be due to a confirmation bias where participants were looking for evidence within the article that supports their existing beliefs [63].

Our analysis of system logged activity patterns found that participants adopted various pathways through the news website experience. Participants with activity to suggest reading the data visualization and reading the article text below the visualization had higher on average recall of the article content, which complements our analysis of the number of items recalled by position within the article (Figure 5). Possible interactions among the participant logged activity predictors and system design conditions did not significantly improve the model performance, so were left out of the final model (Table 7). Finally, our analysis found a significant and negative effect associated with reading above the data visualization, which may be due to the fact that we observed this behavior among most participant sessions (92.8%).

In response to RQ2, our analysis found that the system design conditions affected retention of information presented in the article. Specifically, participants assigned to the Annotation condition recalled fewer of the information items by comparison to participants assigned to the All condition. Our multiple linear regression analysis demonstrates that the negative and significant effect of the discussion system design remains even after accounting for potential confounding factors related to the news article content, as well as participant knowledge, ability, and trust in the reporting.

5 DISCUSSION

Presenting comments as part of a news article is not standard practice. Online commenting sections have become less common at news websites, as many news outlets have opted to shut down their discussion forums and redirect the conversation about news to social media (e.g., Twitter, Facebook) [23, 39, 47, 79]. News outlets that host commenting sections tend to create visual distance between articles and discussion, such as by hiding the comments by default and positioning the discussion interface at the end of an article. At a fundamental level, this study challenges that standard by demonstrating that decisions related to the presentation of comments in the user interface can promote desirable outcomes in the reading experience.

As a motivating case, the study involved a scenario where moderators want to use design elements to highlight comments related to data visualizations in news articles. Prior CSCW research has found that comments about data visualizations can offer feedback for journalists [36] and contain personal stories, new sources of information, and insights that stem from data presented in an article [43, 59]. How does presenting comments in a news article affect the ways that readers engage with and retain information about news? Our analysis reveals a trade-off between user experience

system designs that promote participant engagement with an online discussion (e.g., time spent, comments generated, positively recommending an article) and participant recall of information presented in an article (i.e., knowledge retention [5, 18, 19, 22, 51, 89]).

As outlined in the sections that follow, the findings illustrate the various considerations for integrating discussion content into a news article. These include deciding where and how to represent discussion content related to news articles. While making these design decisions, participant characteristics should also be considered, such as existing knowledge, ability, and feelings of trust in a news source. Additionally, our findings offer implications for content moderation systems [30]. Moderating content from the *back end* of a discussion involves identifying problematic content to reject [13, 40] and valuable content to recommend in a discussion [20, 65]. Our findings offer insights about how moderation systems might assist with the integration of recommended discussion content into the *front end* of online news articles.

5.1 Considering the trade-offs of making discussion more salient

For the purpose of this study, we define "salience of discussion content" as the degree to which design elements separate and integrate content from the discussion into the user experience of reading news articles online. This conceptual definition was operationalized as curating, prompting, and annotating recommended comments to an article (section 3.1.3). In this section we review the findings in terms of specific decisions involved with adjusting the salience of discussion content in the "complex information space" [68] of a news website.

If the preferred outcome for people reading news online is to promote discussion about specific topics in an article, then our findings suggest that Annotations are more effective than the other strategies (i.e., All, Curated, Prompts). Similar to existing research about online discussion in an educational setting [12], our analysis found that participants in the Annotations condition were more active in the commenting system and contributed comments that add relevant topics to the discussion, by comparison to the baseline. In the Annotations condition, participants contributed comments that offer useful feedback for journalists [36] and engage with various aspects of the data visualizations [43, 59] (section 4.1).

If the preferred outcome for people reading news online is to enhance their knowledge retention of topics in an article, then Annotations are less effective than other system design strategies. *Information foraging theory* [68] may offer an analytic lens to interpret how participants experienced each of the design conditions. For example, in the baseline condition (i.e., "All") participants were presented with at least two "patches" of information: i.e., the news article and discussion. In the Curated and Prompts conditions, the "scent" associated with the discussion patch was enriched by using structural elements (e.g., tabs) and conversational prompts. While the same highlighted comments were included in each treatment, participants who were exposed to the Annotations may have viewed the comments as enriching the news article patch, specifically the data visualization, and then shifted their attention towards the discussion.

As evidence of this shift in attention, our analysis found that people in the Annotation condition recalled fewer information items located below the visualization, as well as overall (Table 7). This result is similar to existing research about how navigational systems that encourage users to take a linear path through a complex information space contribute to higher levels of knowledge retention, by comparison to a non-linear path [26]. This result is also similar to findings reported by Balestra et al. [7], that participants exposed to socially-constructed annotations in a consent form felt less confident in their understanding of the materials.

An alternative hypothesis is that our study design was not *social* enough for participants to benefit from the socially-constructed annotation design pattern, as each participant experienced the news website by themselves in a one-shot experimental design. Prior research about online collaboration

has found that the "doubly-linked" navigational structure of socially-constructed annotations can be useful for team-driven analysis tasks (e.g., reviewing budget documents [44], exploring data [92]). Rather than investigate how individual readers engage with and recall information from a news website, future research should consider opportunities for groups of people to read news together. For example, community moderators at the New York Times website rely on regular commenters to help fact-check online discussions [55]; future research could explore how this ad-hoc group activity may be formalized into a team-driven task.

5.2 Paying attention to how participant characteristics relate to desired outcomes

Our findings illustrate how personal characteristics, such as existing knowledge, ability, and trust in the credibility of the news source, play into what people remember about online news articles. It is important to note that some types of information in a news article are difficult for people to recall, in general. While our research team carefully developed the twenty-four closed-response questions for this study (presented in Appendix A), the item-level analysis presented in Tables 8 & 9 indicate that some information was significantly easier to recall than other information. Additionally, our linear regression analysis of article knowledge retention, presented in Table 7, found that the information in Article 2 [70] was significantly easier to recall than information presented in Article 1 [67] (Est=0.48, $p < 0.001$). News outlets could adopt a similar approach as described in the Methods (section 3.3.2) to anticipate information items in an article that may be difficult for people to understand.

However, people have different levels of existing knowledge and ability, which can significantly affect their information recall. While the discussion system design conditions in our study significantly affected knowledge retention, our analysis found that a participants' general retention ability is among the most consequential factors. For every one standard deviation increase in assessed retention ability, predicted article knowledge retention increases by 0.27 ($p < 0.001$). A design implication is that news outlets might use publishing protocols, like *PersaLog* [1], to present different experiences of a news article and online discussion based on known and inferred information about a readers' level of prior knowledge and ability, as a way to promote reading comprehension and specific competencies, such as civic, data, and digital literacy.

Some information processing behaviors can also impair knowledge retention. Our findings confirm existing research about how selective scanning in particular can detrimentally affect knowledge retention [26, 89]. Selective scanning was not uncommon, nearly a third of participants selectively scanned the news article (27.0%, N=93) and the discussion thread (31.5%, N=106). Additionally, selective scanning the discussion had a (marginally) significant negative affect on knowledge retention (Est=-0.29, p=0.059). While existing research about selective scanning has relied on self-reported measures [26, 89], our findings demonstrate that system logged activity data could be used to monitor and respond to this behavior. For example, publishing protocols, like *IdylL* [16], that streamline the process of accessing activity data to create interactive online news articles, could be used to recognize when people begin to scan an article and then suggest topics for them to read or discuss, in-situ and in real time.

The act of selective scanning may be similar to how people using social media can slip into a state of mindlessly scrolling news content without remembering what they had read (called "dissociation"). In a controlled experiment, Baughan et al. [8] explore how system designs that increase the salience of mindless scrolling can disrupt periods of dissociation in social media (e.g., by highlighting usage statistics, setting reminders to move onto another activity). While mindless scrolling is widely considered a symptom of social media addiction, Baughan et al. [8] discuss how some forms of disassociation can serve various cognitive purposes and even yield creative results.

Future research should explore ways of inviting readers to reconnect with articles after periods of disassociation at a news website.

5.3 Moderating content for the front end of a news website

Community moderators at news outlets balance a variety of considerations when deciding whether to reject, approve, and highlight a discussion comment [55]. Our analysis found that presenting comments as part of a news article raises new considerations about how to highlight user-generated content on the front end of a news website.

Simply highlighting a comment can affect reader opinions about reporting. Balestra et al. [7] found that participants exposed to socially-constructed annotations on a consent form felt less trusting of the research organization requesting their consent, than participants in the control condition without annotations. Our analysis found the opposite. Participants exposed to the treatment conditions by comparison to the baseline were more likely to agree (and less likely to disagree) that "other people should read this article to learn about climate change." Pirolli and Card [68] suggest that people follow "scents" as they search a complex information space for content that is credible [68], in a news website discussion these scents can include comment tone [17], commenter expertise [69], and whether journalists are active in the thread [82]. In our case, surfacing the online discussion around data visualizations and recommending them as "Highlights" contributed to participant willingness to recommend the article.

Highlighting a comment with a badge like the "NYT Pick" [54], is a relatively simple user-experience by comparison to complex organizations of user-generated content, such as argument maps [3], pro-con lists [45], thread summaries [62, 96], and the design conditions central to our research. To facilitate complex organizations of user content at scale, moderators could leverage crowdsourcing systems to coordinate the online community of news readers [2]. As an example based on the findings, conversational prompts appear to have boosted engagement, while not reducing knowledge retention. Participants in the Prompts and All (baseline) conditions retained similar levels of information (Figure 5). Participants in the Prompts condition also spent more time in the "Highlights" tab than participants in the Curated condition, though not as much as the Annotation condition (Table 4). Including conversational prompts in news articles can have multiple benefits for an online discussion, such as higher levels of inclusive and rational commenting and lower levels of toxicity [97]. Crafting a good conversational prompt is not an easy task for a moderator [60, 66]; however, existing CSCW research demonstrates that newcomers to an online discussion can be scaffolded through a crowd-writing process [27] of generating conversational prompts from existing discussion comments [57].

Presenting comments as part of a news article is not standard practice, but it could be. Existing news publishing platforms (e.g., *CommentIQ* [65], *VizFlow* [87]) and protocols (e.g., *IdylL* [16], *PersaLog* [1]) could be leveraged to help moderate discussion content in the body of an online news article. Our research demonstrates that doing so can offer new opportunities for people to benefit from the news reading experience.

6 LIMITATIONS

A key limitation is that the study involved a one-time controlled online experiment, rather than investigate the system design conditions over an extended period of time and with real news readers at a news website with a vibrant online community. Although the reported demographics of participants in the study are similar to other studies that have involved university students [34], the article topics related to climate change may have elicited different behavior in ideologically diverse online communities [11, 52]. The majority of study participants trust climate science and feel that there is a moral obligation to address global warming.

Additionally, our measures of participant engagement are relatively crude by comparison to higher fidelity measures, such as eye tracking and screen recording. These alternative research methods may offer a more precise understanding of how participants navigated and experienced our news website stage; however, they are unlikely to change our general findings related to the discussion system design conditions. Future research should investigate more novel system design strategies, by applying a "talk aloud" study protocol to gain a deeper sense of the experience.

Another limitation is that the user experience of the discussion system design conditions could be improved. In this study, we carefully designed each of the conditions so that they were distinct in specific ways, e.g., adding highlighted comments to a new tab, using a discussion prompt to group related comments, using annotations to create a bridge connecting the data visualization with the prompts. While not the most exciting user experience, consistency in the design decisions enabled us to make relatively fair comparisons among the conditions.

Finally, our choice of a controlled, simulated one-time discussion experience leaves open questions about engagement in long-term and ongoing discussions. A more social experience of the study materials may affect engagement. For instance, in a multiple-worlds online discussion experiment McInnis et al. [58] found that participants post more and longer comments when other people are synchronously adding to the thread, but fewer and shorter comments when the thread is less active. In our study, participants were not able to "discuss" the topics with each other, so we also do not know how these designs might have contributed to discussion as it may have naturally unfolded over time. Future research should investigate how the discussion system design conditions affect participant engagement and knowledge retention in-the-wild at online news websites and through multiple news cycles.

In order to facilitate the news website experience in this study, our research team generated a lot of content by hand (see Materials section 3.1). Presenting discussion content on the front end of a news website involves deciding what types of content to highlight [20, 65], how to represent that content (e.g., [3, 45, 96]), and how to facilitate the content moderation work within the fast-paced news cycle [55]. Our findings offer some insights toward making these decisions.

7 CONCLUSION

In this paper, we explore how strategies to elevate comments about data at news websites can affect participant engagement and knowledge retention. We found that participants exposed to our design conditions spent more time engaging with the discussion, posted more data relevant comments, and were more likely to recommend the news articles by comparison to a baseline condition. While participants in the Annotation condition had higher engagement metrics, they recalled less information than participants assigned to our baseline. The findings illustrate a design trade-off and suggest new considerations for integrating discussion content into a news article. Additionally, we found that participant characteristics, such as existing knowledge about a topic, ability to retain information, and trust in the credibility of a news source, also play into what people remember about news. Finally, our findings offer insights about future moderation systems that assist with the integration of discussion into the reading experience of news online.

ACKNOWLEDGMENTS

The research was funded by the National Science Foundation (#2009003). The online experiment platform and study materials were created by Ziwen Zeng, Yiwen Hou, Salma El ghayate, and Seongwook Jang. The paper benefited from many thoughtful suggestions during the peer-review cycle. Additionally, Molly Q Feldman and Stephen MacNeil helped to clarify the study contributions. Finally, we want to thank the UC San Diego Design Lab, as the idea for the research emerged at a lab meeting during an energetic debate about whether online commenting has value.

REFERENCES

[1] Eytan Adar, Carolyn Gearig, Ayshwarya Balasubramanian, and Jessica Hullman. 2017. PersaLog: Personalization of News Article Content. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017). https://doi.org/10.1145/3025453

[2] Tanja Aitamurto. 2016. Crowdsourcing in Open Journalism: Benefits, challenges, and value creation. *The Routledge Companion to Digital Journalism Studies* (11 2016), 185–193. https://doi.org/10.4324/9781315713793-19

[3] Lucas Anastasiou and Anna De Liddo. 2021. Making Sense of Online Discussions: Can Automated Reports help? (5 2021), 1–7. https://doi.org/10.1145/3411763.3451815

[4] Ashley A. Anderson, Dominique Brossard, Dietram A. Scheufele, Michael A. Xenos, and Peter Ladwig. 2014. The "Nasty Effect": Online Incivility and Risk Perceptions of Emerging Technologies. *Journal of Computer-Mediated Communication* 19, 3 (4 2014), 373–387. https://doi.org/10.1111/jcc4.12009

[5] Alyssa Appelman and Paul Bolls. 2011. Article Recall, Credibility Lower with Grammar Errors. *Newspaper Research Journal* 32, 2 (3 2011), 50–62. https://doi.org/10.1177/073953291103200205

[6] Jaime Arguello, Brian S. Butler, Elisabeth Joyce, Robert Kraut, Kimberly S. Ling, Carolyn Rosé, and Xiaoqing Wang. 2006. Talk to me: foundations for successful individual-group interactions in online communities. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 959–968.

[7] Martina Balestra, Orit Shaer, Johanna Okerlund, Madeleine Ball, and Oded Nov. 2016. The effect of exposure to social annotation on online informed consent beliefs and behavior. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, San Francisco, 900–912.

[8] Amanda Baughan, Mingrui Ray Zhang, Raveena Rao, Kai Lukoff, Anastasia Schaadhardt, Lisa D. Butler, and Alexis Hiniker. 2022. "I Don't Even Remember What I Read": How Design Influences Dissociation on Social Media. *Conference on Human Factors in Computing Systems - Proceedings* (4 2022). https://doi.org/10.1145/3491102.3501899

[9] W. Lance Bennett and Steven Livingston. 2018. The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication* 33, 2 (4 2018), 122–139. https://doi.org/10.1177/0267323118760317

[10] John Bowers and Jonathan Zittrain. 2020. Answering Impossible Questions: Content Governance in an Age of Disinformation. *Harvard Kennedy School Misinformation Review* (1 2020). https://doi.org/10.37016/mr-2020-005

[11] Cameron Brick and Sander van der Linden. 2018. Yawning at the apocalypse. *The Psychologist* 5, 1 (2018), 30–35.

[12] Bernheim A J Brush, David Bargeron, Jonathan Grudin, Alan Borning, and Anoop Gupta. 2002. Supporting Interaction Outside of Class: Anchored Discussions vs. Discussion Boards. In *Proceedings of Computer Supported Collaborative Learning*. 425–434.

[13] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. CrossMod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (11 2019), 1–30. https://doi.org/10.1145/3359276

[14] Gina Masullo Chen and Yee Man Margaret Ng. 2016. Third-person perception of online comments: Civil ones persuade you more than me. *Computers in Human Behavior* 55 (2 2016), 736–742. https://doi.org/10.1016/j.chb.2015.10.014

[15] Kevin Coe, Kate Kenski, and Stephen A. Rains. 2014. Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. *Journal of Communication* 64, 4 (8 2014), 658–679. https://doi.org/10.1111/jcom.12104

[16] Matthew Conlen and Jeffrey Heer. 2018. IdylL: A markup language for authoring and publishing interactive articles on the web. *UIST 2018 - Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (10 2018), 977–989. https://doi.org/10.1145/3242587.3242600

[17] Lindsey Conlin and Chris Roberts. 2016. Presence of online reader comments lowers news site credibility. *Newspaper Research Journal* 37, 4 (12 2016), 365–376. https://doi.org/10.1177/0739532916677056

[18] Prabu David. 1996. Role of Imagery in Recall of Deviant News. *Journalism & Mass Communication Quarterly* 73, 4 (12 1996), 804–820. https://doi.org/10.1177/107769909607300404

[19] Prabu David. 1998. News Concreteness and Visual-Verbal Association Do News Pictures Narrow the Recall Gap Between Concrete and Abstract News? *Human Communication Research* 25, 2 (12 1998), 180–201. https://doi.org/10.1111/j.1468-2958.1998.tb00442.x

[20] Nicholas Diakopoulos. 2015. The editor's eye: Curation and comment relevance on the New York Times. In *CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing*. Association for Computing Machinery, Inc, New York, NY, USA, 1153–1157. https://doi.org/10.1145/2675133.2675160

[21] Nicholas Diakopoulos and Mor Naaman. 2011. Topicality, time, and sentiment in online news comments. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. 1405–1410.

[22] Leen d'Haenens, Nicholas Jankowski, and Ard Heuvelman. 2004. News in Online and Print Newspapers: Differences in Reader Consumption and Recall. *New Media & Society* 6, 3 (6 2004), 363–382. https://doi.org/10.1177/1461444804042520

[23] Justin Ellis. 2015. What happened after 7 news sites got rid of reader comments. *Nieman Journalism Lab* (9 2015).

[24] Dima Epstein and Gilly Leshed. 2016. The Magic Sauce: Practices of Facilitation in Online Policy Deliberation. *Journal of Public Deliberation* 12, 1 (2016).
[25] Bassey Etim. 2014. A Comment's Path to Publication. *The New York Times* (4 2014).
[26] William P. Eveland and Juliann Cortese. 2004. How Web Site Organization Influences Free Recall, Factual Knowledge, and Knowledge Structure Density. *Human Communication Research* 30, 2 (4 2004), 208–233. https://doi.org/10.1111/j.1468-2958.2004.tb00731.x
[27] Molly Q Feldman and Brian James McInnis. 2021. How We Write with Crowds. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (1 2021), 1–31. https://doi.org/10.1145/3432928
[28] Rolf Fredheim, Alfred Moore, and John Naughton. 2015. Anonymity and Online Commenting: The Broken Windows Effect and the End of Drive-by Commenting. In *Proceedings of the ACM Web Science Conference*. ACM, Oxford, United Kingdom, 11. https://doi.org/10.1145/2786451.2786459
[29] Tong Gao, Jessica Hullman, Eytan Adar, Brent Hecht, and Nicholas Diakopoulos. 2014. NewsViews: An automated pipeline for creating custom geovisualizations for news. *Conference on Human Factors in Computing Systems - Proceedings* (2014), 3005–3014. https://doi.org/10.1145/2556288.2557228
[30] James Grimmelmann. 2015. The Virtues of Moderation. *Yale Journal of Law and Technology* 17, 1 (2015), 42–109. https://heinonline.org/HOL/P?h=hein.journals/yjolt17&i=42
[31] Nir Grinberg. 2018. Identifying modes of user engagement with online news and their relationship to information gain in text. In *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*. Association for Computing Machinery, Inc, 1745–1754. https://doi.org/10.1145/3178876.3186180
[32] Ali Gürkan, Luca Iandoli, Mark Klein, and Giuseppe Zollo. 2010. Mediating debate through on-line large-scale argumentation: Evidence from the field. *Information Sciences* 180, 19 (2010), 3686–3702. https://doi.org/10.1016/j.ins.2010.06.011
[33] Daniel Halpern and Jennifer Gibbs. 2013. Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. *Computers in Human Behavior* 29, 3 (2013), 1159–1168. https://doi.org/10.1016/j.chb.2012.10.008
[34] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world? , 61–83 pages. https://doi.org/10.1017/S0140525X0999152X
[35] Jessica Hullman, Nicholas Diakopoulos, and Eytan Adar. 2013. Contextifier: Automatic generation of annotated stock visualizations. *Conference on Human Factors in Computing Systems - Proceedings* (2013), 2707–2716. https://doi.org/10.1145/2470654.2481374
[36] Jessica Hullman, Nicholas Diakopoulos, Elaheh Momeni, and Eytan Adar. 2015. Content, context, and critique: Commenting on a data visualization Blog. In *CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing*. Association for Computing Machinery, Inc, 1170–1175. https://doi.org/10.1145/2675133.2675207
[37] Luca Iandoli, Ivana Quinto, Paolo Spada, Mark Klein, and Raffaele Calabretta. 2017. Supporting argumentation in online political debate: Evidence from an experiment of collective deliberation. *New Media & Society* 20, 4 (2017), 1320–1341. https://doi.org/1461444817691509
[38] Mi Rosie Jahng. 2018. From reading comments to seeking news: exposure to disagreements from online comments and the need for opinion-challenging news. *Journal of Information Technology & Politics* 15, 2 (4 2018), 142–154. https://doi.org/10.1080/19331681.2018.1449702
[39] Elizabeth Jensen. 2016. NPR Website To Get Rid Of Comments. *National Public Radio (NPR)* (8 2016).
[40] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction* 26, 5 (7 2019). https://doi.org/10.1145/3338243
[41] Quentin Jones and Sheizaf Rafaeli. 1999. User population and user contributions to virtual publics. *Proceedings of the international ACM SIGGROUP conference on Supporting group work - GROUP '99* (1999), 239–248. https://doi.org/10.1145/320297.320325
[42] Dan M. Kahan, Ellen Peters, Maggie Wittlin, Paul Slovic, Lisa Larrimore Ouellette, Donald Braman, and Gregory Mandel. 2012. The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change* 2, 10 (10 2012), 732–735. https://doi.org/10.1038/nclimate1547
[43] Tobias Kauer, Marian Dörk, Arran Ridley, and Benjamin Bach. 2021. The public life of data: Investigating reactions to visualizations on reddit. *Conference on Human Factors in Computing Systems - Proceedings* (5 2021). https://doi.org/10.1145/3411764.3445720
[44] Juho Kim, Eun-Young Ko, Jonghyuk Jung, Chang Won Lee, Nam Wook Kim, and Jihee Kim. 2015. Factful: Engaging taxpayers in the public discussion of a government budget. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2843–2852.

[45] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012. Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. 265–274.

[46] Travis Kriplean, Michael Toomim, Jonathan Morgan, Alan Borning, and Andrew J. Ko. 2012. Is this what you meant? Promoting listening on the web with Reflect. *Conference on Human Factors in Computing Systems - Proceedings* (2012), 1559–1568. https://doi.org/10.1145/2207676.2208621

[47] Suzanne LaBarre. 2013. Why We're Shutting Off Our Comments. *Popular Science* (9 2013).

[48] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 543–550.

[49] Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik Johnston. 2014. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly* 31, 2 (4 2014), 317–326. https://doi.org/10.1016/j.giq.2013.11.005

[50] Seungae Lee, Lucy Atkinson, and Yoon Hi Sung. 2020. Online bandwagon effects: Quantitative versus qualitative cues in online comments sections. *New Media & Society* (10 2020), 146144482096518. https://doi.org/10.1177/1461444820965187

[51] Jose A. León. 1997. The effects of headlines and summaries on news comprehension and recall. *Reading and Writing* 9, 2 (1997), 85–106. https://doi.org/10.1023/A:1007928221187

[52] Stephan Lewandowsky and Klaus Oberauer. 2016. Motivated Rejection of Science. *Current Directions in Psychological Science* 25, 4 (8 2016), 217–222. https://doi.org/10.1177/0963721416654436

[53] John Michael Linacre. 1994. Sample Size and Item Calibration or Person Measure Stability. *Rasch Measurement Transactions* 7, 4 (1994), 328. https://www.rasch.org/rmt/rmt74m.htm

[54] Kat Long. 2017. Meet The New York Times's Super-Commenters. *The New York Times* (11 2017).

[55] Brian McInnis, Leah Ajmani, Lu Sun, Yiwen Hou, Ziwen Zeng, and Steven Dow. 2021. Reporting the Community Beat: Practices for Moderating Online Discussion at a News Website. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (10 2021), 1–25. https://doi.org/10.1145/3476074

[56] Brian McInnis, Dan Cosley, Eric P.S. Baumer, and Gilly Leshed. 2018. Effects of comment curation and opposition on coherence in online policy discussion. In *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work*. https://doi.org/10.1145/3148330.3148348

[57] Brian McInnis, Gilly Leshed, and Dan Cosley. 2018. Crafting Policy Discussion Prompts as a Task for Newcomers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (11 2018), 1–23. https://doi.org/10.1145/3274390

[58] Brian McInnis, Elizabeth Murnane, Dima Epstein, Dan Cosley, and Gilly Leshed. 2016. One and Done: Factors affecting one-time contributors to ad-hoc online communities. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, Vol. 27. https://doi.org/10.1145/2818048.2820075

[59] Brian McInnis, Lu Sun, Jungwon Shin, and Steven P. Dow. 2020. Rare, but Valuable: Understanding Data-centered Talk in News Website Comment Sections. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020). https://doi.org/10.1145/3415245

[60] Alfred Moore. 2012. Following from the front: Theorizing deliberative facilitation. *Critical Policy Studies* 6, 2 (7 2012), 146–162. https://doi.org/10.1080/19460171.2012.689735

[61] Heather Moore. 2008. Talk to the Newsroom: Community Editor. *The New York Times* (8 2008).

[62] Kevin K. Nam and Mark S. Ackerman. 2007. Arkose: Reusing informal information from online discussions. In *GROUP'07 - Proceedings of the 2007 International ACM Conference on Supporting Group Work*. ACM Press, New York, New York, USA, 137–146. https://doi.org/10.1145/1316624.1316644

[63] Raymond S. Nickerson. 1998. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* 2, 2 (6 1998), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

[64] Michaël Opgenhaffen. 2011. The Impact of Online News Features on Learning from News: A Knowledge Experiment. *International Journal of Internet Science* 6, 1 (2011), 8–28.

[65] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting comment moderators in identifying high quality online news comments. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, 1114–1125. https://doi.org/10.1145/2858036.2858389

[66] Walter C. Parker. 2006. Public Discourses in Schools: Purposes, Problems, Possibilities. *Educational Researcher* 35, 8 (11 2006), 11–18. https://doi.org/10.3102/0013189X035008011

[67] Kendra Pierre-Louis. 2018. Greenhouse Gas Emissions Accelerate Like a 'Speeding Freight Train' in 2018. https://www.nytimes.com/2018/12/05/climate/greenhouse-gas-emissions-2018.html

[68] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological Review* 106, 4 (1999), 643–675. https://doi.org/10.1037/0033-295X.106.4.643

[69] Ivanka Pjesivac, Nicholas Geidner, and Jaclyn Cameron. 2018. Social credibility online: The role of online comments in assessing news article credibility. *Newspaper Research Journal* 39, 1 (3 2018), 18–31. https://doi.org/10.1177/0739532918761065

[70] Brad Plumer. 2019. As Coal Fades in the U.S., Natural Gas Becomes the Climate Battleground. https://www.nytimes.com/2019/06/26/climate/natural-gas-renewables-fight.html
[71] Cynthia R. Farina, Dima Epstein, Josiah B. Heidt, and Mary J. Newhart. 2013. RegulationRoom: Getting more, better civic participation in complex government policymaking. *Transforming Government: People, Process and Policy* 7, 4 (2013), 501–516.
[72] Steven P. Reise, Andrew T. Ainsworth, and Mark G. Haviland. 2005. Item Response Theory: Fundamentals, Applications, and Promise in Psychological Research. *Current Directions in Psychological Science* 14, 2 (4 2005), 95–101. https://www.jstor.org/stable/20182996
[73] Dimitris Rizopoulos. 2006. ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses. *Journal of Statistical Software* 17, 5 (2006), 1–25. http://www.jstatsoft.org/v17/i05/
[74] Ian Rowe. 2015. Deliberation 2.0: Comparing the Deliberative Quality of Online News User Comments Across Platforms. *Journal of Broadcasting & Electronic Media* 59, 4 (10 2015), 539–555. https://doi.org/10.1080/08838151.2015.1093482
[75] Dietram A. Scheufele and Nicole M. Krause. 2019. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences of the United States of America* 116, 16 (4 2019), 7662–7669. https://doi.org/10.1073/pnas.1805871115
[76] Elisa Shearer. 2021. 86% of Americans get news online from smartphone, computer or tablet. https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/
[77] Elisa Shearer and Amy Mitchell. 2021. *News Use Across Social Media Platforms in 2020 | Pew Research Center*. Technical Report. Pew Research Center, Pennsylvania, PA. https://www.journalism.org/2021/01/12/news-use-across-social-media-platforms-in-2020/
[78] Klaas Sijtsma and Brian W. Junker. 2006. Item Response Theory: Past Performance, Present Developments, and Future Expectations. *Behaviormetrika* 33, 1 (6 2006), 75–102. https://doi.org/10.2333/BHMK.33.75
[79] Jonathan Smith. 2016. We're Getting Rid of Comments on VICE.com. *Vice News* (12 2016).
[80] Nina Springer, Ines Engelmann, and Christian Pfaffinger. 2015. User comments: motives and inhibitors to write and read. *Information, Communication & Society* 18, 7 (7 2015), 798–815. https://doi.org/10.1080/1369118X.2014.997268
[81] Kim Strandberg and Janne Berg. 2013. Online Newspapers' Readers' Comments - Democratic Conversation Platforms or Virtual Soapboxes? *Comunicação e Sociedade* 23, 1 (2013), 132. https://doi.org/10.17231/comsoc.23(2013).1618
[82] Natalie Jomini Stroud, Joshua M. Scacco, Ashley Muddiman, and Alexander L. Curry. 2015. Changing Deliberative Norms on News Organizations' Facebook Sites. *Journal of Computer-Mediated Communication* 20, 2 (3 2015), 188–203. https://doi.org/10.1111/jcc4.12104
[83] Leona Yi-Fan Su, Dietram A Scheufele, Dominique Brossard, and Michael A Xenos. 2020. Political and personality predispositions and topical contexts matter: Effects of uncivil comments on science news engagement intentions. *New Media & Society* (2 2020), 146144482090436. https://doi.org/10.1177/1461444820904365
[84] Abhay Sukumaran, Stephanie Vezich, Melanie McHugh, and Clifford Nass. 2011. Normative influences on thoughtful online participation. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*. ACM Press, New York, New York, USA, 3401. https://doi.org/10.1145/1978942.1979450
[85] Margaret Sullivan. 2012. Questions and Answers on How The Times Handles Online Comments From Readers. https://publiceditor.blogs.nytimes.com/2012/10/15/questions-and-answers-on-how-the-times-handles-online-comments-from-readers/
[86] Margaret Sullivan. 2015. Change Needed for Commenting That Favors the 'Verified'. https://publiceditor.blogs.nytimes.com/2015/11/20/change-needed-for-commenting-that-favors-the-verified/?searchResultPosition=6
[87] Nicole Sultanum, Zoya Bylinskii, and Zhicheng Liu. 2021. Leveraging text-chart links to support authoring of data-driven articles with vizflow. *Conference on Human Factors in Computing Systems - Proceedings* 17, 21 (5 2021). https://doi.org/10.1145/3411764.3445354
[88] David Tewksbury and Scott L. Althaus. 2000. Differences in Knowledge Acquisition among Readers of the Paper and Online Versions of a National Newspaper. *Journalism & Mass Communication Quarterly* 77, 3 (9 2000), 457–479. https://doi.org/10.1177/107769900007700301
[89] Perry W. Thorndyke. 1979. Knowledge Acquisition from Newspaper Stories. *Discourse Processes* 2, 2 (4 1979), 95–112. https://doi.org/10.1080/01638537909544457
[90] Fernanda B. Viegas, Martin Wattenberg, Frank Van Ham, Jesse Kriss, and Matt McKeon. 2007. Many Eyes: A site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (11 2007), 1121–1128. https://doi.org/10.1109/TVCG.2007.70577
[91] David Manning White. 1950. The "Gate Keeper": A Case Study in the Selection of News. *Journalism Quarterly* 27, 4 (9 1950), 383–390. https://doi.org/10.1177/107769905002700403
[92] Wesley Willett, Jeffrey Heer, Joseph M. Hellerstein, and Maneesh Agrawala. 2011. CommentSpace: Structured support for collaborative visual analysis. In *Conference on Human Factors in Computing Systems - Proceedings*. 3131–3140. https://doi.org/10.1145/1978942.1979407

[93] Gavin Wood, Kiel Long, Tom Feltwell, Scarlett Rowland, Phillip Brooker, Jamie Mahoney, John Vines, Julie Barnett, and Shaun Lawson. 2018. Rethinking engagement with online news through social and visual co-annotation. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. 2018-April. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174150

[94] Benjamin Wright and Mark Stone. 1979. Best test design. *Measurement and Statistics* (1 1979). https://research.acer.edu.au/measurement/1

[95] Cristian Zanon, Claudio S. Hutz, Hanwook (Henry) Yoo, and Ronald K. Hambleton. 2016. An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica* 29, 1 (6 2016). https://doi.org/10.1186/S41155-016-0040-X

[96] Amy X. Zhang, Lea Verou, and David Karger. 2017. Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 2082–2096.

[97] Marc Ziegele, Oliver Quiring, Katharina Esau, and Dennis Friess. 2018. Linking News Value Theory With Online Deliberation: How News Factors and Illustration Factors in News Articles Affect the Deliberative Quality of User Discussions in SNS' Comment Sections:. *https://doi.org/10.1177/0093650218797884* 47, 6 (9 2018), 860–890. https://doi.org/10.1177/0093650218797884

A ARTICLE TEXT AND CLOSED-RESPONSE QUESTIONS

A.1 Article 1: Greenhouse gas emissions

Note that the text presented in the news website stage was a modified version of the article "Greenhouse Gas Emissions Accelerate Like a *Speeding Freight Train* in 2018," written by Kendra Pierre-Louis and published by the New York Times on Dec. 5, 2018. Modifications involved reducing the number of words and changing the data units for consistency. The location of the data visualization in the article text used for this study reflects the relative position of the visualization included within the original article.

A.1.1 Article text. Greenhouse gas emissions worldwide are growing at an accelerating pace this year, researchers said Wednesday. Scientists described the quickening rate of carbon dioxide emissions in stark terms, comparing it to a "speeding freight train" and laying part of the blame on an unexpected surge in the appetite for oil as people around the world not only buy more cars but also drive them farther than in the past — more than offsetting any gains from the spread of electric vehicles. "We thought oil use had peaked in the U.S. and Europe 15 years ago," Dr. Jackson said. "The cheap gasoline prices, bigger cars and people driving more miles are boosting oil use at rates that none of us expected."

Worldwide, carbon emissions are expected to increase by 2.7 percent in 2018, according to the new research, which was published by the Global Carbon Project. The recent rise in global emissions, combined with other factors such as natural temperature fluctuations, could bring those dire consequences a decade sooner, by 2030. The analysis found that the world is on pace to release a record 37.1 gigatons of planet-warming emissions in 2018, led in large part by China, the United States and India. That is roughly 100,000 times the weight of the Empire State Building.

Reducing carbon emissions is central to stopping global warming. Three years ago nearly 200 nations hammered out the Paris Agreement with a goal of holding warming below 2 degrees Celsius over preindustrial levels. For the Paris goals to be met, scientists say, global emissions from power plants, factories, cars and trucks, as well as those from deforestation, would need to swiftly begin declining to zero.

President Trump, however, has vowed to pull the United States out of the accord and has moved to roll back Obama-era regulations designed to limit emissions from vehicle tailpipes and power-plant smokestacks. On Tuesday he wrote on Twitter that the Paris Agreement was "fatally flawed" because its system of voluntary pledges let other countries off the hook, adding that "American taxpayers — and American workers — shouldn't pay to clean up others countries' pollution."

An American withdrawal would represent a serious blow to the pact. The United States, one of the wealthiest nations in the world, is responsible for a third of all human-caused carbon emissions to date, more than any other country. United States emissions are expected to rise this year after several years of declines. Dr. Jackson attributed part of the increase this year to a colder-than-normal winter in some parts of the country and a hotter summer in other parts, which inflated demand for heating and cooling.

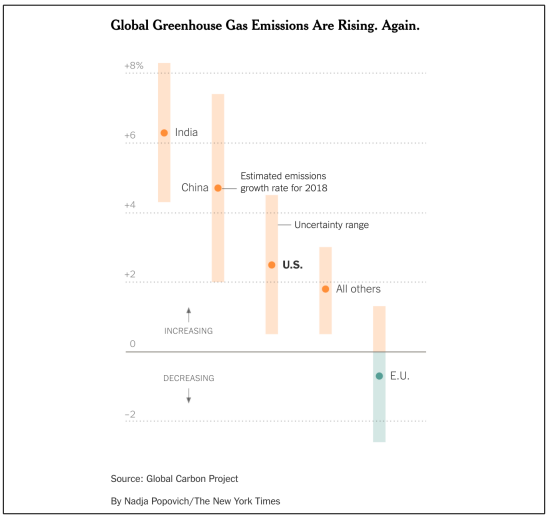

Fig. 6. *Global Greenhouse Gas Emissions Are Rising Again*, Source: Pierre-Louis [67]

China is now the largest emitter of heat-trapping gases. China's emissions are projected to rise 4.7 percent in 2018, the report said. The country is stimulating manufacturing to counterbalance its slowing economy, allowing more coal-based manufacturing that it had avoided in the past, Dr. Jackson said. China is investing heavily in renewable energy, but it is also building new coal-fired power plants at home and planning others in new markets such as sub-Saharan Africa.

In India, a projected emissions increase of 6.3 percent is linked to the country's effort to provide electricity to 300 million people who currently lack it. While many countries have seen an increase in emissions, emissions in the European Union are expected to decline by 0.7 percent this year, but the bloc averaged a 2.2 percent annual decline last decade. The weaker number was attributed to an increase in oil consumption linked to transportation.

The earth has already warmed by 1 degree Celsius above preindustrial levels. The effects of that warming could be seen this year across the United States, from the wildfires that raged in California to the hurricanes that buffeted the Southeast. The warming climate did not cause the individual weather events, but scientists say it created conditions that increased their likelihood. Last year, extreme weather disasters cost the United States a record $306 billion.

Dr. Jackson said the new report was "not good news," but added that it still contained "some glimmers of hope," particularly about air pollution associated with the burning of coal for fuel. "Coal use has dropped 40 percent in the United States, replaced by natural gas and renewables," he said. "That's saving lives as well as helping the climate problem."

A.1.2 Item Response Theory Analysis. Table 8 presents the item level position relative to the data visualization as well as statistics based on the Rasch Model: Difficulty, probability correct, and item-level Goodness-of-Fit.

Article 1: *Greenhouse gas emissions*				*Goodness-of-Fit*	
Item	Difficulty	Position	Pr(correct)	χ^2	Pr(>χ^2)
Q11	-0.44	P8-Below	0.61	19.80	0.05 ·
Q10	-0.04	P3-Above	0.51	12.73	0.32
Q3	0.05	P5-Above	0.49	14.75	0.21
Q2	0.36	P2-Above	0.41	12.91	0.32
Q4	0.39	P6-Below	0.40	5.90	0.85
Q1	0.60	P1-Above	0.35	9.73	0.57
Q7	0.66	P7-Below	0.34	12.36	0.29
Q5	0.69	P6-Below	0.33	6.75	0.83
Q9	1.11	P8-Below	0.25	6.11	0.83
Q6	1.51	P6-Below	0.18	5.81	0.84
Q8	1.51	P2-Above	0.18	18.08	0.05 ·
Q12	1.77	P9-Below	0.15	11.21	0.30
N: 167 (assigned Article 2 in the news website stage)					

Table 8. Item-level statistics for a Rasch Model of participant ability to recall information from the text of Article 1: Greenhouse gas emissions, sorted by item difficulty (from low to high). The table includes information about the item paragraph number (P#) and position relative to a data visualization (i.e., Above, Below) as well as the item difficulty, probability, and Fit. The Rasch Model is based on item response data collected during the post-survey for participants randomly assigned Article 2 during the online discussion simulation. Item-level Goodness-of-Fit indicates that Q11 and Q8 may not fit the model well; however, the overall model fits the data reasonably well based on a χ^2 test of 200 bootstrapped data-sets (p = 0.07).

A.1.3 Closed-response Knowledge Retention Questions.

Q1. The article identifies numerous potential causes for the unexpected 2018 surge in the appetite for oil. What option, if any, was NOT mentioned in the article?
- *Answer*: People are consuming more meat.

Q2. The article described the quickening rate of carbon emissions in stark terms, comparing it to a speeding freight train, likely to increase by what percentage in 2018?
- *Answer*: 2.0-2.99%.

Q3. Based on your memory of the article, fill in the blank: "The United States, one of the wealthiest nations in the world, is responsible for [BLANK] of all human-caused carbon emissions to date, more than any other country"
- *Answer*: 1/3 (third).

Q4. As of 2018, what country was the largest emitter of heat-trapping gases?
- *Answer*: China.

Q5. According to the article, China's greenhouse gas emissions are projected to increase by what percentage in 2018?
- *Answer*: 3.0% or more.

Q6. What reason does the article offer to explain China's recent rise in greenhouse gas emissions?
- *Answer*: More coal-based manufacturing.

Q7. What reasons does the article offer to explain India's recent rise in greenhouse gas emissions?
- *Answer*: Efforts to provide electricity to people who lack access.

Q8. Data presented in the article suggests that the world is on pace to release a record 37.1 gigatons of planet-warming emissions in 2018, led in large part by China, the United States and India. What is 37.1 gigatons relative to the weight of the Empire State Building?

- *Answer*: About 100,000 times.

Q9. According to the article, the earth has already warmed by how much above pre-industrial levels (degrees Celsius)?
- *Answer*: 1 degree Celsius.

Q10. The Paris Agreement is an international partnership of countries with the common goal to hold the earth's global temperature at what temperature above pre-industrial levels?
- *Answer*: 2 degrees Celsius.

Q11. According to the article, approximately how much did extreme disasters cost the United States in 2018?
- *Answer*: $300 billion.

Q12. The article offers some glimmers of hope for curbing greenhouse gas emissions, for example, coal use has dropped, being replaced by natural gas and renewable sources, by what percentage in the United States?
- *Answer*: 40%.

A.2 Article 2: U.S. Energy generation

Note that the text presented in the news website stage was a modified version of the article "As Coal Fades in the U.S., Natural Gas Becomes the Climate Battleground," written by Brad Plumer and published by the New York Times on Jun. 26, 2019. Modifications involved reducing the number of words and changing the data units for consistency. The location of the data visualization in the article text used for this study reflects the relative position of the visualization included within the original article.

A.2.1 Article text. Some large utilities are now planning to sharply cut their coal and gas use in favor of clean and abundant wind and solar power, which have steadily fallen in cost. But in the Southeast and other regions, natural gas continues to dominate, because of its reliability and low prices driven by the fracking boom. Nationwide, energy companies plan to add at least 150 new gas plants and thousands of miles of pipelines in the years ahead. A rush to build gas-fired plants, even though they emit only half as much carbon pollution as coal, has the potential to lock in decades of new fossil-fuel use right as scientists say emissions need to fall drastically by midcentury to avert the worst impacts of global warming.

In some states, policymakers are now pushing to leave gas behind to meet ambitious climate goals. Since 2005, most power companies have lowered their carbon dioxide emissions significantly, in large part by shifting from coal to gas. Coal plants have become uncompetitive with other kinds of energy generation in much of the country, despite the Trump administration's efforts to save them by rolling back federal pollution regulations.

Last fall, in North and South Carolina, a pair of utility companies filed plans with state regulators to continue retiring coal plants and largely replace them with more than 9,500 megawatts of new natural gas capacity by 2033. The utilities also plan to add a smaller amount of solar capacity, about 3,600 megawatts, over the same time frame. One challenge with using more solar power is finding a way to supply electricity when the sun isn't shining. Although large lithium-ion batteries can store solar energy for less-sunny hours, batteries still haven't reached the point where they're as cheap or effective as gas power, which can run at all hours.

Opponents of the plan have urged state regulators to push the utility companies to reconsider. These disputes are popping up in states around the country. Over the last decade, groups like the Sierra Club have tried to persuade utilities and regulators that they could save money by retiring coal and shifting to a cleaner mix of gas and renewables. Now they're running the same playbook

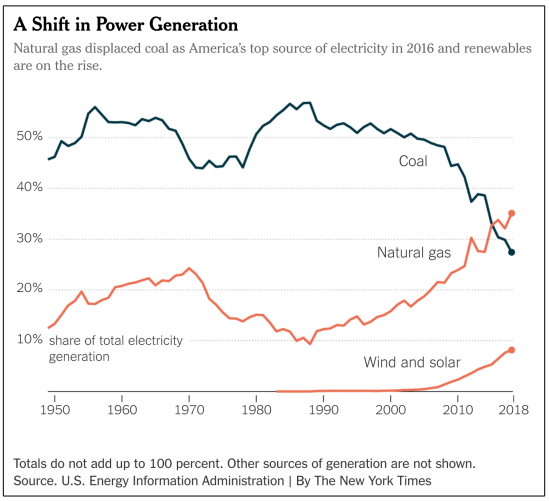

Fig. 7. *A Shift in Power Generation*, Source: Plumer [70]

against gas, arguing that the costs of wind, solar and batteries have declined so drastically that it's time to stop building new gas plants, too. But, so far the results have been mixed.

However, renewable energy sources are gaining in several states. Last year in Indiana, the Northern Indiana Public Service Company, or Nipsco, opened bidding to outside energy developers and found that adding a mix of wind, solar and batteries would be cheaper than building a new gas plant to replace its retiring coal units. (The company will keep its older gas plants online to fill in gaps when wind and solar aren't available.) Doing so, the utility estimated, would reduce its emissions 90 percent below 2005 levels by 2030.

Officials at Nipsco noted that the utility had advantages that others might not have: Its territory sits near land that's ripe for wind development, making it easier to build new turbines close by without the need for lots of costly new transmission lines. In short, the right mix of energy sources might look very different for utility companies just 100 miles away.

Indeed, things look very different nearby in the vast regional grid known as PJM that serves 65 million people from Ohio to New Jersey. There power plants compete in a largely deregulated market and companies are expected to build over 10,000 megawatts of new gas plants by 2024 to take advantage of cheap natural gas from the nearby fracking boom in Ohio, Pennsylvania and West Virginia.

State legislatures are increasingly weighing in on which energy sources get built. To date, 29 states have enacted laws that require their utilities to get a certain fraction of their power from wind and solar. Now, some states are going further. Over the past year, California, Colorado, Maine, Nevada, New Mexico, New York and Washington have all passed laws aimed at getting 100 percent of their electricity from carbon-free sources by midcentury, which would eventually mean phasing out conventional gas plants.

Getting to 100 percent carbon-free power will likely require new technology that can supplant natural gas as a cost-effective backup fuel. Some possibilities include burning clean hydrogen instead of gas in power plants, developing techniques that enable carbon produced by gas plants to be captured and stored underground, advanced nuclear power or the invention of new energy storage techniques. Perfecting that technology would likely require big new investments in research and support from policymakers.

A.2.2 Item Response Theory Analysis. Table 9 presents the item level position relative to the data visualization as well as statistics based on the Rasch Model: Difficulty, probability correct, and item-level Goodness-of-Fit.

Article 2: *U.S. Energy generation*				*Goodness-of-Fit*	
Item	Difficulty	Position	Pr(correct)	χ^2	Pr(>χ^2)
Q4	-1.37	P3-Above	0.80	7.15	0.73
Q3	-0.97	P3-Above	0.73	19.82	0.01 *
Q6	-0.59	P9-Below	0.64	6.67	0.79
Q9	-0.53	P8-Below	0.63	12.12	0.38
Q7	-0.35	P1-Above	0.59	16.03	0.12
Q1	-0.20	P1-Above	0.55	30.34	0.01 *
Q5	0.12	P5-Below	0.47	6.80	0.83
Q2	0.15	P2-Above	0.46	3.22	0.99
Q10	0.20	P3-Above	0.45	9.61	0.61
Q11	0.29	P4-Below	0.43	12.59	0.37
Q12	0.77	P8-Below	0.32	10.61	0.55
Q8	1.03	P7-Below	0.26	12.26	0.43
N: 169 (assigned Article 1 in the news website stage)					

Table 9. Item-level statistics for a Rasch Model of participant ability to recall information from the text of Article 2: U.S. Energy generation, sorted by item difficulty (from low to high). The table includes information about the item paragraph number (P#) and position relative to a data visualization (i.e., Above, Below) as well as the item difficulty, probability, and Fit. The Rasch Model is based on item response data collected during the post-survey for participants randomly assigned Article 1 during the online discussion simulation. Item-level Goodness-of-Fit indicates that Q3 and Q1 may not fit the model well; however, the overall model fits the data reasonably well based on a χ^2 test of 200 bootstrapped data-sets (p = 0.08).

A.2.3 Closed-response Knowledge Retention Questions.

Q1. In the Southeast and other regions, natural gas has continued to dominate the energy sector. Nationwide, energy companies plan to add at least how many new gas plants in the next several years?
- *Answer*: 150.

Q2. By shifting from coal to gas many power companies were able to lower their carbon emissions significantly. When did this shift begin?
- *Answer*: 2005.

Q3. What units of energy does the article use to compare the capacity of each energy source?
- *Answer*: Megawatts.

Q4. Why is gas power still more effective than solar power, according to the article?
- *Answer*: All of the above.

Q5. The article offers several examples of how utility companies might integrate more renewable energy sources in the short-term, such as adopting a mix of gas and renewable sources. What other recommendation did the article offer?
- *Answer*: All of the above.

Q6. Moving to a 100% carbon-free power plan would require new technology that can supplant natural gas as a cost-effective backup fuel, such as burning clean hydrogen. What new technologies and strategies does the article discuss as possibilities?

- *Answer*: All of the above.

Q7. According to the article why is natural gas still dominating in the Southeast region?
- *Answer*: All of the above.

Q8. There has been a fracking boom in Ohio, Pennsylvania, and West Virginia. According to the article, approximately how many people does the PJM utility company serve?
- *Answer*: More than 60 million.

Q9. State legislatures are increasingly weighing in on which energy sources get built. According to the article, how many states have enacted laws that require their utilities to get a certain fraction of their power from wind and solar?
- *Answer*: 20-29 states.

Q10. According to the article, which source of energy is currently the most cost effective?
- *Answer*: Natural gas.

Q11. Organizations like the Sierra Club have advocated for a reduction in coal-based dependence, by encouraging utility companies to adopt what strategies?
- *Answer*: A mix of gas and renewables.

Q12. According to the article, which state has NOT passed laws aimed at getting 100% of their electricity from carbon-free sources by midcentury?
- *Answer*: None of the above.

Received April 2021; revised November 2021; accepted March 2022