Combating Problematic Information Online with Dual Process Cognitive Affordances

Md Momen Bhuiyan

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Applications

Sang Won Lee, Co-chair
Tanushree Mitra, Co-chair
Kurt Luther
Michael A. Horning
Nitesh Goyal

June 20, 2023
Blacksburg, Virginia

Keywords: automatic affordance, reflective affordance, design, information consumption, misinformation, filter bubble

Combating Problematic Information Online with Dual Process Cognitive Affordances

Md Momen Bhuiyan

(ABSTRACT)

Dual process theories of mind have been developed over the last decades to posit that humans use heuristics or mental shortcuts (automatic) and analytical (reflective) reasoning while consuming information. Can such theories be used to support users' information consumption in the presence of problematic content in online spaces? To answer, I merge these theories with the idea of affordances from HCI to into the concept of dual process cognitive affordances, consisting of automatic affordance and reflective affordance. Using this concept, I built and tested a set of systems to address two categories of online problematic content: misinformation and filter bubbles. In the first system, NudgeCred, I use cognitive heuristics from the MAIN model to design automatic affordances for better credibility assessment of news tweets from mainstream and misinformative sources. In TransparencyCue, I show the promise of value-centered automatic affordance design inside news articles differentiating content quality. To encourage information consumption outside their ideological filter bubble, in NewsComp, I use comparative annotation to design reflective affordances that enable active engagement with stories from opposing-leaning sources. In OtherTube, I use parasocial interaction, that is, experiencing information feed through the eyes of someone else, to design a reflective affordance that enables recognition of filter bubbles in their YouTube recommendation feeds. Each system shows various degrees of success and outlines considerations in cognitive affordances design. Overall, this thesis showcases the utility of

design strategies centered on dual process information cognition model of human mind to combat problematic information space.

Combating Problematic Information Online with Dual Process Cognitive Affordances

Md Momen Bhuiyan

(GENERAL AUDIENCE ABSTRACT)

Over the last several decades, billions of users have moved to the internet for everyday information gathering, allowing information flow around the globe at a massive scale. This flow is managed by algorithms personalized to each users' need, creating a complicated trio of producer-algorithm-consumer. This has resulted in some unforeseen challenges. Bad information producers takes the advantage of system to promote problematic content, such as, false information, termed as misinformation. Personalized algorithms have created filters of what people see oftentimes isolating them from diverse perspectives of information, creating a distorted perception of reality. Augmenting the online technology infrastructure to combat these challenges has become crucial and the overall goal of this thesis. Cognitive psychologists theorize that two cognitive processes are at play when people consume information, also known as dual process theories. Can we design new tools to combat these challenges by tapping into each of these processes? In this thesis, I answer this question through a series of studies. In each of these studies, I combine this theory from psychology with design guides from Human-Computer Interaction to design socio-technical design. I evaluated each of these systems through controlled experimentation. The result of these studies informs ways we can capitalize on users' information processing mechanism to combat various types of problematic information online.

Acknowledgments

This thesis would not be possible without the support of so many unbelievably talented people. You have helped me be a better person and a better learner. Therefore, I dedicate this thesis to you:

- My mom, my late dad, and my brother, who have been always there for me
- My advisors Sang Won Lee and Tanushree Mitra, who supported and inspired me both in good times and bad
- My dissertation committee members: Mike Horning, for helping me make sense of newsmaking from the journalistic side; Kurt Luther, for piquing my interest from my first HCI course with you and your quick feedback whenever I needed; and Nitesh Goyal, for your encouragement and bringing in the industry perspective in our discussion. Your feedback during this thesis has been invaluable
- All the lab members from both Social Computing Lab at UW and Echolab. Notably, Mattia Samory, an incredible mentor; Shruti Phadke, although we didn't collaborate, I received your help more times than I deserved; Prerna Juneja, for your amazing attention to detail; Vartan Kesiz Abnousi, for all the late-night laughs before deadlines; Carlos Bautista Isaza, for softening my workload; the rest of the graduate students: Brian, Deepika, Kristen, Neelesh, Saloni, Andy, Amber, Daniel(E), Daniel(D), Danny, Dash, Emily, Muskan, Priya, Robin, Rodney, Ruipu, Tausif, and Viral; and the undergraduate team for all your help with my research
- Mentors and Friends from the HCI community, Aakash Gautam, Lindah Kotut, Jane Im, Tianyi Li, Shuo Niu, So Yeon, Sukrit Venkatagiri, and Yan Chen

- My mentees Donghan, Marx, Jelson, Liling, Stephen, Stephanie, and Hayoung for being such great mentees
- My collaborators in and outside VT: Connie Moon Sehat from Credibility Coalition, Amy Zhang from MIT, Kelsey Vick, from department of Communication
- All my friends, classmates, seniors and juniors here at Virginia Tech
- Finally, all the participants of my studies, from the industry, online and the local community

Contents

List of Figures xvii

List of Tables xxiii

1 Introduction 1

1.1 Problematic Information Online: Misinformation and Filter Bubbles 2

1.2 Affordance 4

1.3 Dual Process Cognitive Affordances 5

1.4 Designing Affordances against Misinformation and Filter Bubbles 7

1.5 Designing Automatic Affordances to Distinguish Content Credibility 8

1.5.1 NudgeCred: Supporting News Credibility Assessment on SocialMedia Through Nudges 9

1.5.2 TransparencyCue: Designing Transparency Cues in Online News Platforms to Promote Trust 10

1.6 Designing Reflective Affordances to Raise Awareness on Filter Bubble 12

1.6.1 NewsComp: Facilitating Diverse News Reading through Comparative Annotation 12

1.6.2 OtherTube: Facilitating Content Discovery and Reflection By Exchanging YouTube Recommendations with Strangers 14

1.7 Thesis Contributions 15

1.8 Thesis Outline 16

2 Background and Literature Review **18**

2.1 Problematic Information Typology: Misinformation and Filter bubble 18

2.2 Dual Process Theories on Information Processing 20

2.3 Affordances in HCI 22

2.4 Designing Against Problematic Information 23

2.4.1 Designing for Distinguishing Information Credibility 23

2.4.2 Designing for Reflection on Algorithmic Filters 25

3 NudgeCred: Supporting News Credibility Assessment on Social Media Through Nudges **27**

3.1 Introduction 27

3.2 Related Work 31

3.2.1 Nudges to Steer Human Behavior 31

3.2.2 Heuristic Cues for Credibility 31

3.2.3 Factors Affecting Credibility Perception: Partisanship, Attitude towards Politics, and Media 33

3.3 Formative Study 34

3.4 Designing NudgeCred 36

3.4.1 Design Guides . . . 36
3.4.2 Outlining the Design . . . 39
3.5 Study 1: Evaluating Impact on Perceptions of Credibility in a Controlled Setting . . . 43
3.5.1 Method (Study 1) . . . 44
3.5.2 Results (Study 1) . . . 49
3.6 Study 2: Field Deployment . . . 52
3.6.1 Method (Study 2) . . . 53
3.6.2 Results (Study 2) . . . 54
3.7 Discussion . . . 59
3.7.1 RQ1: Effect of Nudges on Credibility . . . 59
3.7.2 RQ2: Influence of Political Partisanship on Nudge Effects . . . 60
3.7.3 RQ3: Influence of Political Cynicism and Media Skepticism on Nudge Effects . . . 61
3.7.4 Opportunities in Designing News Credibility Nudges . . . 62
3.7.5 Challenges in Designing Nudges with Heuristics . . . 62
3.8 Implications And Opportunities for Designing Credibility Nudges . . . 63
3.9 Limitations and Future Work . . . 68
3.10 Conclusion . . . 69

4 TransparencyCue: Designing Transparency Cues in Online News Plat-

forms to Promote Trust **70**

4.1 Introduction 70

4.2 Related Work 74

4.2.1 Defining Transparency in Journalism 74

4.2.2 Existing Transparency Practices in Journalism 75

4.2.3 Designing for Trust in News Through Information Disclosure 75

4.2.4 Effect of Transparency on Perception of Trust 77

4.3 Interviewing News Consumers and Journalists Using a Scenario 78

4.3.1 Developing a Scenario 78

4.3.2 Recruitment 82

4.3.3 Interview Procedure and Analysis 84

4.4 News Consumer Perspective 86

4.4.1 RQa: What aspects of journalistic practice do news consumers want disclosed within news articles as transparency cues? 86

4.4.2 RQb: What should designers consider in promoting transparency cues for news consumers? 92

4.5 Journalist Perspective 94

4.5.1 RQa: What aspects of journalistic practice do journalists want disclosed within news articles as transparency cues? 94

4.5.2 RQb: What should designers consider in promoting transparency cues for journalists? 97

4.6 Discussion . . . 99
4.6.1 Comparing the Perspectives of News Consumers and Journalists . . . 100
4.6.2 Design Suggestions Based on Existing Journalistic Practices . . . 102
4.6.3 Considering Design Issues . . . 107
4.6.4 Limitations . . . 113
4.7 Conclusion . . . 114

5 NewsComp: Facilitating Diverse News Reading through Comparative Annotation 117

5.1 Introduction . . . 117
5.2 Background and Related Works . . . 121
5.2.1 The Need for Multiperspective News Consumption . . . 121
5.2.2 Barriers to Multiperspective News Consumption Online . . . 122
5.2.3 Designing for Information Consumption through Comparison . . . 123
5.2.4 Annotating Using the Crowd and its Effect . . . 124
5.3 Formative Study For Designing *NewsComp*: Think-Aloud Interviews . . . 125
5.3.1 Inaccurate Algorithmic Annotation . . . 125
5.3.2 Annotating on Google Drawings . . . 126
5.3.3 The NewsComp Interface and How It Works . . . 128
5.4 Evaluation Study . . . 131
5.4.1 Article Selection . . . 132

5.4.2 Measuring Credibility, Quality, Current Event Knowledge, Media Literacy . . . 134
5.4.3 Recruitment . . . 135
5.4.4 Procedure . . . 136
5.4.5 Participant Pool . . . 136
5.4.6 Gold Standard Generation . . . 137
5.5 Results . . . 138
5.5.1 RQ1: How well do users perform comparative annotation? . . . 139
5.5.2 RQ2: How does comparative news annotation affect users' perceptions of credibility and news quality? . . . 146
5.6 Discussion . . . 147
5.6.1 RQ1: Annotation Performance . . . 148
5.6.2 RQ2: The Effect of Engaging through Annotation . . . 150
5.6.3 Applications of Annotated Data . . . 151
5.6.4 Merging Articles Into One and Testing Effects . . . 151
5.6.5 Implications for Comparative Annotation Task Design . . . 152
5.6.6 Limitations . . . 153
5.7 Conclusion . . . 154

6 OtherTube: Facilitating Content Discovery and Reflection By Exchanging YouTube Recommendations with Strangers 155

6.1 Introduction . . . 155
6.2 Related Work . . . 158
6.2.1 Supporting Diverse Content Discovery Online Through Recommendations . . . 158
6.2.2 Social Comparison and Self-Presentation . . . 159
6.3 OtherTube: Design and Implementation . . . 159
6.3.1 Creating an Anonymous Profile . . . 160
6.3.2 Sharing and Removing YouTube Recommendations . . . 162
6.3.3 Interacting with Strangers' Recommendations . . . 163
6.4 Study Deployment . . . 165
6.4.1 Recruitment . . . 165
6.4.2 Procedure . . . 166
6.4.3 Participants . . . 167
6.4.4 Data Collection . . . 167
6.4.5 Method of Analysis . . . 170
6.5 Results . . . 173
6.5.1 (Content Discovery) RQ1. How do users discover content by browsing recommendations personalized for strangers? . . . 173
6.5.2 (Interaction and Engagement) RQ2. What factors affect users' interaction and engagement with recommendations personalized for strangers? 176

6.5.3 (Self-Presentation) RQ3. How do users present themselves when sharing recommendations with strangers? . . . 180
6.5.4 (Self-Reflection) RQ4. How does browsing recommendations personalized for strangers facilitate self-reflection? . . . 182
6.5.5 (Learning about others and the algorithm) RQ4. How does browsing recommendations personalized for strangers facilitate understanding? 185
6.6 Discussion . . . 188
6.6.1 RQ1 & RQ2: Content Discovery, Interaction, and Engagement . . . 188
6.6.2 RQ3: "Profile Work" for Self-Presentation to Strangers . . . 189
6.6.3 RQ4: Reflecting on Oneself and Others . . . 191
6.6.4 Design Implications . . . 192
6.6.5 Ethical Considerations . . . 193
6.6.6 Limitations . . . 193
6.7 Conclusion . . . 194

7 Discussion **195**

7.1 Design Implications . . . 195
7.1.1 Affecting Stakeholder Power Dynamics . . . 195
7.1.2 Addressing Adversarial Manipulation . . . 196
7.1.3 Designing for Long Run . . . 197
7.1.4 Choosing Between Cognitive Processes . . . 197

7.2 Cognitive Affordances and Attitude Formation: Moderators 199
7.3 Applying Affordances in Other Scenarios 200
7.3.1 Against Other Problematic Content 201
7.3.2 For Certain User Groups 202
7.4 Ethical Considerations for Cognitive Affordances 202

8 Conclusion **204**
8.1 Future Work 205

Bibliography **206**

Appendices **271**

Appendix A NudgeCred **272**
A.1 Example Tweets Used in Study 1 272
A.2 Study 2: Interview Questionnaire 272

Appendix B NewsComp **275**
B.1 Thinkaloud Interviews Questionnaires 275
B.2 Articles Used in the Deployment 276
B.3 Effect of User Characteristics 277
B.4 RQ2: Mixed-Effects Models 277

Appendix C OtherTube **278**

C.1 Distribution of Users Who Passed Eligibility Criteria and Signed Up 278

C.2 Need for Reflection and Insight Questionnaire 278

C.2.1 Need for Self-Reflection 278

C.2.2 Insight 279

C.3 Semi-Structured Interview Questions 279

List of Figures

1.1 NudgeCred Interface. 9

1.2 NewsComp system. 13

1.3 OtherTube embedded inside the YouTube homepage. 14

3.1 Three types of interventions (marked by blue arrows) currently employed by Twitter to tackle misinformation. Tweet (a) with a link to proper authority regarding COVID-19, (b) with a warning, and (c) removed. Here, both (a) and (b) are examples of nudges. Around the beginning of our work (July 2018), only (c) was operational. Twitter added others later. 28

3.2 Our nudge design: [Top] A decision tree shows the intervention logic and [Bottom] three nudge designs. (a). The *Reliable* nudge on a tweet from CNN Breaking News without questions in its comment thread. (b). The *Questionable* nudge is applied to a tweet with questions from Fox News, a mainstream media outlet. (c). The *Unreliable* nudge is activated on a tweet from `100PercentFedUP.com`, an extremely biased, non-mainstream website. The numbers indicate: (1) a change in background, (2) a tooltip message shown when hovered over, (3) a button to open a survey questionnaire for users to rate the credibility of the news tweet, and (4) a button to show more questions in the comments. 29

3.3 Screenshot of how clicking on the survey button would pop open the five-item credibility questionnaire. 45

3.4 Distribution of demographics, political ideology, political cynicism, and media skepticism in our participants in Study 1. 46

3.5 Shows interaction effects between user groups and nudge types in Study 1. The numbers inside the brackets are the effect sizes, Cohen's *d*. 49

3.6 Types of nudges based on transparency and mode of thinking. This figure emulates Figure 1 by Caraban et al. [70]. This work lies in the bottom-right quadrant. 64

4.1 Our scenario with three transparency features. Here, feature "a" corresponds to source characteristics conveying the expertise of the author; features "b" and "c" are message characteristics showing, respectively, crucial details about the event and the reporting style. In feature "c," reporting style includes whether the article is high or low in summary news lead (SNL) or inverted pyramid style reporting, the proportions of first- and secondhand accounts, the proportions of direct and indirect quotes, and the number of claims made. 80

5.1 A Google Drawings board used for think-aloud interviews. Similar to the high-fidelity interface, two articles are presented side by side here. Users can use all the available tools to link similar statements or highlight dissimilar statements that contain important information which should be included in the other article. 126

5.2 NewsComp Interface showcasing features with random annotations. ❶ Annotation instructions in two steps: find and connect similar statements, and answer if a statement with no corresponding, similar statement is important to include in the other article. ❷ Toolbar to finalize a connection by providing a rationale ❸ A solid arrow representing a connection already created ❹ A dashed arrow indicating that the connection creation tool is active ❺ A list of connections including deletion buttons ❻ The importance question in step 2. 129

5.3 Study design showing the experimental conditions for each of the four participant groups. Here, C and T respectively represent control and treatment groups; the number of participants is given in parentheses. Because we used four articles, we had two control groups (C1–2) and two treatment groups (T1–2). Article E_XP represents an article about event X from a source with political leaning P (L for left, R for Right). Articles with $X = 1$ were about immigration, while those with $X = 2$ were about abortion. For example, E_2R indicates a news article about abortion from a right-leaning source. The E_1 pair had high contrast, while the E_2 pair had low contrast. In the study, we randomized the order/position of the articles for each participant. 131

5.4 Graphs showing the distribution of participant demographics across the treatment and control groups. 137

5.5 Distribution of connection making by users. White and red dots respectively represent the average and experts' annotation. 138

5.6 Distribution of importance detection by users. White and red dots respectively represent the average and experts' annotation. 139

5.7 (a & b) User agreements on incorrect and correct annotations. (c & d) We filtered annotations by the number of concurring users to see how annotation performance changes as the threshold moves. Here, for connection making and importance detection, the F1 scores peak at five (55%) and six (41%) users, respectively. 140

5.8 Distribution of recall and precision for connection-making and importance detection divided into low/high CEK users (top), and low/high VML users (bottom). 141

5.9 Annotation counts by coded rationales divided into correct and incorrect annotations. The numbers over the bars represent the ratio of correct to incorrect annotations within each code. 143

5.10 False positive detection with OLS using the top 50 TF-IDF words in users' responses. Here, we listed only words with significant coefficients. For example, when users' mentioned "quote" in a rationale, the annotation was less likely to be erroneous. On the other hand, when users mentioned the general nature of the event ("lawsuit" in this example), the annotation was more likely to be erroneous. The model effect sizes (R^2) were 0.34 and 0.22, respectively. . . . 144

5.11 Interaction effects of groups and articles. We only found a marginal interaction effect (p=0.052) for credibility score on articles regarding immigration (c). 144

6.1 How OtherTube works. Each day, OtherTube collects YouTube recommendations when users access the YouTube homepage. Users have until the end of the day to remove items they do not want to share. Users can browse recommendations collected from others as late as the previous day. 160

6.2 OtherTube Options page. ❶ Avatar builder ❷ Shared demographic info ❸ How the profile will appear to others. 161

6.3 OtherTube Browser action page. ❶ Collected videos with options to remove from the shared set. 161

6.4 OtherTube embedded inside the YouTube homepage. ❶ Show or hide the embedded content. ❷ Browse different strangers or different recommendation sessions from the current stranger, and pin the current stranger. ❸ The stranger's profile. ❹ YouTube recommendations collected from this stranger. ❺ Link to a daily survey. ❻ The user's own YouTube recommendations, which OtherTube collects. 163

6.5 Demography of the participants in the study. 168

6.6 Distribution of participants' need for self-reflection and insight, bucketed for ease of understanding. 169

6.7 Mean with 95% CI of participants' responses to the daily survey Likert items from the first and last days of the study. We also performed a Mann-Whitney U test comparing responses on the first and last days. In (b), * indicates $p < 0.05$; in (f), . indicates $p < 0.10$. 173

A.1 Sample tweets used in Study 1 without the interventions. The examples include reliable, questionable and unreliable tweets from left-/center-/right-leaning sources. Here, there is a mix of politically contentious (e.g., immigration, racism and LGBTQ+) and not so contentious issues (e.g., flood and national security). 273

C.1 Demography of the participants who passed screening criteria and signed up for the study. 279

List of Tables

1.1 Transparency cues based on the suggestions from the journalists and news consumers. 11

3.1 Example sources in our *mainstream news* category. 41

3.2 Example *non-mainstream news sources* and their categories of reporting inaccuracy. The tooltip messages read: "This source is considered unreliable because it promotes <InaccuracyType>". 41

3.3 Example news sources and their political biases. 43

3.4 IRR of the five-item questionnaire on credibility in the formative study. . . . 45

3.5 Items used in measuring political cynicism and media skepticism. We used a five-point Likert scale (Strongly Agree – Strongly Disagree) with a "Don't know" option. 45

3.6 Mann-Whitney U test results for Study 1. Here, 'n' denotes the number of tweets rated in each condition. Avg. Cred. is the mean of 'n' credibility scores; * $p < 0.05$, *** $p < 0.001$. 49

3.7 Regression models on the credibility score. The base model contains nudge type, user group, control variables and the interaction between user group and nudge type. The politics and media model adds users' political ideology, media skepticism and political cynicism variables to the base model. The 3-way interaction model further includes the interactions of nudge type, user group and other variables with significant main effects in the politics and media model (Gender, Interest in the Tweet, Ideology and Media Skepticism). 50

4.1 Demography of our journalist pool. "Journ." here stands for "journalist." . . 82

4.2 Professional background of our journalist pool. In the network column, "freelance" indicates that the journalist is not associated with an organization. Note that we aggregated the roles of journalists and their associations to news networks and audiences to ensure anonymity, as required by the IRB. Knowledge of network affiliation and role would have been enough to reveal the identities of several participants. 82

4.3 Demography of our news consumer pool. We asked demographic questions at the beginning of each interview (see Appendix A for the list of questions). Note that some of the participants chose not to specify a political affiliation. * This participant was a former journalist. 83

4.4 Theme summary split by when each theme emerged—before or after showing the scenario to the participants. Note that while responses aligned with certain themes emerged both before and after scenario exposure, the themes' positions in this table are dictated by when they emerged most commonly. . 86

4.5 Design suggestions summarized according to two criteria: implementation requirements and consensus or disagreement among participants. Here, the requirements of access to (and knowledge of) an organization's resources (and protocols), such as internal/external databases of prior corrections, conflicts of interest, and behind-the-scenes materials could make it difficult for a third party to implement the design cues. The two right-most column suggest both within-group and between-group disagreement among our participants. As an example, the Behind-the-scenes Cue, discussed in section 4.6.2, could be hard to implement without access to organizations' materials. Some of the journalists disagreed as to the feasibility of disclosing portions of this information (e.g., televising meetings). Another example is the Author Expertise Cue (section 4.6.2) discussed by both groups with some disagreement from journalists due to its (e.g., years in reporting) negative impact on new journalists, which could be hard to implement due to the required access to organizations' protocols. Additionally, we provided sample questions designers could use to build transparency cues. 116

5.1 Questionnaires used in the study. Credibility and quality questions were asked after reading or annotating. (I) means these items were inverted for analysis. The correct responses appear in boldface. The CEK questionnaire contains multiple-choice questions, while the VML, credibility, and quality questions are 5-point Likert items. The VML and CEK items were presented in the pre-survey. 133

5.2 Coding scheme for annotation rationales. 142

5.3 Themes in users' responses to a question asking what they noticed about the two articles overall. Note that while an example response may belong to multiple themes, only the portion relevant to the listed theme is presented in bold. 145

6.1 Daily Survey Questions 169

6.2 Mixed effects negative binomial model for daily click count on Another Persona in Figure 6.4. In this model, user is a random-effects variable. User demography, their browsing habits, their need for self-reflection, the day when the clicks were counted, and the number of unique stranger data sets available to browse on OtherTube on the day when the clicks were counted are fixed-effects variables. The estimated negative binomial regression coefficient β is the difference in the logs of expected counts of the response variable due to a one-unit change in the predictor variable. 176

B.1 Linear models of recall and precision for connection-making and importance detection with user characteristics as predictors. 275

B.2 Mixed-effects regression on quality and credibility score using the interaction of experimental condition and articles. 276

Chapter 1

Introduction

Over the last couple of decades, online information feeds like the social feed on Facebook or Twitter, the newsfeed on Google News or CNN, and recommendation feed like YouTube or Reddit have become the place where users go every day and browse information served to them. Each of these feeds is generated by the algorithms based on different properties [164]. For example, Facebook feeds take into account the virtual network the user has built with their friends, family, acquaintances, other users, and other interests. Research shows that YouTube algorithm utilizes users' interests and watch history to generate recommendations [99]. News sites like CNN or Google News are also ramping up their use of algorithms on their feeds to provide more personalized experience to their readers [167]. Overall, these online information feeds play an important role in shaping public opinion.

With more and more people using online feeds for daily information consumption, a seemingly unimportant misleading content can be spread widely and have a significant impact. For example, when social feeds allow each individual to act as information creators, many named and unnamed sources with dubious credibility can easily flood the information space with misinformation [505]. Or, when the recommendation feed generated by YouTube heavily applies personalization, it creates an algorithmic filter where users are often exposed to limited viewpoints [341]. Due to the limited capacity of the users to attend to these issues on their feeds, design interventions are required to support users in online information consumption. In this dissertation proposal, I apply the theory of dual process information

consumption to develop design guides that can help combat these rising issues in online news feeds.

1.1 Problematic Information Online: Misinformation and Filter Bubbles

A typical user parses through a complex web of information online every day. They encounter information on social media, news platforms, entertainment platforms, and many others. Users go online for various purposes including out of curiosity, to learn, to get entertained and such. However, users have a limited bandwidth to pay attention to all the complexities and problematic content on online news feeds where they need external assistance. In online setting, various types of problematic content exist, such as, misinformation, hate speech, conspiracy theories, filter bubble, and algorithmic injustice. I further discuss the definition of problematic content and its typology in more details in Section 2.1. Among all types of problematic content, misinformation and filter bubbles are two major concerns that I aim to address in this thesis. Below, I will first introduce the advent of each of these issues.

Now, socio-technical systems and micro-blogging platforms like Facebook, Twitter, and YouTube on the web allow a direct path from the producers (e.g., journalists, content creators, and influencers) to the consumers. Often, information feed generated by social algorithms on these platforms comes from a vast number of sources. Unlike the past, when much of the curation of information were led by mainstream news outlets for TV media, information from any point of the internet can flow into these feeds, passing through very few filters. As the difference between producer and consumer of information gets more blurry day by day, it changes the way users become informed, debate, and form their opinions.

Such an information environment can foster confusion, and encourage speculation, rumors, and mistrust. We see increasing evidence of such mistrust on various public issues every day. For example, Covid-19 related misinformation has been so widespread, it causes confusion for many people [495]. In some cases, increasing exposure to unsubstantiated rumors may become more mainstream, like QAnon conspiracy [391]. Therefore, there has been an increase in public scrutiny on social platforms and their role in filtering and advancing harmful content online. While platforms like Facebook and Twitter put significant effort into stopping and spreading such misinformation through algorithmic filters, technological solutions often have a limited effect on the overall spread of misinformation. Most often, misinformation spread before platforms can identify whether something is misinformative because algorithms are not sophisticated enough to identify human-generated problematic information. In many cases, these algorithms catch up, only after misinformation has already spread. Furthermore, sources of misinformation also adapt as technological solutions designed against them improve at identifying them. Overall, these preventive measures built into algorithms have their limitations. On top of that, platforms' efforts to curb misinformation often get scrutinized by the public due to the concern that it threatens the circulation of free speech [108].

With the Internet and social media, a common belief is that it is supposed to increase the number of available viewpoints and perspectives available, leading to a very diverse pool of information However, some have also argued that it could also lead to people joining groups of their choice who share their world view and cut themselves off from any information that might challenge their beliefs [439]. Indeed, some research suggest that personalization algorithms used by online platforms such as Facebook and Google show users perspectives aligned with what they prefer and filters out contents with viewpoints they differ from [341]. These personalization algorithms typically prioritize and filter information based on a user's

prior interaction with the system on similar interests [113], because it helps platforms' goal to improve engagement [151]. This process could lead users to receive biased information. In particular, for political information, it might lead to the situation where the user never sees contrasting or opposing viewpoints on a political or moral issue. Pariser et. al. termed this situation as a "filter bubble" where users will not even know what they are missing in their feeds [341]. Filter bubbles have been widely criticized for their negative effect on users' autonomy of choices. In response, researchers are developing algorithms to combat those bubbles by promoting diversity [200, 491]. However, these developments are often at odds with platforms policy for increasing engagement [15]. Overall, both the problem of misinformation and filter bubbles are pervasive problems in online news feeds and hard to resolve by simply the algorithms.

1.2 Affordance

To address this concern, we need to help users recognize problematic content. In this thesis, I use the concept of affordance from HCI to create designs that enable people to recognize content. The idea of affordances was first established by Gibson in "The Ecological Approach to Visual Perception", based on his observation on human interactions with objects in the physical world [160]. To Gibson, "affordances of the environment are what it offers the animal, what it provides or furnishes, either for good or ill". Gibson exemplifies this with an example of surface of the earth and what it affords, "If a terrestrial surface is nearly horizontal (instead of slanted), nearly flat (instead of convex or concave), and sufficiently extended (relative to the size of the animal) and if its substance is rigid (relative to the weight of the animal), then the surface affords support." Gibson's "ecological" approach means that to understand either the properties of the surface or the perception of the animal, one

must also consider the environment created by the relationship between the two. Norman took Gibson's idea to the field of human-computer interaction and the research of user experience [159]. Where Gibson was concerned with the theories of human perception and interactions in nature, Norman's focus was on the creation and use of the affordances in HCI design. So, in *The Design of Everyday Things*, when Norman popularized the use of the word "affordance", he was describing a characteristic of a designed product that informs the user of a function [329]. However, Norman's early work did explore the functionality and use of physical objects, including doors, switches, keyboards, and clocks. But as these methods were extended to digital interfaces, the focus remained on optimizing for interface usability.

1.3 Dual Process Cognitive Affordances

While Norman popularized the concept of affordances, there has been some debate around it, including Norman saying some of his readers misuse the term in HCI. Norman says that the unqualified term affordance refers to real affordance, which is about the physical characteristics of a device or interface that allow its operation, whereas many HCI and usability researchers use it without qualification to refer to perceived affordance, which is about characteristics in the appearance of a device that give clues for its proper operation [331]. To address this concern, Hartson came up with four types of affordances based on the role an affordance plays: cognitive, physical, sensory, and functional [191]. To him, Norman's perceived affordance was actually cognitive affordance. As Hartson defines, "A cognitive affordance is a design feature that helps, aids, supports, facilitates, or enables thinking and/or knowing about something. As a simple example, clear and precise words in a button label could be a cognitive affordance enabling users to understand the meaning of the button in terms of the functionality behind the button and the consequences of clicking on it". Now,

to help users recognize problematic content, I take this concept of cognitive affordances that enable thinking about something and extend it by looking into the underlying process that goes during such thinking.

This is where *Dual Process Theories of Mind* comes in. Cognitive and social psychologists have argued about the idea of dual process cognition over several decades based on their work on thinking, reasoning, decision-making, and social judgment [129]. They argue that there are two separate cognitive systems underlying thinking and reasoning. Throughout this work, we will be referring to these two processes as automatic and reflective modes of cognition. Automatic process is universally shared by many animals and is the process often used during information consumption. This process allows people to use mental shortcuts as heuristics to quickly make judgments about a piece of information. On the other hand, reflective process is deliberate and requires imagining new scenarios to analyze before making a judgment about a piece of information. A brief discussion on the evolution of these theories is further discussed in the Related Work section. To illustrate, let's take the example of a user coming across a piece of information and its source. When consuming this piece of information, the reader may use the appearance of the source as a shortcut to judge whether the piece of information is credible or not. Or, before judging the credibility of a piece of information, they might consider their existing knowledge around the topic of that information and reflect on the possibility of the truthfulness of the information. To assist users during information processing, we can design new affordances that enable the use of either or both of these processes. I will term these together as *Dual Process Cognitive Affordances*, where Automatic Cognitive Affordance enables using automatic thinking and Reflective Cognitive Affordance enables using reflective thinking. In short, I will call these Automatic Affordance and Reflective Affordance throughout the text.

1.4 Designing Affordances against Misinformation and Filter Bubbles

In this work, I use the idea of designing dual process cognitive affordances against the online problematic information domain. To illustrate this idea, let's take the example of the structure of a social media platform, its existing affordances, and how we can reimagine new cognitive affordances. A fundamental aspect of such platforms is the social interaction feature which allows users to interact in various ways, one-to-one, one-to-many, or many-to-many fashion. Allowing users to create posts and allowing others to leave comments in response to a post is an example of a many-to-many interaction feature. Generally, this comment section may not have been built with any idea of perceiving the credibility of the information in a post. However, what I argue here is that we can augment this interaction feature with new design affordances that will enable users to judge the credibility of the information in a better manner. For example, an algorithm can be devised to analyze the comments to check whether these comments identify the limitations of the piece of information conveyed in the post. If this analysis is presented to the users, it will afford, that is, enable a user to use the analysis as a basis for making (not necessarily) quick judgments. Now this is an example of designing automatic affordances. We can also augment this interaction feature with a reflective affordance design. Consider the example of someone posting a news article from a certain news source. Now, let's imagine a design affordance that allows readers to see how a different news source frame the same story with a different title. This affordance could enable users to compare and make a reflective judgment of the credibility of the news item. Notice that the purpose of devising affordances in the context of designing against problematic information is goal centric, instead of usability centric, that is, unlike Norman's initial intent. This is where I diverge a little from Norman when talking about affordances.

To summarize, the goal of this thesis is to present how we can help users become aware of problematic information on their online information feed through meaningful design adoption. To do that, I marry the dual-process theories from social psychology with the idea of affordance design from Human-Computer Interaction (HCI) to develop a systemic approach to designing dual process cognitive affordances for assisting users in information consumption online. Below, I outline how I designed four systems using these affordances against misinformation and filter bubbles. These four systems are asked based on four major online scenarios where users typically face the problem of misinformation and filter bubbles. Two of these scenarios are related to misinformation and designing automatic affordances against it, while the other two are related to filter bubbles and designing reflective affordances against it.

1.5 Designing Automatic Affordances to Distinguish Content Credibility

As mentioned earlier, automatic cognitive routes are effortless where people typically use shortcuts (e.g., attractiveness of a source) based on past experiences. Some communication theorists have suggested that design affordances online spaces can help users apply such shortcuts or heuristics (for automatic processing) to distinguish misinformative content from more credible ones by acting like a cue [435]. Therefore, in another word, automatic affordances can also be called design cues. Throughout this work, I will use the terms automatic affordance and design cues interchangeably. Now, to help users distinguish misinformative content using automatic processing, I explore two questions:

(RQA) How can we use design automatic affordance to help users distinguish reliable and

unreliable information sources?

(RQB) How can we design affordances to help users distinguish degrees of trustworthiness among reliable information sources?

I answer these research questions through two studies described below. In the first study, called NudgeCred, I designed affordances to help users distinguish reliable content from mainstream news sources and unreliable information from fringe misinformative sources. In the second study, I utilize the concept of transparency as a means to design automatic affordances to separate different degrees of trustworthiness within reliable mainstream news sources.

1.5.1 NudgeCred: Supporting News Credibility Assessment on SocialMedia Through Nudges

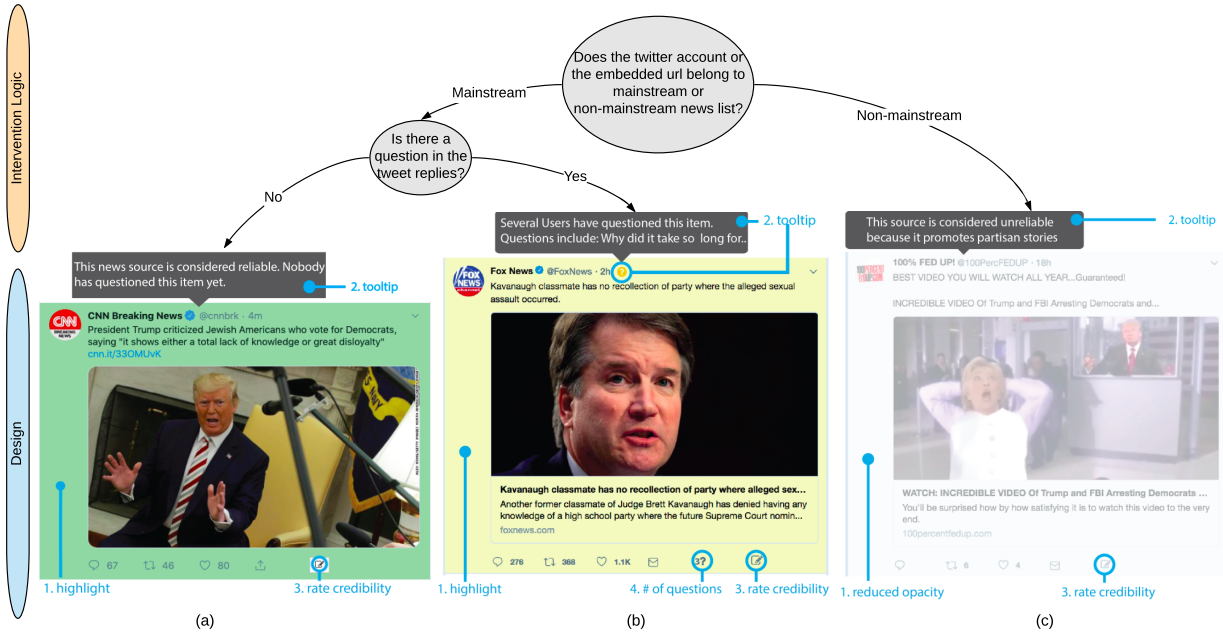


Figure 1.1: NudgeCred Interface.

When it comes to misinformation on online news feeds, social media feeds are the ones with the most impact [9]. Therefore, in designing automatic affordances or cues, I address the

first research question on helping users distinguish misinformation from credible information through a system called NudgeCred. Here, I apply communication theories on heuristics that people often use to automatically judge content credibility to design new cues to *nudge* users into better credibility evaluation. Specifically, I use two heuristics in this design: authority heuristic and bandwagon heuristics. Authority heuristic allows users to differentiate news content from mainstream sources and non-mainstream sources. Bandwagon heuristic further helps users to differentiate news stories with questionable or controversial or ambiguous reporting from news stories without such questions. Figure 1.1 summarizes how this design works. To examine its effectiveness, I conducted two studies—a controlled experiment in Section 3.5 and an in-the-wild experiment in Section 3.6—by recruiting the US nationally representative population sample. My analysis reveals significant evidence that design cues based on heuristics in NudgeCred can help users distinguish content credibility. This design is extensible through the use of other heuristics. However, in the second study, participants also showed concern that design cues built around bandwagon heuristics, i.e., relying on the annotation from other users or the crowd, can be problematic in practice and we continue on this concern in the next work. Chapter 3 reveals the details of this work.

1.5.2 TransparencyCue: Designing Transparency Cues in Online News Platforms to Promote Trust

To answer our second research question, I examine what values stakeholders desire when judging between reliable news sources. For news platforms, like `cnn.com` or `Google News`, such values could be based on journalistic values or news consumers' human values. In this work, I look into how to adopt one such value shared by both groups, i.e, the value of transparency in news reports through affordance design. To incorporate it in our design I conducted an interview study. In this study, I interviewed both news consumers and

Transparency Cues (Section)	Example Questions	Implementation Requirements	Suggested By	Disagreed By
Newsworthiness Cues (6.2.1)	Which news values does this report represent (e.g., conflict, sensationalism, eliteness and entertainment)? To what degree?	Requires access to organizations' news values	News consumers	-
Fairness Cues (6.2.2)	Who are the named parties in this article? To what degree does this report represent each political affiliation (left/center/right)? Did the reporter receive comments from all contacted parties?	Requires knowledge of organizations' procedures for getting comments	News consumers	-
Presence of Evidence Cue (6.2.3)	Does the report cite an authoritative source of evidence? Is there ambiguity in how sources are represented?	Requires access to official source materials (might be openly available)	Both groups	-
Anonymous Source Cue (6.2.4)	Does this report contain anonymously sourced information? Why was the information not available without anonymity? How did the reporter verify the information? How acceptable is the verification material?	Requires knowledge of organizations' procedures for anonymous sourcing	News consumers	Some journalists
Fact-check Cue (6.2.5)	Has any internal/external entity fact-checked this information? Who fact-checked it (with links)?	Requires access to internal/external fact checkers and their procedures	Both groups	-
Correction Cue (6.2.5)	Have there been any corrections to this report? Why? How were the corrections framed?	Requires knowledge of organizations' correction protocols	Both groups	Some journalists
Author Expertise Cue (6.2.6)	What skills does the reporter have? What is the reporter's educational background? How objective has the reporter been in past reports?	Requires comprehensive knowledge of journalists' reporting history	Both groups	Some journalists
Behind-the-scenes Cue (6.2.7)	Does this report contain any behind-the-scenes details? How did the reporting process occur over time?	Requires access to behind-the-scenes materials for a report	Journalists	Some journalists
Conflict of Interest Cue (6.2.8)	Does this report cover any entity with which the reporter/organization has a conflict of interest? How does the reporter/organization deal with such conflicts?	Requires access to news organizations'/journalists' financial information	Journalists	-

Table 1.1: Transparency cues based on the suggestions from the journalists and news consumers.

journalists to understand how transparency can be realized on news platforms to promote their perception of the trustworthiness of news items there. My analysis revealed a set of transparency-centered cues, that is, automatic affordances and their design considerations including the cases where stakeholders' have conflicting desires. A summary of this result is show in Table 1.1. Similar, to the value of transparency, other journalistic values (such as, accountability or journalistic freedom) can be considered in future works. I present this work in chapter 4.

1.6 Designing Reflective Affordances to Raise Awareness on Filter Bubble

Compared to *automatic* processing, people use *reflective* or a slow and effortful route for analytical tasks. In HCI research, a large number of works explore the space for design for reflection in many application areas, such as behavioral change, personal informatics, and mental health [411, 421]. Extending this approach to combat filter bubbles, I examine the design of reflective affordances that may allow users to apply this mode of thinking to understand their filter bubbles and discover content outside such bubbles. Here, I ask:

(RQC) How can we design reflective affordances to facilitate news reading outside their filter bubbles?

(RQD) How can we design reflective affordances to facilitate content discovery outside their interest filter imposed by recommender systems?

For this purpose, I built two systems to help answer these two questions, outlined below. For RQC, I designed a system called NewsComp that enables reflection by seeing the side-by-side contrast of news stories from opposing-leaning sources. To address RQD, I designed OtherTube which enables content discovery outside of users' existing interests on YouTube by swapping recommendation feed with a stranger.

1.6.1 NewsComp: Facilitating Diverse News Reading through Comparative Annotation

In a news environment, a vast majority of news consumers tend to consume news stories from their preferred source, which typically puts a particular political spin on most news stories.

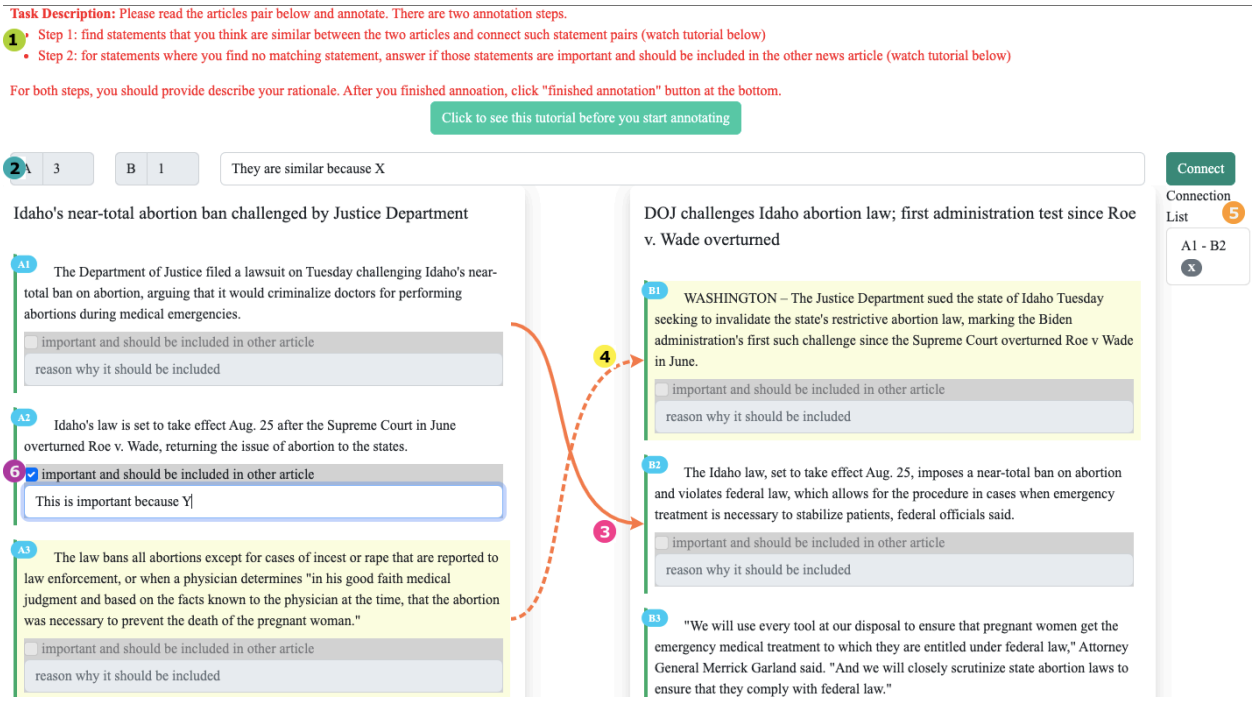


Figure 1.2: NewsComp system.

Thus, news consumers often see one-sided views on news stories. How can we address this challenge of informing readers of multiple perspectives on any news event? For this purpose, I built NewsComp, a system affording users to contrast news stories from multiple perspectives by showing a side-by-side view of news stories from opposing-leaning sources. Figure 1.2 shows the interface. This interface enables readers to engage with opposing perspectives by allowing them to annotate both similarities and dissimilarities between two stories. Using this interface, I conducted a study to examine whether using this system impacted users' attitudes toward news articles. Results reveal that between high-contrast news articles, NewsComp affects users' credibility perception. Furthermore, the annotation activity also led users to recognize various differences across articles, such as how journalists place information in an article differently, how different news articles emphasize different viewpoints, and how different news articles use various linguistic markers to attract their audience. I have explained the details of this work in chapter 5.

1.6.2 OtherTube: Facilitating Content Discovery and Reflection By Exchanging YouTube Recommendations with Strangers

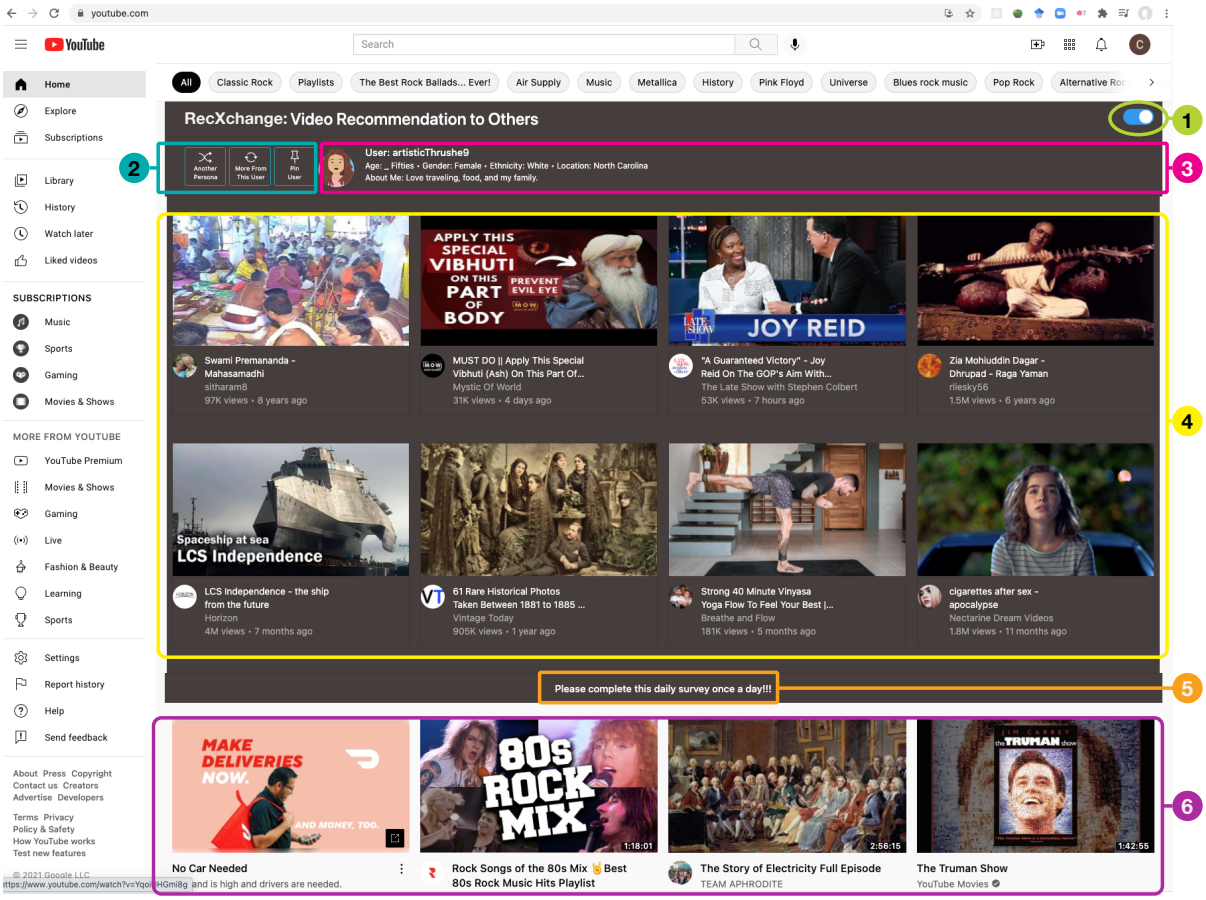


Figure 1.3: OtherTube embedded inside the YouTube homepage.

To help user understand their filter bubbles on non-news setting like for YouTube, I designed a system called OtherTube. YouTube's recommender algorithm primarily takes users' prior engagement into account to provide recommendations [99]. To help user realize algorithmic filters on their interests, OtherTube allows users to exchange their YouTube recommendations with strangers. To help further reflect, OtherTube allows users to share their persona with demographic details and their interests with other users. Figure 1.3 shows the interface. I have conducted a study to examine what users realize about their own YouTube recom-

mendations and what type of content they discover by seeing others' recommendations by using this system. Our results show promising results. We find that users reflect on their own interests by comparing against others' interests when using OtherTube. I have further explained this system in chapter 6.

1.7 Thesis Contributions

The focus of this dissertation is to examine how to combat problematic information online by empowering users through the means of design. To that end, I make three main contributions:

- First, I introduce the concept of dual process cognitive affordances, consisting of automatic and reflective affordances, by marrying the idea of cognitive affordances with dual process theories on cognition

- Second, using the concept of automatic and reflective affordance, I introduce four novel systems I built called NudgeCred, TransparencyCue, NewsComp, and OtherTube

- Third, I test the effectiveness of these systems against two types of problematic content, misinformation and filter bubbles

These contributions inform researchers, design practitioners, platforms and the users about the application of dual process cognitive affordances in the problematic information setting in online space.

1.8 Thesis Outline

In Chapter 2, I provide relevant background around the dual-process theory and existing approaches to address misinformation and filter bubbles. Here, I present how this work fits into the existing literature.

In Chapter 3-6, I present the systems and studies I conducted to answer our four research questions.

- Chapter 3 introduces NudgeCred system with design cues for using two heuristics as automatic affordance. It also includes two studies to answer particular research questions catered to the design. I present the analysis of those studies and discuss future research directions.
- Chapter 4 shows the details of the interview study I conducted to find transparency features to design automatic affordance including the recruitment of a diverse pool of news consumers and journalists, the analysis of responses, and a comparison between the two groups' responses.
- Chapter 5 shows our design of NewsComp to contrast and reflect on opposing news sources. Here, I present the deployment of the system to examine the effectiveness of the design and the performance of the users in comparative annotation tasks.
- Chapter 6 explains our design of OtherTube by exchanging recommendation feed between strangers to reflect on recommendation filter. This chapter explains how I conducted a 10-day study with this design and how the use of the tool impacted users' perception of their YouTube recommendation bubble.

Chapter 7 presents the design implications from the four affordances I designed, the potential opportunity to apply the design in different scenarios, and a discussion on ethical

considerations for affordance design. In chapter 8, I conclude by proposing future research on design-based approaches to tackle problems in online news feeds.

Chapter 2

Background and Literature Review

This thesis presents a set of systems built around the concept of dual-process cognitive affordances to address the problematic information, such as, misinformation and filter bubbles. Below, I provide background on problematic information, dual process theories, and HCI research on designing affordances. Then I present contemporary works on approaches to address these problems and how our work fits into those works.

2.1 Problematic Information Typology: Misinformation and Filter bubble

Problematic information has been used in various context online, especially in the last seven years, i.e., 2016 US Presidential Election [285]. Researchers have used the term problematic information in both broad and limited context. For example, some use it only regarding misinformation and disinformation [210]. However, in this work, I will use it to cover "a broad range of information created by a human or non-human entity that can cause any type of harm, intentional or unintentional towards an individual or the society in general". From human-generated side, this definition includes misinformation, disinformation, conspiracy, propaganda, hate speech, extremism, and others. This definition include problematic machine behavior such as filter bubbles, echo chamber, discrimination, and behavior modifi-

cation through psychological manipulation (e.g., addiction and radicalization) from algorithmic biases, misclassification and misuse [33, 413]. Since this thesis focuses on misinformation and filter bubbles, below I will focus on these two.

Within the academic investigation, various terms and definition have been used to identify misinformation. Misinformation in general has been described as a new types of information disorder connected to infostorm or infoglut or information overload [17, 185, 245]. Early scholarships tended to narrower definition focusing on particular topics of misinformation, such as misleading health information, governmentally organised propaganda or Wikipedia hoaxes [131, 249, 485]. Some scholars would distinguish between misinformation and disinformation by the intent of the producer [249], while some may use it interchangbly [138]. There are two primary types of distinction scholar make, one group consider disinformation as a subset of misinformation, with the intentional spreading [466]. The other group consider disinformation as the opposite of misinformation, i.e., disinformation is intentional and misinformation being non-intentional production and spread [133]. However, in this work, I will consider the core premise of all these definitions, i.e., misinformation is "false or misleading information", independent of the intent. The reason behind this choice is that in online space the intention can frequently vary. Here, one person may share a piece of false information as a joke where the next person re-sharing the same piece of information can do it with an intent to mislead their peer. Furthermore, this definition is more useful in the context of designing intervention where the design practitioner do not need to consult the source to determine whether an intervention is needed or not.

The term filter bubble was coined by Pariser et. al. in his book titled "Filter Bubble" [341]. Pariser calls it, "A world constructed from the familiar is a world in which there's nothing to learn ... (since there is) invisible autopropaganda, indoctrinating us with our own ideas". A predecessor of this concept can be found in the idea of cyberbalkanization, stemming

from the limits created by bounded rationality [62]. As Alstyne et. al. put it, "[on the internet] Regardless of how fast data scrolls across the screen, absorption is bounded. In the limit, people must choose some information contacts over others. Filters, even sophisticated electronic filters, must be selective in order to provide value. Thus, certain contacts, ideas, or both, will be screened out" [470]. Over the years, online platforms have developed better algorithms to personalize and to selectively curate information shown to a user based on their their location, their interaction behavior, and history [57]. Consequently, users become isolated from information that disagrees with their own viewpoints, effectively being trapped in their own cultural or ideological bubbles [341]. Here, algorithms continuously learn from the feedback of the user, termed as feedback loop, which gate keeps opposing view, causing filter bubble [51]. Though impact of this problem is often considered in the context of political ideological isolation, both cyberbalkanization and filter bubble defines isolation in terms of any interest internet users may have. In this work, we subscribe to this broader definition, instead of the political ideological isolation.

2.2 Dual Process Theories on Information Processing

Dual Process theories were developed over multiple decades by cognitive psychologist. The two processes that occur during information processing has been referred by different names including automatic and controlled [414], Implicit and Explicit [378], heuristics and systematic, System 1 and System 2 [428], and reflexive and reflective [268]. The automatic mode of cognition is generally considered as more universal cognition shared between humans and any other animals [128]. It includes instinctive behaviors that are innately programmed or learned over through evolution. Dual-process theorists generally agree that automatic processes are rapid and parallel in nature, although there are deviations that consider these

processes as sequential.

Dual-process theorists suggest that the reflective mode of cognition has evolved much more recently [128]. This mode of thinking is slow and makes use of the central working memory. The psychology of memory systems has a large focus on this conceptualization [153]. Despite its limited capacity and slow speed of operation, reflective mode permits a person to use hypothetical thinking that cannot be achieved by the automatic mode of cognition. Here, hypothetical thinking refers to constructing mental models or simulations of future possibilities. While we may oftentimes decide our actions based on past experience, doing what has worked well in the past, reflective thinking allows us to consider new avenues.

In social psychology, dual-process theories of social cognition emerged in the 1980s [80]. In this domain, research particularly focuses on the automatic and unconscious processing of social information in such domains as stereotyping, and attitude change [142]. They posit that cues in a persuasion context can lead the user to make some associations between the cue and the information [435]. Petty and Cacioppo introduced the elaboration likelihood model (ELM) which suggests that such cues can result in an attitude formation through peripheral information processing [355]. Another well-known model, the heuristic-systematic model (HSM) makes use of dual processes where systematic processing refers to the use of analytical consideration on the judgment-relevant information, and heuristic processing relies on mental shortcuts to judgment rules, i.e., heuristics that are already stored in memory [80]. For example, an attractive source in an advertisement can promote a positive while a superficial association between the source and the product. For another, long information can trigger the "length implies strength" heuristic [435], leading to the conclusion that the message is credible without taking into consideration what the message says. This is contrasted with the more cognitively effortful information processing. In such a case, a user may pay significant attention to evaluating message content rather than peripheral characteristics

like the attractiveness of the source or font color. Overall, dual process theories provide the basis that if design patterns can trigger these cognitive routes in a particular fashion it can help users consume information. In particular, in this thesis, we carefully craft design cues to trigger automatic cognition that helps users associate content credibility. To trigger a reflective mode of thinking, we propose design affordances that allow contrast between content within and outside of the filter bubble.

2.3 Affordances in HCI

As mentioned earlier, affordances originally proposed by Gibson, was introduced to HCI by Norman in the late 1980s [161, 330]. Over the years, this idea became a key concept in HCI. Initially, Norman described affordances as "the perceived or actual properties of the thing, primarily those fundamental properties that determine just how the thing could possibly be used" [330]. To Norman, "When affordances are taken advantage of, the user knows what to do just by looking: no picture, label, or instruction is required." [330]. Scholar have identified various properties and taxonomy of affordances. Scholars identified three types of affordances depending on the relationship between affordances and perceptual information: visible, hidden, and false affordances [154]. Affordance theory has also been applied in interaction design by categorizing affordances by levels of cognitive control [471] McGrenere et. al. argued for separating affordances from their perception, and Norman concurs [293, 327]. Some argued for mediated action perspective on affordances, where technology affordances are "possibilities for human action mediated by cultural means conceived as a relational property of a three-way interaction between the person, mediational means, and environment" [219]. Hartson came up with the idea of cognitive affordance to distinguish affordances that help certain cognitive function from Norman's real affordances, or as he calls it physical

affordances [191]. In this work, I build on top of this concept proposed by Hartson et. al. by connecting cognitive affordances with Dual Process theories on cognition.

Despite the introduction, Norman did not provide any recipe for how to design affordances in his seminal work [330]. Though, Norman did recommended some practices in design of affordance, such as, considering cultural constraints, conventions, using metaphors and coherent conceptual model [328]. Later, scholars have introduced some processes for designing affordances, such as, considering artifact-user interaction in affordance design [280]. Affordance-based approach has been utilized in various aspects of HCI research, including product development, interaction design, usability, robotic agent design[155, 288, 498]. In online setting, affordance has also been used various context, such as, social media[477]. In this work, I apply affordance in the context of problematic information online.

2.4 Designing Against Problematic Information

HCI designers have long been working on design-based solutions to problems like information credibility judgment and filter bubbles. Below, I outline these works and where the works presented in this thesis stand.

2.4.1 Designing for Distinguishing Information Credibility

At present, social media platforms are taking three approaches to combat misinformation—removing misinformation, reducing their reach, and raising awareness [389, 393]. The first line of action falls under the practices of crowdsourced (in-house and community-driven) and technology-assisted moderation [193, 198, 419] by enforcing established community guidelines [60, 393]. The second approach involves reviews from fact-checking services followed by

downranking [64, 277] and the application of warning/correction labels [312, 394, 467]. The third approach largely focuses on contextualizing misleading content through design interventions, such as providing source transparency [132, 423, 445, 455], prioritizing content from trusted authorities [192, 388], and showing related news stories from various sources [276]. Some of these interventions also target particular issues (e.g., voting [395]) or particular interactions (e.g., message forwarding [433]).

Aside from these platform-led efforts, researchers have also taken up the challenge of designing tools to aid in assessing information credibility. These works span several approaches, including fact-checking systems, interventions, media literacy programs and games [76, 114, 255, 279, 387]. There are multiple scholarly efforts for computationally assessing content credibility [75, 181, 286, 309, 374]. There are some scholarly works on establishing appropriate credibility signals for online content, as well as on designing guides for labeling manipulated media [207, 397, 500]. Some works examine particular crowd-led credibility labeling, including ratings by partisan crowds and the relationship between ratings from crowds and experts [36, 42, 308, 351].

Scholars have employed multiple types of messages as interventions against misinformation, including theory-centered messages [82], warning messages [68, 352], corrective messages [152, 213, 361], and opposing argument messages [94]. Studies examined the efficacy of interventions in various genres of news, including public health [354, 361] and politics [352]. Some research examined the effectiveness of interventions across countries [178]. Others examined effects for interventions across time by offering real-time correction versus delayed retraction [152, 213]. Real-time correction tools utilize various strategies, including mining databases of well-known fact-checking websites (such as, Snopes and PolitiFact) or crowdsourcing fact-checking. Paynter and colleagues looked into how to strategize corrections by combining several existing techniques (e.g., salience of a graphical element in a warning);

they call this approach "optimized debunking" [348]. Some suggest that while corrections can be effective, they can also backfire by inadvertently provoking users into attitude-consistent misperceptions [350]. However, others were unable to replicate such backfiring effects [494]. Warnings about information credibility have been more successful than corrections and are not prone to the same backfiring effect [53]. This work builds on providing warnings through design cues that could trigger automatic cognition by associating design cues with the credibility of the content.

2.4.2 Designing for Reflection on Algorithmic Filters

Reflection has received significant attention in HCI works. Existing research has shown promise of reflection in various areas such as education [170, 214, 253, 461], health or wellbeing [149, 174, 398, 454], and self-knowledge or personal informatics [16, 123, 261, 265, 266]. Several models have been proposed around personal informatics systems including Li et. al.'s Stage- Based Model of Personal Informatics Systems [265], Epstein et al.'s the Lived Informatics Model of Personal Informatics [123] and Niess et. al.'s Tracker Goal Evolution Model [324]. Purpose of many of these systems is to facilitate behavior change [35, 93, 283], especially through goal-setting [244, 304]. In a similar vein, designing for reflection around online social space has also received some attention [28, 105, 264]. Some of the prior work on social media revolves around personal informatics [105, 264]. For example, some works track affective expressions in users social activities (e.g., posts and comments [105, 231]), thus providing insight into their personal behavior.

While existing systems for reflection typically have an explicit metrics with desirable behavior changes—e.g., eating healthily, regular exercise, enhancing productivity, the current approach of personal informatics or self-tracking may not work well for reflecting upon their

filter bubbles and discovering new content. In this case, previous works that involves social elements for reflection can be effective for users to understand their position through projecting oneself onto others. There have been works focusing on *interpersonal informatics*, helping users understand their social network and how that influence their behavior [31]. For example, Feustel et. al. examines reflection using cohort data from multiple sources [136]. Different properties of data may stimulate reflection including showing invisible information, allowing to compare, revealing ambiguity and providing multiple perspectives [310]. For example, some previous work shows promise of providing data for comparison [469]. However, use of recommendations or works against filter bubble in this space is sparse. Among the existing works, some used personalized recommendation as a tool to trigger reflection on a particular artifact [242, 333]. Some systems also used visualization of unfiltered and curated feeds to improve users understanding of recommender systems by reflecting on them [126, 127]. Our work builds on these prior works to help users reflect on their own feed and escape from their filter bubbles.

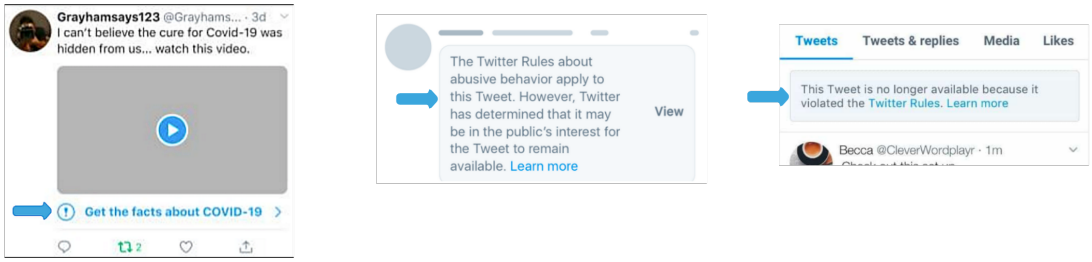
Chapter 3

NudgeCred: Supporting News Credibility Assessment on Social Media Through Nudges

3.1 Introduction

Social media platforms have witnessed an unprecedented rise in misinformation around public issues (e.g., COVID-19 and the 2016 US Presidential Election [8, 495]) [1]. To tackle, they have responded by experimenting with various design interventions [394, 467]. These include attaching warning labels and links to show additional context from trusted sources (see figure 3.1 for an example). With an intent to raise awareness and lead users to investigate the veracity of a news item [396], such design interventions can act as *nudges*—a choice-preserving technique to steer behavior [442]. Nudges differ from methods such as real-time corrections that often act as mandates and can backfire [152]. Furthermore, nudges can overcome the scale issue that many real-time systems face who rely on limited number of expert fact-checkers and quality crowd workers [24]. Despite the benefits, there is little empirical evidence whether nudge-based socio-technical interventions affect users' perception of credibility of online information. Furthermore, what complicates such investigation is that

[1] part of this chapter appears in [43]



(a) (b) (c)

Figure 3.1: Three types of interventions (marked by blue arrows) currently employed by Twitter to tackle misinformation. Tweet (a) with a link to proper authority regarding COVID-19, (b) with a warning, and (c) removed. Here, both (a) and (b) are examples of nudges. Around the beginning of our work (July 2018), only (c) was operational. Twitter added others later.

people with strong ideological leaning may resist nudges [449]. Existing works, again, lack empirical evidence of the effects of ideological leaning on nudges regarding perception of credibility. Therefore, considering the constraints under which these nudges may or may not work is crucial. This paper does just that.

For the purpose of our investigation, we design nudges with heuristic cues, i.e., mental shortcuts that people often use to judge credibility [300]. The choice of heuristic cues over reflective ones reduces cognitive burden on users, given the immense amount of content users see online [128]. Incorporating design guides from nudge and heuristics literature [435, 458], we built NudgeCred which operationalizes three design nudges—*Reliable*, *Questionable* and *Unreliable*. Devised with two heuristic cues—the authority of the source and other users' opinions—each of our nudges designates a particular level of credibility of news content on social media. These two heuristics comprise both external authoritative sources of information and social interactions of the crowd. Among the three nudges, both *Reliable* and *Questionable* are applied to information originating from mainstream sources on Twitter, while *Unreliable* makes posts from non-mainstream sources less visible (see Figure 3.2). Here, *Questionable* and *Reliable* differentiate between mainstream news items that raised questions in their Twitter replies compared to those that did not. Questioned items are

highlighted in yellow to warn of the potential controversy in evolving news stories from mainstream media, while those lacking questions are highlighted in green, signifying their reliability. By directing users' attention to the two heuristics, NudgeCred assists users in making meaningful news credibility evaluations.

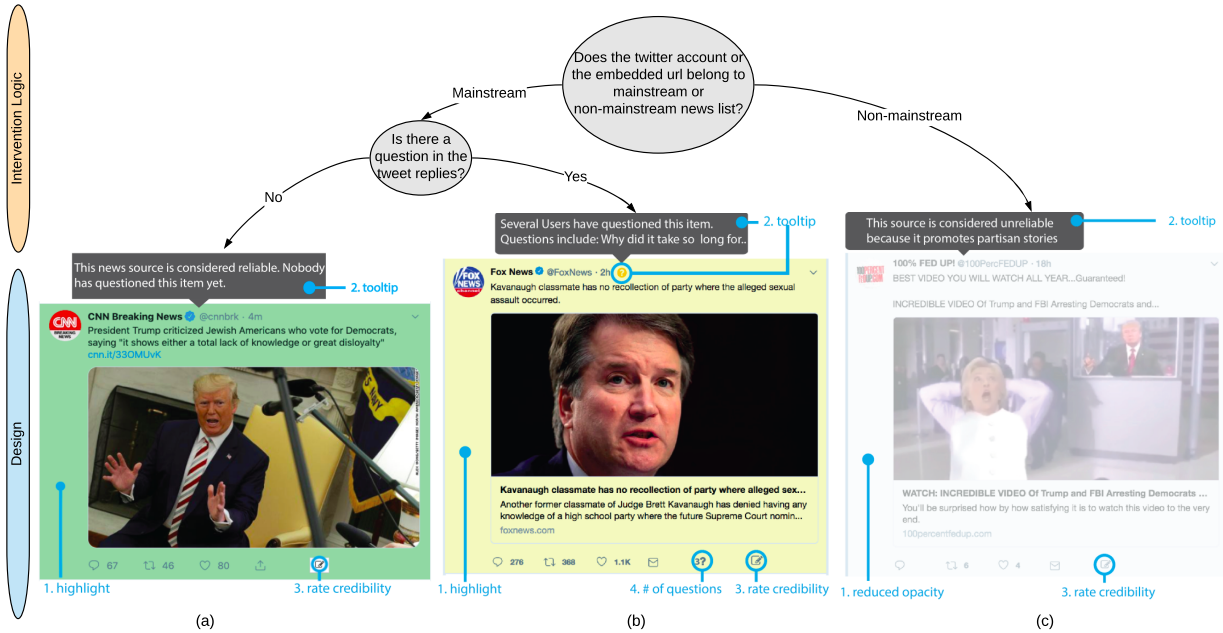


Figure 3.2: Our nudge design: [Top] A decision tree shows the intervention logic and [Bottom] three nudge designs. (a). The *Reliable* nudge on a tweet from CNN Breaking News without questions in its comment thread. (b). The *Questionable* nudge is applied to a tweet with questions from Fox News, a mainstream media outlet. (c). The *Unreliable* nudge is activated on a tweet from `100PercentFedUP.com`, an extremely biased, non-mainstream website. The numbers indicate: (1) a change in background, (2) a tooltip message shown when hovered over, (3) a button to open a survey questionnaire for users to rate the credibility of the news tweet, and (4) a button to show more questions in the comments.

To demonstrate our nudge-based approach, we built NudgeCred as a Chrome extension [2]. We followed an iterative design process. We first tested our initial design of NudgeCred by conducting a formative study with 16 university students and 36 Amazon Mechanical Turk workers. Results from the formative study helped us refine our design and suggested three

[2]How NudgeCred works: https://www.dropbox.com/s/2mt4tpdxebccokt/nudgecred_cropped.mp4

confounds—political ideology, political cynicism, and media skepticism—that may restrict impacts on users' credibility perceptions. We then conducted two sets of experiments using our final design: Study 1, a controlled experiment to examine the impact of the nudges with a representative US population ($n = 430$); and Study 2, a qualitative field deployment with Twitter users ($n = 12$) to gain insight into how we can improve NudgeCred's design. Analyzing users' credibility responses from Study 1 revealed that the *Unreliable* nudge significantly reduced users' perceptions of credibility for non-mainstream news sources. For *Questionable*, users in the treatment group rated news tweets with questions as less credible than the users in the control group, and those without questions as more credible. We did not find any effect of users' political ideology, media skepticism, or political cynicism on the effects of nudges. These outcomes suggest that NudgeCred worked irrespective of these confounds. Results from our field deployment (Study 2) show that NudgeCred improved recognition of news content and elicited attention towards all three nudges, particularly *Questionable*. Participants also suggested additional design considerations, such as incorporating heuristics that users would trust, applying nudges to share buttons, and using nudges to distinguish news genres and biases. To conclude, we offer design directions for news credibility nudging by exploring transparency-mode of thinking nudge categories, other heuristics and nudging methods from prior literature. Overall, our contributions include:

- A novel approach using heuristic cues to nudge users towards meaningful credibility assessment of news items in social media.
- A quantitative evaluation of this approach by examining users' credibility perception while considering three confounds—political ideology, political cynicism, and media skepticism.
- A qualitative understanding of the opportunities and challenges of this approach in

designing credibility nudges.

3.2 Related Work

3.2.1 Nudges to Steer Human Behavior

The concept of *nudges* has been frequently used to steer civic behavior for achieving important societal goals [182, 441, 443]. The idea stems from behavioral economics and psychology, which define it as a "choice architecture" that encourages citizens to act in a certain way while allowing them to act in other ways, thereby being a favorable alternative to imposing mandates [212]. Such approaches have been highly effective in areas such as environmental protection, financial regulation, and anti-obesity policy [182, 441, 443]. In online settings, technology-mediated nudges have been applied for such purposes as encouraging better password management, improving mobile privacy, and encouraging frugal shopping [25, 30, 218]. Comparatively, nudges regarding online news is getting traction recently in works such as Pennycook and colleagues' "accuracy nudge" (an accuracy reminder) and Nekmat's "fact-check alert nudge" [319, 354], who investigated impact of nudging on misinformation sharing intention. This work not only extends existing line of research by employing heuristics to assist credibility judgment, but also shows a method of devising such heuristic cue design.

3.2.2 Heuristic Cues for Credibility

Cognitive psychologists have long argued that when information overload occurs—as it typically does in online social media environments—humans turn to the cognitively effortless route of peripheral processing [80, 355]. While existing HCI works focus extensively on reflective processing, we use automatic or peripheral processing in this study [2]. Periph-

eral processing means that they depend on heuristic cues, such as the attractiveness of the source or the font color, to evaluate message content [355]. Communication researchers, in response, have offered a well-established list of technology-mediated heuristic cues, highlighted in the MAIN model, which attempts to explain how certain cues influence users' perception of credibility in online contexts [435]. The MAIN model suggests that four technological affordances influence perceptions of credibility: **M**odality, **A**gency, **I**nteractivity, and **N**avigability. These technological affordances trigger various peripheral cues by which users then judge the credibility of online content. For example, the agency affordance focuses on how users perceive source information in computer-mediated contexts. Often, the perceived agent or source of authority can be the machine, the user themselves, or the perceived authors of information on a particular website. For online news, agency is often attributed to the message's source, and these sources can trigger the *authority* heuristic—the perception that the source is an expert on the subject matter [437]. Similarly, information surrounding a message, such as ratings and recommendations, may also provide contextual information. For example, when a group of users likes or shares a news article on social media, the action signals that the group deems the information trustworthy. This signal, in turn, can influence users' perception of the information's credibility while serving as a *bandwagon* heuristic [435, 436]. In summary, the space of all possible heuristics under the four affordances is vast, offering us numerous possibilities for designing credibility nudges. Among these heuristics, we utilize *authority* and *bandwagon* heuristics in our nudge design. We discuss the remaining design possibilities later (see section 3.8).

3.2.3 Factors Affecting Credibility Perception: Partisanship, Attitude towards Politics, and Media

Scholars have found numerous factors that may influence information credibility. Based on findings from our formative study (discussed in section 3.3), we contextualize our research on three behavioral confounds—partisanship, attitude towards politics, and attitude towards media. Historically, scholars have failed to reach a consensus on the role of partisan bias in credibility perception. Earlier work suggested that users perceive unbiased information as more credible compared to one-sided information [13, 349]. Compared to this result, other research found that users would perceive news conforming to their own attitudes as more credible than unbiased news [88, 278, 301]. For this reason, users with strong partisan biases might even resist nudges on attitude-challenging news items, rather than be influenced. Sunstein hypothesized that a considerably large number of people evaluate nudges based on whether they approve of the underlying political objective, naming this "partisan nudge bias" [443, 449]. Hence, we made sure to test our nudge design across a population with balanced partisan affiliations, allowing us to measure the effects of partisanship.

Similar to users' partisan attitude, users' media skepticism can influence their perceptions of the credibility of mainstream and non-mainstream content. Media skepticism is "the feeling that the mainstream media are neither credible nor reliable, that journalists do not live by their professional standards, and that the news media get in the way of society rather than help society" [463]. Scholars have found that media skepticism is negatively associated with exposure to mainstream media and positively associated with non-mainstream media [462]. Media is also generally blamed for its role in enhancing institutional distrust by depicting most governmental policies negatively and causing cynicism towards politics [69]. Studies have also demonstrated that users with high media skepticism and political cynicism rated

citizen journalists as more credible than mainstream ones [71]. Drawing from these works, we examine and provide the first empirical evidence of the effects of political ideology, media skepticism, and political cynicism on credibility nudge.

3.3 Formative Study

We designed NudgeCred in an iterative fashion. Initially, we built a prototype and conducted a pilot study to evaluate it [3].

Method We built our prototype as a Chrome extension with two types of nudges (described in section 3.4.2); namely, *Questionable* (tweets highlighted in yellow, which indicate caution) and *Unreliable* (tweets that are less visible). This extension would alter the Twitter homescreen in real-time when users visit them. As mentioned in Figure 3.2, users could click on a survey button added by the extension. Clicking the survey questionnaire button would open a pop-up overlay comprising our study measurements for credibility. We discus them further in section 3.5.1 (refer Figure 3.3 to see how it looked). With this setup, we conducted a study with 52 participants from Amazon Mechanical Turk (n=36) and the university (n=36) [41]. For recruitment from the university, we used a university portal available for participant recruitment. For the MTurk users, we used MTurk portal with some filtering conditions, such high rate of work acceptance (>95%), over 18 years of, US resident and familiarity with Twitter. In a pre-study survey, we collected users' demographic details, such as gender and political leaning. Participants were divided into two groups of treatment (seeing tweets with nudges) and control (not seeing any nudge). In a 2-week study period, we asked our participants to rate the credibility of three to five

[3] All of our studies have been approved by our Institutional Review Board.

tweets from their Twitter feeds every day. We did so by reminding them everyday to spend around 20 minutes on Twitter by completing an MTurk HIT. Afterwards, we reached out to 16 users—8 control and 8 treatment users—to get feedback on our design where 8 of them finally agreed. In all studies, we compensated our participants adhering to Federal minimum wage requirements ($7.25).

Result In our study, we hypothesized that users in the treatment group would rate tweets with both *Questionable* and *Unreliable* nudges as less credible compared to users in the control group. A Mann-Whitney U test on the credibility ratings showed that our hypothesis was true for *Unreliable* nudge (avg. cred. (Control) = 0.51, avg. cred. (Treatment) = 0.43 and $Z = 210236, p < 0.001$, Cohen's $d = 1.291$). However, we found the opposite for *Questionable* nudge, i.e., the treatment group rated those tweets as more credible than the control group (avg. cred. (Control) = 0.67, avg. cred. (Treatment) = 0.71 and $Z = 502140, p < 0.001$, Cohen's $d = 0.188$). Furthermore, in our post-hoc analyses, we found that for Republican users the effects of nudges were not significant.

To make sense of the discrepancies in our quantitative result, we conducted interviews followed by a thematic analysis. We identified three themes in the interviews. First, when asked which news organization users follow, participants showed a trend of interest in ideologically aligned news sources. While a majority of Democrats mentioned mainstream sources (e.g., CNN, NBC, the New York Times, and the Washington Post), most Republicans named a mixture of mainstream and non-mainstream sources (e.g., the Wall Street Journal, Fox News, Joe Rogan, and Candace Owens). This trend led us to assume that our intervention may be less effective if it contradicts users' political stances. Second, we found several hints that *cynicism towards politics* and *media skepticism* can influence the impact of nudges. For example, one participant suggested that he prefers news without biases which mainstream

media does not do anymore. Another (Republican) participant expressed frustration that she had to stay away from discussing politics on social media, as she often ran into arguments with others. If Republicans are indeed more skeptical of mainstream media on the whole, and also equally mistrusting of social media platforms, then our intervention could be perceived as yet another attempt by social media to integrate ideologically motivated interventions into their news feeds. Therefore, we decided to examine whether these sentiments of media skepticism and political cynicism adversely affect the interventions. Third, consistent with our quantitative result, we found the opposite of the expected reaction to the *Questionable* intervention. For example, a participant responded: *"I found that these tweets [with Questionable intervention] seem ... more accurate than things that I normally read"*. This conflicting reaction may have stemmed from the lack of a clear hierarchy, i.e., the absence of nudges on more credible news tweets. Subsequently, we revised our design with a third nudge called *Reliable* (tweets highlighted in green to indicate reliability). These findings suggest that our initial prototype did not adequately support better news credibility judgments by users, and informed us to consider three confounds (users' political ideologies and attitude towards politics and media) in evaluating our system.

3.4 Designing NudgeCred

3.4.1 Design Guides

To design nudges with heuristic cues, we employ design guides from two strands of literature: the nudge perspective and the heuristic perspective.

Nudge Perspective

To design effective nudges, the literature suggests two primary aspects to consider: the mode of thinking involved (automatic vs. reflective) and the degree of transparency (transparent vs. non-transparent) [186].

Mode of Thinking: Cognitive psychologists developed *dual process* theories, a set of psychological theories for understanding human decision-making. These theories describe two main modes of cognition: *automatic* and *reflective* [128]. The automatic mode is fast and instinctive. It uses prior knowledge or past repeated behavior and minimal cognitive capacity to decide on actions. Reflective thinking, on the other hand, is slow and effortful. It uses greater cognitive capacity to make a goal-oriented choice by critically examining the effects of choices before selection.

Transparency: Scholars introduced epistemic transparency (i.e., whether users would understand the purpose of a nudge) to divide existing nudge designs into two categories: transparent and non-transparent [186]. Thaler and Sunstein adopted transparency as a guiding principle for nudges [451]. This is because of the concern that a designer may manipulate people into their own preferred direction using systems for behavioral changes.

Using the combination of these two dimensions, Hansen and Jespersen grouped existing nudges into four categories: reflective transparent, automatic transparent, reflective non-transparent, and automatic non-transparent [186]. In designing technology-mediated nudges for credibility, we pick one quadrant from these categories: *transparent* nudges with the *automatic* mode of thinking. We chose the automatic mode as it requires less cognitive effort to process information, especially given the information overload in social media and the instant nature of media consumption. Scholars in the past argued that use of automatic mode over reflective mode for design could address two potential problems–lack of motivation

and lack of ability–that typically restrain users from performing tasks such as critically evaluating credibility [2]. Furthermore, our design does not prevent users from critically reflecting on the news content. We chose the *transparent* design to explicitly reveal the motives behind it. We later discuss the potential for nudge designs in the remaining three quadrants (see section 3.8).

Heuristic Perspective

This work applies heuristics to design nudges for social media in order to enhance users' perceptions of the credibility of news. Cognitive psychologists have proposed models of how effective heuristics work [239]. One of the models, called *Fast and Frugal Heuristics* suggests that users should be able to make inferences using "fast, frugal, and accurate" heuristics when faced with environmental challenges (e.g., information overload) [458]. According to Todd et. al., simple heuristics work when they follow two principles: they exploit the structure of the environment and are robust. In social media, structures include sources of news items, popularity (indicated by the number of shares or replies), and the way that information is organized by time and personal interactions. They argued that heuristics that exploit existing structured information can be "accurate without being complex" [458]. Another success criteria for heuristic design is the robustness of the decision model. A computational strategy utilizing a limited set of information can yield more robustness [458]. Employing these principles, our design includes only two heuristics, outlined below. These heuristics seem useful to users to investigate misinformation [156].

- *Authority Heuristic*: We limit the source of news to a handful of known organizations followed by a binary classification of the organizations.

- *Bandwagon Heuristic*: We utilize the conversational structure (or replies) of the envi-

ronment as an indicator of credibility assuming a skew in the reply distribution.

3.4.2 Outlining the Design

Our design of NudgeCred is built on the idea of applying subtle heuristic cues in certain contexts in social media. It is powered by three types of socio-technical interventions—*Unreliable*, *Questionable*, and *Reliable*. Using the principles of fast and frugal heuristic design, our design uses a two-level decision tree with two heuristics (see figure 3.2). They are triggered based on whether a news tweet originates from an official authority. Thus, the first step of our tool design relies on the *authority heuristic*. Communication scholars have long argued that revealing the official authority of content results in applying the authority heuristic in credibility judgments [435]. We apply the authority heuristic by differentiating between mainstream and non-mainstream news tweets. We do not apply nudges to tweets that do not come from mainstream and non-mainstream sources. We opt to use source-based credibility annotation due to the challenging nature of annotating article-level credibility. While we may flag some accurate articles from non-mainstream media in this method, other work demonstrated that this number could be few (14%) compared to the accuracy (82%) [412]. For such false-positives, users still have the opportunity to fact-check them.

To flag inaccurate content from mainstream media, we apply another criteria of whether someone replied to a mainstream tweet with a question. Literature suggests that such questions, depending on users' prior knowledge of the subject matter, can instigate curiosity and motivate them to investigate to a varying degree [272]. We employ this property by showing the number of questions as a *bandwagon heuristic* ("if others question this story, then I should doubt it, too") on mainstream news tweets. Thus, our study had three nudges:

mainstream news tweets that did not have questions raised about their content (*Reliable*), mainstream news tweets that had questions raised about their content (*Questionable*), and non-mainstream news tweets (*Unreliable*). To employ epistemic transparency in the design, our design includes a tooltip with the reasoning behind each nudge. Figure 3.2 shows the overall design. Before delving into each intervention, we propose a classification of news sources that enables our application of the *authority heuristic*.

Classifying Authority of News

Our nudge-based interventions work on two types of news sources: mainstream and non-mainstream. In journalism and communication literature, the term "mainstream media" lacks an official definition. For the purposes of our study, **mainstream news sources** are sources recognized as more reliable in prior scholarly work. Opting for a heuristic approach, we use such existing literature to create a reliability measure which may later be replaced; this is not our primary contribution. In this approach, the first two authors iteratively collected a list of mainstream news websites by referring to two prior works, including Pew Survey and NPR [305, 410]. Next, we refined our list with the help of our in-house journalism and communication media expert by referring to the most circulated and the most trusted news sources [87, 305], subsequently removing a news aggregator (Google News) and a local source (AMNewYork). Table 3.1 shows a sample. Every source on our final list of 25 news sources follows standard journalistic practices in news reporting [18].

For **non-mainstream news sources**, we refer to `Opensources.co`, a professionally curated list of websites known to spread questionable content [338]. Each source in this list is categorized based on its level of information reliability (e.g., 'extreme bias,' 'rumor,' 'clickbait'). The curators manually analyzed each source's domain-level characteristics, reporting and writing styles before assigning a particular category. From this list, we remove the 'politics'

Mainstream Source		
The Economist	CNN	The Blaze
New York Times	NPR	BBC
Washington Post	MSNBC	Fox News
Chicago Tribune	WSJ	Politico
New York Post	Newsday	NY Daily

Table 3.1: Example sources in our *mainstream news* category.

Website	Category	Inaccuracy Type Message
abcnews.com.co	Fake news	misinformation
breitbart.com	Extreme Bias	partisan stories
americantoday.news	Rumor Mills	rumor
infowars.com	Conspiracy Theory	conspiracy
rt.com	State News	state propaganda

Table 3.2: Example *non-mainstream news sources* and their categories of reporting inaccuracy. The tooltip messages read: "This source is considered unreliable because it promotes <InaccuracyType>".

and 'reliable' categories to retain sources which were explicitly labeled as promoting unreliable news, a total of 397 sources spanning 10 categories. Table 3.2 shows a sample from this list. We do not intervene in the rest of the sources that do not fall into these two categories. Using this notion of mainstream and non-mainstream news sources, we apply three nudges.

Three Nudges

The ***Unreliable*** nudge detects whether a tweet comes from an unreliable authority. Our design applies this nudge by examining whether a tweet from a user's feed originates from a non-mainstream news site and subsequently reduces the item's opacity, rendering it harder to read. We call these tweets unreliable non-mainstream tweets (T_U) [See (c) in figure 3.2]. To instigate epistemic transparency, *Unreliable* provides an explanation of its action through a tooltip message: "This source is considered unreliable because it promotes <InaccuracyType>." Table 3.2 shows the list of <InaccuracyType> messages based on the source and its category in *opensources.co*.

The ***Questionable*** nudge is applied to mainstream news tweets (T_Q) when at least one question is raised about the information in the corresponding Twitter reply thread[4]. Prior studies suggest that less credible reports on an event are marked by questions and inquiries [309, 503].

[4]Though Twitter has recently rolled out a threaded reply structure, at the time of the study, it did not exist. Thus, we only took direct replies into account.

To detect questions, our algorithm is kept intentionally simple. Our algorithm looks for "?" mark to identify questions—a simple but transparent method that is understandable by users. It is worth noting that our focus is not to develop the most sophisticated algorithm to detect questionable news, rather testing the effectiveness of nudge in credibility assessment. In that regard, using the simple heuristic serves its role and have benefits in simplicity and transparency for users to understand. While investigating advanced natural language parsing methods to identify relevance of the questions to the news article or more advanced machine learning techniques to detect questions is worth looking into [260, 496], such investigations would require significant work, perhaps amounting to a separate full contribution. Hence we leave those as future paths to pursue. Instead our approach works as a minimum baseline to identify questions. To make users aware of these questioned mainstream tweets, the *Questionable* nudge is applied by changing the background color of the tweet to yellow while showing the number of questions (see (b) in figure 3.2). By showing this number, we promote a collective endorsement that multiple users have doubts about this news [436]. Additionally, a tooltip message offers transparency by explaining the reason behind the nudge activation. For T_Q, the tooltip message follows the format: "Several users have questioned this item. Questions include: <first reply tweet with a question>" (e.g., for a tweet containing a news report with missing details, such as time of an event, a reader may have replied to ask: "When did this happen?"), thus directing further attention to other users' comments in an effort to stimulate the *bandwagon heuristic*. The bandwagon effect has been demonstrated to be powerful in influencing credibility judgments [237].

The ***Reliable*** nudge is triggered when the source of the news tweet is an official, mainstream source and was not questioned in the replies; specifically, reliable mainstream tweets (T_R) were emphasized with a green background highlight (see figure 3.2 (a)). A tooltip message is formatted for T_R as follows: "This tweet seems more reliable. Nobody has questioned this

Source Type	Twitter Account	Political Bias
Mainstream	CNN Breaking News	Left
Mainstream	NY Post	Right
Mainstream	Politico	Center

Source Type	Twitter Account	Political Bias
Non-mainstream	Daily Kos	Left
Non-mainstream	Breitbart	Right
Non-mainstream	Zero Hedge	Conspiracy

Table 3.3: Example news sources and their political biases.

item yet." The colored highlights and the corresponding tooltip messages create contrast within the mainstream news tweets, helping users navigate them better. Note that we included this nudge in response to the findings from the formative study.

3.5 Study 1: Evaluating Impact on Perceptions of Credibility in a Controlled Setting

To evaluate our design, we conducted two studies. Study 1 evaluates impact on credibility perception in a controlled setting while Study 2 is a field deployment. In Study 1, we examine three research questions on the effect of nudges on users' credibility perceptions in a controlled setting simulating a Twitter feed with multiple tweets for each nudge.

RQ1. *Can heuristic-based design nudges on an online social news feed help users distinguish between reliable, questionable, and unreliable information?*

RQ2. *Do users' partisan attitudes affect their responses to credibility nudges?*

RQ3. *Do users' political cynicism and media skepticism affect their responses to credibility nudges?*

3.5.1 Method (Study 1)

Selecting news tweets for a controlled environment

For this study, we simulated a Twitter feed with a fixed set of tweets which would be shown to every user. To simulate a realistic Twitter feed, we selected news sources from our previously compiled list of mainstream and non-mainstream sources, and then selected several tweets from each source (see Table 3.1 and 3.2). We used a balanced approach to select sources with left-wing, centrist, and right-wing biases. For bias categorization, we used `mediabiasfactcheck.com`, a source used in other scholarly works [228, 295]. For each news source under each bias category, we first found the source's Twitter handle. We retained the Twitter accounts that had the greatest numbers of followers[5] (a mark of popularity of the news source on Twitter). Table 3.3 shows sample Twitter accounts with their perceived political biases. For each source, we selected the tweet within the last 48 hours that had the highest number of shares. With three nudges working across three political leaning categories, our feed comprised 9 tweets (3 political leanings $\times$ 3 nudges). Appendix A.1 shows several tweets from this list. To add variation, we created another set of 9 tweets using the second most followed Twitter accounts from our list of news sources, resulting in a second feed for our controlled environment. Users were randomly shown one of the two feeds, totaling 9 tweets in each case. To evaluate our RQs, we needed to measure users' credibility perception, political ideology, political cynicism, and media skepticism.

[5]We excluded Fox News' Twitter handle due to its inactivity for several months until the time of the study.

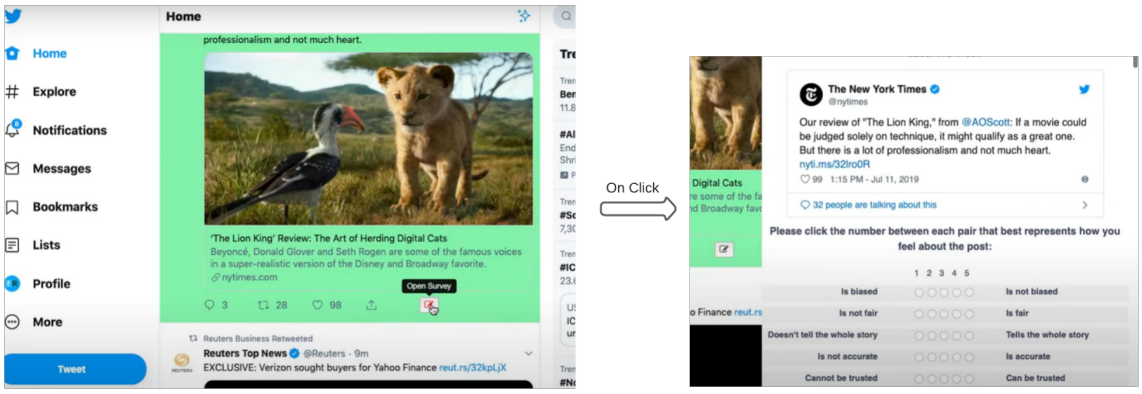


Figure 3.3: Screenshot of how clicking on the survey button would pop open the five-item credibility questionnaire.

Item	IRR
Is/not biased	0.83
Is/not fair	0.79
Does/not biased tell the whole story	0.80
Is/not accurate	0.79
Can/not trusted	0.79

Table 3.4: IRR of the five-item questionnaire on credibility in the formative study.

Pol. Cyn.	1. Elected officials put their own interests ahead of public's interest
	2. It seems like politicians only care about special interests
Med. Skep.	1. The media provide accurate information
	2. The media provide trustworthy information
	3. The media deal fairly with all sides
	4. The information provided by the media needs to be confirmed

Table 3.5: Items used in measuring political cynicism and media skepticism. We used a five-point Likert scale (Strongly Agree – Strongly Disagree) with a "Don't know" option.

Measuring News Credibility, Political Ideology, Political Cynicism & Media Skepticism

We used a five-item questionnaire by Meyer et. al. [303] to measure users' perceptions of credibility for every news tweet (see Figure 3.3). In our formative study, we found this measure had a high Cronbach α ($\alpha = 0.95$) and individual inter-item correlations (see Table 3.4), showing a high level of internal consistency. To capture partisan attitudes, we survey participants for their political ideology on a seven-point Likert scale ranging from "strong Republican" to "strong Democrat". We survey participants on media skepticism and political cynicism using a validated questionnaire from journalism scholarship (see Table 3.5) [71]. For both variables, we average the responses across questions and use the median to create a

binary response variable with values "low" and "high." Note that we had to revert the first three media skepticism questions before averaging.

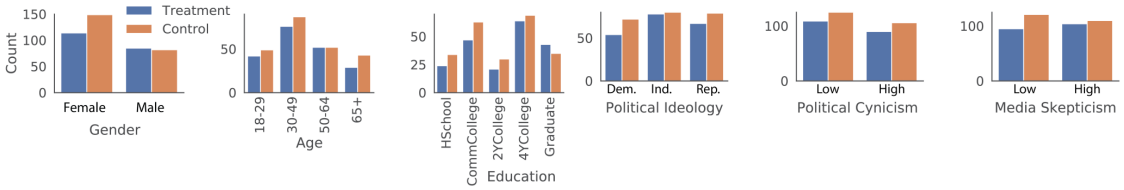


Figure 3.4: Distribution of demographics, political ideology, political cynicism, and media skepticism in our participants in Study 1.

Recruitment

Our study participants were recruited starting the third week of July 2019 and spanning a period of three weeks. We required three qualifications for user participation: (1) age of 18 or older, (2) US resident, and (3) familiarity with Twitter. This choice of US population was purposeful due to the difficulty in measuring our confounds across global population. Users' political leaning has different meanings in different countries (e.g., political left-right are different in the US and Europe). Similarly, levels of skepticism/cynicism might vary by country. We focused on US-population due to the availability of well-established measurements for our confounds from the communication literature [71, 303]. We recruited 430 users from *Qualtrics*, well-balanced by partisan affiliations. Figure 3.4 shows their demographics. This sample is mostly balanced across gender, age, and education, with a slight skew toward females.

Study Procedure

We presented our participants a set of tweets collected right before the start of recruitment. We chose this approach because studies have shown that there is a lag in terms of the

amount of time media coverage takes to influence public opinion, with some exceptions (e.g. mass shootings) [385]. As a result, we anticipated that participants would be least likely to be familiar with the most current tweets. Participants were randomly assigned to either the treatment or the control group, with a quota check to ensure balanced allocation across political ideology. To counter order effects, we presented tweets in random order. We added attention checks–questions with options reversed–right before the Twitter feed. Taking recommendations from Qualtrics, we also discarded participants who spent less than six minute to respond–the half of the median time spent by users in a soft-launch of 50 users. Participants saw each tweet and answered the questions for that item before scrolling down to the next one. This approach reflects a natural setting of modern social media sites, where users browse feeds in real-time, click links in-situ, and the same post usually do not appear again at a later time point. To reduce response bias, we framed the questions to ask for credibility of the items (e.g., how do you feel about the post?) instead of the effects of nudges (e.g., how does the nudge affect your perception?). The unexpected effect on *Questionable* in our formative study suggests a lack of response bias.

Method of Analysis

We initially perform mean comparison with Mann-Whitney U-test. However, note that each user saw multiple tweets with each intervention. To model such repeated measurements for the same intervention, we further use a mixed-effects logistic regression.

$$\mathbf{y} = X\beta + Z\mathbf{u} + \epsilon \tag{3.1}$$

In Eq. 3.1, the response variable (credibility score) ($\mathbf{y}$) is the dependent measure for our experiment. While fixed effects (X) are the independent measure, random effects (Z) are the variables repeated in multiple observations; that is, tweets. The residual (ϵ) is the error in fitting. Finally, β and $\mathbf{u}$ are the coefficients of fixed and random effects, respectively.

We used an **R** implementation of a linear mixed-effects regression model, *lme4.lmer*, on our dataset [52].

Dependent Variable: Our dependent measure is the credibility score, a continuous variable computed by averaging the five credibility question responses (see Figure 3.4) followed by standardization. We perform robustness checks by rerunning mixed-effects ordinal logistic regressions on each of the five credibility questions. We find no significant differences in the resulting model coefficients, suggesting sufficiency in modeling the credibility score as continuous.

Independent Variables: The independent variables related to RQ1 include main effects and interaction effects derived from the two experimental conditions: users' group (control or treatment) and intervention type (T_R, T_Q, T_U). For RQ2 and RQ3, we examine three variables, including political ideology, political cynicism, and media skepticism. For the sake of analysis, we map political ideology, measured on a seven-point Likert scale, to three groups consisting of Democrats, Independents, and Republicans. Following a prior scholarly work [71], we used the median score across all the questions in each variable (political cynicism and media skepticism) to split the participants into two groups. In our representative US sample, the median for media skepticism was 2.75 ($\alpha = 0.72$, $M = 2.48$, $SD = 0.98$) and the median for political cynicism was 4.00 ($r = 0.53$, $M = 4.09$, $SD = 0.81$). Similar to Carr et al., we considered values greater than the median as high on that category and vice versa [71]. In other words, media skepticism of 3.00 would be labeled high media skepticism, while 2.75 would be labeled low media skepticism. We also include the political leanings of the news sources used in our tweet selection procedure as an additional independent variable.

Control Variables: Prior studies indicate that the level of interest in a news story can influence users' credibility assessment [299]. Therefore, we include participants' interest in a tweet as a control variable, measured on a five-point Likert scale ranging from low to high

	Control Avg. Cred.	Treatment Avg. Cred.	Bet. Subj. MWU-test (Cohen's d)
T_R	0.62	0.67	187488.0(0.162)***
T_Q	0.58	0.55	198180.5(0.072)*
T_U	0.46	0.37	171763.0(0.296)***
n	693	597	

Table 3.6: Mann-Whitney U test results for Study 1. Here, 'n' denotes the number of tweets rated in each condition. Avg. Cred. is the mean of 'n' credibility scores; * $p < 0.05$, *** $p < 0.001$.

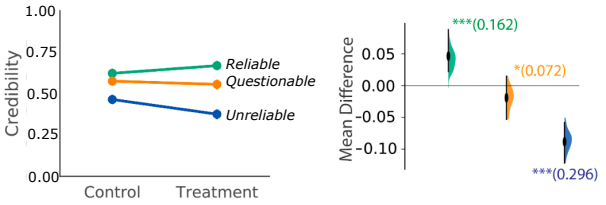


Figure 3.5: Shows interaction effects between user groups and nudge types in Study 1. The numbers inside the brackets are the effect sizes, Cohen's d.

interest. Other control variables include users' demographics, such as gender, age, education, and Twitter usage frequency.

3.5.2 Results (Study 1)

RQ1: Effect of the Nudges

For RQ1, initially we investigated our data using mean comparison. Table 3.6 shows the mean values and Mann-Whitney U test results of our experiment and Figure 3.5 shows corresponding interaction plots. Users in the treatment group rated the credibility of non-mainstream news tweets (T_U) significantly lower than did users in the control group, suggesting the effectiveness of our intervention (Z=171763, p<0.001, Cohen's d=0.296). Additionally, treatment users rated mainstream news tweets without questions (T_R) as more credible than corresponding control users (Z=187488, p<0.001, Cohen's d=0.162). Our participants showed significant decrease in their rating of mainstream news tweets with questions (T_Q) (Z=198180, p<0.05, Cohen's d=0.072).

Our experimental setting with each user rating multiple tweets prompted us to further analyze the data using a series of mixed-effects regression models. Table 3.7 shows this analysis. To determine the effects of the experimental conditions, our base model includes

	Base Model		Politics & Media Model		3-Way Interaction Model	
	β	SE	β	SE	β	SE
(Intercept)	0.29***		0.27***		0.32***	
Control Variables						
Gender (Male)	0.03*	0.05	0.03*	0.05	0.04	0.07
Education	-0.00	-0.01	-0.01	-0.03	-0.01	-0.04
Age	0.00	0.01	0.01	0.03	0.01	0.04
Social Media Usage	0.01	0.01	0.01	0.02	0.01	0.02
Interest in the Tweet	0.09***	0.45	0.09***	0.45	0.08***	0.36
Experimental condition						
Type(Mainstream-Question)	-0.03	-0.05	-0.03	-0.05	-0.07	-0.11
Type(Non-mainstream)	-0.16**	-0.26	-0.16**	-0.26	-0.23**	-0.38
Group(Treatment)	0.04**	0.07	0.04**	0.07	0.08	0.13
Type(Mainstream-Question):Group(Treatment)	- 0.06***	-0.08	- 0.06***	-0.08	-0.11*	-0.14
Type(Non-mainstream):group(Treatment)	- 0.10***	-0.13	- 0.10***	-0.13	- 0.21***	-0.26
Politics and Media						
Ideology(Democrat)			0.03*	0.05	0.01	0.01
Ideology(Republican)			-0.03*	-0.04	-0.03	-0.05
Political Cynicism(Low)			0.01	0.02	0.01	0.02
Media Skepticism(Low)			0.05***	0.08	0.06**	0.11
Account Leaning(Conspiracy)			0.07	0.08	0.06	0.07
Account Leaning(Left)			0.09	0.14	0.09	0.14
Account Leaning(Right)			0.00	0.01	0.00	0.00
Experimental Condition x Other Variables						
Group(Treatment):Gender(Male)					-0.08*	-0.11
Type(Non-mainstream):Interest in the Tweet					0.02*	0.11
Type(Non-mainstream):Group(Treatment):Gender(Male)					0.09*	0.08
Type(Non-mainstream):Group(Treatment):Interest In a Tweet					0.03*	0.10
Adj. R^2 (Marg./Cond.)	.310/.483		.347/.494		.358/.508	
N = 3870					* p<.05, ** p<.01, *** p<.001	

Table 3.7: Regression models on the credibility score. The base model contains nudge type, user group, control variables and the interaction between user group and nudge type. The politics and media model adds users' political ideology, media skepticism and political cynicism variables to the base model. The 3-way interaction model further includes the interactions of nudge type, user group and other variables with significant main effects in the politics and media model (Gender, Interest in the Tweet, Ideology and Media Skepticism).

news source type, group assignment, and their corresponding interactions. We find that tweet type—mainstream or non-mainstream—is strongly correlated with a tweet's credibility score. Non-mainstream tweets are generally rated less credible than mainstream news sources with a small effect size ($\beta = -0.16$, $p<0.01$ and Cohen's $f=.07$). This result suggests that users can differentiate between mainstream and non-mainstream tweets even without our nudges. However, treatment users (those who received nudges) generally rated tweets as slightly more credible than control users ($\beta=0.04$, $p<0.01$, Cohen's $f=0.04$). Nevertheless, there is an interaction effect between tweet type and user group. Treatment users scored non-mainstream tweets lower than control users with a medium effect size ($\beta = -0.10$, $p<0.001$ and Cohen's $f=0.16$). Treatment users also rated mainstream tweets with questions as less

credible than did control users ($\beta = -0.06$, $p<0.001$, Cohen's $f=0.16$). The decrease in the credibility perception scores of both mainstream questioned tweets (T_Q) and non-mainstream ones (T_U) suggests that our nudges can help users consume these news items as less credible, thereby answering RQ1.

RQ2: Effect of Political Ideology

To answer RQ2, we examine the political ideology variable in our politics and media regression model. Politically Independent users serve as the point of reference for this variable. We find that Democrats generally rated all tweets slightly higher in credibility than did Independent users ($\beta=0.03$, $p<0.05$, Cohen's $f=0.06$), whereas Republicans rated them slightly lower than Independents ($\beta = -0.03$, $p<0.05$, Cohen's $f=0.06$). Due to this main effect, in our 3-way interaction model, we further explore whether political ideology had any interactions with the three nudges and the users' group assignments. However, we find no significant interaction. Therefore, nudges changed users' credibility perceptions irrespective of their political leanings. We discuss these findings later.

RQ3: Effect of Political Cynicism and Media Skepticism

To answer this RQ, we examine two variables (political cynicism, and media skepticism) in our politics and media regression model. Between media skepticism and political cynicism, only media skepticism had a significant effect, where users with lower media skepticism rated tweets as more credible ($\beta=0.05$, $p<0.001$, Cohen's $f=0.10$). In our 3-way interaction model, we further explore whether media skepticism had any interactions with our key variables of interest—treatment and control groups, and the three nudges. We do not find any significant interaction effects. Therefore, nudges changed users' credibility perceptions irrespective of

their attitudes towards politics and media. We elaborate on these findings in the Discussion section.

Effect of Control Variables

We examine whether user demographics and users' interest in a news story had any effect on how they rated the credibility of the news tweet. Across all three models, the effects exerted by our control variables are consistent. Independent of whether a user was assigned to the control or treatment group and independent of the type of news (whether mainstream or non-mainstream) they saw, users provided higher credibility scores when they were interested in a story, with a large effect (base model effects:β=0.09, p<0.001, Cohen's f=0.5). Among demographic variables, male users rated tweets as more credible with a small effect size (base model effects: β=0.03, p<0.05, Cohen's f=0.03). The remaining demographic variables did not show any significant effect [6].

3.6 Study 2: Field Deployment

To gain insights into how we can understand and improve the current design nudges for credibility, we conducted a qualitative study. For this purpose, we recruited participants to use NudgeCred on Twitter for five days. This process allowed us to evaluate it in a more ecologically valid environment than Study 1 [59]. Below, we describe the process.

[6] Additionally, we examined whether there was any learning effect compounded from seeing multiple nudges. To do so, we added the order (from 1 to 9) of the tweets in which participants evaluated their credibility and its interaction with nudge type and user group in our regression models. We found no significant effect of the order.

3.6.1 Method (Study 2)

Recruitment

To recruit users for this study, we used Twitter advertising, following a prior strategy [207]. We devised several targeting mechanisms to promote the advertisement to our desired group, including age range (>=18), language (English), location (USA) and whether users followed top mainstream and non-mainstream news accounts in our list. Initially, we were not successful in getting responses from broader regions, so we iteratively revised the location to target nearby states for effective recruitment. Additionally, we promoted the advertisement within our academic network on Twitter. From 50 interested participants, we recruited 12 participants for the study by filtering our spams and users with less than 100 followers . Overall, our participant group consisted of 5 females and 7 males with an average age of 29.5 (std. dev. = 6.5) and a political tilt towards Democrats (Democrat = 6, Independent = 4, Republican = 2).

Procedure

Followed by an informed consent process, we instructed users to install the NudgeCred browser extension. To promote a natural setting, we encouraged users to use Twitter during the five-day study as they normally would. However, it is possible that Twitter's news feed algorithm may not have surfaced news items on their feed each time. Hence, we also encouraged users to visit the Twitter profiles of some news sources each day to ensure that users experience how NudgeCred works. After five days, we asked them to fill out a post-study survey on demographic details followed by an interview in a semi-structured manner (see Appendix A.2 for the interview questions). To facilitate their responses, we asked users to walk us through their news feed during the interview. Each participant received a $30

gift card for participating.

3.6.2 Results (Study 2)

We analyzed the interview data using a grounded theory approach [430]. The first author transcribed the audio from the interviews and analyzed the data to come up with a set of themes. These themes were discussed and refined with the other authors. Below, we present our final set of themes.

NudgeCred facilitates more conscious news consumption

Most of our participants (9/12) provided positive feedback on their overall experience with NudgeCred, referring to the design and application. Some participants (U1,U6,U9) particularly mentioned that they liked the bandwagon heuristic with the questions in replies.

> *"It [NudgeCred] quickly highlights. So you know what to look for. Especially when it's a question mark, I do actually open up the comment section and read what the question is."* ***(U9)***

Others liked it because it served as an educational tool to "train the user to thoughtfully think about" news (U4) or because it did not add overhead to their current Twitter experience (U2). Overall, users reported two phenomena. We describe them below.

Improved Recognition of News Content and News Genres: One of the impacts that participants (5/12) mentioned was the perceived difference in the amount of news content in their feed compared to their prior experience. For example, they perceived that there was more news content in their feed than before. NudgeCred even helped some participants pay more attention to the types of content that news sources produced.

> *"It [NudgeCred] really just told me that NPR produces more articles that are opinionated [the participant was referring to an Op-Ed] on Twitter than I thought."* ***(U1)***

The article labeled as *Questionable* made them realize that it was an op-ed article, rather than news.

Attention towards Bandwagon Heuristic: Out of the three nudges, participants (7/12) noticed the *Questionable* tweets highlighted in yellow the most, and the number of questions below them.

> *"I noticed the one [question icon] at the bottom more ... most people who use Twitter a lot ... 9 out of 10 are more likely in tune with the replies/retweets. Usually I use those as a sign of popularity."* ***(U1)***

> *"If I see it [a tweet] as yellow ... I do get the information that this is either ... a lot of people don't like the news article or this article might have controversial or incorrect facts."* ***(U10)***

While users, as U1 indicated, may see traditional retweet or reply numbers as an indicator of popularity, one participant (U10) correctly pointed out the bandwagon cue as an indicator of controversy. Thus, nudges imitating existing designs on social media can be useful.

Overall, these phenomena support that our nudges can improve users' news perception in two ways: (i) with an overall impression of total news items on users' feeds broken down based on the reliability of sources, facilitating better perception on its genres; and (ii) with individual attention towards particular news items.

Concerns in Using Heuristics for Nudging News Credibility

Interviews also revealed two concerns regarding our nudge design. We discuss these concerns below.

Trust Issues with Heuristics: A majority of our participants (7/12) questioned the use of bandwagon heuristic to differentiate *Reliable* and *Questionable* news items. Because audience composition can vary by the source and the topic of a news item, and influence bandwagon heuristic, they were concerned about its disparate impact. One participant pointed out that followers of the New York Times (NYT) are comparatively more diverse than followers of Fox News. Consequently, audiences with an opposing stance on a news report from the NYT may question it. In contrast, Fox News, having a homogeneous audience which mostly supports its reporting, may hardly question their reports. Therefore, our bandwagon heuristic would be skewed based on the audience of a report.

> *"NYT [The New York times] and MSM [mainstream media] in general have a lot more reactions from skeptical readers given the current administration. And to some, the color-coding and the number of questions may indicate that the news is subjective or "fake" when you compare it with other outlets such as Fox News that have fewer reactions on Twitter and a more homogeneous audience."* ***(U7)***

These responses suggest that even though a user may understand the bandwagon heuristic, the heuristic itself may have some shortcomings, which makes it challenging for the user to trust it as a metric for gauging credibility.

Adverse Effects of Nudges: Our participants (2/12) suggested two adverse effects of the nudges. One participant (U11) proposed that users may use the bandwagon heuristic based on like-minded questions as a justification for their attitude-consistent belief in politically opposing news, thus promoting a **confirmation bias**.

> *"If I agree with you and you are questioning an article that I questioned as well ... since you personally agree with me, it confirms my bias against that piece of information."* ***(U11)***

A similar effect has been suggested by scholars in the past, who show that rating mechanisms on online platforms (e.g., Facebook "likes") may guide users' news selection process [298]. If such effects exist, users may become more polarized. Compared to confirmation bias stemming from the existence of a nudge, the **absence of nudges**, mentioned by U4, could also have a harmful effect.

> *"I think the ones that I subconsciously ignore are the ones that have no color at all. If there aren't any flags ... no color blocks, I am more inclined to assume that the content is valid."* ***(U4)***

This participant suggested that the absence of nudges creates an illusion of validity of content without nudges. Indeed, recent research points out the same phenomenon when false reports are not tagged, resulting in a false sense of being validated [353]. One way to address this concern is, again, to be transparent about not being nudged with an additional tool tip message for the news items that are not nudged.

Overall, our participants' concerns suggest that designers need to evaluate two aspects of nudges: (i) How trustworthy the design components of the nudges are (ii) Whether the presence and absence of nudges adversely affect users.

Opportunities for Credibility Nudge Design

In addition to differentiating news credibility, we asked participants what other functions they would like in NudgeCred. Participants suggested improvements in three directions.

Extending the News Source List: Our participants were concerned on the limited set of news sources we considered. They (5/12) suggested that they often see misinformation from non-news entities, including their acquaintances. To allow more news identification, some (2/12) asked us to include local news sources. With our participants following diverse sources for information, our limited news source list was naturally inadequate for them.

Indicating News Genres and Reporting Biases: Suggestions from the participants included distinguishing opinion items from news (3/12) and indicating bias in a report as a nudge (2/12).

> *"Give a notification that say what I am seeing is an op-ed rather than straight facts."* ***(U4)***

> *"Is it possible to state that this news article is biased towards this particular claim?"* ***(U10)***

A recent survey shows that about 50% US adults "are not sure what an op-ed is" and that about 42% of respondents perceive that opinion and commentary are often posed as news in most news articles [367]. Therefore, a significant share of the population may appreciate having nudges that differentiate op-eds from news as well as other indicators of bias stems. Incorporating such attributes in nudging might help users better determine the credibility of news content.

Curbing Misinformation Sharing: To prevent the sharing of misinformation, some participants (2/12) proposed implementing nudges on share (or retweet) buttons.

> *"[when someone clicks the share button] If there is a notification that says this source is not credible then people would be less likely to share it."* ***(U2)***

Research indicates that about 59% of links shared on Twitter have never been clicked [146], i.e., users often share news items without reading them. If nudges can help user determine the unreliability of news from misinformative sources, they might be less likely to share them.

In summary, our participants proposed improvements to our nudge design in three key areas: (i) improving existing classifications by extending the source list, (ii) expanding news classifications of nudges in alternate areas, and (iii) targeting users' interactions with news on social media.

3.7 Discussion

Below, we elaborate on our results, starting with each research question from Study 1, followed by opportunities and challenges suggested by the participants in Study 2.

3.7.1 RQ1: Effect of Nudges on Credibility

Our regression analyses in Study 1 revealed that users' credibility ratings were considerably different between the treatment (nudged) group and the control group. While other nudge designs have proven effective in reducing sharing intentions of misinformative content, their effectiveness was shown on a particular news genre, such as COVID-19 and HIV [319, 354]. In contrast, we examined the effects of our nudges on a wide variety of popular news items surfacing over multiple days, thus offering a more generalized result. Our intervention provides a less authoritative approach that gives users simple but transparent information for them to make their own judgments. News feeds, as they typically present limited information on social media, have few features to distinguish content quality. Tweets from bots, friends, mainstream news, and non-mainstream sources are all given equal weight in terms of

visual representation in any feed, making it difficult for users to sift through them. Though people are capable of identifying misinformation, social media makes it challenging to make informed analytical judgments [352]. Our results suggest that users might appreciate it if social media sites provide tangible signals to work through this clutter, which is further exemplified by participants' suggestion to differentiate news and op-eds in Study 2.

Apart from facilitating better perceptions of the credibility of news, NudgeCred may also act as a "translucent system" [124]. The theory of social translucence posits that we should aim for systems that make online social behavior *visible* to facilitate *awareness* and *accountability*. Note that our participants in Study 2 suggested improved recognition of particular types of news content on their feed and were more aware of what they were seeing on their feeds. Our nudges on news content that are liked or shared by users' followers or friends could also have similar impacts, wherein users become more *aware* of their peers' news consumption behaviors. When their peers like/share misleading news, one may hold the peers *accountable* by pointing out the misleading content. Besides, after seeing the nudges on unreliable content, users may restrain themselves from sharing such content and reflect on their sharing habits.

3.7.2 RQ2: Influence of Political Partisanship on Nudge Effects

Our regression results suggest that NudgeCred changed users' perception of credibility irrespective of their political views; that is, there were no interaction effects between political characteristics and the effects of interventions. This result is consistent with recent studies showing the success of interventions in limiting sharing intentions of misinformation, irrespective of users' political affiliation [387, 494]. Although some prior literature argue that citizens may view nudges as partisan policies and may not support nudges when they con-

flict with users' partisan preference [443], other scholars suggest that this behavior can be countered by reducing partisan cues in nudges [449]. We incorporated this suggestion in our design by showing nudges on news content from all political leanings and nudging news content in both directions (Reliable, Questionable, and Unreliable). However, in practice, users tend to typically follow news based on their partisan preferences [241]. In such a setting, users who follow only alternative fringe sources may see mostly *Unreliable* nudges triggered on their reports and perceive NudgeCred as partisan. One potential design solution is to show the similar news item from reliable sources, with the same partisan view, to help them understand the alternatives.

3.7.3 RQ3: Influence of Political Cynicism and Media Skepticism on Nudge Effects

In our study, we did not find any impact of political cynicism or media skepticism on nudge effects. This convincing nature of our nudges—that nudges worked irrespective of users' prior media skepticism and political cynicism—is promising. Our result for media skepticism is aligned with a recent work where media skepticism did not affect nudge effects [319]. Research suggests that media skeptics, despite significant exposure to alternate sources, still seem to have moderate to high consumption of mainstream news [464]. Therefore, our nudges could improve news credibility assessment of both mainstream and non-mainstream sources by skeptics in the wild. Scholars suggest that exposure to fake news mediated by belief in its realism increases political cynicism [32]. Thus, if nudges can reduce belief in fake news, it could help mitigate increasing cynicism towards politics. Furthermore, our nudges can be utilized as an alternative to censorship by social media, thus helping mitigate the concern that social media apply censorship in a disparate manner across different political affiliation, as raised by participants in our formative study.

3.7.4 Opportunities in Designing News Credibility Nudges

Our field deployment of NudgeCred showed several opportunities in designing credibility nudge in the future. First, participants' attention to the bandwagon heuristic reveals how designers can utilize existing Twitter infrastructure in their design. Though the impact of the bandwagon effect in collaborative filtering has been discussed in the literature [436], it has been underutilized in a news credibility context. Our study suggests that applications similar to ours can act as valuable markers of information credibility. Second, participants seeking nudges on a wider set of sources (e.g., news and non-news sources), and alternate types of taxonomies (e.g., news and op-eds) suggests their need for tools to differentiate information reliability. Comparatively, nudges on tweet actions (e.g., retweet and share) may play a stronger role in curbing the spread of misinformation, as research indicates that sharing without thinking is often weaponized for this purpose [484]. For example, when users click on the share button, activating a "timer nudge"—a visual countdown to complete the action—could make users rethink before they share [481].

3.7.5 Challenges in Designing Nudges with Heuristics

Our final evaluation presented several challenges in designing credibility nudges. First, participants showed their skepticism towards the selection of heuristics (e.g., bandwagon heuristic) in design. Though the bandwagon heuristic can be robust and hard to falsify, in an open forum such as Twitter, it is open for manipulation. Perhaps, as one of the participants suggested, feedback on the validity of the question may be helpful. Still, problems may also exist with feedback. For one, due to partisan attitude, feedback wars, similar to edit wars in Wikipedia, might result [434]. Additionally, due to frequent updates and the scale of social media content, feedback from volunteers might be scarce on most items. Perhaps a system

designer can show the distribution of feedback by users' leanings to reduce the effects of feedback wars and solicit user contributions by promoting items with scarce feedback. Second, participants' concerns with the bandwagon heuristic promoting confirmation bias might be an extension of prior findings that users tend to prefer selecting information (including political information) consistent with their preexisting beliefs [236]. However, scholars have shown that the extent of confirmation bias in information selection, particularly in news selection, is small [190]. In the event of confirmation bias, we can computationally de-bias the heuristic cue, as in prior works [3]. Lastly, audience misperceptions of non-nudged content being credible indicate an additional challenge in design. This effect has also been demonstrated in a recent work [353]. One way to solve this problem would be to add nudges to all content. Aside from these challenges, one participant (U7) pointed out that *"News might change and the user will not see the update when more legitimate questions are added to the replies"*. As user interactions accumulate over time, the number of questions in replies could change, wherein the same tweet would be categorized as *Reliable* at first and *Questionable* at a later time. This change in nudge category stemming from our choice of the bandwagon heuristic could imply an inconsistency in nudge design. System designers can incorporate delayed intervention mechanisms and inform users of this cold-start issue. Overall, these challenges inform designers about considerations for designing nudges with heuristics.

3.8 Implications And Opportunities for Designing Credibility Nudges

We built NudgeCred by fusing the theoretical background on nudges with theories underpinning credibility evaluation. Researchers have the opportunity to explore the nudge categories we built our design around, experiment effectiveness of other heuristics and utilize alternate

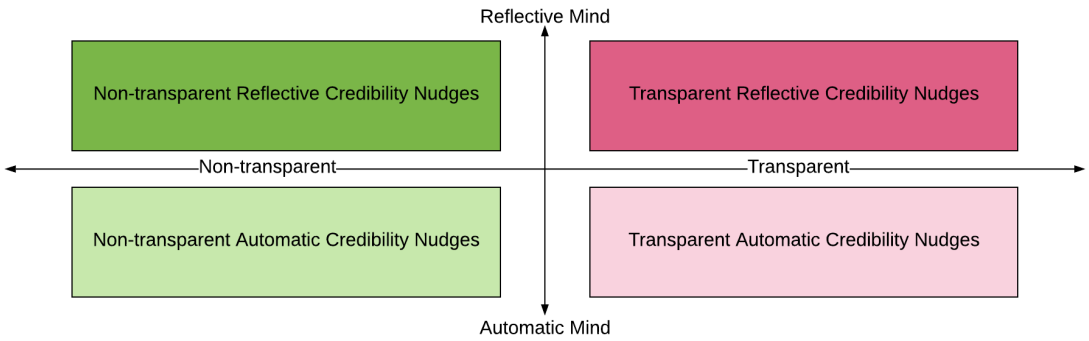


Figure 3.6: Types of nudges based on transparency and mode of thinking. This figure emulates Figure 1 by Caraban et al. [70]. This work lies in the bottom-right quadrant.

nudging method. Below, we elaborate a few possibilities around these areas while discussing potential considerations.

Exploring Additional Nudge Categories Referring to the four categories of nudges, divided along the two dimensions of *transparency* and *mode of thinking* (automatic or reflective), we have illustrated how NudgeCred resides on one of the four quadrants—*transparent automatic* (see Figure 3.6). Nudge theorists describe transparent nudges with reflective and automatic mode as, respectively, reflective decision making and instinctive decision making [186]. Technology-mediated nudging research focus more on transparent reflective quadrant [70]. For example, Facebook's "Related Articles" feature to tackle misinformation exemplifies such a nudge design [422]. Twitter's blue check "verified" marker on profiles is another example of a transparent automatic credibility nudge. Between reflective and automatic mode of thinking, each has their own benefit. For example, while nudges with automatic mode of thinking can be diminished over time, reflective mode of thinking can educate users and have a lasting impact. On the other hand, design considering reflective mode of thinking requires additional considerations such as motivating the users and assisting in reflection [2]. For example, to motivate users we can show statistics of how often users

misread a news tweet as a nudge and to assist them in reflection this nudge can include the list of the most common mistakes. For non-transparent quadrants, scholars propose reflective ones as a manipulation of behavior and automatic ones as a manipulation of choice, in both cases without users' awareness. An example of non-transparent automatic nudging could be showing the average reading time of an article which could prompt users to think shorter articles as less detailed and less credible. Comparatively, research show fewer work in non-transparent reflective quadrant [70]. Due to their deceptive nature, designers need to consider ethical consideration while experimenting on non-transparent quadrants. Overall, this is the start of work in this space; much research needs to be done in cataloging and evaluating news credibility nudges along the other dimensions spanning the four quadrants.

Exploring Alternate Heuristics for Nudging Credibility Our nudge design is based on two heuristics under the **A**gency affordance, drawn from the MAIN model's list of technology-mediated affordances affecting credibility judgment. Designers have the opportunity to explore the other heuristics in their design. For example, designers could offer design cues to distinguish news items with/without video footage and prompt a *realism heuristic* from **M**odality affordance. Or, design cues distinguishing news tweets with interactive content (e.g., interactive 3D footage or charts) could prompt *interaction heuristic* from **I**nteractivity affordance. For **N**avigability affordance, designers can prompt *browsing heuristic* by providing additional hyperlinks to journalists' past activities (e.g., MuckRack profile [376]) besides news items.

Examining Alternate Nudging Method While the original proposers of the concept did not lay out a fixed method for creating successful nudges [409], Caraban et. al. recently devised six broad categories for 23 nudging methods used in prior HCI works, namely,

`Facilitate` (e.g., *default choice* and *hiding*), `Confront` (e.g., *remind consequence* and *provide multiple viewpoint*), `Deceive` (e.g., *add inferior alternative* and *deceptive visualization*), `Social Influence` (e.g., *enable comparison* and *public commitment*), `Fear` (e.g., *reduce distance* and *scarcity*) and `Reinforce` (e.g., *ambient feedback* and *subliminal priming*) [70]. Under these categories, NudgeCred utilizes two methods from two categories. First, it works as a `Facilitate` category by facilitating decision making through a combination of color-coding and translucent *hiding* method. Second, it operates as a `social influence nudge` category with *enabling social comparison* method through the use of number of questions asked by other responders. Technology-mediated credibility nudges can utilize other methods from this classification. For example, similar to NewsCube [342], designers can use `confront` category by *offering multiple viewpoint* method on news items. Or, flashing keywords around news items (e.g., "reliable" or "questionable") utilizing *subliminal priming* method under `Reinforce` category [356], could affect users' credibility perception of those items.

Designing Against Adversarial Agents. Our approach to be transparent suggests that adversarial entities, upon knowing algorithmic detail, can manipulate the system. For example, by commenting on a factual mainstream news with a question, they can create a misperception of it being questionable. This is a problem that most online platforms struggle with–balancing between being transparent about their algorithm while safeguarding against adversaries. Indeed, platforms like Reddit while publishing their ranking algorithms include "some fuzzing" to avoid manipulation of their voting system [91]. Hence, some opaqueness in the algorithm might be desirable. At the same time, platform developers could also misuse this fuzziness in the algorithm for their own benefit, such as to drive engagement [187]. Indeed, Twitter's content moderation practice has been controversial in the past, in some cases resulting in reversal of moderation decisions [273]. Similar controversy could arise re-

garding nudging policy. Therefore, designing nudges requires consideration for the multiple stakeholders of a social platform—platform developer, consumers and news producers. Designers would need to consider the degree to which a nudge would be resistant to adversarial attacks by each stakeholder. For example, crowds' question based bandwagon heuristic has a high-level of susceptibility of manipulation by the consumers compared to the platform developers and news producers. On the other hand, our authority heuristic is more susceptible to manipulation by the platform designers compared to the news producers and consumers. Overall, a potential solution to this problem would be creating a collaborative standard authorized by all stakeholders. For example, Facebook has already created a third-party oversight board for content moderation [49]. A similar strategy can be applied to determine nudging criteria.

Considering Shortcomings of Nudging as a Design Approach Despite its success, nudging has its own shortcomings. Prior literature proposes that nudges may have diminished effects in the long term for two reasons: (i) nudges relying on automatic cognition may fail to educate the user, and (ii) prolonged exposure may have various effects such as transform nudge effects into background noise [262, 444], invoke a feeling of intrusiveness [482], and reduce users' perception of autonomy [263]. Such effects may lead to unforeseen consequences. For example, a default choice nudge promoting vaccination that reduced users' perception of autonomy resulted in users unsubscribing from the program [263]. In our case, if users repeatedly encounter their favored political news source labeled as *Questionable* or *Unreliable*, they could become averse to the design. However, designers can apply several strategies to counter this problem. On one end, they can alter the design over time or prompt users to change the intervention settings over time [247]. As an alternate, they can also choose to deploy reflective nudges which are less susceptible to the diminishing effect. A potential problem with the altering design is that news consumers may need to re-learn

how nudges operate. Regardless, designers would first have to understand the rate at which nudge effects diminishes, a direction for future research.

3.9 Limitations and Future Work

Our work is not without limitations. Our Study 1 was conducted in a controlled setting as opposed to in the wild. However, we see two reasons that suggest that our results demonstrating the utility of nudges in credibility assessment could extend to naturalistic Twitter setup as well. First, research shows that self-reported measures pertaining to social media usage correlate with observed behavior in the wild [177, 353]. Second, large-scale survey-based research on nudges pertaining to news on social media, show that nudges affect related attitude, such as sharing intention of misinformation [319, 354]. Our second limitation relates to choice of population. Because we tested variables (e.g., political ideology and media skepticism) that has different meaning across countries, we had to limit our experimental population to the US. To generalize our findings to the global population, future research could replicate our study in the context of each country. Third, our recruitment had limitations that are characteristic of any online service-based recruitment. Though we may not have obtained the true nationally representative sample in the US, research suggests that Qualtrics provides reasonably representative US sample (approximately 6% deviance from the US census) around demography such as gender, age, race and political affiliation [197]. Overall, a large-scale Twitter deployment might reconcile these concerns in the future. We initially attempted to do so by contacting a large number of Twitter users, without much success due to a lack of platform-level cooperation (Our research account was repeatedly blocked). While a 2015 study had successfully piggybacked on Twitter's infrastructure to run large-scale recruitment efforts on the platform [173], we were unable to do so, despite fol-

lowing similar strategy. We anticipate that changes to Twitter's policies may have prevented us from running large-scale recruitment on the platform [208].

3.10 Conclusion

In this study, we provide evidence that a nudge-based design that directs users' attention to specific social cues on the news can affect their credibility judgments. We used three nudges: *Reliable*, applied to mainstream news tweets without questions in replies; *Questionable*, applied to mainstream news tweets with questions in replies; and *Unreliable*, applied to non-mainstream news tweets. Our experiment suggests that users who saw tweets with *Reliable* nudge as more credible, and tweets with *Questionable* and *Unreliable* nudges as less credible compared to the control users. Moreover, our nudges were not affected by users' political preferences, political cynicism, and media skepticism. Through interviews, we found evidence of how nudges can impact users' news consumption and how the current design can be improved. This research proposes further exploration of nudge-based system design approaches for online platforms.

Chapter 4

TransparencyCue: Designing Transparency Cues in Online News Platforms to Promote Trust

4.1 Introduction

While growth in communication technology has connected news consumers to diverse sources of information, it has also prompted a steady decline in public trust in the mainstream media, as well as the exploitation of public opinion through the spread of "fake news" or misinformation [27, 78] [1]. A partial cause of the larger issue of distrust in mainstream media might be attributed to news consumers' inability to distinguish quality journalism from deceptive content; this inability is the focus of this study. To this end, scholars suggest adopting greater level of transparency within the news-making process [289, 359, 407]. This shift from a "trust me" to a "show me" journalism has the potential to promote trust among readers, especially by protecting against erroneous and deceptive sources of reporting [246, 311]. Online news platforms have the potential to offer novel interface designs to support such an increase in transparency [48, 89].

Despite calls for transparency in news, surprisingly little is known about how to promote

[1] part of this chapter appears in [44]

transparency to increase trust in news articles, especially from a design perspective. In this respect, designers in news organizations need to consider the competing perspectives of stakeholders; for example, journalists may want to protect certain information, whereas news consumers may expect it to be revealed. Furthermore, journalists and consumers have to consider organizational policy. Although some prior scholarly work has explored transparency practices in journalism [79, 175, 220], these works discuss them without a particular focus on design and without considering the perspectives of both stakeholder groups. Some recent work has promoted transparency standards for news websites [369], but they focus largely on site-level measures instead of article-level measures. As discussions to support transparent journalism are just starting to gain traction in HCI communities (see the CHI'19 workshop on the topic [4]), we have yet to identify effective design principles that can guide the design of these user interfaces.

This work bridges the gap between communication literature and HCI design in terms of presenting transparency in news articles. We examine the perspectives of two stakeholders—journalists and news consumers—to inform designers of ways to embed transparency in news articles. In particular, news organizations both large and small with internal design teams (e.g., The New York Times [457]) can benefit from the findings of this work. Additionally, news aggregators (e.g., Google News) and other third-party news providers (e.g., `AllSides.com`) may also choose to adopt our design suggestions on how transparency should be practiced in digital news articles. Communication scholars have suggested that people take into consideration certain information characteristics when evaluating whether a piece of information is trustworthy. People assess the *source* of the message (e.g., the author's expertise) and the *message* itself (e.g., the quality of the message), and they use these insights as meaningful cues to evaluate the overall trustworthiness of a piece of information [299]. Combining the concept of design cues with transparency, we argue that disclosing certain

source- and *message*-level transparency cues within news articles can be an important step towards designing trustworthy online news websites [139, 220, 435]. Overall, we ask:

RQ. *How can designers utilize transparency cues to promote trust in digital news articles while considering the perspectives of journalists and news consumers alike?*

RQa. *What aspects of journalistic practice do news consumers and journalists want disclosed within news articles as transparency cues?*

RQb. *What should designers consider in promoting transparency cues for news consumers and journalists?*

To answer these questions, we interviewed journalists and news consumers. Our journalist pool consisted of 15 journalists with diverse reporting backgrounds who worked in local, national, and international newsrooms. For our news consumer interviews, our pool comprised 16 citizens from both local and online communities with novice to savvy news-reading behaviors. We complemented our interviews with a scenario-based design approach [72, 202]—a user-centered design approach that employs prototype mockups and descriptions to help end-users envision a future system and how they will use it. We provided our interview participants with a scenario: a prototype mockup developed with two sets of transparency cues. These cues disclose underlying journalistic practices pertaining to the *message* being reported and the *source* reporting that message (see figure 4.1). Stimulated by our scenario, participants recommended various transparency cues capturing source and message characteristics to address the reasons for underlying distrust, with some additional transparency cues to reveal the process of reporting. Many of these suggestions concern information which is fundamental to high-quality journalistic news reporting.

Our analysis revealed that news consumers identified two main areas in which increased transparency could improve trust; namely, the level of objectivity in a report and the qual-

ity of the evidence that underlies the claims in a report. For objectivity, our analysis showed that designers have the opportunity to incorporate transparency cues to reveal biases in news selection and framing. For evidence, our news consumers asked for transparency in evidence presentation, anonymous sourcing, fact-checking, and correction. Meanwhile, our journalists proposed transparency in evidence presentation, the reporting process, and personal/organizational conflicts of interest (e.g., gifts journalists might receive and stakes in other business organizations owners might have). Both groups suggested the need for transparency cues to highlight evidence in news reporting, such as by including verification materials. The scenario design also prompted participants to point out additional considerations, such as designing for ease of comprehension, reducing bias while presenting transparency indicators, and including appropriate details within the transparency feature (such as displaying both the number and nature of corrections to an article). Other suggestions include designing markers for distinguishing reporting quality within news items, and for distinguishing between news and non-news items. When we compare the responses between our two stakeholder groups, we see conflicting suggestions that would require trade-offs in design. For example, trade-offs exist between being fully transparent and preserving autonomy by placing limits on transparency. Similar trade-offs exist between opting for simple features to ease comprehension and including sophisticated features to illustrate appropriate nuances. We discuss how designers can utilize existing journalistic practices in designing transparency cues catering to our participants' areas of interest, and thus, how they can consider the perspectives of both stakeholders in their designs. However, to draw a more complete picture, future work would need to consider news organizations and sources (people) themselves as stakeholders of news. Furthermore, while participants responded with consistent themes in the interviews, without properly controlled experiments, our results are not evidence of efficacy of the design ideas. Still, our results have implications for design practitioners in highlighting appropriate journalistic practices on news websites through transparency and for

design researchers in future investigations of balancing the constraints imposed by stakeholders, considering remaining stakeholders, and conducting controlled experiments to evaluate efficacy.

4.2 Related Work

4.2.1 Defining Transparency in Journalism

With rapid improvements in communication technology, the news-gathering process has shifted from a model of verification (the principal value being getting things right and exercising caution against getting things wrong) to a model of assertion, where the principal value is to get the news out fast, even at the expense of rigorous verification [224]. This shift may result in journalists making more mistakes, thereby directly affecting news consumers' trust in journalism [19]. Some scholars argue that transparency (or openness) could be a mechanism for building both trust and accountability [10, 358]. In this work, we define transparency as this notion of openness about the news makers themselves, their journalistic routines and practices, and their decision-making processes [465]. Prior work shows that one of the ways journalists practice transparency is through disclosure of information, known as *disclosure transparency* [220]. Disclosure transparency can be realized in different ways, such as how story assignments are decided over meetings, disclosing verification process, linking source material for a report, and acknowledging and correcting errors [195, 256]. When discussing transparency in this work, we utilize this principle of disclosure transparency.

4.2.2 Existing Transparency Practices in Journalism

The earliest examples of transparency practices in journalism date back to the 1930s, when bylines containing the names of reporters responsible for an article were first introduced [405]. Practices have evolved from early practices like the use of ombudsman columns (which goes back to the 1960s) to newer ones, such as inviting users to experience the journalistic process firsthand (e.g., attending news meetings) [100, 112]. For a comparison of transparency practices across newsrooms, see [459]. Prior scholarly works looked at various aspects of transparency in online news. Examining three online newspapers (The New York Times, The Guardian, and Dagens Nyheter), scholars found that each of these organizations implemented different sets of transparency features [220]. Revers found wide adoption of transparency in social-media-aided reporting [381]. Offering transparency features across organizations is a comparatively recent phenomenon. For example, some social media sites have begun offering transparency in news content (e.g., through a corrections and ethics policy) [132, 423, 445, 455]. Despite significant work, research seems to look at transparency in isolation, from either journalists' or news consumers' perspectives alone. We bridge this gap in existing work by exploring both journalists' and news consumers' views on a problem scenario.

4.2.3 Designing for Trust in News Through Information Disclosure

HCI research has been considering trust as a driver for design for a considerable amount of time in such areas as e-commerce, remote work, and digital currency [55, 95, 382, 399]. Generally, these works use two concepts of trust: trust being mediated by technology for individual-to-individual relationships and technology being the object of trust [97, 179]. Existing works support that openness (or transparency) influences trust [238]. In this work, we explore the trust relationship between individuals and information. In online informa-

tional systems, particularly for news, trust is often considered synonymous with credibility [140, 448, 460], and we adopt this concept in our work. Communication literature further guides us around the relationship between information disclosure and trust.

In communication literature, trust has typically been considered a subset of the larger concept of credibility, where individuals who "trust" a given speaker (i.e., the source) or piece of information (the message itself) also deem it more credible [299]. Studies have suggested that when people evaluate the trustworthiness of online information, they consider a number of characteristics. These may include evaluations of who conveyed the information (the source), how it was said (the message), and where it was said (medium or channel characteristics) [230, 299]. However, research also indicates that most people don't spend a tremendous amount of time engaging in close analysis of online content, often relying instead on surface-level features, such as the quality of a website's design or how easy it is to use [139]. Several theories have expanded this observation and shown that individuals have only a limited number of cognitive resources to evaluate information online; therefore, they often rely on cognitive heuristics to evaluate information [128, 435]. By cognitive heuristics, we mean the mental shortcuts individuals use to ignore some information on a site while paying attention to other information when evaluating the trustworthiness of the claims presented. On any given website, various design features serve as cues to trigger a heuristic. For example, Facebook uses "likes" as a cue to trigger the "popularity heuristic" of a given post [54].

Several researchers have attempted to explain the role that these feature cues play in helping people evaluate information. Fogg's Prominence-Interpretation Theory suggests that people evaluate the trustworthiness of a website based on feature cues that attract users' attention the most [140]. For example, only if a user notices a feature cue (such as an indicator of an author's expertise) would that cue have an impact on the user's overall perception of the source's credibility. Sundar's MAIN Model also suggests that credibility is conveyed online

by four broad website characteristics that serve as their own cues to users for denoting the credibility of information [435].

Drawing upon the notion that feature cues can be used as ways to trigger heuristic evaluations of a given piece of content, we explore the design of a series of disclosure transparency cues that aid users in evaluating the trustworthiness of a given piece of online news. Literature on the subject does not provide any exhaustive list of features for disclosure. Therefore, when designing our problem scenario, we selected a few features mentioned in existing literature that could influence users' trust. For example, studies have shown that a source's apparent expertise has a significant positive impact on users' trust in it [302]. Similarly, message-level characteristics, such as how crucial pieces of information in an article are emphasized or framed can improve the article's level of transparency and cue readers to evaluate its claims meaningfully. In Section 4.3.1, we explain in greater detail how we used these characteristics to develop our scenario.

4.2.4 Effect of Transparency on Perception of Trust

Though scholars suggest a positive association between transparency and perceptions of credibility [195, 221, 246], when investigated, effects of transparency on news consumers' perceptions of a news article's credibility seem somewhat ambiguous. While one prior study suggests that a single transparency feature had almost no effect on readers' perception of news credibility [223], a recent study supports the idea that multiple transparency indicators can improve an audience's perception of news articles' credibility [103]. News consumers may also prefer certain features over others (e.g., hyperlinking source material over describing how an article was framed) [222]. Taken together, there seems to be some effect of transparency on trust; however, proper sets of features must be constructed, and their construction is the

focus of this work.

4.3 Interviewing News Consumers and Journalists Using a Scenario

We studied our research question through interviews. For this purpose, we built a scenario to stimulate our participants. Below, we describe the scenario and our recruitment process.

4.3.1 Developing a Scenario

One of the main challenges in designing transparency attributes in news websites to build trust is obtaining stakeholders' reactions to a system that has not yet been built. In such cases, scholars have suggested opting for an alternate approach, called a scenario-based approach, to facilitate users' reactions [72]. A scenario not only provides concrete examples to stimulate users' responses, but it also provides a holistic view of future possibilities. Furthermore, scenario narratives can be made flexible and adapted to expand a user's imagination. In his seminal work on scenario-based design, Jack Carroll suggests that such an approach motivates a more integrative problem analysis compared to traditional requirement gathering techniques [392]. We exploited this approach to build an example news article with multiple transparency cues. Our scenario is powered by disclosing two sets of transparency cues: one that captures the characteristics of the *message* being reported, and one that shows the quality and expertise of the *source* reporting that message. Source- and message-based cues are fundamental for signaling the underlying credibility of a report [435], and are thereby key to enhancing users' trust in journalistic practices. Figure 4.1 shows the scenario we developed. The left side of this figure contains a news article, and the right side contains the

transparency cues. Below, we describe how we adopted and implemented each of the design characteristics shown.

Source Credibility Cues

Prior literature indicates that source expertise is one of the primary indicators that people use to evaluate the credibility of a news article. In this context, users may consider sources from two perspectives: sources can be taken to be either the organizations that employ journalists or the journalists themselves. In our design, we disclose three indicators of a journalist's expertise, including experience in years, number of retractions, and domain of expertise. Here, we contextualize the number of retractions by presenting it together with the total number of articles written by the journalist. For example, 2/100 indicates that the author had 2 retractions out of the 100 articles she wrote. In Figure 4.1, we show this information in the upper-right corner as part of feature "a". We used this terminology (features "a," "b," and "c") during our interviews to simplify references to specific features for our participants.

Message Credibility Cues

To give readers cues about the quality of an article, we turn to existing journalistic practices in article writing. We rely on cuing users with the presence of two types of attributes inside a report: attributes identifying the degree of completeness of crucial information in a report (as shown in feature "b") and how information is presented within the report in terms of its use of language characteristics that could indicate bias (as shown in feature "c").

Cuing the Presence of Crucial Information. As early as the mid 19th century, traditional journalism has followed the practice of formatting a story by first presenting the

Figure 4.1: Our scenario with three transparency features. Here, feature "a" corresponds to source characteristics conveying the expertise of the author; features "b" and "c" are message characteristics showing, respectively, crucial details about the event and the reporting style. In feature "c," reporting style includes whether the article is high or low in summary news lead (SNL) or inverted pyramid style reporting, the proportions of first- and secondhand accounts, the proportions of direct and indirect quotes, and the number of claims made.

crucial information about an event, called main event descriptors [362]. These main event descriptors answer fundamental questions about an event, and these questions are sometimes referred to as 5W1H: *who, what, where, when, why*, and *how*. To give users cues as to the presence or absence of these descriptors, we highlight each of them in our design with different colors, as shown in Figure 4.1 in the section named "News Snips" (or feature "b"). This feature was inspired partially by Wikipedia's Infoboxes [490], and it isolates the most important information within a news report. These descriptors can also be calculated computationally—multiple prior works have calculated the 5W1H descriptors by applying natural language processing techniques [203, 267, 326, 479, 480].

Cuing Reporting Style. To give readers cues as to how a report is written, we resort to identifying two aspects of standard journalistic practice in a report: how information is presented and how it is attributed to its source. Journalistic practice in structuring news

articles varies, though a handful of patterns are commonly used [189]. In our design, we focus on "hard news," or news that involves political, economic, and social issues [269]. This is also a prevalent type of news in the misinformation domain, and these articles tend to follow a particular pattern. Prior literature suggests that journalists traditionally format hard news by organizing information in a top-down style, starting with the most important information before presenting supporting information in diminishing order of prominence [362]. This format is also known as the "inverted pyramid style," and it is a common structure in content from news agencies like the Associated Press, whose content is frequently syndicated to many national and international outlets. We examine the degree to which a report follows this standard of writing through a metric called the summary news lead, or SNL [125]. In Figure 4.1, the left side of feature "c" visualizes this metric as high or low SNL, indicating the degree to which a report follows this style.

Another crucial property of an article is the sources underlying it. Traditionally, journalists divide sources into two categories: firsthand accounts and secondhand accounts. Firsthand accounts are retellings of reporters' direct observations of an event, while secondhand accounts come from authoritative sources with direct knowledge of an event. While both first- and secondhand accounts are useful, firsthand accounts are more accurate [297]. To give users cues as to the nature of sources used in an article, we identify the proportion of claims based on first- and secondhand accounts.

Bias reduction is yet another standard journalism practice. To identify the presence of bias or subjectivity in a report, we examine the framing of quotes in reported secondhand accounts. Reporters can quote secondhand accounts either directly or indirectly [38]. They may also introduce their own biases in the process by using framing that alters perspectives [172, 271]. For example, a journalist may describe a secondhand account using the word "claim" to express doubt in the quoted statement. In prior works, natural language processing has

Journ.	Gender	Exp. (yrs)	Journ.	Gender	Exp. (yrs)	Journ.	Gender	Exp. (yrs)	Journ.	Gender	Exp. (yrs)
J1	Male	7	J5	Female	20	J9	Male	5	J13	Male	5
J2	Male	11	J6	Male	5	J10	Female	9	J14	Male	6
J3	Female	3	J7	Female	3	J11	Female	7	J15	Male	10
J4	Female	3	J8	Male	35	J12	Female	3			

Table 4.1: Demography of our journalist pool. "Journ." here stands for "journalist."

been applied to detect these kinds of cues [400, 425]. Both observation types and biases are shown to the right of the article as feature "c" in Figure 4.1.

4.3.2 Recruitment

Using this design, we interviewed 15 journalists from various news organizations and 16 general news consumers recruited from the local community and online communities on Reddit. Our interviewee pool primarily resided in the United States, with the exception of two international journalists. We recruited journalists through a combination of personal contacts and social media (Twitter and Facebook). We recruited a majority of our journalist pool ($n = 11$) from personal contacts. Due to the recent focus on misinformation in US politics, we also searched for journalists covering politics and misinformation on Twitter through keyword searches of Twitter bios containing a combination of the following terms: "journalist" and another word from "misinformation," "disinformation," and "politics." After manually verifying the collected Twitter profiles, we compiled a list and reached out to eight journalists via email. Out of the eight, two eventually participated in an interview. On

Role	Network	Audience
Anchor, editor, reporter (corporate, court, foreign, misinformation, politics), technology research	ABC, AP, Athens Messenger, Buzzfeed, CBC, Daily Reflector, freelance, KBZK, MotherJones, News4SA, Toledo Blade, Roanoke Times, WSJ	Local/metro, national, international

Table 4.2: Professional background of our journalist pool. In the network column, "freelance" indicates that the journalist is not associated with an organization. Note that we aggregated the roles of journalists and their associations to news networks and audiences to ensure anonymity, as required by the IRB. Knowledge of network affiliation and role would have been enough to reveal the identities of several participants.

News Cons.	Gender	Profession	Political Orient.	Leaning
U1	Male	Attorney	Democrat	Conservative
U2	Male	Real estate agent (veteran)	Libertarian	Conservative
U3	Male	Residential service	Republican	Conservative
U4	Male	Educator (retired)	Democrat	Liberal
U5	Female	Director (local theater)	Democrat	Liberal
U6	Male	Construction management (ex-police)	Libertarian	Conservative
U7	Male	Software developer	Independent	Moderate
U8	Female	Graduate student (neuroscience)	Democrat	Moderate
U9	Male	Manager (food service)	Libertarian	Moderate
U10	Male	Undergraduate student (geography)	Independent	Moderate
U11	Male	Undergraduate student (aerospace)	Democrat	Liberal
U12	Female	Financial advisor	–	Liberal
U13	Female	Management consulting	–	–
U14	Female	Special education teacher	–	–
U15	Female	Civic organization (ex-nurse)	Independent	Liberal
U16*	Female	Civic organization	–	–

Table 4.3: Demography of our news consumer pool. We asked demographic questions at the beginning of each interview (see Appendix A for the list of questions). Note that some of the participants chose not to specify a political affiliation. * This participant was a former journalist.

Facebook, we joined an invite-only Facebook group of international journalists (IJNet[2]), which comprises a global network of more than 90,000 media professionals.We posted a recruitment advertisement in this group and were able to recruit 2 international journalists. Through these combined techniques, our final journalist group consisted of 15 journalists, with 2 reporters residing outside the US. Our 15 journalists (hereinafter referred as J1–J15) had varying journalism experience, including such roles as anchor, editor, reporter (corporate, court, foreign, misinformation, politics, social media) and technology research. Their network affiliations included ABC, AP, Athens Messenger, Buzzfeed, CBC, The Daily Reflector, freelance, KBZK, MotherJones, News4SA, Toledo Blade, Roanoke Times, and WSJ. The distribution of journalists' genders was even. Their experience in the industry ranged from 3 to 35 years, and they served audiences at varying levels, including local/metro ($n = 7$), national ($n = 6$) and international ($n = 2$). Tables 4.1 and 4.2 show the demography and professional background of our journalist pool. To preserve journalists' anonymity, we do not report their roles and organizations individually.

[2]https://www.facebook.com/IJNet/

For news consumer recruitment, we reached out to the local Chamber of Commerce and several local institutions (e.g., theaters, schools, and civic organizations). Additionally, we advertised on Reddit using multiple approaches. We posted on two subreddits dedicated to surveys, `r/favors` and `r/samplesize`. We also sent direct messages to the top 10+ most active members (during the month when we were recruiting) in news and politics subreddits, such as `r/politics`, `r/news`, `r/worldnews`, and `r/inthenews`. We resorted to sending direct messages because the moderators of these subreddits did not allow us to post recruitment advertisements. we reached over 50 users through direct messages sent from a research account. We selected the most active members of each community based on the number of comments users in those subreddits posted that month, which we calculated using Bigquery[3]. We also advertised through several mailing lists at our university. Overall, among our pool of news consumers, we had 2 university students, 2 Reddit users, and 12 participants from local institutions. We offered our news consumers $15 gift cards for participating in our interviews. Table 4.3 shows the demography of our pool of news consumers. Considering users' political leanings (the rightmost column), our news consumer pool seemed fairly balanced. We took political leaning into consideration because it affects users' perceptions of the trustworthiness of various news sources [215].

4.3.3 Interview Procedure and Analysis

Our semi-structured interviews with participants took place mostly over Zoom video call software, though eight interviews with news consumers took place in person. A typical interview lasted 40–70 minutes. Each interview was recorded and transcribed verbatim by the first author for subsequent analysis. Using the grounded theory method [81], the first and

[3] https://bigquery.cloud.google.com

second authors open-coded all the interviews and came up with an initial set of themes [4]. In this phase, these authors read all the transcripts multiple times to determine codes. Next, these two authors discussed the codes with the other authors to resolve inconsistencies and determine relationships between the codes through axial coding (e.g., combining news selection and framing into objectivity in section 4.4.1). After repeating this process multiple times, we revised the codes into a final set of themes. We discuss those final themes as they pertain to our research question in the next section.

To summarize, our interviews with the participants (journalists and consumers) revolved around transparency factors related to trust through a series of questions pertaining to reasons for distrust of media (*In your lifetime, do you think the news industry has changed? In what way?*), how emergence of fake news affects news consumption (*As you are probably aware, the subject of fake news is now a popular topic. Has this focus on fake news impacted your news consumption in any way?*), and through a discussion on possible design improvements (after viewing the design scenario, *Would there be any other interactive-type features that you would think could be beneficial?*). See Appendix A for the full list of questions. For each participant pool, we focused our analysis on two sub-RQs: (RQa) What aspects of journalistic practice do news consumers and journalists want disclosed within news articles as transparency cues? (Sections 4.4.1 and 4.5.1) and (RQb) What should designers consider in promoting transparency cues for news consumers and journalists? (Sections 4.4.2 and 4.5.2). However, participants' responses also varied before and after seeing the scenario. The answers to the sub-RQs are summarized in Table 4.4.

[4]While some work in CSCW and HCI provides inter-coder reliability for the grounded approach [292], it is typically neither appropriate nor required due to a lack of a predefined coding scheme [66]

	Pool	(*Section*) **Before Showing the Scenario**	(*Section*) **After Showing the Scenario**
RQa. What aspects of journalistic practice do news consumers and journalists want disclosed within news articles as transparency cues?	News Consumers	(4.4.1) Transparency in objectivity of news selection and framing (4.4.1) Transparency in four aspects of evidence: presentation, sourcing, verification, and correction	(4.4.1) Transparency in author expertise
	Journalists	(4.5.1) Transparency in reporting process	(4.5.1) Transparency in evidence presentation (4.5.1) Transparency in conflict of interest
RQb. What should designers consider in promoting transparency cues for news consumers and journalists?	News Consumers	(4.4.2) Designing for easy comprehension	(4.4.2) Hyperlinking without taking away from the article (4.4.2) Reducing bias considering consumer cynicism
	Journalists		(4.5.2) Presenting complicated details with features (4.5.2) Contrasting attributes between organizations (4.5.2) Two dimensions for distinguishing article quality

Table 4.4: Theme summary split by when each theme emerged—before or after showing the scenario to the participants. Note that while responses aligned with certain themes emerged both before and after scenario exposure, the themes' positions in this table are dictated by when they emerged most commonly.

4.4 News Consumer Perspective

Here, we outline results from our news consumers divided into themes that emerged in our analyses. We contextualize them with respect to our two sub-questions regarding attributes to disclose as transparency and design issues to consider. Note that we redacted some entities (e.g., name of a news organization) in our quotes for the sake of anonymity.

4.4.1 RQa: What aspects of journalistic practice do news consumers want disclosed within news articles as transparency cues?

Analyzing the responses to our interview questions, we found three aspects of reporting regarding source and message to consider for transparency design that may improve trust in

news media. We discuss each aspect below.

Transparency in objectivity of news selection and framing

We identified three notions in our interviews that collectively suggest that the lack of trust in the media is partially derived from a perception that the media is losing its ability to be objective in its reporting. Specifically, news consumers highlighted that current practices of heightened polarization, opinion-laden reporting, and sensationalism in headlines have increased mistrust. In combination, these themes show a lack of objectivity in two aspects of news reporting: *news selection* and *news framing*. Here, news selection refers to the process of selecting stories to cover, and news framing refers to how reporters make certain information more salient in an article [106, 489]. We present these themes below.

Many of our news consumers (9/16) emphasized that readers' perceptions of polarization in the media and bias in news stories are two reasons behind distrust of the news. Some of this bias is exhibited in **news selection**, or what is prioritized in news coverage. News consumers illustrated their view through examples, such as conservatives feeling that left-leaning news organizations prioritize negative coverage of right-leaning politicians and issues, and vice versa.

> "*The British tanker gets taken [captured] in the Straits of Hormuz [by Iran]. But that's not the headline. The headline is Trump call Ocasio [a Democratic congresswoman] bad name and told her to go back to wherever she came from.*" — *U2*

> "*Like I said earlier, they want to get a certain story. And then they build the facts around what they want, how they want to portray the story, instead of actually just reporting the news.*" — *U12*

On top of news selection, news consumers (5/16) often saw news articles loaded with opinion and complained that news organizations **frame** information based on a particular agenda. If readers are unable to differentiate between opinion and news, they may run into the pitfall of considering everything opinion and consequently distrusting journalistic news.

> "*And even within the same paper, I think it's easy to lose sight of whether it's an opinion piece or it's an analysis. That's another thing they (news papers) say, "it's news analysis." But the analysis is loaded with opinion.*" — *U5*

> "*I'm just giving an example. An earthquake happened. And they show the one building that fell down and say it's destruction, and that's not really what happened.*" — *U12*

Some of our news consumers (4/16) also noted a lack of objectivity in the form of **unrepresentative headlines**. Given the relatively recent influx of alternative news sources, traditional organizations are trying different tactics to seize the attention of consumers. Since many people read only headlines, news organizations often attempt to gain consumers' attention by writing headlines that are considered clickbait.

> "*Clickbait is a big problem. Even highly reputable sources are resorting to sensationalist headlines at times to grab attention. And obviously, sensationalism has always sold. "If it bleeds, it leads" has been a line for decades.*" — *U7*

Taken together, these findings suggest that news consumers are concerned that modern journalism is losing its capacity for objectivity. Some see this lack of objectivity in partisan reporting, while others observe it in how reporters choose their words. Some readers also take issue with headlines that either express a partisan view or engage in some form of trickery to entice users to click and read. These concerns could be a byproduct of "agenda

setting"[107]. Prior research has argued that, as part of "agenda setting", news organizations often push a certain narrative by selectively publishing a subset of news stories and focusing on particular pieces of information within a given story [119, 291, 404]. Taking this aspect into consideration, system designers can design cues to distinguish high-quality articles from low-quality articles (discussed in sections 4.6.2 and 4.6.2).

Transparency in four aspects of evidence: presentation, sourcing, verification, and correction

Next to objectivity, our news consumers expressed concerns with how current journalistic practice uses facts and evidence in reporting. Participants focused on four areas relevant to this topic: (1) the presence of evidence details, (2) the use of anonymous sources, (3) decreasing levels of fact-checking, and (4) hard-to-find corrections. Below, we discuss these points regarding evidence.

The presence of evidence details was one important aspect for consumers (4/16) in regard to trusting the legitimacy of news reporting. According to our news consumers, links to source materials and information on all named parties in an article would help users better determine the trustworthiness of an article.

> "*Maybe if it has statistics and similar stuff in it... where that information come from? If you know where the source came from, you can kind of test the legitimacy.*" — *U10*

News consumers (5/16) also referred to one particular aspect of sourcing—the rise in **reports based on anonymous sources**—as a problematic trend in news media and a reason for doubting the trustworthiness of such reports. Some news consumers suggested that much political reporting now relies on anonymous sources, while others pointed out how biases in anonymous sources compel them to doubt the veracity of such reports.

> "*Certainly, anytime anyone quotes an anonymous source, it's a suspect. [Based on] The background of that anonymous source, what biases do they have?*" — *U14*

It has long been a common practice in journalism to issue retractions and corrections when journalists have reported something inaccurately. Some news consumers (2/16) expressed concern that modern journalism either seems to **check facts** less frequently, or that journalists often don't **correct** themselves when they make mistakes in reporting. These concerns about a lack of visible corrections and retractions could indicate a lack of accountability in reporting.

> "*I think the biggest problem with any media or any news now in any format is that when they screw up, they won't own it. They'll dance around and point fingers somewhere else.*" — *U6*

Consequently, news consumers (4/16) mentioned that they were now becoming more reliant on checking facts through other means, such as comparing news stories against related content and using fact-checking services.

> "*I guess, I try to verify things by looking at multiple sources. If a bunch of different sources are all saying the same thing, then that seems more likely to be accurate. And I use websites that are specifically set up for fact checking purposes, like political things.*" — *U8*

In this vein, some users (5/16) suggested features to compare journalistic performance. For example, they expressed a need to include information on journalists' past reporting trends, such as hyperlinks to past articles and reviews of the accuracy or political stance present in previous coverage by a given journalist.

> "*Provide a link to maybe a list of their other articles. So, you can see perspectives on them.*" — *U9*

Collectively, our pool of news consumers felt that journalists should present more details on evidence. They expressed concern over the increasing reliance on anonymous source–based reporting and decreasing stress on error correction when journalists make mistakes. They also noted a wider availability of news sources that often complicate ascertaining what is fact. These concerns are not unwarranted. Research shows an increase in anonymous source use in news reports [287, 373] and a significant number (48%) of factual inaccuracies in US news articles [364]. Though there is an increase (twofold, between 1997 and 2007) in the number of corrections over time [320], newspapers still fail to correct a large number of mistakes [281, 418]. Transparency regarding each of these aspects can be designed into a news web page (discussed in sections 4.6.2–4.6.2).

Transparency in author expertise

After seeing our design, news consumers (5/16) asked for several author-centric details to promote transparency. They suggested multiple transparency features, such as outlining the biography of the journalist and including their educational background (*"where they're from, where they went to college"* - *U1*) and other organizations they have worked in (*"... nice to have a way to expand expertise to see all news organization names they worked in."* - *U7*). The disclosure of such information could inform consumers about reporters' expertise. To some news consumers, trusting a journalist is easier than trusting an institution (*"I don't think you should really trust one news organization. You should look at the reporters that you feel [are trustworthy]"* - *U1*). This suggestion is not surprising, given how existing research suggests that expertise is a primary dimension of credibility; background information about a reporter might be a manifestation of such expertise [299]. In that case, cues that expose the expertise

of the organization may also affect news consumers' perceptions of how a story was created, and, in turn, result in greater trust. Indeed, prior research showed that providing additional information about journalists, such as photos and bios of reporters, fosters greater trust in an organization [102]. Our participants provided some nuance to these findings by suggesting specific types of background information, such as journalists' professional and educational background. It seems that respondents found most of these suggestions desirable because the information may help individuals determine whether journalists are capable of reporting objectively on a given topic. For example, consumers might question "where they're from" in order to understand how coverage of a story might be affected by the journalist's pre-existing biases. Considering these suggestions, designers can implement transparency cues that are indicative of reporters' objectivity. We discuss this later in Section 4.6.2.

4.4.2 RQb: What should designers consider in promoting transparency cues for news consumers?

Designing for easy comprehension

Several news consumers (4/16) preferred information displayed statistically, since it is easy to comprehend. More specifically, they asked for design that presents information in a simplified manner. Oftentimes, they used nontechnical terms to convey this idea.

> "*Give me a report, ABCD... like accuracy... like 3/4 key attributes that you're looking for in a reporter [The participant is asking for statistics on a set of easy-to-comprehend evaluation criteria.]*" — *U2*

Having too much information could itself be distracting, as pointed out by some of the news consumers (3/16). Consequently, this suggestion might be coming from a tendency to

reduce cognitive load. Taking these problems into consideration, designers could implement transparency features as simple markers and give users the ability to switch them on and off.

Hyperlinking without taking away from the report

Though several news consumers (5/16) suggested hyperlinking as a way to provide more information, some (2/16) insisted that hyperlinks take users away from the primary news and might thus distract readers.

> "*If there is even one more click to get to that information [more transparency detail]... I feel like losing tons of people. Because we're all just so lazy or busy.*" — *U8*

This suggestion to reduce distractions might be another reference to reducing cognitive load in getting information. To address this problem, web pages could provide previews of each hyperlink whenever a user hovers over them. For example, Wikipedia has already implemented such a preview feature to provide context without the cost of context switching [340].

Reducing bias considering consumer cynicism

Throughout the interviews, our news consumers (7/16) showed cynicism towards news media in general. Referring to biases in news reporting, news consumers expressed a reluctance to trust any information. Consequently, some of them (3/16) suggested that transparency features should come from a third-party reviewer.

> "*I still watch the news. I want to be informed. But I don't take everything for face value.*" — *U12*

> "*Potentially not even in their own words, in a third party words... like an outside reviewer that judges their way of what they've written previously and puts them on a political scale.*" — *U7*

This suggestion for outside oversight in transparency indicators implies that news consumers suspect news organizations of making improper claims of transparency. If readers cannot trust the transparency indicators, they will have no effect. To design against cynicism, designers may apply de-biasing techniques to modify either the environment or the decision-maker [316]; in our case, these are the design and the news consumer, respectively. As participant U7 mentioned, setting up a (bipartisan) third-party authority might be useful as a modification to the environment. Alternatively, designers can pursue nudging techniques, such as considering the opposite (e.g., asking news consumers why a statistic is inappropriate) [1, 274].

4.5 Journalist Perspective

In this section, we outline the themes from journalists as we did with those from news consumers with respect to our sub–research questions about attributes to disclose for transparency and design issues to consider.

4.5.1 RQa: What aspects of journalistic practice do journalists want disclosed within news articles as transparency cues?

Our journalists suggested transparency in two specific areas: characteristics of news organizations and the reporting process. Overall, they discussed three features for transparency: (i) providing evidence to support an article, (ii) emphasizing the process of reporting, and

(iii) noting the reporter or organization's conflicts of interest. Below, we elaborate on their definition of trust and the areas where they suggested increased transparency.

Transparency in evidence presentation

A majority of our journalists (10/15) brought up evidence as one key area for improving trust in news. Some of them (5/15) suggested that publishing unedited evidence material alongside a report can help news consumers to fact-check the report. We also found that news consumers preferred such transparency around evidence.

> "*I would say... give out documents for open access. I am for opening that up so that people can check it up... some leverage for the people to do fact checking on the journalist's work.*" — *J13*

Providing source material can empower news consumers to verify information themselves, thus streamlining their capability to fact-check while holding journalists more accountable. For example, transparency cues that highlight a position (e.g., a page number or time code) within a source (e.g., a document, audio recording, or video) can help news consumers identify discrepancies between the source material and an article based on it. Studies have found that the use of hyperlinks for evidence on news sites can increase perceptions of news credibility, drive users to seek out more information, and drive longer engagement [486]. We further discuss how designers can utilize this suggestion in section 4.6.2.

Transparency in the reporting process

Our scenario suggestions also prompted several journalists (4/15) to think about other ways newsrooms could show news audiences their reporting processes, such as by providing behind-the-scenes details. Such contextualization could include how journalists went about making

decisions while writing a story, thus revealing the underlying journalistic process of news reporting.

> "*I think as soon as the meeting ends, I, as a reporter, can go to Facebook or social media, and say, this is what I'm doing today, this is what I'm doing right away at 9 o'clock in the morning, this is why I do this. what do y'all think about it?*" — *J9*

Journalists' suggestions to show the processes that produce their reports seem to be an attempt to show professional practices in their newsrooms compared to others, and to allow consumers to compare the quality of the journalistic standards of various organizations. Some research suggests that explaining elements of the reporting process, such as verification procedures, increases news consumers' trust in a given article [321]. Research also points out that compared to author-related attributes, such as a journalist's biography, showing evidence or a behind-the-scenes verification process more significantly enhances users' trust [102]. We discuss what transparency cues designers can utilize regarding behind-the-scenes verification in section 4.6.2.

Transparency in conflicts of interest

Apart from the features related to author expertise and corrections provided in our scenario, several journalists (4/15) suggested conflicts of interest as a consideration for transparency. They suggested that both journalists and organizations can be more transparent in presenting their conflicts of interest, such as relationships with the people that journalists are reporting on (*"... reporting mechanism for meetings when you go to different parties and lobbying events."* - *J12*) or the financial background of a journalistic organization (*"The key would be if it is a ... for-profit/non-profit organization. If for-profit, is there an easy way to to show profitability? funding-wise where it is coming from."* - *J2*). Such details may contextualize

the perspective in a report and help consumers identify any potential bias in coverage. We further discuss how designers can follow existing practices to show transparency on conflicts of interest in section 4.6.2.

4.5.2 RQb: What should designers consider in promoting transparency cues for journalists?

Presenting complicated details with features

After reviewing our scenario with several transparency features (e.g., years in reporting, retractions), a majority of the journalists (8/15) raised concerns about missing complicated details in our design. For example, our journalists suggested that reporting the number of years a reporter has worked in journalism does not necessarily reflect whether the reporter has done high-quality work throughout that time. Conversely, readers could use this metric to discount high-quality coverage from less experienced journalists:

> "*It's like... first time I am getting in the news business... writing a story... reporting my first story... and somebody immediately rejects it because they see that the person only had one year of professional [journalistic] experience.*" — *J1*

Several journalists (4/15) similarly suggested improving the retraction metric by including additional disclosures as to the nature of retractions and corrections, ranging from "small mistakes," like minor technicalities or spelling errors, to "severe mistakes," like factual errors.

> "*Those two things (corrections) could have been the address of a building for example, which is meaningless. Or they could have been like just fundamental facts about a story. Like she said, x happened and it didn't happen.*" — *J14*

These suggestions imply that designers may inadvertently create a false sense of accuracy if they do not lay out these additional nuances. To address this issue, designers need to handle these details appropriately, with statistical evidence.

Contrasting attributes between organizations

While discussing our feature set in the scenario, several journalists (4/16) suggested that some transparency features (e.g., corrections) regarding journalistic practices might not be comparable between organizations. Though designers can promote transparency to compare practices within and between organizations, our journalists were particularly concerned about inter-organization comparisons. They reasoned that organizations often vary in their practices and standards, making it harder to conceptualize a fair comparison of such practices. For example, referring to the number of retractions in the scenario, some journalists mentioned diverging practices across organizations: Some organizations might issue correction in an article for minor errors, while others may not. Consequently, such comparisons could lead readers to draw mistaken conclusions.

> "*It depends on the situation. [For minor framing issues,] some news organization might change that wording but not issue a correction. Others will issue a correction and change the wording. Others won't do anything at all.*" — *J10*

Due to lack of standards in practices across organizations, designers may seek to create their own standards for presenting this contrast. For instance, they could set criteria for retractions and penalize institutions that do not follow them. If designers apply such methods to standardize these practices, news consumers should be informed about the standards.

Two dimensions for distinguishing report quality

From their domain knowledge, our journalists (4/15) showed interest in differentiating the quality of a journalistic report along two dimensions: (i) quality within news items and (ii) quality between news and non-news items. With respect to differentiating quality within a report, our journalists suggested markers to signal original and derivative work (*"...that's good because it will show if it's an original piece of work, as opposed to a derivative."* - *J2*) and markers to denote whether the story is a breaking news report (*"So a pop-up-like disclaimer that this is what we're doing right now, we are trying to collect the information."* - *J4*, speaking about a piece of breaking news). For differentiating along the news/non-news dimension, journalists suggested markers that indicate news versus opinion, news versus satire (*"...maybe something that scrapes the website, looks at the 'about me,' looks for satire and flags that."* - *J12*), and credible news versus misinformation (*"has it been, like, highly distributed and shared across channels that are sort of questionable?"* - *J2*). Designers need to make news consumers aware of the relationship between each attribute and the corresponding dimension. Considering news consumers' desire for simplicity, designers may seek to prioritize one dimension over others. For example, designers might consider distinguishing news versus opinion, as research indicates that the majority of US adults are poor at differentiating between the two [306].

4.6 Discussion

Our news consumers and journalists suggested several aspects of news coverage where transparency cues can be helpful, as well as a set of design issues to consider. Below, we first compare the two groups (section 4.6.1), followed by discussions of how designers can leverage aspects of transparency in existing organizational practices (section 4.6.2) and of design

issues (section 4.6.3).

4.6.1 Comparing the Perspectives of News Consumers and Journalists

Agreement on presenting evidence and differences in reporting aspects for transparency

News consumers and journalists alike agreed that presenting evidence details can effectively improve transparency in news reporting. In terms of disclosing the reporting context for transparency, our analysis shows that news consumers focus on a given report with only passing interest in the reporter, while journalists' suggestions encompassed four main areas: the report, the reporter, the reporting organization, and the reporting process. While our news consumers were often interested in indicators of bias in a story, our journalists showed openness to sharing information, such as revealing conflicts of interest and reporting procedures to address concerns around bias.

Conflict between simplicity and nuance in design

Recall that several news consumers asked for designs that simplify information. Contrary to that, our journalists' objections to some of our design features also point out concerns that simple metrics may be insufficient without additional context. From a technological standpoint, our journalist pool seemed more savvy and provided technical details (e.g., "heatmap" and "pop-up") when discussing feature designs. Comparatively, some of our news consumers offered suggestions in more general terms (e.g., "accuracy"). Given this tension between the two stakeholder groups, designers would have to be conscious of trade-offs between simplicity

and nuance in the design of these features. For example, they may provide specific, nuanced information in some cases (e.g., severity of corrections) while excluding other details. When such a level of nuance is required, designers can engage the audience by adding summary markers as a means of digging deeper into more general statistics.

Conflict between transparency and autonomy

Several journalists (5/15) felt that some of the transparency features were impractical due to ethical, professional, or corporate boundaries. Some mentioned their own stances on protecting the confidentiality of sources, especially anonymous sources like whistleblowers, fearing backlash. The pervasiveness of anonymous sources in political reporting may reinforce this stance. Others mentioned that news organizations were unlikely to support some of the suggestions, such as televising meetings, from fear of exposing trade secrets. However, there have also been instances in the past where an organization, such as the NYT, televised its editorial meetings [456]. These instances suggest that organizations could be open to these kinds of transparency practices.

> "*Will that [putting a live camera in a morning meeting] ever happen? I don't think so. Because there's just a lot of stuff that we talked about in that meeting that's sort of, like, behind-the-curtain stuff. And we talk about like, should we do the story about the school? Yeah! Because our demo[graphic] is 24- to 44-year-old moms who care about the school system.*" — *J9*

Prior research suggests that transparency, which creates a limited form of accountability [143], can facilitate severe scrutiny and restrict journalistic autonomy [10]. As news consumers seek greater disclosure, organizations might impose constraints on full transparency due to concerns regarding secrecy and maintaining an autonomous corporate im-

age. Designers will have to address this tension between stakeholders that arises from these competing values. To address this issue, designers could ask at least three questions when implementing a transparency feature in a news article: *Does this feature violate any policy of the organization? Does this feature violate the privacy expectations of a source referenced in the article? Do news consumers desire this transparency feature?* If the answer to either of the first two questions are yes, designers would have to revise the feature. To this end, understanding the policy norms of news organizations, as well as the privacy expectations of sources referenced in a news report, might be an important consideration for research.

4.6.2 Design Suggestions Based on Existing Journalistic Practices

Both of our participant groups suggested several aspects of transparency for reporting. What design cues can HCI designers build around these aspects? We propose that these cues be developed on the basis of existing practices in news organizations. These design suggestions are summarized in Table 4.5.

Newsworthiness cues for news selection bias

Considering objectivity, our news consumers referred to two areas where bias is injected in the production of news: news selection and framing. Journalists select stories to cover using criteria for *newsworthiness*, also known as *news values*, based on their desire to appeal to public interests [67]. Though early research proposed such news values as timeliness (how recently the event occurred), proximity (how close to the audience the event took place), conflict, and sensationalism [147, 415], journalism has evolved to consider additional news values, such as eliteness (presence of an entity with great societal power), exclusivity, and entertainment [188]. Empirically, scholars have found that conflict and eliteness are the

strongest predictors of newsworthiness [56]. Existing transparency practices in the media often broadly specify news selection on a site-level basis. For example, ProPublica offers a mission statement that says "to expose abuses of power..." [372]. Compared to such site-level practices, designers could offer transparency cues explaining the newsworthiness of each article. For example, they can identify the criteria of newsworthiness that a particular story meets.

Fairness cues for framing bias

The fairness concern is studied in communication literature as *framing bias* (i. e., levels of opinion within a given report). Scholars propose that journalists raise the salience of specific information to prime the target audience into thinking in a particular way [119, 325]. Such framing is often found in coverage of congressional candidates [117, 416], tax policies [121] and racial issues [148]. As one news consumer suggested, left-leaning organizations often cover right-leaning politicians by framing their point of view negatively. Research also suggests that journalists can be unaware of their own biases in how they select certain words, and in how they omit or decide to include certain details [250, 251]. Overall, prior research indicates that a skew in fairness exists in the coverage of certain news areas. To differentiate balanced news stories from ones that are skewed, designers can promote design cues that indicate the viewpoints or sources covered in a story and how such viewpoints are presented. For example, designers can use computational tools to identify all named parties (both persons and organizations) in a report and detect the author's or organization's stance towards each of them [21, 254]. Transparency cues could also make consumers aware of the efforts of journalists to get multiple sides of a given story. Journalists might consider disclosing which viewpoint(s) they were unable to cover or did not receive any comment on, as well as the extent of their efforts to obtain such information.

Presence of Evidence Cues

Our news consumers suggested opportunities for transparency cues that highlight good journalistic practices relating to source materials. Designers can construct transparency cues to indicate how sources are presented in a report. For example, design cues can differentiate the existence of source materials, show the timeline of the collection of such materials, and indicate how information from the source materials is presented in a given report. In this respect, designers can utilize computational tools to highlight evidence details. For instance, they might indicate whether all named parties in a report have hyperlinks, whether sources are referenced ambiguously (e.g., "sources said" vs "<name>, a spokesperson for <company>, said"), and contextualize quotes (e.g., when users hover over a quote, a tooltip can show the full paragraph containing the quote).

Anonymous Source Cues

A recent survey suggests that although news consumers realize the numerous reasons for using anonymous sources, they are concerned that news organizations unnecessarily omit justification for this practice [368]. To justify their use according to the Associated Press's 2014 guidelines [366], designers may show transparency cues regarding why the requested information is not available on the record, how reliable the source is, and what resources (e.g., public documents, on-the-record sources, and reactions from those affected by a story) have been used to corroborate information from anonymous sources. In practice, some organizations show their verification procedures for anonymous sources at the site level, instead of at the article level. For example, ProPublica shares how they generally verify anonymous sources without going into specifics for each case [176]. Taking a step further, designers can establish verification standards for anonymous reporting and provide design cues outlining

them on news websites. For example, designers can show the degree of acceptability of various verification materials (e.g., public documents compared to reactions from those affected by a story).

Fact-checking and Correction Cues

Considering the concerns pertaining to fact-checking and corrections, designers could offer transparency cues that detail who fact-checked the information (e.g., reporter / copy editor / editor / third-party fact-checker), quick links to ask ombudsmen to verify the information, a timeline of changes to the article, and statistics pertaining to corrections made by the organization and the author. Computationally, designers can also use third-party trackers (e.g., www.newsdiffs.org) that analyze changes in news reports published by an organization and show what changes were made. Designers could propose cues that highlight how corrections are framed in a report, signifying the degree to which organizations take responsibility for errors in their coverage. For example, when news media organizations include corrections, they often use a range of framing devices, such as "clarification" (evades responsibility) and "apology" (assumes responsibility) [162].

Author Expertise Cues

In addition to article-related aspects, our news consumers suggested several features related to author expertise. In practice, some organizations display biographies of their journalists along with their professional histories (e.g., reporting areas) and personal information (e.g., education) on their websites. Designers could additionally indicate differences, such as desired values in journalism. Existing surveys list several characteristics (e.g., skills, knowledge, and work values) and their importance in the journalism profession [141]. Using this

information, designers can prioritize showing certain information (e.g., active listening skills, which are more important than time management skills). However, as suggested in the interviews, designers would need to consider nonstandard practices across organizations before showing these contrasts. Comparisons within an organization might also not be meaningful in some cases, due to skewed distributions (e.g., the level of education for New York Times reporters [478]). Moreover, comparisons between generations can be tricky due to changes in required skills (e.g., the addition of technological skills in current job postings [141]). Apart from professional skills, HCI designers can also focus on journalists' expertise in being objective, since some news consumers positioned author expertise in relation to objectivity in a report. Transparency in journalists' and their editors' writing skills might be useful in this regard. Furthermore, designers can construct new indicators, such as summaries of fairness in journalists' past reports, signaling their historical patterns of objectivity.

Behind-the-scenes Cues for the Reporting Process

Apart from presenting ample evidence in the report, our journalists proposed greater disclosure of reporting processes or behind-the-scenes details as another avenue of transparency. In the past, some organizations have used such practices as inviting community members to meetings to disclose these details. Some recent examples in which behind-the-scenes details are disclosed include podcasts on how investigative journalists track their sources, and tweets by journalists containing their daily notebooks [47, 240]. Designers can adopt these details when creating news websites. For example, by leveraging diverse sources of information (e.g., social media posts and organizations' internal trackers), designers can present details that correspond to a given report in a timeline. Furthermore, HCI designers might also consider building systems to support journalists in sharing their work in progress and collaborating with the community [166].

Conflict of Interest Cues

Transparently disclosing an organization's conflicts of interest may be useful in providing news consumers with insight into the financial holdings of a company, which might impact its reporting. For example, many consumers aren't aware that ABC is owned by The Walt Disney Company, while CBS is owned by ViacomCBS, a subsidiary of National Amusements. McChesney has extensively documented the conflicts of interest that arise with news organizations when they are forced to cover issues that impact their parent companies [290]. While currently not prevalent in journalism, some organizations engage in practices detailing funding sources at the site level. For example, several organizations, including The Economist [118], NJ Spotlight [322], and ProPublica [371]) list all of their donors. Some others, such as The Wisconsin Center for Investigative journalism [492], provide extensive details that include donation amounts. However, a handful go so far as to highlight when a donor is mentioned in a story (e.g., Texas Tribune). Informed by existing practices, designers could promote design cues at the article level whenever donors are mentioned.

4.6.3 Considering Design Issues

Conflicting priorities: value-sensitive design

Throughout the study, we noticed several conflicts between our two stakeholder groups. For example, news consumers prefer designs that are quick and easy to comprehend, while journalists prefer designs that provide sufficient nuance in transparency attributes. Similarly, journalists would like to retain their autonomy, while news consumers want more accountability on the organizational level. To address these conflicts, HCI designers can utilize an existing design approach known as *value-sensitive design* [144]. This approach suggests that designers should not consider conflicts as "either/or" situations, but rather as constraints

on the design space [145]. Depending on the nature of the conflict, designers might be able to balance some tension using existing HCI design principles. For example, *social translucence* is a design principle that addresses the accountability-autonomy conflict [124]. To illustrate, in the case of putting live cameras in news meetings, designers could incorporate *social translucence* by showing the video while redacting sensitive content through appropriate methods, such as blurring and muting. Such details could lead to increased visibility of journalists' activity while imposing limited accountability on the journalists' work, without sacrificing privacy. When balance between conflicts is not feasible, designers could discuss a workable design space with the stakeholders. For example, news consumers may spare transparency in some aspects (e.g., the identity of an anonymous source) while requiring it in others (e.g., how many times a particular anonymous source was used).

Organizations' openness to transparency

Some of our participants were wary of whether organizations would be open to implementing site-level features (e.g., *"I don't foresee it being adopted broadly by news organizations voluntarily. Because I think the commercial ones, especially, would see it as just a cost with no benefit."* - *U8*). Prior scholarly work has revealed organizations' unwillingness to increase transparency regarding certain aspects, including their methods and motives [79, 246]. However, there has also been a general trend of news organizations adopting greater transparency. In this respect, some news organizations have been opting for an in-house, ad hoc framework of transparency. For example, Axios created a "bill of rights" for news production that prioritizes transparency surrounding editorial ethics policies for sources and article corrections [22]. Meanwhile, a large number of national and international organizations ($n > 200$) are adopting third-party frameworks on transparency (e.g., Trust Project) [369]. In its 37 transparency indicators, the Trust Project largely focuses on site-level features, with only

a handful of article-level transparency indicators, such as disclosing corrections and distinguishing between news and other kinds of material (viz., opinion, satire, advertising) [370]. Comparatively, our work illuminates ways in which listed organizations can transform site-level features into more specific, article-level features. For example, as mentioned earlier, instead of providing a comprehensive list of donors, designers can offer transparency whenever a donor is mentioned in a report. Article-level transparency may especially help with the concern that news consumers may not notice transparency features [447]. In general, this work offers guidance for priorities, as well as design issues to consider for transparency.

Standardizing organizational practices for contrast

Our discussion with journalists emphasizes the need for a set of transparency metrics that can be compared across news organizations. A precursor for such an implementation is a standardization of practices across organizations. If consensus develops around a standard that is common to a group of stakeholders, it may guide organizations to follow that standard when reporting news. Considering recent efforts in standardizing transparency practices (e.g., the Trust Project), creating credibility standards (e.g., W3C Credible Web Community Group [476]), and creating transparency markup standards (`schema.org`[402]), standards in organizational practice assist can these efforts. Notably, past platforms like news aggregators had to work with external websites and tools (e.g., Politifact and the Share the Facts widget [452]) to identify fact checks. Designers can address this problem by transmitting transparency details as metadata and making it easier for site designers across websites to show transparency without added hassle. In addition, standardization of organizational practices could create further barriers to fringe sources' constant production of false stories. For instance, if providing source material in a report is mandatory per transparency

standards, it may place a burden on the creators and maintainers of alternative and fake news websites, therefore slowing their pace of publication.

Feasibility of transparency cues in practice

In addition to the complexity arising from the lack of standards in practices among organizations, designers would also have to comply with the priorities of the organizations for which they work. As we mentioned before, apart from dedicated design teams in news organizations (e.g., the design team at the NYT [457]), other third-party news entities could also adopt our results by providing "transparency as a service" through either a browser extension or a dedicated website. For example, `Allsides.com` boasts of showing news from the left-center-right perspective [14]. Here, `Allsides.com` may have a priority in being transparent about biases in left-center-right news sources. For such third-party entities, designers may have to consider their access to organizational resources in implementing certain features (e.g., prior reporting history). To this end, we summarized the implementation feasibility for our design suggestions based on access to an organization's resources in Table 4.5. Here, behind-the-scenes details are only available directly from organizations, while cues for presence of evidence can be computationally detected. There are certain details that fall in between that can be managed with secondary services. For example, information about the author's reporting history or retractions could be sourced from third-party databases, such as Muck Rack [375]. Additional complexity may arise whenever subjective ratings for transparency cues have to be computationally scored. For example, designers may seek to compute how well a news report fits with newsworthiness criteria, and this information could be subjective to the audience. Crowdsourcing of such subjective ratings may be a possible solution to this issue. Other works have shown promise in crowdsourcing subjective ratings (e.g., credibility) [42].

Considering news consumers' changing values

Apart from the specific design issues of production-side aspects, our study indicates that designers would have to take into account the changing habits of news consumers, such as increasing polarization and cynicism. For such changing values, designers can personalize transparency cue design. For instance, some news consumers (U3) preferred earlier journalistic norms ("Cronkite era" reporting) such as objectivity; that is, reporting the facts only while leaving the interpretation and implications up to the audience. For these news consumers seeking fact-only news, designers could focus on certain transparency features (e.g., presence of evidence cues) that highlight facts in an article. On the other hand, there were news consumers (U1 and U9) who considered such an understanding of objectivity outdated [390] and instead asked journalists to be open about their biases. For those who asked for transparency on biases, designers could show features (e.g., conflict of interest cues) that disclose this information. In either case, future work focusing on how each design affects trust for readers with varying political stances could provide further guides for designers, especially since prior scholarly works suggests that design interventions may or may not affect perceptions of trustworthiness depending on users' political ideologies [387, 494].

Resiliency considerations to avoid manipulation

Some news consumers were concerned about the trustworthiness of transparency cues. Considering that transparency features can also be exploited by bad actors, these features must be made resilient to manipulation. In our study, each design cue has different levels of resilience to manipulation. While bad actors can easily manipulate numbers (e.g., an author's experience in years), it is harder to exploit cues that require further documentation with proof (e.g., for-/nonprofit organizations' legal documents, police reports as evidence). Sev-

eral steps can be taken to further prevent exploitation of these cues. Interdependent cues can be devised, such that manipulation of certain attributes can reveal inconsistencies or defects in the cues. For example, the number of reports and percentage of corrections by an author have a dependency that can signal manipulation of each individually, but not in unison. Resiliency can also be achieved through a third-party authority who verifies this data (e.g., an ombudsman and third-party fact-checkers). Eventually, frequent audits may also help keep the system in check, albeit at the cost of resources and time. However, it is important to note that bad actors with enough resources (e.g., state-funded actors) may still manipulate highly resilient features.

Can transparency do harm?

The suggestion that transparency can improve trust is quite complicated in practice. As O'Neill argues, despite increasing transparency, trust has receded [335]. According to her, while transparency may help when there is prior deception, increasing transparency may also "produce a flood of unsorted information and misinformation that provides little but confusion unless it can be sorted and assessed." As our journalists noted, some features could create a false equivalence between reporters from different organizations. Given the level of skepticism in news media [169], more transparency could cause consumers to become more skeptical about news—perhaps even about factual reports. To counter, our results offer particular priorities to consider (e.g., news selection and framing) in a structured fashion (e.g., through proper comparison across organizations). There are also considerations for transparency around specific areas. As Cunningham notes, "To assume that we can know what someone thinks by identifying their personal traits, habits and predilections is a dangerous notion, and really has nothing to do with clarity." [101] Our journalists remarked that transparency around sources could be harmful. Some sources prefer to stay anonymous for

a variety of reasons, such as preventing backlash from the organization. Some journalists also prefer to keep sources anonymous out of necessity, so as not to lose potential sources of information on future events. Given these constraints, future research could be directed towards quantifying the benefits and drawbacks of each transparency cue.

4.6.4 Limitations

In this study, we explore a subset of the larger issue that is growing distrust in the media; that is, how transparency through design cues can contribute to improving perceptions of trust for news articles. Our work is limited in focus, as we prioritize the perceptions of two particular groups of stakeholders: news consumers and journalists. While news organizations may play a much larger role in what gets adopted on their sites, our intent is to inform HCI designers and these organizations of the recommendations from our two stakeholder groups. This work is also limited in its perspective on designing features, to the exclusion of examining and setting organizations' policies. It is worth noting that we did not consider journalists from alternative media (e.g., Breitbart, One America News Network, and Newsmax), who are known for their particular (and perceivably strong) partisan views. Journalists in such organizations with significant viewership may view these questions of trust and transparency differently, and they may need separate attention in future works.

Furthermore, while we used a scenario-based design to elicit feedback on which transparency cues can act as trust indicators, we can not definitively predict how the cues may affect users' trust without conducting a controlled user study on a functional system, given the ambiguous effect found in past experiments [103, 223]. However, similarly to one of the experiments, the addition of multiple transparency features to a news item has the potential to affect users' perceptions of credibility [103]. Still, taking these design considerations into

account, we have to be careful to test whether such a design could simply overwhelm news consumers, resulting in an undesirable opposite effect.

Result-wise, our study suffers from some limitations common in qualitative studies. First, we had a limited number of participants. Furthermore, we could only interview people who agreed to participate. Therefore, a self-selection bias exists in our study. Moreover, we reached a majority of our participants by a combination of convenience and snowball sampling strategies. Though these are acceptable strategies in social science research [39], their use adds to the limitations of our study. Even so, we tried to address this problem by reaching out to a variety of sources. For example, in our news consumer pool, we had a mix of news-savvy and non-savvy participants with diverse backgrounds. Additionally, as we progressed through our interviews, very few new themes emerged in the final interviews of each group. Therefore, our results may have reached empirical saturation despite the limited sample size. Second, our participant population is mainly US-based, with some exceptions; therefore, our findings might not be generalizable to other countries. Indeed, trust in the media varies by country [383]. Since we conducted the study during the latter years of the Trump administration, there might be some period-specific effects that influenced the perceptions of both groups of participants (e.g., the concern over objectivity in news selection). Future studies in other regions and time periods might be able to address this limitation.

4.7 Conclusion

In this work, we examined how designers can adopt transparency features as indicators of trust on news websites. We explored these questions in a dual-perspective setting, interviewing journalists and news consumers with a scenario-based approach. Our results imply

that HCI designers can offer indicators of trustworthiness through transparency cues that reflect objectivity and evidence in news articles, authors' expertise, the process of reporting, and personal and organizational conflict of interest. Both groups agreed on some cues while offering differing views on others.

Transparency Cues (Section)	Example Questions	Implementation Requirements	Suggested By	Disagreed By
Newsworthiness Cues (6.2.1)	Which news values does this report represent (e.g., conflict, sensationalism, eliteness and entertainment)? To what degree?	Requires access to organizations' news values	News consumers	-
Fairness Cues (6.2.2)	Who are the named parties in this article? To what degree does this report represent each political affiliation (left/center/right)? Did the reporter receive comments from all contacted parties?	Requires knowledge of organizations' procedures for getting comments	News consumers	-
Presence of Evidence Cue (6.2.3)	Does the report cite an authoritative source of evidence? Is there ambiguity in how sources are represented?	Requires access to official source materials (might be openly available)	Both groups	-
Anonymous Source Cue (6.2.4)	Does this report contain anonymously sourced information? Why was the information not available without anonymity? How did the reporter verify the information? How acceptable is the verification material?	Requires knowledge of organizations' procedures for anonymous sourcing	News consumers	Some journalists
Fact-check Cue (6.2.5)	Has any internal/external entity fact-checked this information? Who fact-checked it (with links)?	Requires access to internal/external fact checkers and their procedures	Both groups	-
Correction Cue (6.2.5)	Have there been any corrections to this report? Why? How were the corrections framed?	Requires knowledge of organizations' correction protocols	Both groups	Some journalists
Author Expertise Cue (6.2.6)	What skills does the reporter have? What is the reporter's educational background? How objective has the reporter been in past reports?	Requires comprehensive knowledge of journalists' reporting history	Both groups	Some journalists
Behind-the-scenes Cue (6.2.7)	Does this report contain any behind-the-scenes details? How did the reporting process occur over time?	Requires access to behind-the-scenes materials for a report	Journalists	Some journalists
Conflict of Interest Cue (6.2.8)	Does this report cover any entity with which the reporter/organization has a conflict of interest? How does the reporter/organization deal with such conflicts?	Requires access to news organizations'/journalists' financial information	Journalists	-

Table 4.5: Design suggestions summarized according to two criteria: implementation requirements and consensus or disagreement among participants. Here, the requirements of access to (and knowledge of) an organization's resources (and protocols), such as internal/external databases of prior corrections, conflicts of interest, and behind-the-scenes materials could make it difficult for a third party to implement the design cues. The two right-most column suggest both within-group and between-group disagreement among our participants. As an example, the Behind-the-scenes Cue, discussed in section 4.6.2, could be hard to implement without access to organizations' materials. Some of the journalists disagreed as to the feasibility of disclosing portions of this information (e.g., televising meetings). Another example is the Author Expertise Cue (section 4.6.2) discussed by both groups with some disagreement from journalists due to its (e.g., years in reporting) negative impact on new journalists, which could be hard to implement due to the required access to organizations' protocols. Additionally, we provided sample questions designers could use to build transparency cues.

Chapter 5

NewsComp: Facilitating Diverse News Reading through Comparative Annotation

5.1 Introduction

News media often produces content that is significantly biased in favor of a particular ideology, especially on contentious topics [199, 313, 438], and news consumers are affected by such biases [111] [1]. Therefore, developing an informed opinion on a subject requires critically consuming news content from multiple sources. While the internet gives users access to news from multiple sources, when given choices, people tend to choose content that aligns with their viewpoints due to confirmation-seeking tendencies [84, 323, 323, 431, 440]. Furthermore, the task of engaging with multiple perspectives is not easy and probably not performed equitably by all users [209, 420]. One potential solution to this problem could be to use experts (i.e., journalists) to combine news items on an event from varying sources into a single story. However, a limited number of experts would likely find it difficult to manage the volume of news stories generated by news outlets from around the world. On the other hand, studies have shown that crowdworkers' output can be significantly correlated with experts'

[1] part of this chapter appears in [46]

in some annotation tasks [11, 12, 42, 63, 446]. Building on such results, this work explores whether crowdsourcing could be a viable approach to combining news articles from varying sources. For a lay user, such a crowdsourcing task can be broken down into two aspects of comparative annotation: (i) finding similarities and (ii) finding important disparities. These annotations can be useful to both news consumers and fact-checkers, whether professional or crowdsourced (e.g., BirdWatch[2]). For everyday news consumers, merged articles can provide balanced perspectives on news events. Second, fact-checkers can use similarity/dissimilarity annotations to validate claims through multiple sources or trace the origin of specific statements. Besides, a by-product of any annotation task is that performing the task could also affect the annotators attitude towards the content, in our case, the news articles or the issue at hand. In this work, we ask:

RQ1. *How well do users perform comparative annotation?*

RQ2. *How does comparative news annotation affect users' perceptions of credibility and news quality?*

Here, we use a simplified notion of comparative annotation: statements that are similar and statements that are dissimilar but important. Using the concept of comparative news annotation, we developed and tested NewsComp (see the interface in Figure 5.2): a prototype that allows readers to compare and annotate similar and contrasting statements between only two news articles. NewsComp has two components: (i) a comparative or side-by-side view of two articles from different sources, and (ii) an annotation tool. Specifically, the annotation tool allows performing the two annotation tasks: (i) identifying similar statements across a pair of articles and connecting them with lines (❸ in Figure 5.2), and (ii) identifying disparate statements (statements with no similarity) from each article that are important

[2]https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation

and should be included in the other article in the pair (6 in Figure 5.2). To design the tool, we conducted a series of think-aloud formative studies with Google Drawings to observe the annotation process. During those interviews, we noticed users considering different criteria for annotation. For example, some considered only the content in each statement, while others considered the underlying themes behind the statements. Informed by the think-aloud sessions, we ask annotators to provide the reasoning behind their annotation (e.g., 2 in Figure 5.2).

To answer our research questions, we conducted a between-subjects experiment with NewsComp in a controlled environment. We recruited 109 participants using Facebook advertising, which allowed us to recruit users from a large and diverse pool. Participants were randomly assigned to either the treatment or the control group. For the study, we used two pairs of articles on two contentious topics: immigration and abortion. To generate gold standards for the sake of comparison, we recruited two experts from the university's Department of Communication (one of whom had five years experience as a journalist) and asked them to perform annotation and rate the articles. Our experts found different degrees of similarity between the two pairs of articles; specifically, the pair of immigration articles had high dissimilarity (high contrast), while the pair of abortion articles pair was highly similar (low contrast). During the study, users in the treatment read and annotated a pair of articles on the same topic, and then responded to a questionnaire designed to address our research questions (related to perceptions of credibility and quality). Meanwhile, users in the control group read a pair of articles on separate events without adding annotations and responded to the same questionnaire. We analyzed the extent to which news consumers' annotations matched experts' annotations, the impact of article topic and users' news expertise (knowledge of current events, perceived value of media literacy) on annotation quality, the reasons behind annotations, and how the treatment group's article perception compared to

the control.

Regarding RQ1, we found that users performed poorly on both annotation tasks. However, they had higher precision in finding similarities than in identifying important statements among the disparate statements. We also found that filtering out annotations based on the number of users who annotated an item can rule out some false positives in finding similar statements, thus improving their collective F1 score. In our study, ruling out annotations made by fewer than five or six users produced the highest F1 score. Users with low current event knowledge made more annotations and had higher recall. We also found that while annotating statement similarity, users provided different types of criteria, such as seeing connections when two statements discuss the same person, location, date, quote, or other information. Among statements with no similarity, when annotating if a statement is important and should be included in the other article, users sometimes marked a statement important if it provided clarification or elaboration on other statements or if it provided a missing perspective. Furthermore, we found that both generic words (e.g., "quote" and "similar") and article-specific words (e.g., "lawsuit") mentioned in the rationales can differentiate incorrect annotations from correct ones. Perhaps such generic words in rationales can be used to filter out false positives in annotations on articles on different topics. Comparing the articles, annotators also saw differences in perspectives presented, information placement, depth of detail, amount of factual/opinion statements, empathetic presentation, and use of inflammatory language. Perceptions of NewsComp itself were mixed, though skewed more towards positive than negative. Regarding RQ2, we found that the treatment group's credibility ratings were significantly different compared to the control group's for high-contrast articles. For low-contrast articles, users in both groups performed similarly. There were no significant effects on perceptions of quality. Overall, this study indicates that we can leverage the comparative annotation mechanism to engage users in reading multiple

perspectives. However, since users produce annotations with high error rates, creating tools to assist in annotation could help reduce errors.

We discuss applications for annotated data, such as developing a holistic view of an event from multiple news sources, teaching machines to discern article quality, training machine learning algorithms to generate better annotations, and assisting fact-checkers in their work. We conclude with implications for the design of future comparative annotation tasks, such as modularizing into subtasks, providing supporting features to reduce load, and supporting co-annotation by multiple users.

5.2 Background and Related Works

In this section, we begin by providing some background on media bias and multi-perspective online news consumption. Thereafter, we discuss related research on designing annotation tools for making sense of information and the effects of such annotations.

5.2.1 The Need for Multiperspective News Consumption

While news articles should ideally follow established journalistic practices, various forms of biases and inaccuracies are injected into articles during the content production process. This begins in the information gathering stage, where journalists must select events and related facts from sources. In doing so, news publishers can influence which topics readers perceive to be relevant by selectively reporting on topics of their choosing [403]. Next, journalists include and exclude information from sources (e.g., press releases, other news articles, and studies), shaping the perspective on the event. In the writing phase, journalists make stylistic choices which may reflect their view of the news item, thereby producing biased coverage.

For instance, journalists may introduce bias through the use of labeling ("a senator" vs. "a Republican senator") and word choice ("illegal alien" vs. "undocumented immigrant"). Such methods allow journalists to promote a particular interpretation of a topic [120].

Research suggests that a majority of news consumers are affected by media bias [111, 248, 317] in different ways [117, 403]. Such bias can influence voting or election outcomes [110, 117, 317]. Furthermore, media bias promotes polarization in public opinion, especially on contentious topics [438]. Some scholars argue that media bias challenges the pillars of American democracy [217, 499]. Overall, these works point to the need to consume news from diverse perspectives to deal with biases in the media.

5.2.2 Barriers to Multiperspective News Consumption Online

Lazarsfeld et al. introduced the two-step flow model of communication, referring to the two gatekeeping stages that occur before an individual forms an opinion on a subject: first by news organizations, and then by opinion leaders in the individual's social circle [258]. Even though the internet has democratized access to information, including news, news consumption in the internet age still seems to follow the two-step flow model in communication in at least two ways: news selection and consumption [86, 258, 424]. First, personalization algorithms act as filters for content selection; thus, they perform a gatekeeping function similar to that of opinion leaders in the pre-internet age [341, 424, 432]. Second, pervasive, echo chamber–esque news comment sections tend to promote opinions from opinion leaders with views aligned with users' own beliefs [86, 211]. One problematic aspect of this internet-based, two-step communication is that users may not be aware of the second gatekeeping stage, given that algorithmic effects are often hidden, and partisan biases of opinion leaders in comment sections may also be obscured by anonymity [96, 424]. Even when readers become

aware of content with a political slant opposed to their own, they may lack the motivation to consume that content that due to political polarization and confirmation-seeking tendencies [29, 77, 109, 137, 234, 323, 440]. Indeed, some research has found that while people might read more content when using diverse content selection tools, this leads primarily to an increase in the amount of content consumed, not the diversity of the content [84]. One reason for this outcome could be individual differences between diversity-seeking and challenge-averse people; challenge-averse people may tend not to consume diverse content [314]. To address this bias, some prior works developed mechanisms to promote diverse news selection through design tools, such as NewsCube [342, 343, 344, 345], or through nudges to read alternative viewpoints [406]. Though these prior works demonstrated improved exposure—that is, clicks or visits to news sites with diverse political slants—there is a gap in our understanding of whether design tools can encourage critical engagement and whether such engagement affects users' perceptions of the news. Furthermore, these tools do not ensure that people read articles on the same events from politically diverse sources. We aim to bridge this gap by bundling pairs of articles on the same event from differing perspectives in a comparative annotation interface to test engagement and its effects.

5.2.3 Designing for Information Consumption through Comparison

Scholarship on reasoning, comprehension, and learning outlines different mechanisms in understanding information, whether users learn from data or the structure of information [20, 225, 474]. Sometimes, reading multiple sources alone can help change a reader's mental model of a subject [58, 427]. Comparison can further help people recognize common features shared across items or identify features that distinguish them [85, 225, 473, 474]. Some suggest that a comparison mechanism allows users to create broad concept categories

by grouping similar concepts in either a bottom-up approach (clustering) or a top-down approach (assigning existing categories) [501]. In HCI research, creating design elements or affordances for easier information consumption is not new. For example, interactive elements allow users to choose where to go or what to read next [130]. Such design elements can assist readers in constructing a cognitive model to support a thorough understanding of a news event. This construction of a news schema is supported by providing signals—layouts, visual elements, and textual structures—to news readers that meet their expectations for news [229]. For example, a newspaper reader's understanding of certain affordances (e.g., section labels, such as "opinion") may assist them in contextualizing and understanding the information [450]. It may even boost recall significantly [357]. Building on a similar idea, we design a comparative interface where pairs of articles are displayed side by side to facilitate the comparison process for users.

5.2.4 Annotating Using the Crowd and its Effect

In educational settings, annotation has long been used to boost reading comprehension, critical thinking and meta-cognitive skill improvement. Many online annotation tools have been developed over the last decade, including Gibeo [34] and HyLighter [259]. Much of the research on annotation focuses on effects in classroom settings and often takes the form of collaborative annotation [332]. On annotation tasks, prior research also suggests that users' performance may vary with demographic characteristics, political biases, task complexity, and subject matter [23, 42, 194, 301, 307]. Some research indicates that annotation technology could improve users' effectiveness and efficiency in information-related tasks, such as search tasks [226]. Informed by such outcomes, we explore the effectiveness of comparative annotation in identifying content quality in a news consumption setting.

5.3 Formative Study For Designing *NewsComp*: Think-Aloud Interviews

To design an interface where users can simultaneously compare news articles, we began with a set of think-aloud interviews [3]. We used two types of prototypes during this phase: a high-fidelity interface powered by algorithmic similarity metrics and a Google Drawings board. By high-fidelity, we mean an interactive prototype with a working front-/back-end. During this phase, we conducted a total of 10 think-aloud interviews with four members of our research groups (none of whom are authors of this paper) and six undergraduate students with different majors (communication, political science, and computer science) and levels of news consumption expertise. We used a separate set of participants for interviews with each prototype. These interviews revealed several aspects for consideration in our design. Below, we discuss how our interview process evolved and summarize the insights we gathered.

5.3.1 Inaccurate Algorithmic Annotation

This phase consisted of six interviews conducted with an interface we designed to support comparison through similarity scores obtained from a state-of-the-art sentence transformer and its semantic sentence matcher[4]. All six participants mentioned that the algorithm's similarity annotations were inaccurate. This effect may result from differences in how users and algorithms compute similarities. Whereas a human can take a statement, event, surrounding sentences and other contextual aspects into consideration while finding similarity, algorithms are likely to prioritize word similarity.

[3] All of our studies were approved by the institutional review board at our university

[4] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

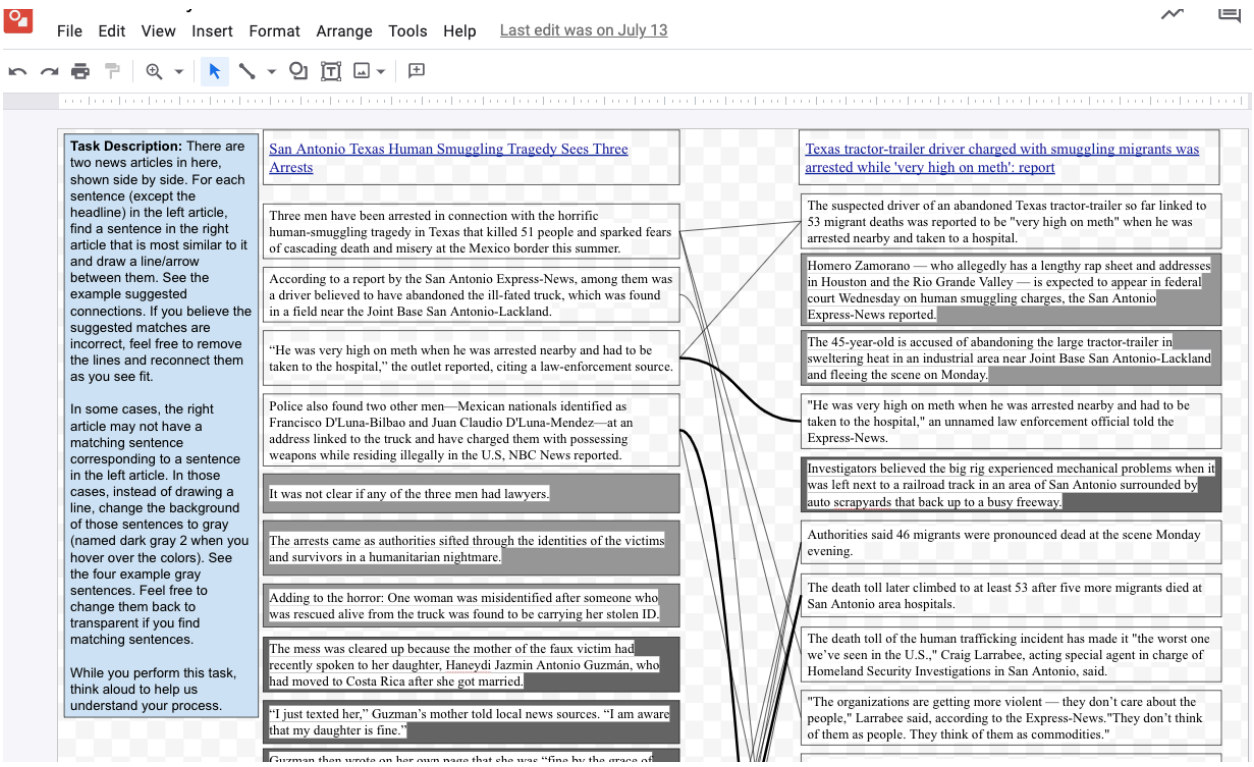


Figure 5.1: A Google Drawings board used for think-aloud interviews. Similar to the high-fidelity interface, two articles are presented side by side here. Users can use all the available tools to link similar statements or highlight dissimilar statements that contain important information which should be included in the other article.

5.3.2 Annotating on Google Drawings

Next, we moved towards asking users to perform the annotation task using a Google Drawings board. Here, we laid out a pair of news articles side by side on the drawing board by segmenting them into sentences (see Figure 5.1). Then, we asked our participants to perform two annotation tasks, one after another: (i) find similar statements between the two articles and draw lines between them, and (ii) revisit statements without corresponding, similar statements to see if they convey important information that should be included in the other article. Since users reported inaccuracies in algorithmic statement matching, we refrained from providing machine-generated annotations as suggestions. Figure 5.1 shows a screenshot of the interface and the annotations provided by one of the participants. As in

the prior interviews, we recorded participants' actions. These tasks took longer compared to the previous interviews, as users iterated over each statement multiple times to add annotations. After completing the task, we asked semi-structured interview questions to clarify the participants' actions. Our observations and the participants' answers helped us identify several considerations for our study, outlined below.

Criteria for Finding Similarities: Content and Underlying Theme

After the annotation task sessions, we asked participants to elaborate on the criteria they used to find similarities. From their responses, we found two similarity criteria: content and underlying theme. Though one participant mentioned structural position (e.g., the lede in a news article) as a criterion, none of the other participants mentioned it. In our final deployment, we asked users to provide rationales for why statements were similar.

Considerations for Finding Important Information Present in Only One Article

When asked about how participants chose statements conveying important information that was worthy of inclusion in the other article, they mentioned two considerations. The first of these considerations was whether a statement fit the narrative of the other article. Participants suggested that a statement in article A should only be labeled "important" if it fits the narrative in article B and provides important context missing from article B. Such missing information might include statements detailing what happened after an event or how something happened. Even when a statement did not fit within the narrative of the other article, some suggested that such a statement should still be included (*"This task is difficult, because the two articles are focusing on different narratives ... the other article does a bad job at portraying them as such [smuggled people being seen as inhuman]. Therefore, I think*

that bringing the human cost displayed here into the other article would be helpful." - P3). Participants mentioned that any information among the dissimilar statements that could be inferred from other statements or the context of an article did not need to be included. In light of these nuances in the reasoning behind answering the questions, the result suggests that participants were more critically engaged in reading and comparing the two articles during this exercise than they were during the exercise that presented ML-recommended results. Therefore, in our final design, we asked users to provide rationales behind annotations when finding something important to be included in the other article.

Readers Have Varying Expertise in Identifying Similarities

During these interviews, we noticed that readers' different levels of expertise in reading news and knowledge on the topic led to different annotations. During two interviews with participants who had less news expertise, we presented another participant's thematic connection annotations and asked if the interviewees could understand the original annotator's intentions. Neither participant was able to explain the annotator's intentions. These findings led us to one of our research questions; specifically, how user characteristics relating to news expertise affect comparative annotation.

Overall, both studies revealed to us that a comparative annotation task could strengthen users' engagement with news content, and we implemented such a task in our final design for NewsComp.

5.3.3 The NewsComp Interface and How It Works

Based on the findings from our two formative studies, Figure 5.2 shows the final NewsComp interface we implemented. At the top of the page, instructions for the tasks are laid out

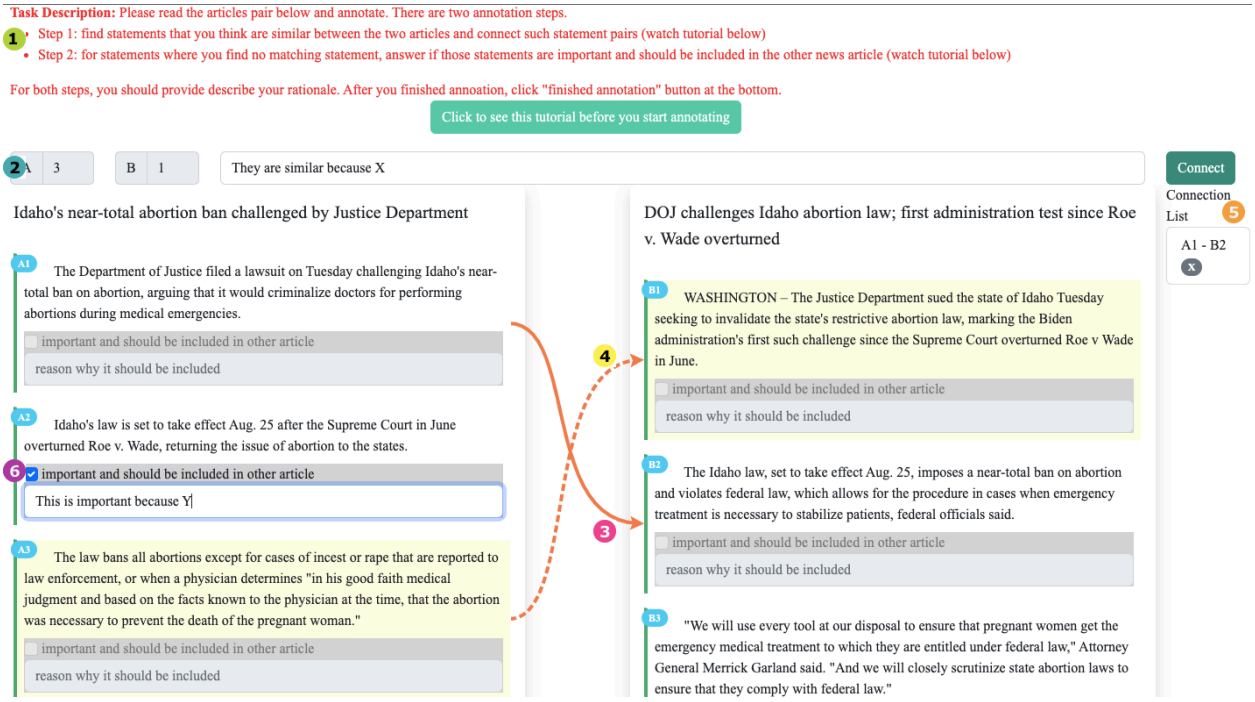


Figure 5.2: NewsComp Interface showcasing features with random annotations. 1 Annotation instructions in two steps: find and connect similar statements, and answer if a statement with no corresponding, similar statement is important to include in the other article. 2 Toolbar to finalize a connection by providing a rationale 3 A solid arrow representing a connection already created 4 A dashed arrow indicating that the connection creation tool is active 5 A list of connections including deletion buttons 6 The importance question in step 2.

(1). The task asks the user to read the pair of articles and perform two steps: (i) find similar statements within the articles and create links between them, and (ii) check if a statement with no corresponding similar statement in one article is important and worthy of inclusion in the other article. To help users understand how to perform the two steps, there is a button below the task description that opens a video/GIF showing a tutorial of both steps. In our deployment, we made a point of reminding users to watch the tutorial before proceeding. A toolbar below the taks description (2) helps users perform the first step. Specifically, when users link two statements, the toolbar shows the statements that are highlighted (in yellow) and provides a text box where they can supply a rationale for the connection right before finalizing the annotation. Below the toolbar, two news articles are presented side by side (as in our initial interface) with the article title at the top, followed by statements segmented exactly as in the original article. To limit preexisting biases, there are no links to the canonical source, nor any reference to the authorship. The interface also hides any nontextual components (i.e., videos and images). To begin connecting statements, users click the two statements to select them. Selected statements are highlighted with a yellow background. To mark a statement as worthy of inclusion in the other article, each statement contains a checkbox (6). There is also a text box below this checkbox for users to provide a rationale for marking a statement as important. When a user connects two statements, the checkbox and text box for the statement are programatically disabled and grayed out. After selecting a statement from each article, a dashed arrow representing a pending connection appears (4). When a user finalizes a connection by filling in a rationale and clicking "connect," the dashed arrow (⇢) changes to a solid arrow (➡) to represent a confirmed connection (3). In addition, a list of connections appears to the right of the articles (5) to allow users to delete connections they have created. Users can delete a connection by clicking the cross button next to it. After finishing both tasks, users scroll to the bottom of the page and click a button to confirm they have finished the annotation tasks.

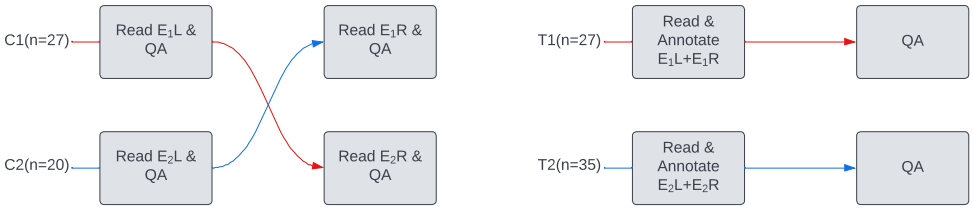


Figure 5.3: Study design showing the experimental conditions for each of the four participant groups. Here, C and T respectively represent control and treatment groups; the number of participants is given in parentheses. Because we used four articles, we had two control groups (C1–2) and two treatment groups (T1–2). Article E_XP represents an article about event X from a source with political leaning P (L for left, R for Right). Articles with $X = 1$ were about immigration, while those with $X = 2$ were about abortion. For example, E_2R indicates a news article about abortion from a right-leaning source. The E_1 pair had high contrast, while the E_2 pair had low contrast. In the study, we randomized the order/position of the articles for each participant.

This system was built with a React front end with Bootstrap CSS, and a Flask back-end server with a MySQL database. To draw the connection lines, we used the Leader-Line[5]. To scrape news articles, we used NewsPaper[6]. For the formative study, we used Sentence Transformer[7] to generate sentence embedding and calculate sentence pair similarity score.

5.4 Evaluation Study

Using NewsComp, we examine two research questions:

RQ1. *How well do users perform comparative annotation?*

RQ2. *How does comparative news annotation affect users' perceptions of credibility and news quality?*

To answer, we conducted a between-subjects experiment in a controlled environment using

[5] https://github.com/anseki/leader-line
[6] https://newspaper.readthedocs.io
[7] https://huggingface.co/sentence-transformers

two pairs of news articles. We created a separate interface for the control users. In the control interface, only one article is shown at a time. Figure 5.3 shows the study design for our experiment with four experimental groups: two treatment and two control groups. Each group read two news articles. While the treatment group was able to view two articles on the same topic simultaneously, the control users read articles on different topics sequentially to account for any learning effects from recall and comparison. All four groups read stories from two sources with different political leanings. We randomized article location (left or right) for the treatment group and article order (first or second) for the control group to account for any ordering effect.

5.4.1 Article Selection

For our study, we picked two politically contentious topics (immigration and abortion), where reading content from diverse perspectives can be beneficial. The topics were chosen from recent news coverage at the time of the user studies. For each topic, we chose articles published at least two weeks prior to deployment to limit possible recall effects. Pairs were selected by finding two articles from politically opposed sources under the same story bundle on Google News. When choosing article pairs, we picked pairs with different levels of similarity and difference. Since the article pair on abortion(E_2) had more similarities than differences, we categorized the pair into the *low-contrast* category. On the other hand, the pair on immigration (E_1) had more apparent differences than similarities, so we categorized it into the *high-contrast* category. This categorization was confirmed by our experts' gold standard annotations (see 5.4.6), which identified more than 50% of the article text as similar in the low-contrast pair while identifying less than 25% of the text as similar in the high-contrast pair. The selected articles (E_1L, E_1R, E_2L, E_2R) are reproduced in Appendix B.2.

Credibility [303]	i) It is biased (I) ii) It is not fair (I) iii) It doesn't tell the whole story (I) iv) It is not accurate (I) v) It cannot be trusted (I)
Quality [468]	i) It shows multiple viewpoints ii) It has information on causes and consequences iii) It provides balanced viewpoints
Current Event Knowledge (CEK) [282]	i) Who is/was Kamala Harris? (a) President (b) **Vice President** (c) **Senator from California** (d) UN Ambassador ii) What does the recent Supreme Court ruling overturning Roe v. Wade entail? (a) **Abortion is not a constitutionally protected right** (b) In Missouri, abortion is legal before 24 weeks (c) All US states allow abortion for rape and incest (d) **There is confusion about abortion rights relating to miscarriage and ectopic pregnancy** iii) In California (a) **everyone, including undocumented individuals, has the right to access their crime report** (b) there are no immigrant detention facilities (c) **state and local police officers cannot inquire about an individual's immigration status during a routine check** iv) How is the Fed responding to the high inflationary economic condition? (a) **Raising the interest rate** (b) Lowering the interest rate (c) Keeping the interest rate the same
Value of Media Literacy (VML) [475]	i) Two people might see the same news story and get different information from it ii) People are influenced by news whether they realize it or not iii) News is designed to attract an audience's attention iv)Writing techniques can be used to influence a viewer's perception v) People should accept information from the news on face value (I) vi) It is the job of citizens to overcome their own biases in consuming news vii) People need to critically engage with news content viii) The main purpose of the news should be to entertain viewers (I)

Table 5.1: Questionnaires used in the study. Credibility and quality questions were asked after reading or annotating. (I) means these items were inverted for analysis. The correct responses appear in boldface. The CEK questionnaire contains multiple-choice questions, while the VML, credibility, and quality questions are 5-point Likert items. The VML and CEK items were presented in the pre-survey.

5.4.2 Measuring Credibility, Quality, Current Event Knowledge, Media Literacy

To address RQ1, we measured two expertise metrics: *current event knowledge* and *value of media literacy*. Here, the value of media literacy differentiates users' general media literacy from their expertise on topics related to our study. To answer RQ2, we use perceptions of article *credibility* and *quality*, and we compare the treatment groups' assessments with the control groups'. Below, we discuss how we measured each metric.

Current Event Knowledge (CEK) and Value of Media Literacy (VML)

To capture users' news-related knowledge, we adapted the Current Event Knowledge measure created by Maksl et. al. [282]. Here, we included questions relevant to the two chosen article topics and some other timely topics (see Table 5.1). To measure users' perceptions of media literacy, we used a prior scale created by Vraga et. al. [475]. To calculate CEK scores, we added 1 point for each right answer and deducted 1 point for each wrong answer. In our study, CEK ranged from -1 to 7, with 4 being the median value. For VML, we average the responses across items. The score for VML ranged from 1 to 8, with 6 being the median. Finally, for both measurements, we use the median score to create a binary response variable with values "low" and "high." For example, users scoring less than 4 in CEK were categorized as low-CEK users and vice versa.

Credibility & Quality

We used a five-item questionnaire by Meyer et al. [303] to measure users' perceptions of credibility for every news item (see Table 5.1). In our study, we found that this measure had high internal consistency (Cronbach's $\alpha = 0.85$), close to the result in Meyer et al. For

news quality detection, we use a modified version of the questionnaire suggested by Urban et al. [468] (see Table 5.1). Similarly to the credibility questionnaire, participants' responses to these questions showed high internal consistency (Cronbach's $\alpha = 0.89$). We measured all of these items on a 5-point Likert scale, from "Strongly Disagree" (1) to "Strongly Agree" (5). Note that scores for the credibility items are inverted for analysis.

5.4.3 Recruitment

To recruit participants for our final NewsComp interface, we used Facebook advertising for two weeks in August 2022. This method allowed us to organically recruit diverse participants from a large pool. We also did limited advertising on news subreddits (such as, `r/politics, r/moderatepolitics, r/news, r/neoliberal, and r/conservative`) through private messages from our research group's Reddit account, reaching about 40 users. Two users responded to these messages. Since the article topics are US-centric, our ads targeted people living in the US with interest in news-related pages. Thus, our study result may not be generalizable beyond the context of the US. The advertisement led users to a pre-survey to sign up for the study. In the pre-survey, we screened users with the following study eligibility criteria: (i) I am 18 years old or over, (ii) I reside in the United States, (iii) I read at least one news article online every day, (iv) My primary language for news consumption is English, and (v) I use a laptop or a desktop for online news reading. Besides these criteria, we also screened out users who failed attention checks, had an IP address outside of the US, or spent very little time (less than half of the median time, which was two minutes) in the pre-survey. Overall, 685 users clicked on the survey, out of which 238 passed the screening criteria. We invited all of these participants to the study in multiple batches. Ultimately, 109 participants completed the study. Participants who completed the study were compensated with $7.50 gift cards for the 30-minute study, in line with the

state's minimum wage.

5.4.4 Procedure

Users who met the screening criteria in the pre-survey filled out the rest of the survey, which contained questions about demography, including gender, age, race, education, and political affiliation, and the two news expertise measures. Within three days of submitting the survey, we invited eligible users to participate in the study via email. In the email, we provided the consent document, instructions for using the interface, and a link to the study website. When users clicked the link to access the study website, they were randomly assigned to one of four groups (two treatment and two control) to ensure a balanced sampling design. Since some people who clicked the link ultimately did not complete the study, the final group sizes are not exactly equal. Recall that after visiting the study website, treatment users were asked to view a tutorial on how to add annotations before reading and annotating articles. After finishing the annotation process, users responded to the credibility and quality questionnaire. The annotation interface, including the articles and annotations, was still visible at this time. In the control condition, there was no annotation task, and participants read only one article at a time. The control users additionally responded to the credibility and quality questionnaire after reading each article (see Figure 5.3).

5.4.5 Participant Pool

Due to screening and self-selection bias, our study participants were not equally distributed in certain demographic dimensions, such as age. Additionally, since one of our aims was to identify how users with different demographic characteristics compare in their annotations in RQ2, we invited more users in the treatment condition. Though our pre-survey had

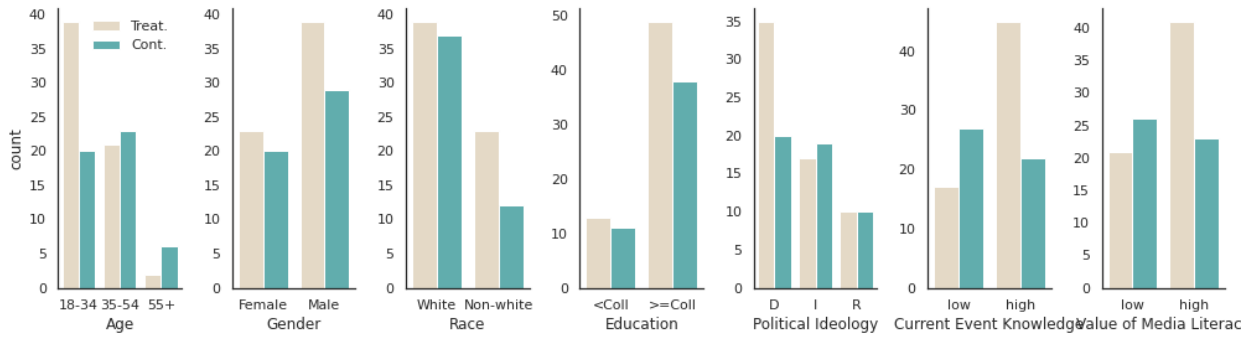


Figure 5.4: Graphs showing the distribution of participant demographics across the treatment and control groups.

a large number of categories for different demographic characteristics, we merged groups with small numbers of respondents for more meaningful differentiation. Figure 5.4 shows this distribution. Here, we grouped two consecutive age groups, merged participants from nonwhite races together, and divided respondents by education into those with any university degree versus those with no degree. Generally, participants were skewed towards younger age groups, male, white, college-educated, and politically left-leaning.

5.4.6 Gold Standard Generation

To compare annotation quality, we used expert-produced gold standards. To obtain gold standards for both the annotations and the perception metrics, we recruited two senior PhD students from the university's Department of Communication for an interview session. Both had past experience in conducting news content analysis research and were familiar with both topics used in our study. One of the experts also worked as a journalist for more than five years. To generate credibility and quality perception scores, both were given the original links to the news articles and asked to rate the RQ1 questions[8]. They were also

[8]This task was performed before generating annotations. We did not ask them to work within the comparative interface, with the assumption that they would rate the articles accurately irrespective of any comparison.

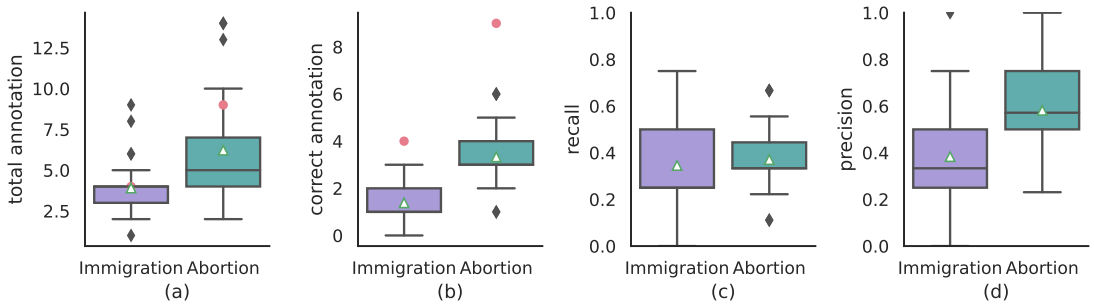


Figure 5.5: Distribution of connection making by users. White and red dots respectively represent the average and experts' annotation.

allowed to do any outside research they wished. After rating their perceptions, we provided the article pairs in a Google Drawings board and asked them to add annotations, much like the process from our think-aloud interviews. After adding annotations independently, the expert annotators met with each other to resolve any conflicts. Through this method, we built consensus around our gold standard annotations.

5.5 Results

To answer both research questions, we compared users' annotation and perception responses against the expert-produced gold standards. For this purpose, we performed a series of analyses involving mean testing, analysis of variance, and regression. For free-form text responses (specifically, the annotation rationales), three authors performed thematic coding (see supplemental document for data and codes). Below, we outline the results.

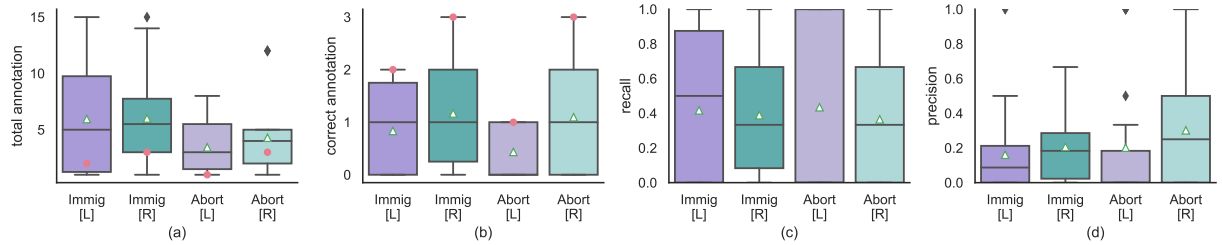


Figure 5.6: Distribution of importance detection by users. White and red dots respectively represent the average and experts' annotation.

5.5.1 RQ1: How well do users perform comparative annotation?

Performance on Connection-Making

Figure 5.5 shows the distribution of total connections users made, correct connections made, their recall, and precision relative to the gold standard. The median number of connections between articles fell below the gold standard for both article pairs, as shown in Figure 5.5(a). Between the two article topics (immigration and abortion), users on average made more connections—correct or otherwise—between the abortion articles (the low-contrast article pair). Furthermore, users' precision was significantly better on the abortion articles than on the immigration articles (Mann-Whitney U = 136.5, $p < 0.001$). However, we did not find any significant difference in recall.

Performance on Importance Detection

Figure 5.6 shows the distribution of total importance annotations users made, correct importance annotations, recall, and precision relative to the gold standard. As shown in Figure 5.6(a), the median number of importance annotations was consistently above the gold standard. Between the two article pairs, users on average annotated more items as important—correctly or otherwise—in the immigration articles (the high-contrast article

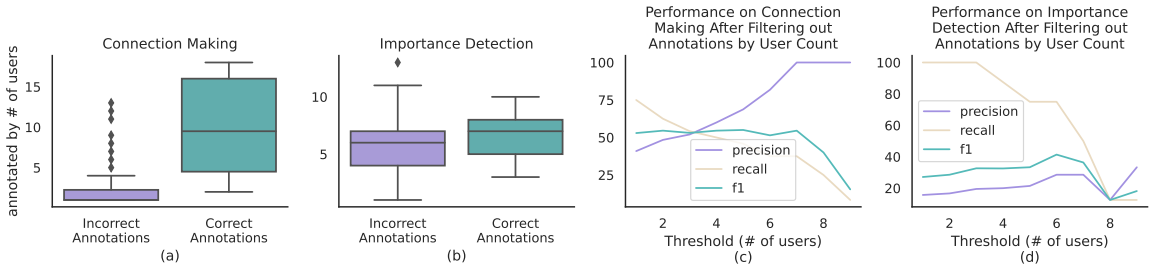


Figure 5.7: (a & b) User agreements on incorrect and correct annotations. (c & d) We filtered annotations by the number of concurring users to see how annotation performance changes as the threshold moves. Here, for connection making and importance detection, the F1 scores peak at five (55%) and six (41%) users, respectively.

pair). Though users' recall was high due to the large numbers of importance annotations added, their precision was low, with the median per article being less than or equal to 0.25. Comparatively, for connection-making annotation task, users' median precision and recall are higher than these median for importance annotations.

Annotation Agreement

Next, we examined how users agreed on annotations among themselves by plotting the count of users annotating each item. Figure 5.7 shows the distribution of this analysis. Here, we differentiated between agreement on correct and incorrect annotations. For the connection-making task (Figure 5.7(a)), we found that the annotation count for correct items was significantly higher than the count for incorrect items (Mann-Whitney U = 517.0, $p < 0.01$). However, for importance detection (Figure 5.7(b)), the corresponding counts did not differ significantly. Furthermore, we also observed some outliers (high agreement in some annotations) in the connection-making annotation task not made by the experts. In Figures 5.7(c) and (d), we examine how annotation performance changes by filtering annotations by the number of concurring users. Overall, a threshold of five users produces the highest F1 score (55%) for the connection task, while peak performance (41%) occurs at

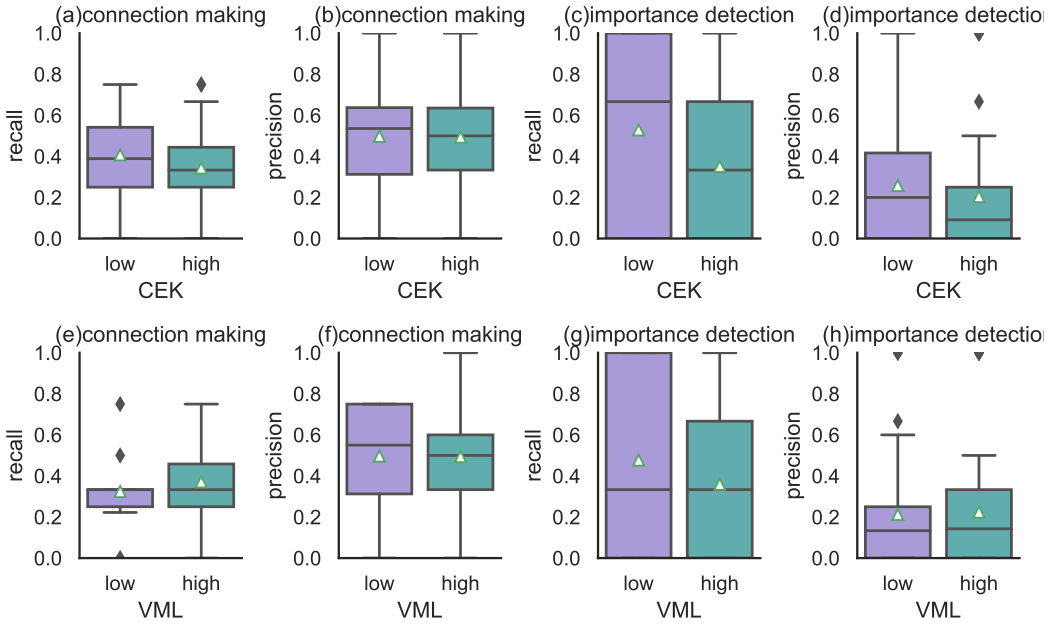


Figure 5.8: Distribution of recall and precision for connection-making and importance detection divided into low/high CEK users (top), and low/high VML users (bottom).

a threshold of six users for the importance detection task.

Effect of News Expertise

We investigated whether levels of news expertise affect users' annotation performance with NewsComp by performing Mann-Whitney U tests on precision and recall scores (for both connection-making and importance detection). Figure 5.8 shows the distribution of those scores divided by two news expertise criteria, Current Event Knowledge (CEK) and Value of Media Literacy (VML) perception. Here, only for recall scores on the importance detection task (Figure 5.8 (c)), we found significant differences in values between low and high CEK (Mann-Whitney U $= 810.5$, $p < 0.05$). None of the other tests detected a statistically significant difference. Furthermore, we modeled these variables against user characteristics with a series of linear models (M1–M4 in Table B.1 in Appendix B.3). These models were

	Code	Definition	Example Response
Connection	Empty (20%)	Empty or N/A response	
	Person (19%)	Mentions that both statements refer to one or more persons involved in an event (not including quotes)	*They are similar because they both mention **the owners of the truck***
	Location (1%)	Mentions that both statements refer to the same location where an event occurred	*This excerpt shows where the truck was found and both gave **an identical location** ...*
	Date (2%)	Mentions that both statements refer to a single date when an event occurred or will occur	*Both statements note that the ban will take effect on **August 25th***
	Quote (9%)	Mentions that both statements reference either the same quote or different quotes from the same person	*They are similar because both highlight **a quote from Becerra (HHS Secretary)** insisting ...*
	Information (48%)	Mentions that both statements contain the same information describing the what, why, or how of the event	*Similar because [both] discuss **Medical Treatment and Labor Act***
Importance	Empty (29%)	Empty or N/A response	
	Important (15%)	Mentions that a statement is important without providing a reason	*[because it is an] **important part** of the news*
	Clarification (43%)	Mentions that a statement clarifies or elaborates on other statements	*The **statement in the other article** from Becerra (HHS Secretary) is **confusing**.*
	Missing (8%)	Mentions that a statement presents a perspective missing from the other article	***No statement from Lawrence** in the other article*
	Factual (4%)	Mentions that a statement is factual and not an opinion	***Facts here.** It isn't opinion being interjected into a news story.*

Table 5.2: Coding scheme for annotation rationales.

similarly significant ($\beta = 0.35, p < 0.01$ in M1). However, since the model effect sizes (R^2) are low (0.10), there may be confounding variables not accounted for in these models affecting the outcome. It is therefore difficult to make any strong claims in this regard, and we instead leave this to future experiments.

Reasons behind the Annotations

Three of the authors thematically coded the rationales provided by the participants during annotation. Each author performed initial coding and discussed the results with the others to agree upon a code book. Then, the first author coded the responses accordingly and the others checked the final codes. Participants annotated 250 connections and 305 important statements. Table 5.2 shows the coding scheme we developed for each annotation task with sample responses matching the code. There are six codes for connection-making and five codes for importance detection (including one in each for empty responses). Here, the codes in the connection-making task offer potential answers to the 5W1H questions (Who,

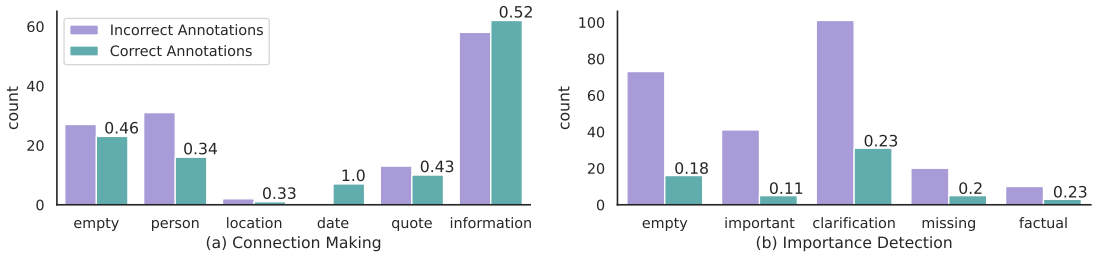


Figure 5.9: Annotation counts by coded rationales divided into correct and incorrect annotations. The numbers over the bars represent the ratio of correct to incorrect annotations within each code.

What, Where, When, Why, and How). For example, the code "Person" answers the *who* aspect of the event, while the code "Information" answers a combination of what, why, and how questions. For importance detection, users sometimes claimed that a statement was important without explaining the reason for this assessment. In other cases, users mentioned that a statement was important because it clarified or elaborated on existing statements, or because it provided an account from a missing perspective. Figure 5.9 shows the count of each rationale in terms of the developed codes, divided into correct and incorrect annotations. For connection-making (Figure 5.9(a)), we found that a majority of users identified similarities when the same information was presented in both articles, followed by mentions of the same person. For importance detection(Figure 5.9(b)), many responses were coded into the clarification and elaboration categories, followed by the empty response category. Figure 5.9 suggests that the rationales are not distributed proportionally for correct and incorrect annotations. Notably, users appear more likely to make errors in certain cases. For example, for importance detection (Figure 5.9(b)), the ratio of correct and incorrect "important" annotations shows that these annotations are more likely to be mistaken than others. Although we performed regression on the codes to differentiate correct and incorrect annotations, the model effect sizes were very low for both models ($R^2 < 0.05$).

Since differentiating correct and incorrect annotations by codes did not work well, we fit two

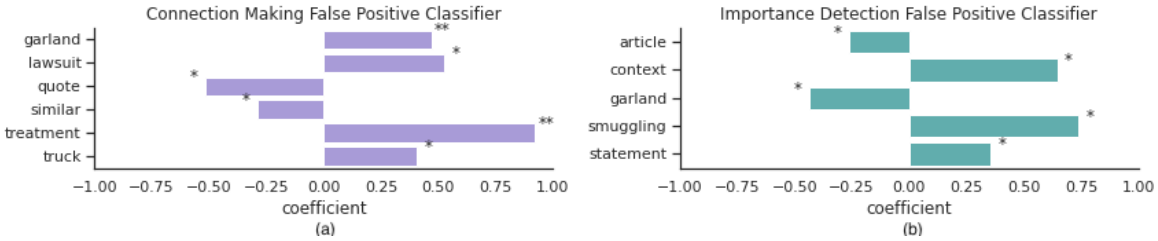


Figure 5.10: False positive detection with OLS using the top 50 TF-IDF words in users' responses. Here, we listed only words with significant coefficients. For example, when users' mentioned "quote" in a rationale, the annotation was less likely to be erroneous. On the other hand, when users mentioned the general nature of the event ("lawsuit" in this example), the annotation was more likely to be erroneous. The model effect sizes (R^2) were 0.34 and 0.22, respectively.

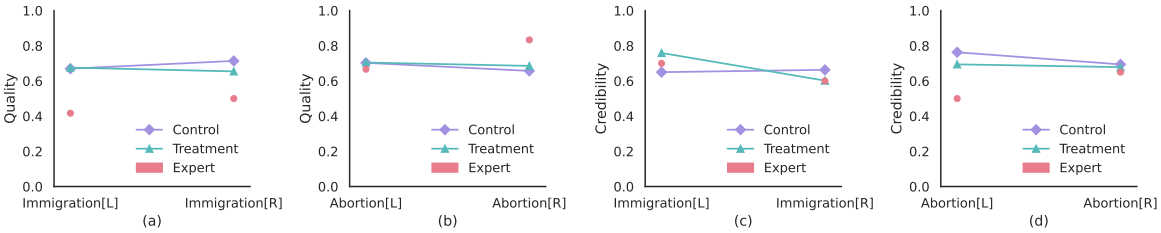


Figure 5.11: Interaction effects of groups and articles. We only found a marginal interaction effect (p=0.052) for credibility score on articles regarding immigration (c).

models to predict incorrect annotations (false positives) for both the connection-making and importance detection tasks using the top 50 TF-IDF [9] text features from users' responses on the rationals. Figure 5.10 shows the text features with significance for this analysis. Examining the words with significant coefficients, we can see that some words are generic (e.g., "quote", "similar", "context"), while others are article-specific (e.g., "Garland" (the current attorney general), "lawsuit", "smuggling"). These differences suggest that such generic words in rationales can be used across articles to differentiate false positives from true positives, while specific words may not be usable. We discuss these results further in section 5.6.1.

[9]TF-IDF stands for term frequency–inverse document frequency, a statistic representing how important a word is to a document in a collection of documents

Theme(n)	Example Response
Perspectives and Biases (22)	They were taking **different sides** of the equation and putting forward different thought processes
Information Placement or Depth (14)	The right article was **less descriptive** and focused more on the restrictions and not the case. It was definitely telling the story from one point of view. The article on the left was very informative and unbiased
Factuality or Opinions (16)	Article B provided responses from the Idaho government, whereas Article A did not include **commentary** from Idaho, but instead from Texas which did not seem relevant
Empathetic Reporting (5)	There were more **humane aspect** in the article on the left
Inflammatory Language (3)	Article B seemed to be making the issue out to be more **controversial** by going back and forth between perspectives more frequently

Table 5.3: Themes in users' responses to a question asking what they noticed about the two articles overall. Note that while an example response may belong to multiple themes, only the portion relevant to the listed theme is presented in bold.

Comparative Perception between Article Pairs

After the annotation task, in addition to asking about credibility and quality perception, we asked users what they noticed when comparing the two articles ("Comparing the two articles, what else did you notice about how each portrayed the issue?"). Analyzing the responses, we found five themes, summarized in Table 5.3. Notably, more than one fourth of the participants remarked on informational placement or depth (16/62), perspectives and biases (22/62) and factuality/opinions (16/22). Some also noticed empathetic news reporting (5/62) and the use of inflammatory language (3/62).

Perception of the Tool

Apart from the task-specific questions, we also asked annotators about their perceptions of the NewsComp tool. Overall, perceptions of the tool were split among positive (28/62), neutral (21/62), and negative (15/62) sentiment. The reasons behind negative sentiment included the lengthy nature of the task (3/15), difficulty in performing annotation (8/15), and confusion regarding the instructions (4/15). While it may have been hard in the beginning, users quickly learned how to use the tool (*"It was a bit confusing to learn how to use the tool, but it was easy to use once I played around with it."* - *U1*). Improvements to the

tool design could potentially address these issues. For instance, during connection-making, a search tool could assist users with finding similar statements quickly. Users also suggested improvements such as allowing them to set the weight of the connected lines, change the colors of lines, and see how others annotated a statement.

5.5.2 RQ2: How does comparative news annotation affect users' perceptions of credibility and news quality?

To answer this question, we performed a two-way ANOVA on the response variables, credibility scores, and quality scores. To calculate each score, we first summed item scores for the questionnaire on each score (credibility and quality) and standardized them on a [0, 1] interval. Then, we performed the two-way ANOVA on the group and article interaction. Figure 5.11 shows the result of this analysis. We did not find any significant difference in quality perceptions. We also performed a one-sample *t*-test on users' quality perception responses against those of the experts. Though we did find some similarity for the abortion articles (Figure 5.11(b)), in the case of the immigration articles (5.11(a)); that is, the high-contrast pair, the difference between the users' and experts' ratings was significant. This result suggests that neither group performed well on the quality question, especially for the high-contrast article pair.

Next, we analyzed the interaction of credibility with the articles. In Figure 5.11(c), our analysis found that the interaction between the experimental group and the article is marginally significant for the articles on immigration with a close to moderate effect size ($F(1) = 3.83$, $p = 0.052$, Cohen's $f = 0.22$). We did not find any significant effect for articles on abortion. Note that when we fit a mixed-effects regression model on the same data additionally considering repeated measure, we found significant interaction effects on the high-contrast

articles (see Appendix B.4). Moreover, when we look at the articles on abortion, despite the experts not performing comparative annotation, their perceptions of one article (from a left-leaning source) were significantly lower than those of the other (from a right-leaning source). During our discussion, we found that experts used certain criteria to arrive at this assessment [10]. On the other hand, the comparative annotation task may have given users the impression that both articles are highly similar (leading to similar ratings) and discouraged them from examining other differences too closely. This result suggests that there is room for improvement in task designs for comparative annotation. We discuss further implications in section 5.6.2.

5.6 Discussion

Through our experiment, we found that users generally perform poorly in annotation tasks for finding similar statements and identifying important statements among statements with no similarity between two news articles. However, they have better precision in finding similarities than in identifying important statements. We found that when a high number of users find two statements similar, such annotations have a high chance of coinciding with experts' annotations. Furthermore, we found that users with low current event knowledge may perform better annotations. Analyzing users' rationales behind the annotations, we found several reasons for finding similarities (e.g., mentions of the same person or information) and identifying important statements among statements with no similarity (e.g., statements clarifying something or providing a missing perspective). Furthermore, we found that certain words have significant power in differentiating false and true positives. After annotation,

[10]These articles talk about the DOJ's challenge against an abortion restriction law in Idaho. Here, the experts mentioned that the article from the left-leaning source did not mention an opposing perspective, such as that of Idaho's government. Note, however, that the article did cite the state attorney general of Texas, who supports abortion restrictions.

users also mentioned noticing differences in how article pairs represented things, such as perspectives, information placement, information depth, and facts/opinions. In RQ2, we found that annotation tasks may have limited effects on users' perception of news credibility for high-contrast news articles. Below, we discuss the implications of these results.

5.6.1 RQ1: Annotation Performance

Our results indicate that although users perform poorly in general, their performance varies across annotation task types and article pairs, depending on the degree of contrast between the articles in the pair. Compared to the experts, average users seem to find fewer connections and consider more statements worthy of inclusion in the other article in a pair. This difference shows how experts and users differ when reading two news articles. This difference could stem from analytical capabilities; that is, perhaps finding similarities and differences is a task that requires particular expertise relating to news. Or, perhaps low knowledge users are more attentive to the articles. Another reason behind the difference could be users' preconceived biases regarding the media in general [216]. Such perceptions may have influenced users to see fewer connections and more differences. Therefore, one direction for future research may be to examine the rationales behind identified differences by varying the task complexity and user characteristics. For example, we can ask readers to perform small subtasks, such as identifying sources or labeling word choices [184] to test if these influence their perceptions of bias.

Another noteworthy aspect here is that our participant group came from Facebook advertising, not from platforms like Amazon Mechanical Turk or Upwork typically used for crowdsourcing. This suggests that users outside of crowd work platforms can also perform effectively on crowdsourced tasks. In the future, research could look into how well workers

from crowdsourcing platforms and other sources compare in terms of performance.

Our result indicates that crowd annotation in subjective tasks is to a little extent affected by users' backgrounds—in our case, their news expertise, aligning with prior works [116, 205, 386]. Therefore, we can train users using their news expertise as a targeting criterion. Since user performance also varies by task, helping users improve quality on a particular task area might also help. Furthermore, designers can support users in annotation tasks through various interventions. For example, since our TF-IDF models identified some generic words that can distinguish false positives, such data could also be used to provide users with feedback or warnings to improve annotation quality.

We also discovered the effects of comparative annotation on users' overall impressions, leading to differences in perceptions of viewpoints, information attributes (placement, depth, and factuality/opinion), and emotional attributes (empathetic vs. inflammatory language). These differences could impact users' attitudes towards an article. For example, between informational and emotional attributes, understanding which differences impact perceptions of trustworthiness could be one future avenue of work.

To improve users' performance, one option could be through collaboration—learning from each other through social annotation [201]. Indeed, prior research shows that when people see others' annotations, it can persuade them to take certain actions, such as changing ratings when faced with opposing social opinions [98]. Furthermore, research suggests that displaying social information about the annotator, such as their level of expertise, can persuade and build trust [165]. Incorporating social information on other annotators during collaboration may improve learning. In a collaborative environment, we still need to handle annotator bias, since bias from a small group of users could propagate to a larger pool of users and cause unexpected effects. Therefore, examining such collaborative annotations and their impact on user performance is another potential direction of research.

5.6.2 RQ2: The Effect of Engaging through Annotation

While it may not be true in all cases, our results indicate that in cases where there is significant contrast between a pair of news articles, users might be somewhat influenced by comparative annotation tasks. Our work can inform related future works on improving engagement with plural viewpoints through annotations [493]. Compared to works that show visualizing biases alone does not improve perception of bias [426], our work suggests that additional engagement could be helpful. The effect we see may stem from complex information processing that occurs when users engage with competing messages [37]. Since our result did not reveal any universally significant effect, it does point towards the idea that only certain perceptions are affected. Therefore, one direction for future research could include looking into different perception paradigms to further identify the limits of such effects.

Even though we found limited effects on perceptions of credibility, this does not necessarily limit the applicability of comparative annotation. As we saw in section 5.1.6, a user's understanding of the differences between articles could have other impacts. Besides, repeatedly annotating two sources can create certain impressions in the long run. For example, seeing repeated differences in the use of factual statements or depth in reporting could affect users' perceptions of credibility. Furthermore, we can ask whether crowdworkers from such platforms as Mechanical Turk would also remain unaffected by the annotation task. In any case, NewsComp could be purposefully deployed to crowdworkers while also providing general users the option to perform annotation. In such a case, users desiring more ways to engage and community fact-checkers might be more attracted to it. Regardless, there are further uses for the annotated data.

5.6.3 Applications of Annotated Data

Our annotated data could be used in various ways. For instance, it could be incorporated into a system that combines information from multiple sources to provide a holistic view of an event. It is not uncommon in online spaces for information overload to make it harder for people to efficiently consume information [6, 122]. A holistic view could particularly be useful to users in such a scenario, especially for sensemaking purposes [487, 488]. Such a system would mimic strategies humans typically employ to consume information efficiently, such as organizing information by tagging, sorting, and indexing [73]. We could further build upon this by introducing mechanisms for peer-curated information [363] A second potential use of the annotated data would be in training algorithmic models to generate better annotations, which could in turn be used to better curate information for readers. As we saw in our initial think-aloud interviews, people find the accuracy of existing SOTA ML systems insufficient for finding semantic similarities and differences. The annotated data could help to improve such algorithms. A third use of the data is for fact-checkers. Fact-checkers can use annotated information to validate claims through the use of linked statements from multiple sources. They can also use such links to trace the origins of statements. Perhaps a portion of these fact-checking tasks could be delegated to automatic fact-checking algorithms. Furthermore, even crowd fact-checkers (e.g., from Twitter's BirdWatch) could use the annotated data to validate claims.

5.6.4 Merging Articles Into One and Testing Effects

One of the goals of this research on comparative annotation was to combine diverse perspectives into one. With our annotated data, crowd tasks can be designed to accomplish such merging of perspectives. However, there are some considerations for task design in this

process. Take similar statements as an example—if two statements are very similar, a task could ask workers to choose one or the other. On the other hand, if selecting one statement necessarily results in the omission of important information from the other, then the crowd task may also require editing. In the case of merging important disparate statements, as noted in our think-aloud interviews, one important consideration is checking whether a statement fits the narrative of the current article. We can either include this criteria or discard it, which would lead to differences in the outcome, (i.e, the merged article). Taking a step further, this merging process can be extended from article pairs to larger groups of articles. Merging larger groups of articles would require a multistep selection, voting, and reconciliation process. Finally, while we found that the effect of annotation on perception was limited, could merged articles affect users' perceptions of an event differently than articles from a single source? Future research answering such a question would generate new knowledge regarding the utility of comparative annotation.

5.6.5 Implications for Comparative Annotation Task Design

Motivated by users' perceptions of NewsComp, we identified two major issues in the comparative annotation task: the lengthy nature of the task and difficulty in performing the task. Since one of our research questions focused on the impact of performing annotation, our experiment was designed so that users performed a complete annotation task on two articles before responding to the questions. If the annotation impact is not of interest, both of these issues can be resolved. First, we can modularize the tasks by breaking them into small pieces (e.g., making connections between two paragraphs instead of two entire articles), in line with prior research on devising microtasks for complex work [232, 233]. However, could such modularization cause a backfire effect? For example, if an annotator is assigned two dissimilar paragraphs from a pair of broadly very similar articles, could that skew their

perception? This is one potential consideration for designing small, modular tasks.

Second, even if the task is not divided into smaller components, there are other options for improvement. For instance, finding similarities can be made easier through the addition of such features as automatic suggestion and filtering. Here, algorithms can provide automatic suggestions and users can search by keyword to limit the options to choose from.

Third, tasks can be divided for co-annotation to reduce difficulty. For example, one annotator might suggest connections while another annotator votes on the suggestions. In addition, as a tutorial, displaying example annotations from other users could also help resolve some concerns. However, the examples need to be generic enough not to significantly impact users' own future annotations.

Fourth, apart from issues related to task difficulty, there is another issue that will need attention in the future. In the think-aloud interviews and the deployment of NewsComp, we discovered some disconnects in the rationales provided for annotations. Particularly, for connection-making, we did not see use of thematic similarity during deployment. Perhaps regular users may need nudges to identify high-level thematic similarity. Overall, there are ample opportunities for improving the tasks in NewsComp.

5.6.6 Limitations

Our work is not without limitations. First, our study was conducted in a controlled environment which may differ from that of a user's typical news reading sessions. Therefore, some of the observed effects could have been products of the environment. However, we emulated a typical news consumption environment as best we could, from content selection to the design of the interface. Therefore, our results offer some validity that future works can build on. Second, since our study procedure involved signing up for the study and voluntary

completion criteria, some self-selection bias exists, similar to other research in this domain. However, we did advertise on Facebook to find users organically instead of recruiting users from crowd survey platforms, which provided some benefits to the selection process. Third, we conducted the study within a US-centric context, limiting its generalizability. Future research could resolve such issues by conducting similar research with a larger country pool. Finally, the task in the study was a bit lengthy (20 min) relative to tasks that crowd workers typically perform. Though the articles in the study were not excessively long (11-16 sentences), this could still have affected task quality. Future work can further examine how performance varies by task complexity. Overall, our work has certain merits that require further exploration in the future.

5.7 Conclusion

In this work, we examined how well users perform on a comparative news annotation task featuring a pair of news articles, and how the annotation task affects users' perceptions of the articles. Comparing our users' annotations against those of experts, we found that users generally performed very poorly on the annotation task. However, certain information, such as the number of users who made a given annotation and users' rationales behind annotations, can be used to detect incorrect annotations. Furthermore, we found some marginal changes in users' credibility perceptions for certain news articles after completing the annotation process. Our work has implications for designing future comparative annotation systems.

Chapter 6

OtherTube: Facilitating Content Discovery and Reflection By Exchanging YouTube Recommendations with Strangers

6.1 Introduction

Social media and content sharing platforms primarily use algorithms to individualize their feeds and content in order to increase user engagement [104, 134] [1]. These algorithms typically work by predicting what users will be interested in based on their prior interaction histories [99]. This mechanism often ends up limiting the set of content that users are likely to consume, filtering out the vast majority of content available that users could have enjoyed [104]. While the resulting recommended feed may increase user engagement, users may be trapped in a filter bubble—isolation from alternate viewpoints—potentially limiting their choices [40, 341]. However, it is challenging for users to gain awareness of their limited content consumption and to understand others with broad information intercepted by algorithms. Users with low cognitive reflection are especially susceptible to being swayed

[1] part of this chapter appears in [45]

to extreme beliefs [429]. Other research shows that while some people might be aware of the existence of such filters, they take little action against them (e.g., clearing their browsing history, using a browser's "incognito" function, and clicking/liking different posts) [65]. Though algorithmic improvement for diverse recommendations has been an active area of research [200, 491], it still falls short of its goal, resulting in the persistence of filters [61]. Therefore, the limitations of modern recommender systems raise the need for design interventions that can facilitate diverse content discovery, reflection, and understanding.

One potential intervention approach in bursting algorithmic bubbles is to present diverse viewpoints [337, 406] by exchanging recommendations with others and recognizing how one's social media feed is different from those of other users. For example, in the case of YouTube, one way to implement this solution is to show users recommendations that others received; that is, a collection of videos that YouTube's algorithms recommended to other users of the platform. We anticipate that seeing recommendations from strangers may be beneficial for users who are otherwise exposed to a limited set of content. First, knowing the kinds of videos that are recommended to other users can facilitate reflection on one's own tastes and consumption behaviors through social comparison. Prior research suggests that such comparison between peers could lead to improved self-knowledge or reflection [135, 502]. Second, seeing diverse recommendations from strangers could also facilitate discovery of new content, such as content that simply seems interesting, content that specific groups of users watch, and content that a user did not know was available. Lastly, exchanging recommendations may be more effective if users present themselves (or a proxy of their tastes) to strangers, to some extent [26, 408].

The goal of this paper is to explore the idea of exchanging algorithmically mediated recommendations as a way to facilitate content discovery and reflection, and to assess the potential barriers to such approach in self-presentation and content consumption. We accomplish this

by designing, developing, and evaluating OtherTube—a browser plug-in for YouTube—which records the videos recommended to a user from the YouTube homepage and displays them to others. OtherTube allows users to see strangers' YouTube recommendations (see Figure 6.4) as part of the homepage. To facilitate better social comparison, OtherTube also lets users create an anonymous persona (see Figure 6.2) and display it along with the recommended videos. Furthermore, OtherTube lets a user remove recommendations that they do not want to share. More specifically, we aim to answer the following research questions:

RQ1. *How do users discover content by browsing recommendations personalized for strangers?*

RQ2. *What factors affect users' interactions and engagement with recommendations personalized for strangers?*

RQ3. *How do users present themselves when sharing recommendations with strangers?*

RQ4. *How does browsing recommendations personalized for strangers facilitate reflection?*

To answer, we are conducting a 10-day long user study with the plug-in. To recruit a diverse set of users for the study, as opposed to recruiting a more homogeneous group from a university, we are using a Facebook advertisement. In addition, we log all the interactions that happened within OtherTube during that period, such as the number of videos clicked and the number of clicks to see different personas.

Our preliminary analyses show that OtherTube can help some users, but not all, in developing new interests and rediscovering prior ones by seeing strangers' personalized recommendations. Upon viewing others' recommended videos, users were able to understand more about their interests and how unique those interests were. Encountering other users with similar interests also gave users a sense of belonging. Next, I outline the design of OtherTube, our study protocol and preliminary results.

6.2 Related Work

In this section, we briefly review existing research around content discovery and reflection pertaining to recommender systems. With our system supporting self-presentation and comparison with strangers, we also review related literature.

6.2.1 Supporting Diverse Content Discovery Online Through Recommendations

Social recommender systems have become ubiquitous over the last decade, in areas such as social media (e.g., Facebook), e-commerce (e.g., Amazon), video sharing platforms (e.g., YouTube), and recreational services (e.g., Netflix). As recommender systems have become highly accurate in estimating users' preferences [180], it also comes with caveats, such as filter bubbles [341]. These problems prompted inquiry into diversifying users' exposure to differing viewpoints [380]. In this respect, one line of research takes diversification as a quality metric for recommender systems' performance and introduces novel approaches to improve it [74, 200, 257, 294, 491]. However, these methods face challenge as they have to trade-off diversity with accuracy [504]. In parallel, there has also been some research over design-centric approach to address the issue of filter bubble on social platforms [163, 252, 296, 336, 401]. These approaches include design interventions to understand users' own content consumption habit by showing information such as their topic-wise content consumption [401], political leaning of the sources they consume information from [406] and political leaning of their own social network [163]. Some of these approaches also promote viewpoints from alternate perspectives, such as, related content from alternate sources [336] and viewpoints from a user with different political ideology [252]. Our work adapts the approach of showing alternate viewpoints for YouTube by extending a particular demography-based feed exchanging

approach to a stranger-centered feed exchanging one.

6.2.2 Social Comparison and Self-Presentation

The theory behind behavior change leveraging social comparison is not new [135]. In some cases, such comparison could act as a support. In others, social comparison can trigger peer pressure which promotes competition [90, 360]. Prior studies found this mixed effect within the same system [92, 270, 497]. Research also shows that constructing better self presentation for social comparison on sites like Facebook may lead to improved self concept and self-esteem [157, 502]. Motivated by these existing results, we designed OtherTube to allow users to create own persona which can be shared and compared against strangers. While comparison might be easier when information from others are available, it also conflicts with users' need to preserve privacy for certain information [315]. To address this concern, Garbett and colleagues used pseudonyms and avatars, protecting users' identities [150]. Since some research suggest the presence of toxic interactions on YouTube [83, 334], we use a similar approach to anonymize users' self-presentation (using pseudonyms, avatars and generic demographic information) in our design of OtherTube.

6.3 OtherTube: Design and Implementation

To provide an environment that can be integrated into users' YouTube usage, we built OtherTube. OtherTube is a Chrome extension usable across all operating systems; users need only use the Chrome browser to browse YouTube. Our system works by collecting a user's YouTube recommendations—specifically, the top two or three rows of videos—each time a user visits the YouTube homepage. It stores the recommended videos in a database to be

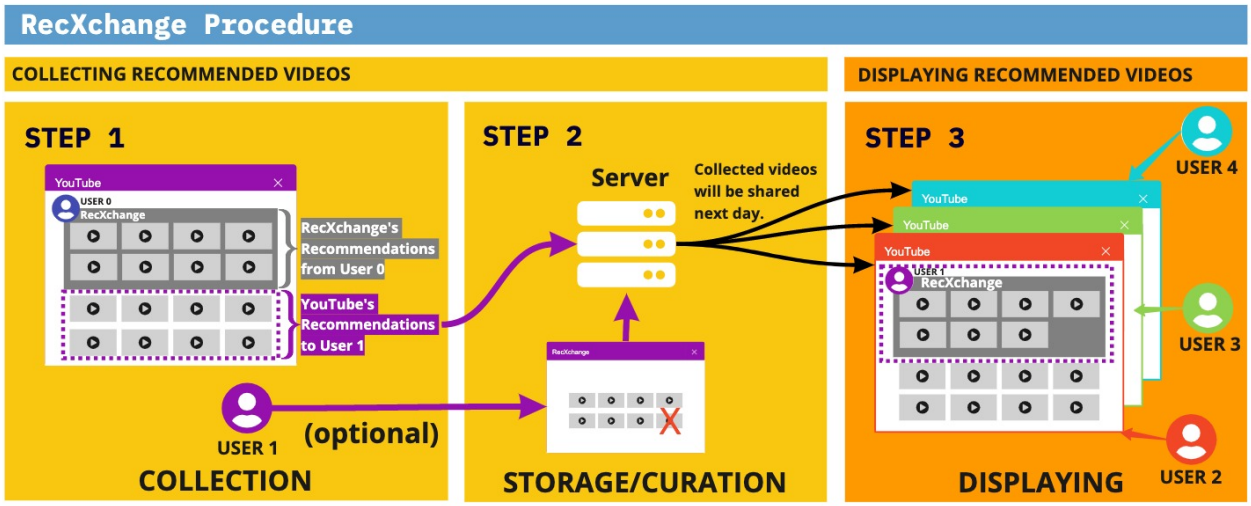


Figure 6.1: How OtherTube works. Each day, OtherTube collects YouTube recommendations when users access the YouTube homepage. Users have until the end of the day to remove items they do not want to share. Users can browse recommendations collected from others as late as the previous day.

shared with strangers from the next day onward. In short, users are given access to strangers' recommended videos in exchange for providing their own recommended videos to strangers. Figure 6.1 demonstrates this process. Additionally, OtherTube provides three main affordances: (a) an option to allow users to create an anonymous profile (Figure 6.2), (b) an option to remove collected recommendations that they may not want to share (Figure 6.3), and (c) an option to choose between browsing strangers' profiles[2] and recommendations from users' own YouTube homepages (Figure 6.4). We describe each of these affordances below.

6.3.1 Creating an Anonymous Profile

To give users extra information about the strangers whose YouTube recommendations they are browsing, OtherTube asks users to create an anonymous profile. For each user, we generate a random screen name—a combination of an adjective, a noun, and a number (e.g.,

[2]Throughout the text, we use the term persona and profile interchangeably.

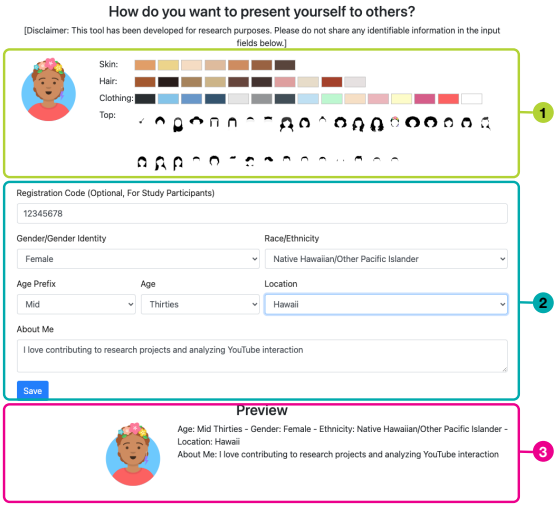


Figure 6.2: OtherTube Options page. 1 Avatar builder 2 Shared demographic info 3 How the profile will appear to others.

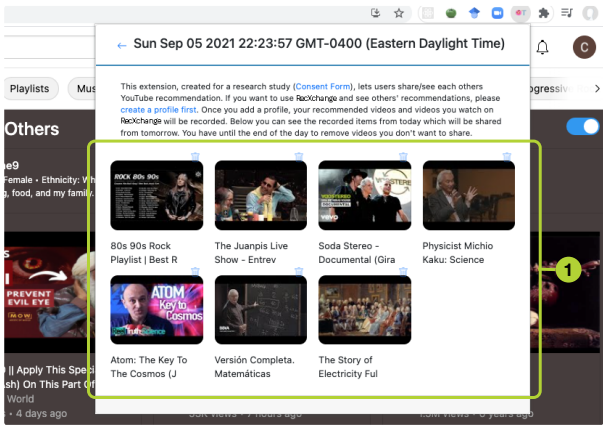


Figure 6.3: OtherTube Browser action page. 1 Collected videos with options to remove from the shared set.

amazedOtter4)—automatically in the back end. Users can choose an avatar for their profile and set several demographic attributes. Figure 6.2 shows the available options. The avatar builder allows users to choose a skin tone, hair color, clothing color, and appearance[3]. For demographic attributes, users have the option to share or not share their age, gender / gender identity, ethnicity, race, and location. To provide anonymity, users can only set their age as a decade-based bucket (e.g., Teen and Twenties) and location by state or province. Per each attribute, users also have the option to enter their own values for some of these attributes. Apart from the demographic details, users have the option to fill out an open-ended "About Me" section in their profile. For example, Figure 6.4 shows a profile with the text "Love traveling, food, and my family" in this section. Furthermore, at the top of the page and the About Me input field, we put disclaimers asking users not to share any personally identifiable details. Users also have the option to update their profiles at any time. Finally, note that OtherTube does not start collecting recommendations from a user or show others recommendations until users have created a profile. This ensures that users

[3]We used a third-party library, *AvataaarsJs*, for the avatar builder

only see collected videos attached to a profile.

6.3.2 Sharing and Removing YouTube Recommendations

Each time users visit the YouTube homepage, OtherTube collects their YouTube recommendations and sends them to a back-end server, which then stores them in a database. Going forward, we will call each visit a *session*. It is worth nothing that the recommended videos are generated by YouTube's algorithm; they do not simply consist of a user's browsing history. This means that the collected videos do not constitute an interaction trace, so the plug-in does not have to monitor a user's entire watch history, which could be perceived as private data. While recommended videos are typically not the one that a user watched, recommended videos are heavily customized for individuals based on their watch history, containing items from topics and content creators they watched previously [99]. It remains unclear if users would perceive algorithmically recommended content as a part of their self-presentation. Such content is neither technically private data nor under users' complete control. We discuss related results in section 6.5.3.

Once a session of recommended videos is stored on the server, the recommendations are shared with other users over the following days. If users do not want to share certain YouTube recommendations with strangers, they have the option to remove collected videos. We facilitate this through a feature of the Chrome extension. When users click on the extension button next to Chrome's address bar, the extension shows a list of sessions sorted by time, with the most recently collected items at the top. While browsing the collected sessions, users can click a remove button in the upper-right corner of each video (the blue trash can icon in Figure 6.3) to remove content that they would rather not share with strangers.

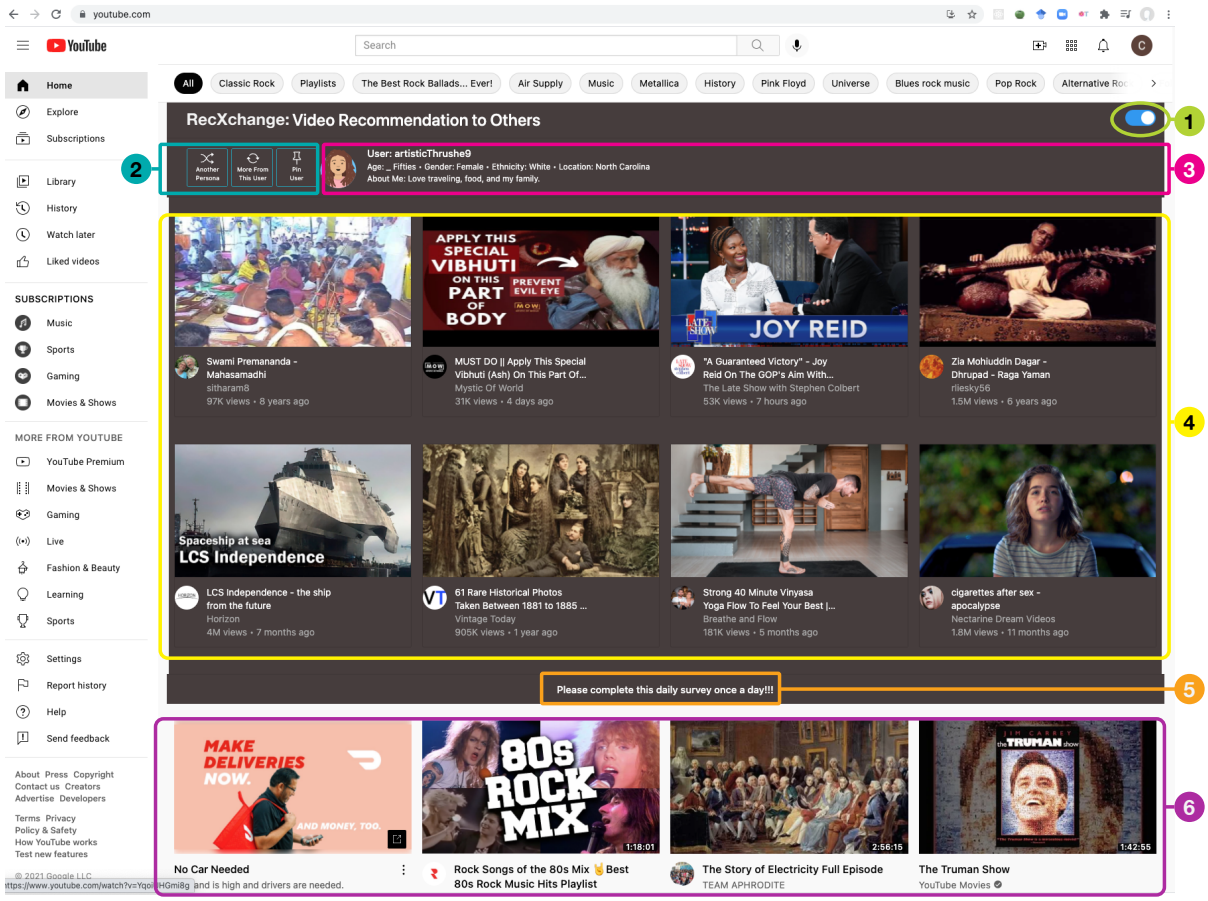


Figure 6.4: OtherTube embedded inside the YouTube homepage. ❶ Show or hide the embedded content. ❷ Browse different strangers or different recommendation sessions from the current stranger, and pin the current stranger. ❸ The stranger's profile. ❹ YouTube recommendations collected from this stranger. ❺ Link to a daily survey. ❻ The user's own YouTube recommendations, which OtherTube collects.

6.3.3 Interacting with Strangers' Recommendations

OtherTube embeds recommendations from strangers, which are collected through the steps described in Sections 6.3.1 and 6.3.2, at the top of the YouTube homepage (see Figure 6.4). Users have an option to hide or show the recommendations using a toggle button in the upper-right corner of the embedded content (Figure 6.4-❶). In the upper-left corner, OtherTube displays three buttons that allow users to browse another stranger's recommendations (Another Persona, browse another recommendation session within the same stranger's persona

(More From This User), and pin this stranger's persona (Pin User) (Figure 6.4-❷). Each time users click Another Persona or More From This User, the server returns a random stranger's recommendations or a random session from the current stranger, respectively. When a user clicks Pin User, a shortcut to the displayed user's profile and recommended video collections is created below the button. We added this button in case a user wants to follow another stranger's recommended videos. To the right of the three buttons, a stranger's profile is shown, consisting of their avatar, their demographic info, and their About Me text. Below the profile, OtherTube shows the current stranger's YouTube recommendations. While YouTube has recently started showing animated previews of each video, to make our implementation easier, OtherTube only shows still images (known as *thumbnails*) of videos, as it was in the past. Below the recommended videos, there is a link to a daily survey which we asked participants to fill out each day during our study (see Figure 6.4). Finally, OtherTube only tracks users' interactions within the plug-in (e.g., clicks on the Another Persona, More From This User, and Pin User buttons; and clicks on videos).

We built the front end of OtherTube using the React and Polymer JavaScript libraries, with Bootstrap CSS for styling. The back end consists of a Flask-Nginx server with a MySQL database for storage. All communication between the front end and back end is encrypted using SSL. After building the tool, we tested it within our research groups and ran a pilot study, fixing technical issues and improving usability. We distributed the extension through the Chrome Web Store.

6.4 Study Deployment

Using OtherTube, we conducted a 10-day-long study[4]. Below, we outline our recruitment method, study procedure, data collection process, and analysis.

6.4.1 Recruitment

For our study, we aimed to recruit participants who use YouTube on a regular basis. In addition, we decided to recruit participants using social media, specifically using Facebook Ads, informed by others' successes in recruiting diverse populations [7, 377, 483]. This advertising technique allowed us to reach a more diverse and targeted demography compared to Amazon Mechanical Turk or dedicated survey sites like Qualtrics [50]. Initially, we ran an advertisement campaign targeting individuals living in the US who are 18 years of age or older, speak English, and are interested in YouTube videos. While limiting our target demographic to those who live in the US would limit our findings, we did not want to have to account for language barriers in exchanging video recommendations. Our goal was to still reach diverse populations in terms of age, gender, and ethnicity. In addition, studying users living in a single nation provides some useful common ground upon which they can relate their interests to those of others (e.g., popular artists and domestic news in the nation). The recruitment campaign was set to run for one week, from July 8, 2021 to July 15, 2021. We spent $315 on the campaign and received 568 responses.

[4]This study was approved by the university's Institutional Review Board.

6.4.2 Procedure

Users who clicked the Facebook advertisement were redirected to a pre-survey to sign up for the study. At the beginning of the pre-survey, we screened users according to several criteria. To be eligible for the study, users had to: (i) be 18 or over, (ii) be currently residing in the United States, (iii) visit YouTube at least once a day, (iv) typically browse YouTube on a laptop or desktop computer (as opposed to mobile-only users), (v) typically use Chrome to browse YouTube, (vi) have English as the primary language of the YouTube content they watch, and (vii) typically start browsing YouTube from YouTube homepage (youtube.com). Out of the users who submitted the presurvey ($n = 568$), 318 were eligible for the study. We invited these participants via email to start the study[5]. The invitations were sent out in two batches: (i) July 19–August 3, 2021, and (ii) August 5–August 18, 2021. Note that while the emails to all participants in a given batch were sent out on the same day, users could have started using the extension on different days, resulting in batch periods exceeding 10 days. Out of 318 invitees, 41 participated in the study by installing the plug-in and filling out the daily survey at least once. We created an instructional document describing how to participate in the study. Users who accepted the invitation had to install the OtherTube extension from the Chrome Web Store. After installing the extension, users had to create profiles. At the beginning of the study, to mitigate the cold start problem, we created a research account so that participants could begin to see embedded recommendations from the first day. For 10 consecutive days, users were asked to use YouTube as they normally would and interact with OtherTube. Each day, they were also asked to submit a daily survey which took about five minutes. We sent reminder emails around 6 P.M. EDT each day to remind users who had not yet submitted the daily survey. Despite the reminders, participants did not consistently submit the survey, leaving us with 356 (8.7 on average) responses instead of

[5]Initially, we prioritized minorities for invites to form a diverse pool. Due to the limited response, we eventually reached out to all participants.

410 (41 participants x 10 days). Upon completion of the study, we invited about half ($n = 19$) of the participants to an interview based on their survey completion rates. Of those invited, 12 participated in the interviews. Each meeting was recorded for analysis. Because one user revealed that they neither followed the study instructions nor had a clear understanding of how the plug-in works, which would have left the user with no context for many of our questions, we ignored this user's response and analyzed the remaining 11 recordings. We compensated study participants with $25 gift cards and interview participants with $15 gift cards, adhering to federal minimum wage requirements.

6.4.3 Participants

Our pre-study survey was mainly designed to filter out ineligible users and create a diverse participant pool by matching demographic quotas. However, we did not completely fulfill this objective due to an inconsistent response to study invites. Figure 6.5 shows the distribution of age, gender, political affiliation, and length of a typical YouTube browsing session among our 41 participants. The participants' demography is balanced by gender. However, it is heavily skewed by race, with only one Black or African American participant despite a sufficient number of Black or African American users signing up for the study (see the contrast in Appendix C.1). This disparity could be caused by hesitancy towards installing tools or hesitancy towards research studies due to past injustices [206, 243]. Finally, by political affiliation, the majority were Democrats.

6.4.4 Data Collection

We collected data from our participants in multiple ways, beginning with the pre-study questionnaire. The questionnaire was followed by interaction traces and daily surveys during

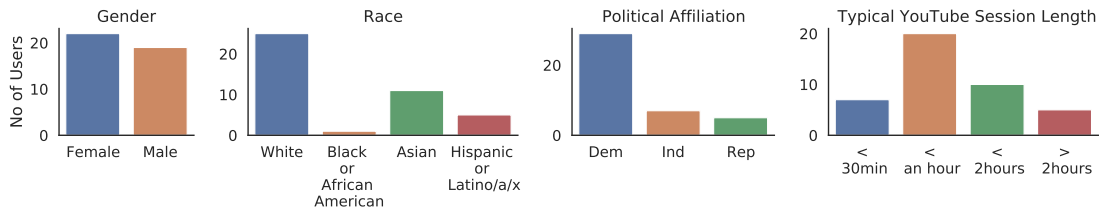


Figure 6.5: Demography of the participants in the study.

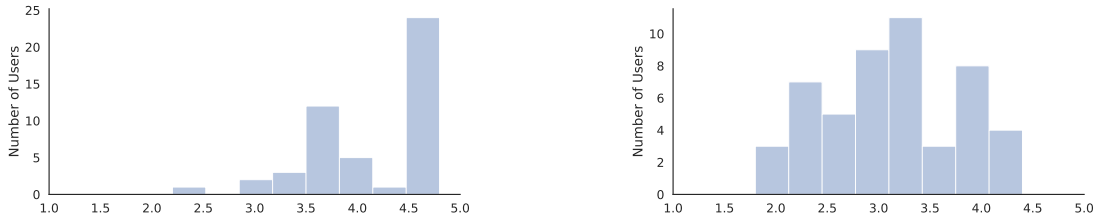
the study, and the post-study interviews came last.

Pre-Study Questionnaire: Need for Self-Reflection and Insight

In the pre-study questionnaire, along with demographic questions, we asked about users' need for self-reflection and insight (see Appendix C.2 for the items), using scales from prior research [183] on 5-point Likert items with responses from "strongly disagree" (1) to "strongly agree" (5). We measured these metrics to see if the self-assessed need for self-reflection and insight would correlate with how users interact with OtherTube. Participants' responses had a high level of consistency for both questions, similar to prior studies (Cronbach α [Need for Self-Reflection]: 0.97, Cronbach α [Insight]: 0.95) [171, 183]. For each trait, we found the mean of the items after inverting items that were phrased in the opposing sense. Figure 6.6 shows the distribution of user responses for need for self-reflection and insight. Here, we see a skew towards a greater need for self-reflection.

Interaction Traces from OtherTube

As mentioned before, we collected data on users' interactions with OtherTube, including clicks to see recommendations from different strangers, clicks to see more from a stranger, clicks to watch videos, clicks to pin strangers, and clicks to remove own recommendations.



(a) Need for self-reflection

(b) Insight

Figure 6.6: Distribution of participants' need for self-reflection and insight, bucketed for ease of understanding.

Question Type	Item
Likert-scale	(a) Today, I saw some videos on OtherTube that caught my attention. (b) Today, OtherTube recommended me videos that I would not have expected in my feed. (c) Today, I discovered some new types of content on OtherTube that I would like to watch. (d) Today, YouTube recommended me some videos that I would not have expected in my feed. (e) Because of using OtherTube, YouTube is recommending diverse videos to me. (f) I feel comfortable sharing my recommended videos with others on OtherTube.
Open-ended	(a) If there were any videos that caught your attention from OtherTube, could you tell us what they were and why you were interested in them? (b) If you removed any recommended videos of yours from the OtherTube plug-in, could you tell us why? (c) While using OtherTube, did you learn anything new about certain populations (different age group, different gender)? If so, what did you learn? (d) While using OtherTube, did you learn anything new about your own taste compared to others? If so, what did you learn? (e) What made you hesitate to watch videos from OtherTube, if any?

Table 6.1: Daily Survey Questions

Daily Survey: Perceptions of Recommendations on OtherTube

In the daily surveys, we asked users Likert scale–based and open-ended questions about their experiences with OtherTube on YouTube. Table 6.1 shows the questions. We included several Likert items on a 5-point scale from "strongly disagree" (1) to "strongly agree" (5) about users' perceptions of OtherTube on YouTube. These questions captured users' perceptions of content discovery each day, whether their own YouTube recommendations were affected by their interactions with OtherTube, and how they felt about sharing their YouTube recommendations with strangers. Through open-ended questions, our goal was to understand participants' reflections on themselves and others, and to learn the motivation behind any actions they performed or chose not to perform (e.g., content removal or hesitation to watch

content).

Interview

After completion of the 10-day study, we interviewed 11 participants. The interviews focused on understanding users' behavior when using the tool, including how they presented themselves to others, how they used the features of the plug-in, and whether using OtherTube encouraged content discovery or facilitated any reflection. We ended the conversations with some usability questions. Additionally, after analyzing each interviewee's daily survey responses, we asked them to elaborate on any points that seemed unclear. For example, because one user mentioned that their interests are "conservative" in their daily survey, we asked them to elaborate on what they meant by that (i.e., whether the user was interested in political conservatism or had conservative viewing habits, preferring to watch the same kinds of videos). This participant reported that "conservative" meant they watch only what they like, and they carefully select what videos to watch. See Appendix C.3 for the complete interview questionnaire.

6.4.5 Method of Analysis

We gathered both quantitative (pre-study survey, Likert scale–based daily survey questions, and interaction traces) and qualitative responses (open-ended daily survey questions and interview responses). Below, we describe our analytical methodology.

Likert Items on the Daily Survey

To gain insight into self-assessment on the statements, we checked the average responses to each Likert item on the daily survey. We also performed Mann-Whitney U tests, a

non-parametric test, to see if using OtherTube produced any changes in their perceptions, comparing the responses from participants' first and last days of using OtherTube. Because not all users submitted the survey on each of the 10 days, the last day might not have been the 10th day for each user.

Interaction Traces

To answer RQ2, we modeled the number of daily clicks on the Another Persona button using negative binomial regression[6]. The independent and control variables for the regression consisted of users' demographics, the need for self-reflection, and insight. Because we recorded the number of clicks on each of 10 days, resulting in repeated measures by each user, we used a mixed-effects regression model. Additionally, because clicks are a count variable, we used negative binomial regression[7]. Because some users did not click the buttons every day and random effects require multiple observations per user, we used users who clicked the button on at least 3 days. Therefore, instead of 410, we had 280 data points from 32 users (an average of 8.8 data points per person) for our model. We used *mixed_model* from the *GLMMadaptive* R package [384]. For the sake of interpretation, we present marginal coefficients instead of fixed effect coefficients in this model[8]. Additionally, we also examined the Spearman rank correlation, a nonparametric test, to assess whether there is any correlation between users' interactions and engagement; that is, clicks on videos, Another Persona, and More from This User. Apart from these tests, we also analyzed other simple statistics.

[6]We did the same thing for the More from This User button, but the results were not significant; therefore, we omitted them from this paper.

[7]Due to overdispersion, we chose a negative binomial model over Poisson regression.

[8]Given the nonlinear link function (*Log*) in our model, random effect intercepts can have a multiplicative effect, not additive, complicating interpretation. Therefore, following Hedeker et al. [196], we extracted the marginal coefficients and their standard errors from the model using the *GLMMadaptive* R package.

Open-ended daily survey responses

We performed thematic analysis on the open-ended daily survey responses. With five open-ended questions, we had 1,780 responses (356 daily survey submissions $\times$ 5 questions), 976 of which were either empty or contained unhelpful responses, like "no" or an incomplete response (e.g., "Yoga videos" for Open-ended-(a) in 6.1). This left us with 804 valid responses. Three researchers performed thematic analysis on this data. Initially, each of the coders came up with their own set of codes. After discussion, we converged on a set of 52 codes. Through discussion, we reduced this subset of 52 codes to a list of 21 themes. Two of the researchers coded a sample (159 items, or 20%) into the set of 52 codes. Coders had almost perfect agreement despite the large number of codes (Cohen's $\kappa = 0.82$) [472]. One of the coders coded the rest of the items. Due to the uneven number of codes associated with each user, as some did not regularly fill out the daily survey, we merged the 10-day codes for each person into a single set. Consequently, as we present our results, "theme (10/41)" means that out of 41 users, 10 users' responses contained at least one response belonging to the theme. Each user may have given such a response anywhere from 1 to 10 times over 10 days.

Interview

Similarly to the open-ended survey responses, we performed thematic analysis on the interview responses. One of the researchers performed an initial analysis and came up with 35 codes from the interviews. Then, this researcher discussed the codes and corresponding quotes with other researchers. After resolving disagreements, we were left with 24 codes from the interviews. As most of them were related to the daily survey themes, we merged the two sets.

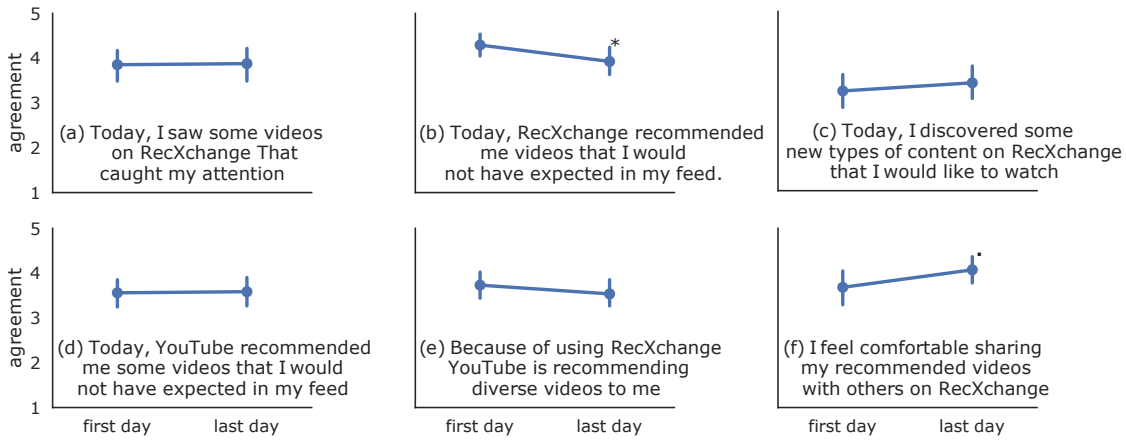


Figure 6.7: Mean with 95% CI of participants' responses to the daily survey Likert items from the first and last days of the study. We also performed a Mann-Whitney U test comparing responses on the first and last days. In (b), * indicates $p < 0.05$; in (f), . indicates $p < 0.10$.

6.5 Results

Based on our research questions, we present results consistent across daily survey themes, interaction trace analysis, and interview themes. Note that the daily survey and interview quotes are presented in the format "U1 (survey)" and "U1 (interview)".

6.5.1 (Content Discovery) RQ1. How do users discover content by browsing recommendations personalized for strangers?

In the daily survey, the majority of participants agreed with the statement (mode: 4, "somewhat agree") saying that they saw some videos on OtherTube that caught their attention. Further analysis shows that this perception did not change between the first and last days. This result is shown in Figure 6.7-(a), where a Mann-Whitney U test shows no significant difference. Our analysis revealed two themes that illustrate how users utilized OtherTube to discover content: (i) users developed new interests, and (ii) users rediscovered content they used to like. We present these themes below.

Users developed new interests

In the daily survey, a majority of our participants ($n = 36/41$) mentioned finding new interests. Users found interests in multiple ways. Sometimes, they found new interests out of curiosity (*"I had no clue what it [a video that caught their attention] was about but just curious on the content" - U1 (survey)*). Other times, new content led to the development of a new interest due to its usefulness (*"... A video about investing for your kids' future. I have a newborn and want to do that for her" - U2 (survey)*). There were also cases where users found new content that fit within their existing interests (*"Relaxing Video Game Music in a Cozy Room (Nintendo 64) [caught my attention]. I enjoy similar relaxing music playlists" - U12 (survey).* Some users also watched content for the sake of exploration (*"I don't normally listen to that [classical music] and it was a welcome change" - U17 (survey)*). Interviewees also responded similarly, with one mentioning how they bumped into new interests on OtherTube. Merely encountering the embedded interface of OtherTube could trigger changes in consumption behaviour.

> "*What I like about it is that some it forced me out of my comfort zone of what I would just do ... It makes you to stop and, like, look and think before you decide to make a different choice.*" — *U17 (interview)*

We also noticed that participants were able to recognize profiles that had similar tastes to their own. (*"I came across a user who had watched videos that were my interest. He was interested in computers and video games. So, I liked those videos too" - U20 (survey)*). This tendency demonstrates not only that one can find interesting videos using OtherTube, but also people that they can connect with, due to the similarity of the content that they watch. OtherTube's additional functions to retrieve more videos from a particular user can be useful to further explore the person's content. Analyzing the number of clicks on More

from This User and the number of clicks on videos users watched, we found a statistically significant positive correlation (Spearman's $\rho = 0.42$, $p < 0.001$). This pattern suggests that finding a profile based on similarities in taste could be a new way to find and watch new content—in essence, subscribing to other viewers, not creators, to follow their content consumption patterns. We discuss these implications further in Section 6.6.1.

Users rediscovered content they used to like

Using OtherTube, some users ($n = 11/41$) found content they used to like (*"old throwback videos from my childhood [caught my attention]" - U11 (survey)*). They also found channels they used to like (*"a video game reviewer, Zero Punctuation, popped up in one of the OtherTube recommendations. So seeing that was pretty nostalgic." - U3 (survey)*).

Apart from these themes, our participant interviews further revealed users' issues with finding strangers with similar interests. For example, while users' About Me details were sometimes useful, they wanted more features, such as filtering users by demographic or interest (*"Their favorite content creator, maybe your favorite video." - U26 (interview)*).

Summary: By browsing OtherTube, participants both found new interests and rediscovered old ones. They found new interests out of curiosity, usefulness, or desire to explore. Items caught participants' attention similarly on both the first and final days of the study. Finally, participants also recognized others who had tastes similar to their own.

	Click Count per Day (Another Persona)	
	β	**std. Err.**
(Intercept)	5.5024***	1.0936
Age	-0.0284*	0.0135
Gender[Male]	0.2592	0.2303
Race[Black or African American]	-1.2113	0.7136
Race[Hispanic or Latino/a/x]	-0.1595	0.3895
Race[White]	0.2526	0.2412
Daily Browsing Length[less than 1 hr]	-0.6087	0.3839
Daily Browsing Length[less than 2 hrs]	-0.1536	0.3628
Daily Browsing Length[greater than 2 hrs]	0.1773	0.4109
Need for Self-Reflection	-0.4149***	0.2108
Insight	-0.0829	0.1730
Day	-0.0453*	0.0185
# of Users Available on the Day	-0.0083	0.009
Dispersion	0.9119	0.1244
log.Lik	-721.8501	
N = 280	* p<.05, ** p<.01, *** p<.001	

Table 6.2: Mixed effects negative binomial model for daily click count on Another Persona in Figure 6.4. In this model, user is a random-effects variable. User demography, their browsing habits, their need for self-reflection, the day when the clicks were counted, and the number of unique stranger data sets available to browse on OtherTube on the day when the clicks were counted are fixed-effects variables. The estimated negative binomial regression coefficient β is the difference in the logs of expected counts of the response variable due to a one-unit change in the predictor variable.

6.5.2 (Interaction and Engagement) RQ2. What factors affect users' interaction and engagement with recommendations personalized for strangers?

To investigate what factors affected users' interactions with OtherTube, we modeled users' click activity on Another Persona over time using a mixed-effects negative binomial model. We considered several factors in this model, including users' demography, their need for self-reflection, their YouTube browsing length, and number of available unique profiles to browse on a given day. Table 6.2 shows the result. We find that users with higher age tend to interact less with OtherTube when browsing different persona on OtherTube ($\beta = -0.03, p < 0.05$). Our users with a greater need for self-reflection also interacted less with OtherTube ($\beta = -0.41, p < 0.001$). This result was particularly interesting because we hypothesized that those who believe that they need reflection would use OtherTube more

than those who did not have such thoughts, given our goal of facilitating reflection behind the tool. Perhaps our design will be more effective and promising for those without an explicit desire for self-reflection. Additionally, as the days passed, users interacted less with the tool ($\beta = -0.05, p < 0.05$). This decrease could have been caused by seeing the same profiles repeatedly, due to the limited number of users in the study (*"[On day 4]i haven't seen many changes. [On day 6] no same as yesterday." - U4 (survey)*). To understand users' engagement with the videos, we analyzed users' clicks on videos. We found that most users ($n = 37/41$) watched at least one video on OtherTube. Overall, participants collectively clicked on 6% or 398 videos out of all the collected videos on OtherTube.

While demographic attributes and other environmental factors can affect interaction, it might have been the content itself that encouraged or discouraged interaction with content from OtherTube. Analysis of users' daily survey responses revealed content-related factors that discouraged users from watching videos; these are discussed below.

Users hesitated to watch content that was not of interest to them

A majority of our users ($n = 25/41$) mentioned that they did not watch content because the content did not relate to their interests (*"I didn't find them [videos I hesitated to watch] interesting" - U40 (survey)*). Some users referred to such content as "boring" (*"seemed boring or predictable" - U35 (survey)*) or "complicated" (*"looked boring, or too techy and complicated" - U36 (survey)*). In the interview, one of the participants further elaborated on how their motivation for browsing YouTube at times affects whether they would engage with certain type of content (*"Sometimes I don't want to be educated. I just want to enjoy. Other people want maybe to learn about the history of Macedonia or something. I don't sometimes want to learn that." - U15 (interview)*).

Users hesitated to watch content that was not helpful

In contrast, content utility was the main factor for some users' ($n = 6/41$) engagement with content in OtherTube (*"I just didn't click the ones that seemed pointless and not helpful to my personal development or learning" - U2 (survey)*). Some expressed hesitation in terms of wasted time (*"They could be a waste of time to watch if I don't like them" - U26 (survey)*).

Users hesitated to watch disturbing and offensive content

Some users ($n = 7/41$) were also hesitant to watch content that seemed disturbing or offensive to them. Various types of content could fall into this category, including horror videos (*"I didn't want to watch horror" - U13 (survey)*), sexually charged videos (*"I didn't want to watch videos that seemed strangely sexually charged, like mouth ASMRs" - U38 (survey)*) and uncomfortable topics (*"I was disturbed by videos in uncomfortable topics I wouldn't want to think about, such as, physiological anomalies and unsolved murder mysteries" - U39 (survey)*).

Users hesitated to watch clickbait-esque content

Clickbait was one reason some users supplied for not watching content ($n = 4/41$). Users used both video titles (*"Titles were terrible and looked like trash" - U16 (survey)*) and thumbnails (*"I was hesitant to watch videos with objectifying thumbnail pictures of women" - U38 (survey)*) to rule out this kind of content.

Some of the barriers discussed so far may reduce the overall effectiveness of our approach, as users may consider the content displayed by OtherTube to consist of low-quality items (e.g., spam, clickbait-esque videos) or even explicit (e.g., profanity) . Future designs could introduce some mechanisms (e.g., automated spam filtering) to mitigate such perceptions.

Users hesitated to watch videos for fear of an interaction effect on their YouTube recommendations

A few of our participants ($n = 4/41$) were hesitant to watch content because they assumed doing so would affect their future YouTube recommendations (*"I saw an interesting video on Netscape, but didn't click given it's from a new channel I'm not familiar with, but also being aware that the YT algorithm is going to try and keep me engaged by literally suggesting more of said video" - U25 (survey)*). Some of our interviewees further revealed that they deal with content that they do not want to watch by blocking it. Similar blocking mechanisms can be added to OtherTube to filter out content that users do not want to see. We leave this for future exploration.

> "*I curated it [YouTube] fairly carefully. Occasionally I would get some random stuff [on YouTube recommendations]. Usually I just click on the little dots [a button on YouTube videos with a drop-down menu to block content users do not want to be recommended] and I'd say "don't recommend this channel"."* — *U39 (interview)*

This issue raises an interesting question of how users' engagement from watching content on OtherTube could affect their future recommendations from YouTube. In light of this, should designers make OtherTube available within a sandbox or incognito viewing mode? Sandboxing may encourage some users concerned about impacting the recommendation algorithm to interact and explore more content.

Summary: Our analyses show that factors such as lower age and lower need for self-reflection positively affected users' daily interaction with OtherTube. However, participants' interaction also subsided as days passed. We additionally found that 90% of our participants watched at least one video on OtherTube. Some content factors also discouraged engagement,

including lack of relevance to users' interests, lack of utility, and offensive or clickbait-esque nature.

6.5.3 (Self-Presentation) RQ3. How do users present themselves when sharing recommendations with strangers?

OtherTube offers two ways for users to present themselves to strangers: (i) a self-built profile and (ii) moderating the YouTube recommendations to be shared with others. Among our participants, few ($n = 5/41$) decided against sharing any of their demographic information, while six participants shared all of the demographic attributes except their location. The rest filled out all the demographic information in their profiles. Most ($n = 30/41$) filled out the About Me section with details ranging from their occupations (e.g., "manager" and "recent college graduate") to generic interests (e.g., photography, traveling, yoga, and music) and more specific ones (e.g., DIY, Hololive videos, and BTS). Our interview revealed that users wanted to change their About Me to explain the reasons behind specific interests (*"[If I were to redo the study,] I would be more specific. My interests may be updated as they changed. So I like to explain to you why I have seen, like, niche content in my recommendations." - U38 (interview)*).

To understand users' moderation practices for self-presentation with respect to YouTube recommendations, we examined users' activity in removing collected videos. We found that eight users removed 104 videos (<1.5% of the total videos shared). Two out of those eight users were responsible for about 80% (84) of the removed items. Therefore, it seems that most participants were not especially concerned about removing content they did not want to share. Both the daily survey and interview responses support the notion that the majority did not show any concern (*"I wasn't going to arrange in any way what videos I was watching*

[or remove videos] because I'm not ashamed of the videos I watch." - U39 (interview)).

Even those who were concerned at first became comfortable sharing over time. In our daily survey, we found that mean of the users' initial responses were close to "neither agree nor disagree" regarding the statement "I feel comfortable sharing my recommended videos with others on OtherTube." However, Figure 6.7(f) shows an increase in perceptions of comfort with sharing over time. We examined the difference in frequencies using a Mann-Whitney U test, and the result was marginally significant ($p = 0.07$). Analysis of the open-ended daily survey responses revealed two further reasons why some users had concerns regarding sharing videos. This result helps us understand how participants self-present through recommended videos.

About a quarter of the participants ($n = 12/41$) did not want to share certain content. Some ($n = 7$) were concerned that certain videos did not represent their interests (*"I removed some for being literally not content that I watch, that your tool seemed to pull from YouTube's suggestions" - U25 (survey)*). Others moderated recommended videos for the sake of other viewers. While participants realized that they were sharing content anonymously (*"[Did not want to share because I am] Self Conscious even though its anonymous haha" - U5 (survey)*), some ($n = 6$) still felt uncomfortable sharing certain content (*"Some [videos] felt excessive, or sorta just weird. And I'm not sure about sharing that." - U25 (survey)*). During the interviews, participants also revealed that they thought their content might not be of interest to others (*"The educational stuff I'm watching is very entertaining. But I'm pretty certain that other people would be extremely bored by that." - U39 (interview)*).

Summary: To present themselves to strangers, participants moderated both their self-built profiles and the videos they shared on OtherTube. While a majority filled out their anonymous profile, some did so only partially or not at all. Participants also revealed rationales for not sharing videos, such as content not matching their interests.

6.5.4 (Self-Reflection) RQ4. How does browsing recommendations personalized for strangers facilitate self-reflection?

Using OtherTube, our participants were able to reflect on themselves from various perspectives. We found the following themes here: (i) understanding of own interests and their uniqueness, (ii) feelings of belonging from seeing users with similar interests, and (iii) feelings of superiority from comparing content. These patterns help us understand the kinds of self-reflection that the system can facilitate. We illustrate these themes below.

Users understood more about their own interests

One of the effects of using OtherTube seems to be users ($n = 14/41$) understanding their own interests, what they like (*"I learn that I have interest in cooking after watching some cooking videos" - U41 (survey)*), and what they do not like (*"I am not interested in animation or anything about computers or how computers work" - U15 (survey)*). Some also realized their preferences in terms of video length (*"I like short and fun videos, nothing too long or too serious." - U16 (survey)*) Others realized that their identity influences their interests (*"I saw my tastes to be very related to my identify [sic] and cultural background" - U29 (survey)*).

After seeing strangers' YouTube recommendations, a majority of our participants ($n = 29/41$) further realized that their interests were unique compared to those of others (*"I learned how unique and distinct my taste and preferences are" - U9 (survey)*). Some thought that their interests on YouTube were very narrow or selective compared to those of others. Some of them defined this narrowness in terms of topics of interest (*"I mainly focus just on gaming while other people hit a lot more genres" - U26 (survey)*).

Seeing others with similar interests gave users feelings of belonging

Users were surprised to find strangers with similar interests (*"I was surprised to see another person also to be having similar tastes as me, it was like seeing my own watch history" - U29 (survey)*). Seeing others with interests similar to their own, many of them ($n = 22/41$) realized that they are not so different from others (*"I continue to learn that we aren't so different. I have always thought that from a gender perspective, but age is really where I feel my eyes are being opened." - U5 (survey)*). Some found commonality among similar demographics (*"People near my age enjoy similar content" - U19 (survey)*). Others found commonality with different demographics (*"I like the same videos as a woman in her 50's which made me laugh because I'm 28" - U2 (survey)*). Some also expressed their alignment in interests with particular demographics over others (*"I learned that I am more likely to share interests with older women than men of my age" - U13 (survey)*). Some users expressed both their uniqueness and their differences as part of being normal (*"I think that I am fairly normal or that weird is normal. Everyone has their own tastes." - U5 (survey)*). To see more from strangers with similar interests, OtherTube provided users with a pin button that allowed them to save other profiles for later perusal. We found that some of the participants ($n = 8/41$) indeed pinned some strangers ($n = 18/41$).

Our interviews further revealed that upon seeing users with similar interests, a few were interested in communicating with them, potentially to recommend them other interests (*"I think it'd be interesting to maybe even have that kind of connection where you're sharing. I don't want to say like you're friending someone on Facebook. But maybe sharing of content ... kind of being able to make a recommendation saying if you like this, maybe you like that." - U17 (interview)*). However, at the same time, others were not interested in communicating for fear of toxic interactions. The quote below illustrates this issue.

> "*But I've been on YouTube for years ... my impression is that overall, YouTube is an extremely toxic place. And the only way I managed to avoid the toxicity is by blocking creators who I find to be like bad people and not reading the comments section. Some of the most innocuous videos will have thumbs down on them for like, I don't know what reason. So I don't know if I want to connect with those people [who watch such videos]... And I thought that the way that you guys were doing it by making it impersonal was kind of good. Because then nobody could spew profanities at each other.*" — *U39 (interview)*

Some users felt superior when comparing their content to others'

Although not many, a few of our participants ($n = 6/41$) also reported feelings of superiority after seeing others' recommendations (*"Apparently, my taste is better than most people's."* - *U16 (survey)*). Some of them attributed this sense of superiority to the lack of variety in others' recommendations (*"I have more tastes than this lady. I seek out more content."* - *U19 (survey)*). Manifestations of superiority go against our goal of understanding others and could be detrimental (*"younger crowds listen to stupid stuff"* - *U6 (survey)*).

Summary: Using OtherTube, participants reflected on themselves and their interests. When participants found others with interests matching their own, they experienced a feeling of belonging, and vice versa. Comparing interests with others, some also felt superior to others.

6.5.5 (Learning about others and the algorithm) RQ4. How does browsing recommendations personalized for strangers facilitate understanding?

After seeing strangers' recommendations, participants expressed what they learned and how OtherTube increased their understanding of others. Some were surprised by the content others watch, while others were surprised by the fact that their preconceived notions about certain demographics did not always match. Users also discovered diversity in strangers' recommendations. We illustrate these themes below.

Users were surprised by some of the content others watch

When browsing their OtherTube feeds, some users ($n = 11/41$) were surprised to see certain content. Sometimes, users were surprised by content when it came from a particular profile or demographic they did not expect it to come from (*"I found it very cool that someone in their 40s was still watching Olivia Rodrigo. Makes me feel better about aging haha" - U5 (survey)*). However, comparing users' responses over time, this surprise seems to have gradually subsided. Indeed, Figure 6.7(b) shows a decrease in perceptions of seeing unexpected content on OtherTube, and this difference was statistically significant ($p < 0.05$). This result may indicate that participants got used to the recommended videos from OtherTube. However, it is worth noting that it is unknown whether this trend would have arisen even with users being exposed to completely new sets of profiles every day; the limited number of profiles available could have contributed to this trend. Still, the mean of their responses in Figure 6.7(b) did not fall below "somewhat agree" (4) on users' last day.

Users were surprised to see that certain stereotypes do not reflect reality

A few of our participants ($n = 6/11$) were surprised to see that some stereotypes did not match reality (*"... It [different interests among people] challenges some stereotypes as I see things on their feed that I wouldn't have expected" - U40 (survey)*). Some found users of certain ages watching videos they would not expect them to watch (*"I continue to be surprised by music recommendations in particular. A lot of older persons getting younger artists and some younger persons recommended classic rock/pop" - U5 (survey)*). Some also were surprised in terms of gender stereotypes (*"I saw a male user watching some homey vlogs and that was a bit surprising" - U8 (survey)*).

Users learned something particular about a person or a demographic, which could be stereotypical

Using OtherTube, about half of our participants ($n = 17/41$) mentioned learning about others' lives by watching the recommendations that strangers received (*"I learnt how you can tell about what is happening in somebody's life by looking at the videos they are watching." - U29 (survey)*). Some participants discovered what a particular demographic might be interested in (*"Men may be more into horror" - U13 (survey)* and *"40s and older watch more health related videos" - U41 (survey)*). While some people experienced a feeling of superiority when learning about others' interests, others had the opposite reaction (*"I learned to respect other's choices" - U29 (survey)*).

Meanwhile, some users ($n = 5/41$) also found that stereotypes about the kinds of content particular demographics prefer matched what they saw (*"... It reaffirms some stereotypes that I had, as some videos are expected" - U40 (survey)*). Users most often mentioned stereotypes about age, particularly for younger people (*"Today was much more in line with expectations.*

A teenager recommended some Pokemon videos, 20s getting anime." - U5 (survey)). Some noticed similar recommendations related to ethnicity (*"I learned than ethnic Americans likely have video interests in their ethnic culture" - U38 (survey)*). Some also employed stereotypes to infer strangers' ages (*"Despite keeping their age hidden, my first profile today was clearly a kid/teen" - U5 (survey)*). One obstacle in understanding more about others could have been that some users kept their demographic information hidden, as allowed by OtherTube (*"several people didn't put their demographics. So I don't know their age group, gender." - U41 (survey)*).

Users discovered diversity in strangers' interests

Some of the participants ($n = 10/41$) liked the diversity of the content they found on OtherTube (*"I like the diversity of participants and the content of the videos" - U11 (survey)*). They found diversity within particular strangers' feeds (*"... even within personas, it's getting harder to pin down a common thread. The recommendations can be pretty diverse." - U3 (survey)*). Users also found diversity across demographics (*"Every age group is different and they all post different material on here." - U4 (survey)*).

Apart from reflecting on self and learning about others, nearly half of our participants ($n = 18/41$) learned something about YouTube content broadly or what is trending on YouTube using OtherTube. Some learned about common interests of people on YouTube (*"They [OtherTube users] all have fun videos" - U11 (survey)*). Others learned of new phenomena on YouTube (*"People are using YouTube as a news source with more regularity" - U18 (survey)*). Some discovered the popularity of certain content genres on YouTube (*"K-pop is massively popular. even more than I imagined" - U17 (survey)*). Some also learned about the popularity of content they already watch on YouTube (*"Realized some content I watch is actually popular." - U25 (survey)*). Some also realized how the YouTube algorithm

can affect their interests (*"When I don't have my own tastes in videos, Youtube shapes my tastes" - U13 (survey)*).

Summary: After using OtherTube for 10 days, participants learned several things about others. Some of these things surprised users, particularly when their preconceived ideas did not match what they learned. Participants realized how diverse people's interests can be on YouTube.

6.6 Discussion

This study presents the utility of exchanging YouTube recommendations with strangers. RQ1 reveals unique ways in which this can facilitate content discovery. RQ2 identifies demographics that might be interested in browsing recommendations catered to strangers and watching videos from them. In RQ3, we found that while few participants put effort into self-presentation, they felt comfortable in setting up an anonymous profile and sharing recommended videos; a few of them, however, did moderate their recommended videos. Finally, we saw how users reflected on themselves and others through RQ4.

6.6.1 RQ1 & RQ2: Content Discovery, Interaction, and Engagement

While YouTube provides some features for finding new content (e.g., search or subscription, exploring trending videos), diversification of YouTube recommendations is still challenging. It is especially hard to recommend diverse content to users based solely on their watch history. As one of our participants noted,

> "*I sometimes want to find new songs ... it [YouTube] just kind of recycles a bunch of songs that I've heard ... In that instance, I actually have to basically open an incognito tab and pretend I'm just a random nobody to get actually novel or interesting music recommendations ... If I were to look for something new, I probably wouldn't be using my existing Google account essentially, rather try pretending to be a new person.*" — *U25 (interview)*

In a way, our design emulates what this user has to do to get novel recommendations. Use of OtherTube shows some promising results, with participants in the study both developing new interests and rediscovering prior interests after seeing content from strangers. As our extension embeds diverse recommendations in available content on YouTube, it could also act as a nudge for people to turn to find new interests. At the same time, our results also illuminate limits in types of diverse content that can be shown. For example, recommending offensive or excessively unfamiliar content could make people more hesitant to watch new kinds of content. Consequently, there is room for improvement in designing systems for exchanging recommendations with strangers, especially in assisting users to find people with similar interests or showing profiles not completely randomly, but based on a few similarities, to prevent adverse reactions from users. We could devise intelligent algorithms based on users' profiles to show content they may like. Additionally, we could computationally filter out spam content, like clickbait videos. Others may need more case-by-case input (e.g., content on unfamiliar topics). Overall, our work lays a foundation in this domain for exchanging recommendations between strangers, and further exploration is still needed.

6.6.2 RQ3: "Profile Work" for Self-Presentation to Strangers

Silfverberg et al. introduced the idea of "profile work," which is the amount of effort people devote to presenting themselves to others [417]. Similarly to their findings, we saw that users

put forth extra effort to moderate the content that would be shared, despite the fact that the content was algorithmically recommended videos—not exact watch histories—and even though their identities were not revealed. However, this inclination was limited to a few users, some of whose efforts even subsided over time based on the result shown in Figure 6.7(f). The implication of this observation is that for some users, such systems can provide features to reduce the effort required to moderate one's self-presentation. For example, in OtherTube, it could have been tiring to remove items one by one per collected session. To improve this experience, designers could provide easy filters for what users do and do not want to share.

With our design centered around strangers, participants also raised questions about the right amount of information to reveal in self-presentation. If users do not present enough information due to privacy concerns, it might be hard for others to differentiate them from bots or spammers. As a result, others may perceive such profiles to be misleading, not credible, or even not authentic. For example, U21's reflection on the avatars in the anonymous profiles made him wonder whether the putative strangers existed in real life.

> "*There were no real pictures. There wasn't any photographs of them on the beach or like a stock photo or something like that ... I'm not sure that I'm really engaging with someone who actually exists and not with algorithm.*" — *U21 (interview)*

Consequently, users may not engage with such profiles. In our interviews, we found some signs of disengagement due to disbelief about certain information presented in profiles (e.g., one user used the word "normy" in his About Me text and claimed to be in his 80s, despite reporting a significantly lower age in the sign-up survey), though this was only the case for one participant. Future work could delve into this tension between anonymity and authenticity.

Finally, while we did not receive complaints about privacy in sharing videos, some of the quotes (e.g., users mentioning that they watch regional/local videos) prompt such a discus-

sion. Here, shared videos could potentially be used to identify users if certain details more specific than those allowed by our anonymous profile format appear regularly (e.g., multiple videos about a certain location or neighborhood, watching a friend or relative's YouTube channel). Existing research already shows how sparse data—unique interests—can be used to de-anonymize users [318]. Further consideration is needed to resolve such issues in practice.

6.6.3 RQ4: Reflecting on Oneself and Others

As our results suggest, OtherTube helps users understand more about their interests and compare those interests with others' on YouTube. This understanding could positively impact users if they experience a feeling of belonging by seeing others with similar interests. Then again, the opposite can also happen; that is, if users are not able to find others with similar interests, they may instead feel disconnected, isolated, or atypical. A potential solution to address the opposite effect is to design the system to suggest recommendations from users with some common ground (e.g., someone from the same demographic group, someone with similar taste) and monitor their interactions (e.g., video clicks). However, pairings of extremely similar users should be avoided, as this may lead users back into their filter bubbles. Understanding the right amount of similarity and dissimilarity from which users can easily relate to others while still learning something is an interesting challenge. Being able to to search or filter users by demography and interests, as opposed to randomly suggesting profiles, may be another option for improving users' experiences with the approach of exchanging recommendations employed in OtherTube.

Our results also suggest potential in creating social connections between strangers with common interests. Future works would need to balance opposing desires (users who want to connect vs. those who do not) for connection in their designs. Perhaps designers can

provide limited options for connecting pairs of strangers. For example, the system could unlock messaging options between two users only if they have multiple shared interests and want to connect [115]. Alternatively, communication can be centered around content, such as by allowing users to recommend videos directly and respond to recommendations in a minimal fashion (e.g., using an emoji). Overall, future works could explore these directions for supporting reflection by exchanging recommendations with strangers.

6.6.4 Design Implications

On platforms such as YouTube, where recommendation algorithms contribute to shaping users' tastes and potentially trap users in filter bubbles for the sake of engagement, our approach of exchanging recommendations shows new ways to expand the scope of content discovery and improve reflection. For content discovery, existing systems already adopt multiple approaches to provide users alternate recommendations to explore. For example, apart from the site's personalized recommendations, *Trending on YouTube* aims to promote exploration into "videos that a wide range of viewers would find interesting" [168]. Meanwhile, services like Spotify suggest content through features like *Discover Weekly*, based on interests from others with preferences similar to users' own [158]. In contrast, our approach does not aggregate interests from any particular group; rather, it promotes exploration into other individual users' interests. Because we use other individuals' YouTube recommendations, our approach is fundamentally unlike YouTube's trending recommendations. As companies such as Alphabet, which owns YouTube, become keen to provide users with more opportunities to curate and expand their interests [227], our approach may provide such an opportunity. Aside from this, our approach also has the potential to fill a gap in the social functionality of systems like YouTube by creating connections between strangers based on similarity in interests. Recommender systems can eventually use such weak social ties to

improve recommendations.

6.6.5 Ethical Considerations

We took several steps in our design and our study to minimize potential harm. First, we collected only data that was needed for the tool and study. These included recommendations from users' YouTube homepages and their activity within the extension: the embedded content on the YouTube homepage, the extension's browser action page, and the extension's options page. Second, we set the extension's content rating to "mature" on the Chrome Web Store to remind users that they could see explicit content in others' YouTube recommendations. Third, we explicitly asked users to not share identifiable information in their profile fields to reduce the risk of an information leak (see Figure 6.2). In case users were uncomfortable sharing any demographic details with others, the extension defaulted to not sharing any information; users manually chose what information to share with others. Similarly, we also let users remove recommended videos if they felt uncomfortable sharing them with others.

6.6.6 Limitations

Our study is not without limitations. First, we had a limited number of participants from the US with limited demographic distribution. Therefore, we cannot account for scenarios in which users live in different nations and speak different languages. Additionally, while many users signed up for the study, only a select few ultimately participated. Therefore, some self-selection bias exists. These users could simply be those who are most open to exchanging recommendations. Furthermore, we recruited users who use the Chrome browser on a desktop or laptop computer (required for the OtherTube extension to work) and who use

YouTube regularly. These criteria likely excluded users who do not use YouTube regularly or who use it on a different platform (e.g., mobile phones). Second, our implementation of OtherTube also introduced some limits to the study. For example, some participants hesitated to watch content because preview clips were unavailable. One interview participant mentioned that they had trouble going back to a profile they had forgotten to pin. This issue could have impacted their overall interaction and engagement with OtherTube. Third, the first few participants from the first batch saw fewer profiles compared to other users; for comparison, the last person in the study had a maximum of 40 profiles to browse. Therefore, users' responses to the daily survey were affected by the limited sample. Future deployment with larger samples could potentially resolve these issues.

6.7 Conclusion

In this work, we investigated the exchange of social recommendations with strangers as a tool to promote content discovery and reflection on social media sites like YouTube. Our investigation revealed how users want to present themselves, as well as factors that affect their interaction and engagement with such a system. Our work has implications for future exploration into exchanging personalized recommendations with strangers.

Chapter 7

Discussion

In our online life, we use various affordances provided by each platform to accomplish different types of cognitive tasks. This work explores design of new affordances to parse the complex web of information in online problematic information space. I have designed four systems that showcase the possibility of leveraging dual process information processing theory to design dual process cognitive affordances. Below I start the discussion with the design implications of these cognitive affordances. Next, I discuss potential direction for applying these affordances, the impact it could bring about on the online information ecosystem. I end with a discussion on ethical implication of design affordances and affecting attitude.

7.1 Design Implications

In this thesis, I have laid out the design and deployment of four novel systems. Through this process, I have collected empirical feedback on considerations for improving such design. Below I summarize the design implications implication.

7.1.1 Affecting Stakeholder Power Dynamics

In TransparencyCue work, we found that different stakeholders had different view when it comes to what affordances should be designed. Whereas news consumers looked for certain

transparency from news organization, there were reluctance among journalists to present a few revealing the underlying journalistic processes. This indicate that certain affordance design can affect the power relationship between different stakeholders, not just between information producers and consumers. Take the power dynamics between a social media platform and its consumers. While platforms tend to dictate what content is shown to users, design affordances against problematic content can disrupt this power relationship. For example, design of OtherTube and NudgeCred could empower content consumers to choose what they want to engage with. In such a case, these platforms may not be open to support third-party intervention on their platform, since it may adversely affect their economic gain. Case in point, recently many platforms including Twitter and Reddit has made it harder to use their API by third parties, by making it significantly expensive [365]. Therefore, while designing affordances against problematic information, future researchers and design practitioner needs to consider how their design could impact the stakeholder power dynamics and how platforms may react to them.

7.1.2 Addressing Adversarial Manipulation

In our NudgeCred study, the simplicity of our algorithm for affordance design brought about questions of adversaries taking advantage of them. This problem is not exactly a new one. Take the case of search engines optimization feature (SEO). Since its introduction, various adversarial agents began taking advantage of SEO features through various means, such as adding more keywords in webpage headers [284]. There are several potential approaches to counter this problem when designing affordances. One of the approaches would be to not use vulnerable information as triggers for affordance design. For example, in NudgeCred design, users' social interaction as a trigger is vulnerable, since this interaction can easily be manipulated. Instead, if filtered social interaction based on users' historical information, we

could have accounted for these vulnerability to a certain degree. On the other hand, if we used a third-party reliable source of information (e.g., fact-checking by snope.com) to detect questionable content, it would be much less vulnerable Another solution to adversarial attack is to make the algorithms behind the affordances very opaque. However, such opaqueness counteracts the transparency principle of designing affordances. In general, whatever means designers employ in designing affordances to counter adversarial manipulation, they still need to continuously monitor for such attacks.

7.1.3 Designing for Long Run

In Othertube work, we saw reduction in user engagement with the affordances. In NudgeCred work, we also discussed the issue of novelty wearing off from the design cues. This is a challenge that impacts most HCI design, not just affordances. Over time, users may get used to a new affordance which could impact their perception of the design as well as interaction with it in different ways. Some may intuitively get used to interacting with an affordance, while other may ignore it like a background noise. Therefore, we need to examine this avenue of research to find out under what condition users would interact with an affordance instead of ignoring them in the long run. For example, a reminder about the functionality of an affordance once in a while could make users interact with it for longer period. A tutorial of how an affordance work in different scenario could also act like such as a reminder. We leave this line of investigation for future research.

7.1.4 Choosing Between Cognitive Processes

Since both types of affordances show some impact on attitude formation, one question that design practitioner may ask is how they should choose to design between designing auto-

matic and reflective affordances. Before responding to this question, let us consider how the cognition process may have worked out for the participants in our study. Now, generally scholars behind dual process theories suggests that these two processes are very distinctive in nature [128]. Keeping that in mind, we build dual process affordances with the assumption that each affordance will lead to use of respective cognitive process, at least for a majority of the users. However, in our study, we did not distinguish whether the effect, that is, the behavior formation came exactly from the use of a particular process. Rather the resulting behavior formation could be product of either of these processes or both of these process, depending on the individual. Take an example quote from NudgeCred where the participant mentioned that when they saw questionable nudge, they thought those news items were controversial. This example seems to indicate the creation of a mental shortcut. This shortcut could have happened just from automatic thinking or from a combination of both automatic and reflective thinking. For the case of combined processing, it could also have played out in different ways. For one, automatic and reflective processing could have happened simultaneously. Since automatic processing is faster, reflective processing would catch up later. Instead of these processes ocurring in parallel, the opposite could have also happened. Automatic process could have finished first, followed by rationalization by the reflective process. So far, I have assumed that both automatic and reflective mode would come to a coherent conclusion, by resolving any conflict if it happens. However, this might not be true in all instances for everyone. Rather, some may see dissonance between these process which may result in more deliberation between these processes later on, although that may not the case for the quote I started with. What this discussion has taught us is that although we may design affordances to elicit certain mental route use, we can not control it, rather the usage of mental routes could play out in different ways depending on individual tendencies.

Now going back to the question of choosing between the processes, my response to this

question is that it depends a lot on the audience. As most people tend to use automatic cognitive function due to the ease of use [128], if a design practitioner want to appeal to the largest number of users, designing automatic cognitive affordance would be the better choice here. However, if the audience appreciates use of reflective thinking, design practitioner should consider reflective affordance design. Overall, a general rule of thumb would be a hybrid approach where practitioners should design both automatic and reflective affordances. This way users can opt to choose either or both.

7.2 Cognitive Affordances and Attitude Formation: Moderators

In this dissertation, I have primarily focused on answering whether dual process cognitive affordances can impact attitude formation. Since we found the evidence of this relationship, the next evident question that arises is: *what are the moderators that impact this relationship?* In our study, we looked for some moderators, such as, user demography, behavioral factors, prior knowledge on information topics, or their values. The reason this list was not comprehensive is primarily due to limitations of standard human-subjects research. Going beyond that, in an information consumption setting, the list of factors, both intrinsic and extrinsic, could be very large as well as diverse. Here, I will discuss a few based on prior literature. In the ELM model, Petty et. al. suggests that motivation behind the information processing is a key factor in this relationship [355]. This motivation could be very different for the scenarios I have designed the systems. Some may go to social media to learn about a news topic or learn about their friends or purely for entertainment purpose. Such motivation could affect the attention mechanism they are willing to deploy during browsing, which in turn could impact the cognitive processing occurring. For example, in NudgeCred,

the motivation of individuals could have greatly impacted individuals' credibility judgment. Another factor that could affect use of a particular cognitive process is the availability of cognitive resources [379]. For example, users might be more alert and engage reflection more frequently at the start of the day compared to the end of the day when their mental load capacity has depleted. Another moderating factor is the expertise of the user on a particular topic of information [5]. Scholars argue that expertise on a topic lead to use of the analytical route on information consumption on that topic [5]. Thus, effect of NudgeCred could have varied by tweet topic. Various individual dispositions could also affect the effectiveness of an affordance [347]. For example, individuals who are more open to diverse information consumption could be affected more by OtherTube tool. Some other factors that could affecting the efficacy of cognitive affordances is the mood of the users [204]. For example, users in positive mood may engage different degree of reflection while using OtherTube. To sum, the list of factors moderating the relationship between affordances and attitude formation is very large. I leave such discussion for the future.

7.3 Applying Affordances in Other Scenarios

In this thesis, I focused on two types of problematic information, misinformation and filter bubble, for designing dual process cognitive affordances. Furthermore, we conducted the deployment of our designs on general US population. However, our affordance design approach has the potential for use to combat other types of problematic information and can be designed targeting any subset of the population. Below, I discuss this opportunity to apply cognitive affordances in these two scenarios.

7.3.1 Against Other Problematic Content

Our design can be adopted against other types of problematic content. To show this, I will start by showing how designs for misinformation be applied against other types of problematic content like hate speech, propaganda, and conspiracy. Lets first look at the algorithm behind NudgeCred. It uses some prior annotation of mainstream and non-mainstream source, and users social interaction like questioning. For the design against hate speech, we will first need a similar annotation of hate speech sources like the hate map from SPLC [1]. Next, we can replace looking for questions by looking for a confirmation of hate speech in user reaction using a lexicon or dictionary of frequently used words in response. However, this is still a simplication and does not consider that keywords appearing in hate speech lexicons is context dependent. For example, using n-word by a black person is not a hate speech while others may use it for that purpose [346]. However, discussion on such nuances is not within the scope of this work. After the detection, we can apply color coding similar to NudgeCred where we can swap the tooltips with information about hate speech. We can follow the same procedure for others as well. Now that we see that design process can be translated, the next question is, would it impact users attitude? This question is more complicated to answer. That is because, while interventions on misinformation and filter bubble have shown promising result, it may not work that well for certain groups. For example, user groups like conspiracists tend to have strong bias in their conspiracy attitude and most prior intervention show lack of any effect [339]. Therefore, before making any claim here, we need empirical evidence, and leave it as a future work.

[1] https://www.splcenter.org/hate-map

7.3.2 For Certain User Groups

Compared to extending the design affordances against other types of problematic content, designing affordances for certain user group is tricky. Take the example of design affordances for user with neuro-divergence. Since their cognitive process differ from neurotypical users, they may not recognize the design cues we built so far. Or even if they do recognize the cues, it could have different impact on their attitude. Furthermore, they may have other issue. For example, reflection is already challening for neurotypical people and could be harder for them [421]. Or, new design affordances can easily trigger sensory overload for them [275]. Therefore, designing cognitive affordances for these users require careful considerations and we leave it up to future designers.

7.4 Ethical Considerations for Cognitive Affordances

Design of affordances against problematic information can bring out some of the concerns of behavior modification. As Selinger critiques, *"Would someone who values their freedom to choose be okay with the idea that their behavior is being modified in ways they are not aware of?"* [409] Inspired by Hansen's work on transparency in nudge design [186], I argue that cognitive affordances can be distinguished into two types: transparent and non-transparent. Here, non-transparent affordances would be manipulative by definition, while transparent affordances could rather be empowering the users instead of manipulating. This empowerment comes from design practitioner disclosing the purpose and mechanism behind an affordance design. Such disclosure removes the said ethical dilemma posed by the question from Selinger. In the four studies conducted, I have followed this principle of transparency (like the tooltip showing the rationale for intervention in NudgeCred). In addition, I would also argue that reflective affordances by design allows analytical thinking among its audience,

and thus not manipulative in nature. Therefore, designing reflective cognitive affordances would be more desirable from an ethical perspective. In addition to thinking of transparency in affordance design, designers also need to think about ethics from the perspective of the type of value is embedded during design. This idea has been explored in many HCI concepts, such as, values-sensitive design, reflective design, and values in design [144, 235, 411]. These research suggest that the value of the designers affect how technologies are imagined, how underlying mechanism behind the design constructed, and corresponding ethical consideration. For example, the type of data being used when devising a cognitive affordance requires corresponding ethical consideration, such as, asking if the data being used is private or not. However, discussion on these considerations relate to each instances of design, instead of the abstract level of cognitive affordance design. Therefore, I leave this discussion up to the practitioners during their design process.

Chapter 8

Conclusion

Online news platforms are increasingly becoming complicated as new sources try to break in and the platforms try to improve engagement. When the HCI community is debating whether to convert the research direction in these domains from a "fix it" to a "burn it down" [453], this thesis still takes the stance of the former by augmenting existing systems with improvements meant to help users. Using dual process theories of mind, I propose designs that can help users tackle the issue of misinformation and filter bubbles on their online news feeds. In NudgeCred, I introduce automatic affordances built on heuristics to assist users in differentiating news content reliability on social media using their automatic mode of cognition. My evaluation shows the promise of this approach working in real-world scenarios. In TransparencyCue, I present a set of design opportunity for automatic affordances to promote stakeholder values like transparency on news platforms through an interview study of news consumers and journalists. In NewsComp, I present a reflective affordance to promote news reading from diverse perspectives. In OtherTube, I used reflective affordance design by swapping recommendation with others to help users become more aware of the interest bubble imposed by algorithmic recommendation bubble and engage in content discovery new interests. Overall, this set of systems may provide much-needed assistance for the users to tackle the issue of misinformation and filter bubbles.

8.1 Future Work

Each of the works presented in this thesis opens up an avenue for a new line of research. NudgeCred shows promise for using alternate heuristics to design automatic affordances. Such systems could be designed based on what different users need, instead of a universal one. Additionally, looking into the effects of such design on users with extreme beliefs in contentious topics could help us understand the limitations of such designs. In TransparencyCue work, I presented a set of design cues where a logical next step would be to conduct a controlled evaluation of how the audience perceives such design cues. Especially, tackling the conflicts in such design could open up direction for compromise in designing such systems. In NewsComp, the performance of the users in annotation creates several challenges for effective adoption of this affordance. If users make more mistakes, it can result in wrong perception. Therefore, this design has to be modified to improve performance. If exchanging recommendations in YouTube through OtherTube helps users realize filter bubbles, it may also present new challenges. For example, at scale, it could be hard to match users against each other. Therefore, sophisticated methods need to be designed to provide improved matching. Future research may also delve into the question of whether these designs provide meaningful guides to the users in the long, which is often an issue with many solutions. Finally, there is also potential to apply the same theoretical grounded approach to designs solutions for other issues in problematic information domain. Overall, these are several avenues for future work beyond this thesis.

Bibliography

[1] Bradley J Adame. Training in the mitigation of anchoring bias: A test of the consider-the-opposite strategy. *Learning and Motivation*, 53:36–48, 2016.

[2] Alexander T Adams, Jean Costa, Malte F Jung, and Tanzeem Choudhury. Mindless computing: designing technologies to subtly influence behavior. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 719–730, 2015.

[3] Gediminas Adomavicius, Jesse Bockstedt, Shawn Curley, and Jingjing Zhang. De-biasing user preference ratings in recommender systems. In *RecSys 2014 Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS 2014)*, pages 2–9. Citeseer, 2014.

[4] Tanja Aitamurto, Mike Ananny, Chris W Anderson, Larry Birnbaum, Nicholas Diakopoulos, Matilda Hanson, Jessica Hullman, and Nick Ritchie. Hci for accurate, impartial and transparent journalism: Challenges and solutions. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, page W22. ACM, 2019.

[5] Joseph W Alba and J Wesley Hutchinson. Dimensions of consumer expertise. *Journal of consumer research*, 13(4):411–454, 1987.

[6] Linda Aldoory and Mark A Van Dyke. The roles of perceived "shared" involvement and information overload in understanding how audiences make meaning of news about bioterrorism. *Journalism & Mass Communication Quarterly*, 83(2):346–361, 2006.

[7] Shahmir H Ali, Joshua Foreman, Ariadna Capasso, Abbey M Jones, Yesim Tozan, and Ralph J DiClemente. Social media as a recruitment platform for a nationwide online survey of covid-19 knowledge, beliefs, and practices in the united states: methodology and feasibility analysis. *BMC medical research methodology*, 20:1–11, 2020.

[8] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, 2017.

[9] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2):2053168019848554, 2019.

[10] David S Allen. The trouble with transparency: The challenge of doing journalism ethics in a surveillance society. *Journalism Studies*, 9(3):323–340, 2008.

[11] Jennifer Allen, Antonio A Arechar, Gordon Pennycook, and David G Rand. Scaling up fact-checking using the wisdom of crowds. *Science advances*, 7(36):eabf4393, 2021.

[12] Jennifer Allen, Cameron Martel, and David G Rand. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in twitter's birdwatch crowdsourced fact-checking program. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[13] Mike Allen. Meta-analysis comparing the persuasiveness of one-sided and two-sided messages. *Western Journal of Speech Communication*, 55(4):390–404, 1991.

[14] AllSides. Allsides | balanced news via media bias ratings for an unbiased news perspective. https://www.allsides.com/unbiased-balanced-news, 2012. (Accessed on 04/08/2021).

[15] Bobby Allyn. 4 takeaways from facebook whistleblower frances haugen's testimony : Npr. https://www.npr.org/2021/10/05/1043377310/

facebook-whistleblower-frances-haugen-congress, 10 2021. (Accessed on 11/13/2021).

[16] Paul André, MC Schraefel, Alan Dix, and Ryen W White. Expressing well-being online: towards self-reflection and social awareness. In *Proceedings of the 2011 iConference*, pages 114–121. ACM, 2011.

[17] Mark Andrejevic. *Infoglut: How too much information is changing the way we think and know*. Routledge, 2013.

[18] AP. Associated press news values and principles. https://www.ap.org/about/news-values-and-principles/, 2018.

[19] Amelia Arsenault and Manuel Castells. Conquering the minds, conquering iraq: The social production of misinformation in the united states–a case study. *Information, Communication & Society*, 9(3):284–307, 2006.

[20] W Brian Arthur. Inductive reasoning and bounded rationality. *The American economic review*, 84(2):406–411, 1994.

[21] Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. Stance detection with bidirectional conditional encoding. *arXiv preprint arXiv:1606.05464*, 2016.

[22] Axios. Axios - editorial ethics policy. https://www.axios.com/about/ethics, 2016. (Accessed on 03/25/2021).

[23] Mevan Babakar. Crowdsourced factchecking, May 2018.

[24] Mevan Babakar. Crowdsourced factchecking: There is a role for crowdsourcing in factchecking but (so far) it's not factchecking. https://medium.com/@meandvan/crowdsourced-factchecking-4c5168ea5ac3, 2018.

[25] Khaled Bachour, Jon Bird, Vaiva Kalnikaite, Yvonne Rogers, Nicolas Villar, and Stefan Kreitmayer. Fast and frugal shopping challenge. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 1459–1460. ACM, 2012.

[26] Mitja D Back, Juliane M Stopfer, Simine Vazire, Sam Gaddis, Stefan C Schmukle, Boris Egloff, and Samuel D Gosling. Facebook profiles reflect actual personality, not self-idealization. *Psychological science*, 21(3):372–374, 2010.

[27] Adam Badawy, Emilio Ferrara, and Kristina Lerman. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 258–265. IEEE, 2018.

[28] Jae-eul Bae, Youn-kyung Lim, Jin-bae Bang, and Myung-suk Kim. Ripening room: Designing social media for self-reflection in self-expression. In *Proceedings of the 2014 Conference on Designing Interactive Systems*, DIS '14, page 103–112, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329026. doi: 10.1145/2598510.2598567. URL https://doi.org/10.1145/2598510.2598567.

[29] Joseph O Baker and Amy E Edmonds. Immigration, presidential politics, and partisan polarization among the american public, 1992–2018. *Sociological Spectrum*, 41(4):287–303, 2021.

[30] Rebecca Balebako, Pedro G Leon, Hazim Almuhimedi, Patrick Gage Kelley, Jonathan Mugan, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. Nudging users towards privacy on mobile devices. In *Proc. CHI 2011 Workshop on Persuasion, Nudge, Influence and Coercion*, pages 193–201. Citeseer, 2011.

[31] Elizabeth Bales and William Griswold. Interpersonal informatics: Making social influence visible. In *CHI '11 Extended Abstracts on Human Factors in Computing*

Systems, CHI EA '11, page 2227–2232, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450302685. doi: 10.1145/1979742.1979924. URL https://doi.org/10.1145/1979742.1979924.

[32] Meital Balmas. When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. *Communication research*, 41(3):430–454, 2014.

[33] Jack Bandy. Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the acm on human-computer interaction*, 5(CSCW1): 1–34, 2021.

[34] Scott Bateman, Rosta Farzan, Peter Brusilovsky, and Gord McCalla. Oats: The open annotation and tagging system. *Proceedings of I2LOR*, 2006.

[35] Eric PS Baumer. Reflective informatics: conceptual dimensions for designing technologies of reflection. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 585–594. ACM, 2015.

[36] Joshua Becker, Ethan Porter, and Damon Centola. The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences*, 116(22):10717–10722, 2019.

[37] W Lance Bennett. Perception and cognition. In *The handbook of political behavior*, pages 69–193. Springer, 1981.

[38] Sabine Bergler. Conveying attitude with reported speech. In *Computing attitude and affect in text: Theory and applications*, pages 11–22. Springer, 2006.

[39] Michael S Bernstein, Mark S Ackerman, Ed H Chi, and Robert C Miller. The trouble with social computing systems research. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 389–398. 2011.

[40] Alessandro Bessi, Fabiana Zollo, Michela Del Vicario, Michelangelo Puliga, Antonio Scala, Guido Caldarelli, Brian Uzzi, and Walter Quattrociocchi. Users polarization on facebook and youtube. *PloS one*, 11(8):e0159641, 2016.

[41] Md Momen Bhuiyan, Kexin Zhang, Kelsey Vick, Michael A Horning, and Tanushree Mitra. Feedreflect: A tool for nudging users to assess news credibility on twitter. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 205–208, 2018.

[42] Md Momen Bhuiyan, Amy X Zhang, Connie Moon Sehat, and Tanushree Mitra. Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria. *arXiv preprint arXiv:2008.09533*, 2020.

[43] Md Momen Bhuiyan, Michael Horning, Sang Won Lee, and Tanushree Mitra. Nudgecred: supporting news credibility assessment on social media through nudges. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–30, 2021.

[44] Md Momen Bhuiyan, Hayden Whitley, Michael Horning, Sang Won Lee, and Tanushree Mitra. Designing transparency cues in online news platforms to promote trust: Journalists'& consumers' perspectives. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–31, 2021.

[45] Md Momen Bhuiyan, Carlos Augusto Bautista Isaza, Tanushree Mitra, and Sang Won Lee. Othertube: Facilitating content discovery and reflection by exchanging youtube recommendations with strangers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022.

[46] Md Momen Bhuiyan, Sang Won Lee, Nitesh Goyal, and Tanushree Mitra. Newscomp: Facilitating diverse news reading through comparative annotation. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2023.

[47] Ricardo Bilton. How one washington post reporter uses pen and paper to make his tracking of trump get noticed » nieman journalism lab. https://www.niemanlab.org/2016/09/how-one-washington-post-reporter-uses-pen-and-paper-to-make-his-tracking-of-trump- 2016. (Accessed on 09/29/2019).

[48] Bernd Blöbaum. Trust and journalism in a digital environment. 2014.

[49] Oversight Board. Oversight board | independent judgment. transparency. legitimacy. https://oversightboard.com/, 2020. (Accessed on 04/05/2021).

[50] Taylor C Boas, Dino P Christenson, and David M Glick. Recruiting large online samples in the united states and india: Facebook, mechanical turk, and qualtrics. *Political Science Research and Methods*, 8(2):232–250, 2020.

[51] Balázs Bodó, Natali Helberger, Sarah Eskens, and Judith Möller. Interested in diversity: The role of user attitudes, algorithmic feedback loops, and policy in news personalization. *Digital journalism*, 7(2):206–229, 2019.

[52] Ben Bolker. lme4 package | r documentation. https://www.rdocumentation.org/packages/lme4/versions/1.1-18-1, 2018.

[53] Toby Bolsen and James N Druckman. Counteracting the politicization of science. *J. Commun.*, 65(5):745–769, 2015.

[54] Porismita Borah and Xizhu Xiao. The importance of 'likes': The interplay of message framing, source, and social endorsement on credibility perceptions of health information on facebook. *Journal of health communication*, 23(4):399–411, 2018.

[55] Paul Bossauer, Thomas Neifer, Gunnar Stevens, and Christina Pakusch. Trust versus

privacy: Using connected car data in peer-to-peer carsharing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[56] Mark Boukes, Natalie P Jones, and Rens Vliegenthart. Newsworthiness and story prominence: How the presence of news factors relates to upfront position and length of news stories. *Journalism*, page 1464884919899313, 2020.

[57] Engin Bozdag. Bias in algorithmic filtering and personalization. *Ethics and information technology*, 15:209–227, 2013.

[58] Ivar Bråten and Helge I Strømsø. Measuring strategic processing when students read multiple texts. *Metacognition and Learning*, 6(2):111–130, 2011.

[59] Marilynn B Brewer and William D Crano. Research design and issues of validity. *Handbook of research methods in social and personality psychology*, pages 3–16, 2000.

[60] Campbell Brown. Introducing facebook news - about facebook. https://about.fb.com/news/2019/10/introducing-facebook-news/, 2019. (Accessed on 09/11/2020).

[61] Lauren Valentino Bryant. The youtube algorithm and the alt-right filter bubble. *Open Information Science*, 4(1):85–90, 2020.

[62] E Brynjolfsson and MV Alstyne. Electronic communities: Global village or cyber-balkans? In *International Conference on Information Systems, December*, pages 16–18, 1996.

[63] David V Budescu and Eva Chen. Identifying expertise to extract the wisdom of crowds. *Management Science*, 61(2):267–280, 2015.

[64] Guido Buelow. Automation plus expert journalism: How full fact is fighting misinformation - about facebook. https://about.fb.com/news/2019/06/inside-feed-full-fact-interview/, 2019. (Accessed on 09/11/2020).

[65] Laura Burbach, Patrick Halbach, Martina Ziefle, and André CaleroÂ Valdez. Bubble Trouble: Strategies Against Filter Bubbles in Online Social Networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11582 LNCS:441–456, 2019. ISSN 16113349. doi: 10.1007/978-3-030-22219-2_33.

[66] Laila Burla, Birte Knierim, Jurgen Barth, Katharina Liewald, Margreet Duetz, and Thomas Abel. From text to codings: intercoder reliability assessment in qualitative content analysis. *Nursing research*, 57(2):113–117, 2008.

[67] Lynette Sheridan Burns and Benjamin J Matthews. *Understanding journalism*. Sage, 2018.

[68] Julie G Bush, Hollyn M Johnson, and Colleen M Seifert. The implications of corrections: Then why did you mention it. 1994.

[69] Joseph N Cappella and Kathleen Hall Jamieson. *Spiral of cynicism: The press and the public good*. Oxford University Press on Demand, 1997.

[70] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 503. ACM, 2019.

[71] D Jasun Carr, Matthew Barnidge, Byung Gu Lee, and Stephanie Jean Tsang. Cynics and skeptics: Evaluating the credibility of mainstream and citizen journalism. *J. Mass Commun. Q.*, 91(3):452–470, 2014.

[72] John M Carroll. *Making use: scenario-based design of human-computer interactions*. MIT press, 2000.

[73] Liz Carver and Murray Turoff. Human-computer interaction: the human and computer as a team in emergency management information systems. *Communications of the ACM*, 50(3):33–38, 2007.

[74] Pablo Castells, Neil J. Hurley, and Saul Vargas. *Novelty and Diversity in Recommender Systems*, pages 881–918. Springer US, Boston, MA, 2015. ISBN 978-1-4899-7637-6. doi: 10.1007/978-1-4899-7637-6_26. URL https://doi.org/10.1007/978-1-4899-7637-6_26.

[75] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proc. WWW*, pages 675–684. ACM, 2011.

[76] Mike Caulfield. Web literacy for student fact-checkers. 2017.

[77] Pew Reseach Center. Political polarization in the american public. *Ann Rev Polit Sci*, 2014.

[78] Pew Research Center. Further decline in credibility ratings for most news organizations, 2012.

[79] Kalyani Chadha and Michael Koliska. Newsrooms and transparency in the digital age. *Journalism Practice*, 9(2):215–229, 2015.

[80] Shelly Chaiken. The heuristic model of persuasion. In *Social influence: the ontario symposium*, volume 5, pages 3–39. Hillsdale, NJ: Lawrence Erlbaum, 1987.

[81] Kathy Charmaz. *Constructing grounded theory: A practical guide through qualitative analysis*. sage, 2006.

[82] Xinran Chen, Sei-Ching Joanna Sin, Yin-Leng Theng, and Chei Sian Lee. Deterring the spread of misinformation on social network sites: A social cognitive theory-guided

intervention. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4, 2015.

[83] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, pages 71–80. IEEE, 2012. doi: 10.1109/SocialCom-PASSAT.2012.55.

[84] Sidharth Chhabra and Paul Resnick. Does clustered presentation lead readers to diverse selections? In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 1689–1694. 2013.

[85] Michelene Chi. Conceptual change within and across ontological categories: Examples from learning and discovery in science. 1992.

[86] Sujin Choi. The two-step flow of communication in twitter-based public forums. *Social science computer review*, 33(6):696–711, 2015.

[87] Cision. Top 10 u.s. daily newspapers. https://www. cision.com/us/2017/09/top-10-u-s-daily-newspapers-2/, 2017.

[88] Russell D Clark III and Anne Maass. The role of social categorization and perceived source credibility in minority influence. *European Journal of Social Psychology*, 18(5): 381–394, 1988.

[89] Mark Coddington. Building frames link by link: The linking practices of blogs and news sites. *International Journal of Communication*, 6:20, 2012.

[90] Sheldon Ed Cohen and SI Syme. *Social support and health.* Academic Press, 1985.

[91] Devin Coldewey. Reddit overhauls upvote algorithm to thwart cheaters and show the site's true scale | techcrunch. https://techcrunch.com/2016/12/06/

reddit-overhauls-upvote-algorithm-to-thwart-cheaters-and-show-the-sites-true-scale?guccounter=1, 2016. (Accessed on 01/03/2021).

[92] Sunny Consolvo, Katherine Everitt, Ian Smith, and James A Landay. Design requirements for technologies that encourage physical activity. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 457–466, New York, NY, USA, 2006. Association for Computing Machinery.

[93] Sunny Consolvo, David W. McDonald, and James A. Landay. *Theory-Driven Design Strategies for Technologies That Support Behavior Change in Everyday Life*, page 405–414. Association for Computing Machinery, New York, NY, USA, 2009. ISBN 9781605582467. URL https://doi.org/10.1145/1518701.1518766.

[94] John Cook, Stephan Lewandowsky, and Ullrich KH Ecker. Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PloS one*, 12(5):e0175799, 2017.

[95] Eric Corbett and Christopher A Le Dantec. Going the distance: Trust work for citizen participation. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13, 2018.

[96] Denzil Correa, Leandro Araújo Silva, Mainack Mondal, Fabrício Benevenuto, and Krishna P Gummadi. The many shades of anonymity: Characterizing anonymous social media content. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[97] Cynthia L Corritore, Beverly Kracher, and Susan Wiedenbeck. On-line trust: concepts, evolving themes, a model. *International journal of human-computer studies*, 58(6): 737–758, 2003.

[98] Dan Cosley, Shyong K Lam, Istvan Albert, Joseph A Konstan, and John Riedl. Is seeing believing? how recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 585–592, 2003.

[99] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 191–198, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340359. doi: 10.1145/2959100.2959190. URL https://doi.org/10.1145/2959100.2959190.

[100] Stephanie Craft and Kyle Heim. Transparency in journalism: Meanings, merits, and risks. *The handbook of mass media ethics*, pages 217–228, 2009.

[101] Brent Cunningham. Skin deep: when'transparency'is a smoke screen. *Columbia Journalism Review*, 45(2):9–11, 2006.

[102] Alex Curry and Natalie Jomini Stroud. Trust in online news. *Center for Media Engagement, University of Texas at Austin*, 2017.

[103] Alexander L Curry and Natalie Jomini Stroud. The effects of journalistic transparency on credibility assessments and engagement intentions. *Journalism*, page 1464884919850387, 2019.

[104] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, and Taylor Van Vleet. The YouTube video recommendation system. *RecSys'10 - Proceedings of the 4th ACM Conference on Recommender Systems*, pages 293–296, 2010. doi: 10.1145/1864708.1864770.

[105] Munmun De Choudhury, Michael Gamon, Aaron Hoff, and Asta Roseway. "moon

phrases": A social media faciliated tool for emotional reflection and wellness. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, pages 41–44. IEEE, European Alliance for Innovation, 2013.

[106] Claes H De Vreese. News framing: Theory and typology. *Information design journal & document design*, 13(1), 2005.

[107] James W Dearing, Everett M Rogers, and Everett Rogers. *Agenda-setting*, volume 6. Sage, 1996.

[108] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, 2016.

[109] Michela Del Vicario, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. Modeling confirmation bias and polarization. *Scientific reports*, 7(1): 1–9, 2017.

[110] Stefano DellaVigna and Ethan Kaplan. The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234, 2007.

[111] Peter M DeMarzo, Dimitri Vayanos, and Jeffrey Zwiebel. Persuasion bias, social influence, and unidimensional opinions. *The Quarterly journal of economics*, 118(3): 909–968, 2003.

[112] Mark Deuze. The web and its journalisms: considering the consequences of different types of newsmedia online. *New media & society*, 5(2):203–230, 2003.

[113] Nicholas Diakopoulos. Algorithmic accountability: Journalistic investigation of computational power structures. *Digital journalism*, 3(3):398–415, 2015.

[114] Nicholas Diakopoulos, Munmun De Choudhury, and Mor Naaman. Finding and assessing social media information sources in the context of journalism. In *Proc. CHI*, pages 2451–2460. ACM, 2012.

[115] Abraham Doris-Down, Husayn Versee, and Eric Gilbert. Political blend: an application designed to bring people together based on political differences. In *Proceedings of the 6th International Conference on Communities and Technologies*, pages 120–130, New York, NY, USA, 2013. Association for Computing Machinery.

[116] Tim Draws, David La Barbera, Michael Soprano, Kevin Roitero, Davide Ceolin, Alessandro Checco, and Stefano Mizzaro. The effects of crowd worker biases in fact-checking tasks. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2114–2124, 2022.

[117] James N Druckman and Michael Parkin. The impact of media bias: How editorial slant affects voters. *The Journal of Politics*, 67(4):1030–1049, 2005.

[118] The Economist. About us | the economist. https://www.economist.com/help/about-us, 2010. (Accessed on 10/11/2020).

[119] Robert M Entman. Framing: Toward clarification of a fractured paradigm. 1993.

[120] Robert M Entman. Framing bias: Media in the distribution of power. *Journal of communication*, 57(1):163–173, 2007.

[121] Robert M Entman, Carole Bell, Cary Frith, and Barbara Miller. Constraining the politics of self-interest: How the media helped to sell the bush tax cuts. In *Research Conference of the National Communication Association, Boston, MA*, 2005.

[122] Martin J Eppler and Jeanne Mengis. The concept of information overload-a review of

literature from organization science, accounting, marketing, mis, and related disciplines (2004). *Kommunikationsmanagement im Wandel*, pages 271–305, 2008.

[123] Daniel A Epstein, An Ping, James Fogarty, and Sean A Munson. A lived informatics model of personal informatics. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 731–742, New York, NY, USA, 2015. Association for Computing Machinery.

[124] Thomas Erickson and Wendy A Kellogg. Social translucence: an approach to designing systems that support social processes. *ACM transactions on computer-human interaction (TOCHI)*, 7(1):59–83, 2000.

[125] Marcus Errico, J April, A Asch, L Khalfani, M Smith, and X Ybarra. The evolution of the summary news lead. *Media History Monographs*, 1(1), 1997.

[126] Motahhare Eslami, Amirhossein Aleyasen, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. Feedvis: A path for exploring news feed curation algorithms. In *Proceedings of the 18th acm conference companion on computer supported cooperative work & social computing*, pages 65–68, New York, NY, USA, 2015. Association for Computing Machinery.

[127] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. " i always assumed that i wasn't really that close to [her]" reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 153–162, New York, NY, USA, 2015. Association for Computing Machinery.

[128] Jonathan St BT Evans. In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences*, 7(10):454–459, 2003.

[129] Jonathan St BT Evans. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, 59:255–278, 2008.

[130] William P Eveland Jr and Sharon Dunwoody. User control and structural isomorphism or disorientation and cognitive load? learning from the web versus print. *Communication research*, 28(1):48–78, 2001.

[131] Gunther Eysenbach. *Credibility of health information and digital media: New perspectives and implications for youth*. MacArthur Foundation Digital Media and Learning Initiative, 2008.

[132] Facebook. How facebook has prepared for the 2019 uk general election - about facebook. https://about.fb.com/news/2019/11/how-facebook-is-prepared-for-the-2019-uk-general-election/, 2019. (Accessed on 09/11/2020).

[133] Don Fallis. What is disinformation? *Library trends*, 63(3):401–426, 2015.

[134] Haiyan Fan and Marshall Scott Poole. What is personalization? perspectives on the design and implementation of personalization in information systems. *Journal of Organizational Computing and Electronic Commerce*, 16(3-4):179–202, 2006.

[135] Leon Festinger. A theory of social comparison processes. *Human relations*, 7(2):117–140, 1954.

[136] Clayton Feustel, Shyamak Aggarwal, Bongshin Lee, and Lauren Wilcox. People like me: Designing for reflection on aggregate cohort data in personal informatics systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–21, 2018.

[137] Morris P Fiorina, Samuel J Abrams, et al. Political polarization in the american public. *ANNUAL REVIEW OF POLITICAL SCIENCE-PALO ALTO-*, 11:563, 2008.

[138] Luciano Floridi. Brave. net. world: the internet as a disinformation superhighway? *The Electronic Library*, 14(6):509–514, 1996.

[139] B. J. Fogg. Prominence-interpretation theory: Explaining how people assess credibility online. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '03, pages 722–723, New York, NY, USA, 2003. ACM. ISBN 1-58113-637-4. doi: 10.1145/765891.765951. URL http://doi.acm.org/10.1145/765891.765951.

[140] Brian J Fogg. Prominence-interpretation theory: Explaining how people assess credibility online. In *CHI'03 extended abstracts on human factors in computing systems*, pages 722–723. Citeseer, 2003.

[141] O*NET OnLine National Center for O*NET Development. 27-3022.00 - reporters and correspondents. https://www.onetonline.org/link/details/27-3022.00, 2020. (Accessed on 10/13/2020).

[142] Joseph P Forgas. Affective influences on attitudes and judgments. 2003.

[143] Jonathan Fox. The uncertain relationship between transparency and accountability. *Development in practice*, 17(4-5):663–671, 2007.

[144] Batya Friedman. Value-sensitive design. *interactions*, 3(6):16–23, 1996.

[145] Batya Friedman, Peter H Kahn, and Alan Borning. Value sensitive design and information systems. *The handbook of information and computer ethics*, pages 69–101, 2008.

[146] Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. Social clicks: What and who gets read on twitter? In *Proceedings of the 2016 ACM*

SIGMETRICS international conference on measurement and modeling of computer science, pages 179–192, 2016.

[147] Johan Galtung and Mari Holmboe Ruge. The structure of foreign news: The presentation of the congo, cuba and cyprus crises in four norwegian newspapers. *Journal of peace research*, 2(1):64–90, 1965.

[148] Oscar H Gandy, Katharina Kopp, Tanya Hands, Karen Frazer, and David Phillips. Race and risk: Factors affecting the framing of stories about inequality, discrimination, and just plain bad luck. *The Public Opinion Quarterly*, 61(1):158–182, 1997.

[149] Feng Gao. Design for reflection on health behavior change. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 379–382, New York, NY, USA, 2012. Association for Computing Machinery.

[150] Andrew Garbett, David Chatting, Gerard Wilkinson, Clement Lee, and Ahmed Kharrufa. *ThinkActive: Designing for Pseudonymous Activity Tracking in the Classroom*, page 1–13. Association for Computing Machinery, New York, NY, USA, 2018. ISBN 9781450356206. URL https://doi.org/10.1145/3173574.3173581.

[151] R Kelly Garrett. Echo chambers online?: Politically motivated selective exposure among internet news users. *JCMC*, 14(2):265–285, 2009.

[152] R Kelly Garrett and Brian E Weeks. The promise and peril of real-time corrections to political misperceptions. In *Proc. CSCW*, pages 1047–1058. ACM, 2013.

[153] Susan E Gathercole. *Short-term and working memory*, volume 9. Psychology Press, 2001.

[154] William W Gaver. Technology affordances. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 79–84, 1991.

[155] William W Gaver. Situating action ii: Affordances for interaction: The social is material for design. *Ecological psychology*, 8(2):111–129, 1996.

[156] Christine Geeng, Savanna Yee, and Franziska Roesner. Fake news on facebook and twitter: Investigating how people (don't) investigate. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[157] Brittany Gentile, Jean M Twenge, Elise C Freeman, and W Keith Campbell. The effect of social networking websites on positive self-views: An experimental investigation. *Computers in human behavior*, 28(5):1929–1933, 2012.

[158] Dave Gershgorn. How spotify's algorithm knows exactly what you want to listen to | by dave gershgorn | onezero. https://onezero.medium.com/how-spotifys-algorithm-knows-exactly-what-you-want-to-listen-to-4b6991462c5c, 2019. (Accessed on 09/09/2021).

[159] Guiseppe Getto, Liza Potts, Michael J Salvo, and Kathie Gossett. Teaching ux: Designing programs to train the next generation of ux experts. In *Proceedings of the 31st ACM international conference on Design of communication*, pages 65–70, 2013.

[160] James J Gibson. 04-JJ Gibson-Ch8-Affordances. *Chapter Eight The Theory of Affordances*, pages 127–136, 1986. ISSN 10725520.

[161] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology press, 2014.

[162] Eytan Gilboa and Uri Paz. Errors and corrections. *The International Encyclopedia of Journalism Studies*, pages 1–5, 2019.

[163] Nabeel Gillani, Ann Yuan, Martin Saveski, Soroush Vosoughi, and Deb Roy. Me, my echo chamber, and i: Introspection on social media polarization. *The Web Conference*

2018 - Proceedings of the World Wide Web Conference, WWW 2018, pages 823–831, 2018. doi: 10.1145/3178876.3186130.

[164] Tarleton Gillespie. The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society*, 167(2014):167, 2014.

[165] Jennifer Golbeck and Kenneth R Fleischmann. Trust in social q&a: the impact of text and photo cues of expertise. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–10, 2010.

[166] Emily Goligoski and Stephanie Ho. 25 ways community members can support your newsroom - global investigative journalism network. https://gijn.org/2019/02/18/25-ways-community-members-can-support-your-newsroom/, 2019. (Accessed on 10/14/2020).

[167] Google. How google news stories are selected - google news help. https://support.google.com/googlenews/answer/9005749?hl=en, . (Accessed on 11/12/2021).

[168] Google. Trending on youtube - youtube help. https://support.google.com/youtube/answer/7239739?hl=en, . (Accessed on 09/09/2021).

[169] Jeffrey Gottfried, Mason Walker, and Amy Mitchell. Americans are largely skeptical of the news media, but say there is room for confidence to improve | pew research center. https://www.journalism.org/2020/08/31/americans-are-largely-skeptical-of-the-news-media-but-say-there-is-room-for-confid 2020. (Accessed on 03/25/2021).

[170] Sten Govaerts, Katrien Verbert, Erik Duval, and Abelardo Pardo. The student activity meter for awareness and self-reflection. In *CHI'12 Extended Abstracts on Human*

Factors in Computing Systems, pages 869–884. Association for Computing Machinery, New York, NY, USA, 2012.

[171] Anthony M. Grant, John Franklin, and Peter Langford. The self-reflection and insight scale: A new measure of private self-consciousness. *Social Behavior and Personality*, 30(8):821–836, 2002. ISSN 03012212. doi: 10.2224/sbp.2002.30.8.821.

[172] Stephan Greene and Philip Resnik. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*, pages 503–511. Association for Computational Linguistics, 2009.

[173] Catherine Grevet and Eric Gilbert. Piggyback prototyping: Using existing, large-scale social computing systems to prototype new ones. In *Proc. CHI*, pages 4047–4056. ACM, 2015.

[174] Andrea Grimes, Brian M Landry, and Rebecca E Grinter. Characteristics of shared health reflections in a local community. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 435–444, New York, NY, USA, 2010. Association for Computing Machinery.

[175] Harmen P Groenhart and Jo LH Bardoel. Conceiving the transparency of journalism: Moving towards a new media accountability currency. *Studies in Communication Sciences*, 12(1):6–11, 2012.

[176] Jason Grotto. How do we verify anonymous sources? - propublica. https://www.propublica.org/article/ask-propublica-illinois-vetting-anonymous-sources, 2018. (Accessed on 10/14/2019).

[177] Andrew Guess, Kevin Munger, Jonathan Nagler, and Joshua Tucker. How accurate are survey responses on social media and politics? *Political Communication*, 36(2): 241–258, 2019.

[178] Andrew M Guess, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. A digital media literacy intervention increases discernment between mainstream and false news in the united states and india. *Proceedings of the National Academy of Sciences*, 117(27):15536–15545, 2020.

[179] Siddharth Gulati, Sonia Sousa, and David Lamas. Modelling trust: An empirical assessment. In *IFIP Conference on Human-Computer Interaction*, pages 40–61. Springer, 2017.

[180] Asela Gunawardana and Guy Shani. *Evaluating Recommender Systems*, pages 265–308. Springer US, Boston, MA, 2015. ISBN 978-1-4899-7637-6. doi: 10.1007/978-1-4899-7637-6_8. URL https://doi.org/10.1007/978-1-4899-7637-6_8.

[181] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*, pages 228–243. Springer, 2014.

[182] David Halpern. *Inside the nudge unit: How small changes can make a big difference*. Random House, 2016.

[183] Kirsi Halttu and Harri Oinas-kukkonen. Human – Computer Interaction Persuading to Reflect : Role of Reflection and Insight in Persuasive Systems Design for Physical Health Persuading to Reflect : Role of Reflection and Insight in Persuasive Systems Design for Physical Health. *Human–Computer Interaction*, 32(5-6):381–412, 2017. ISSN 0737-0024. doi: 10.1080/07370024.2017.1283227. URL https://doi.org/10.1080/07370024.2017.1283227.

[184] Felix Hamborg, Karsten Donnay, and Bela Gipp. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415, 2019. ISSN 14321300.

[185] Pelle G Hansen, Vincent F Hendricks, and Rasmus K Rendsvig. Infostorms. *Metaphilosophy*, 44(3):301–326, 2013.

[186] Pelle Guldborg Hansen and Andreas Maaløe Jespersen. Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation*, 4(1):3–28, 2013.

[187] Karen Hao. He got facebook hooked on ai. now he can't fix its misinformation addiction | mit technology review. https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/, 2021. (Accessed on 04/05/2021).

[188] Tony Harcup and Deirdre O'neill. What is news? news values revisited (again). *Journalism studies*, 18(12):1470–1488, 2017.

[189] Tim Harrower. *Inside Reporting: A Practical Guide to the Craft of Journalism (2012)*. 2007.

[190] William Hart, Dolores Albarracín, Alice H Eagly, Inge Brechan, Matthew J Lindberg, and Lisa Merrill. Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychological bulletin*, 135(4):555, 2009.

[191] H. Rex Hartson. Cognitive, physical, sensory, and functional affordances in interaction design. *Behaviour and Information Technology*, 22(5):315–338, 2003. ISSN 0144929X. doi: 10.1080/01449290310001592587.

[192] Del Harvey. Helping you find reliable public health information on twitter. https://blog.twitter.com/en_us/topics/company/2019/helping-you-find-reliable-public-health-information-on-twitter.html, 2019. (Accessed on 09/11/2020).

[193] Del Harvey and David Gasca. Serving healthy conversation. https://blog.twitter.com/official/en_us/topics/product/2018/Serving_Healthy_Conversation.html, 2018. (Accessed on 09/11/2020).

[194] Naeemul Hassan, Mohammad Yousuf, Mahfuzul Haque, Javier A Suarez Rivas, and Md Khadimul Islam. Towards a sustainable model for fact-checking platforms: Examining the roles of automation, crowds and professionals. Jan 2017. doi: 10.1145/3308560.3316734.

[195] Arthur S Hayes, Jane B Singer, and Jerry Ceppos. Shifting roles, enduring values: The credible journalist in a digital age. *Journal of mass media ethics*, 22(4):262–279, 2007.

[196] Donald Hedeker, Stephen HC du Toit, Hakan Demirtas, and Robert D Gibbons. A note on marginalization of regression parameters from mixed models of binary outcomes. *Biometrics*, 74(1):354–361, 2018.

[197] MS Heen, Joel D Lieberman, and Terance D Miethe. A comparison of different online sampling approaches for generating national samples. *Center for Crime and Justice Policy*, 1(9):1–8, 2014.

[198] John Hegeman. Facing facts: Facebook's fight against misinformation - about facebook. https://about.fb.com/news/2018/05/facing-facts-facebooks-fight-against-misinformation/, 2018. (Accessed on 09/11/2020).

[199] Edward S Herman and Noam Chomsky. *Manufacturing consent: The political economy of the mass media*. Random House, 2010.

[200] Yoshinori Hijikata, Takuya Shimizu, and Shogo Nishida. Discovery-oriented collaborative filtering for improving user satisfaction. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 67–76, New York, NY, USA, 2009. Association for Computing Machinery.

[201] Janette R. Hill, Liyan Song, and Richard E. West. Social learning theory and web-based learning environments: A review of research and discussion of implications. *International Journal of Phytoremediation*, 21(1):88–103, 2009. ISSN 15497879.

[202] Michael A Horning, Harold R Robinson, and John M Carroll. A scenario-based approach for projecting user requirements for wireless proximal community networks. *Computers in Human Behavior*, 35:413–422, 2014.

[203] Lei Hou, Juanzi Li, Zhichun Wang, Jie Tang, Peng Zhang, Ruibing Yang, and Qian Zheng. Newsminer: Multifaceted news analysis for event search. *Knowledge-Based Systems*, 76:17–29, 2015.

[204] Daniel J Howard and Thomas E Barry. The role of thematic congruence between a mood-inducing event and an advertised product in determining the effects of mood on brand attitudes. *Journal of Consumer Psychology*, 3(1):1–27, 1994.

[205] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[206] Travonia B Hughes, Vijay R Varma, Corinne Pettigrew, and Marilyn S Albert. African

americans and clinical research: evidence concerning barriers and facilitators to participation and recruitment recommendations. *The Gerontologist*, 57(2):348–358, 2017.

[207] Jane Im, Sonali Tandon, Eshwar Chandrasekharan, Taylor Denby, and Eric Gilbert. Synthesized social signals: Computationally-derived social signals from account histories. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[208] Twitter Inc. The twitter rules. https://help.twitter.com/en/rules-and-policies/twitter-rules, 2019. (Accessed on 09/13/2019).

[209] Reuters Institute. Overview and key findings of the 2022 digital news report | reuters institute for the study of journalism. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2022/dnr-executive-summary, 2022. (Accessed on 09/11/2022).

[210] Caroline Jack. Lexicon of lies: Terms for problematic information. 2017.

[211] Kathleen Hall Jamieson and Joseph N Cappella. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press, 2008.

[212] Peter John, Graham Smith, and Gerry Stoker. Nudge nudge, think think: two strategies for changing civic behaviour. *The Political Quarterly*, 80(3):361–370, 2009.

[213] Hollyn M Johnson and Colleen M Seifert. Sources of the continued influence effect: When misinformation in memory affects later inferences. *J. Exp. Psychol. Learn. Mem. Cogn.*, 20(6):1420, 1994.

[214] Andrew Johnston, Shigeki Amitani, and Ernest Edmonds. Amplifying reflective thinking in musical performance. In *Proceedings of the 5th conference on Creativity &*

cognition, pages 166–175, New York, NY, USA, 2005. Association for Computing Machinery.

[215] Mark Jurkowitz, Amy Mitchell, Elisa Shearer, and Mason Walker. U.s. media polarization and the 2020 election: A nation divided | pew research center. https://www.journalism.org/2020/01/24/u-s-media-polarization-and-the-2020-election-a-nation-divided/, 2020. (Accessed on 04/14/2021).

[216] Mark Jurkowitz, Amy Mitchell, Elisa Shearer, and Mason Walker. U.s. media polarization and the 2020 election: A nation divided | pew research center. https://www.pewresearch.org/journalism/2020/01/24/u-s-media-polarization-and-the-2020-election-a-nation-divided/, 2020. (Accessed on 09/15/2022).

[217] Daniel Kahneman and Amos Tversky. Choices, values, and frames. In *Handbook of the fundamentals of financial decision making: Part I*, pages 269–278. World Scientific, 2013.

[218] Shipi Kankane, Carlina DiRusso, and Christen Buckley. Can we nudge users toward better password management?: An initial study. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, page LBW593. ACM, 2018.

[219] Victor Kaptelinin and Bonnie Nardi. Affordances in HCI: Toward a mediated action perspective. *Conference on Human Factors in Computing Systems - Proceedings*, pages 967–976, 2012. doi: 10.1145/2207676.2208541.

[220] Michael Karlsson. Rituals of transparency: Evaluating online news outlets' uses of transparency rituals in the united states, united kingdom and sweden. *Journalism studies*, 11(4):535–545, 2010.

[221] Michael Karlsson. The immediacy of online news, the visibility of journalistic processes and a restructuring of journalistic authority. *Journalism*, 12(3):279–295, 2011.

[222] Michael Karlsson and Christer Clerwall. Transparency to the rescue? evaluating citizens' views on transparency tools in journalism. *Journalism Studies*, pages 1–11, 2018.

[223] Michael Karlsson, Christer Clerwall, and Lars Nord. You ain't seen nothing yet: Transparency's (lack of) effect on source and message credibility. *Journalism Studies*, 15(5): 668–678, 2014.

[224] Michael Karlsson, Christer Clerwall, and Lars Nord. Do not stand corrected: Transparency and users' attitudes to inaccurate news and corrections in online journalism. *Journalism & Mass Communication Quarterly*, 94(1):148–167, 2017.

[225] Kenneth A Kavale. The reasoning abilities of normal and learning disabled readers on measures of reading comprehension. *Learning Disability Quarterly*, 3(4):34–45, 1980.

[226] Ricardo Kawase, Eelco Herder, and Wolfgang Nejdl. A comparison of paper-based and online annotations in the workplace. In *European Conference on Technology Enhanced Learning*, pages 240–253. Springer, 2009.

[227] Keen. Keen | expand your interests. https://staykeen.com/about, 2021. (Accessed on 09/07/2021).

[228] Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, 2019.

[229] Damon Kiesow, Shuhua Zhou, and Lei Guo. Affordances for Sense-Making: Exploring Their Availability for Users of Online News Sites. *Digital Journalism*, 0(0):1–20, 2021.

doi: 10.1080/21670811.2021.1989316. URL https://doi.org/10.1080/21670811.2021.1989316.

[230] Spiro Kiousis. Public trust or mistrust? perceptions of media credibility in the information age. *Mass communication & society*, 4(4):381–403, 2001.

[231] Joel Kiskola, Thomas Olsson, Heli Väätäjä, Aleksi H. Syrjämäki, Anna Rantasila, Poika Isokoski, Mirja Ilves, and Veikko Surakka. *Applying Critical Voice in Design of User Interfaces for Supporting Self-Reflection and Emotion Regulation in Online News Commenting*. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450380966. URL https://doi.org/10.1145/3411764.3445783.

[232] Aniket Kittur, Ed H Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456, 2008.

[233] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E Kraut. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 43–52, 2011.

[234] Joshua Klayman. Varieties of confirmation bias. *Psychology of learning and motivation*, 32:385–418, 1995.

[235] Cory Knobel and Geoffrey C Bowker. Values in design. *Communications of the ACM*, 54(7):26–28, 2011.

[236] Silvia Knobloch-Westerwick and Jingbo Meng. Looking the other way: Selective exposure to attitude-consistent and counterattitudinal political information. *Communication Research*, 36(3):426–448, 2009.

[237] Silvia Knobloch-Westerwick, Nikhil Sharma, Derek L Hansen, and Scott Alter. Impact of popularity indications on readers' selective exposure to online news. *J. Broadcast. Electron. Media*, 49(3):296–313, 2005.

[238] Bran Knowles, Mark Rouncefield, Mike Harding, Nigel Davies, Lynne Blair, James Hannon, John Walden, and Ding Wang. Models and patterns of trust. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 328–338, 2015.

[239] D. Koehler and N. Harvey. In D. Koehler & N. Harvey (Eds.). (2004). Blackwell handbook of judgment and decision making (pp. 62–88). Oxford, UK: Blackwell. pages 62–88, 2004.

[240] Sarah Koenig. Season one - serial. https://serialpodcast.org/season-one, 2014. (Accessed on 09/29/2019).

[241] Andrew Kohut, Carroll Doherty, Michael Dimock, and Scott Keeter. Cable leads the pack as campaign news source. *Pew Center for the People and the Press. Retrieved from https://www.pewresearch.org/wp-content/uploads/sites/4/legacy-pdf/2012-Communicating-Release.pdf*, 2012.

[242] Kalliopi Kontiza, Olga Loboda, Louis Deladiennee, Sylvain Castagnos, and Yannick Naudet. A museum app to trigger users' reflection. In *International Workshop on Mobile Access to Cultural Heritage (MobileCH2018)*, Barcelona, Spain, 2018.

[243] Malcolm Koo and Harvey Skinner. Challenges of internet recruitment: a case study with disappointing results. *Journal of Medical Internet Research*, 7(1):e6, 2005.

[244] Elizabeth V Korinek, Sayali S Phatak, Cesar A Martin, Mohammad T Freigoun, Daniel E Rivera, Marc A Adams, Pedja Klasnja, Matthew P Buman, and Eric B

Hekler. Adaptive step goals and rewards: a longitudinal growth model of daily steps for a smartphone-based walking intervention. *Journal of behavioral medicine*, 41(1): 74–86, 2018.

[245] Bill Kovach and Tom Rosenstiel. *Blur: How to know what's true in the age of information overload*. Bloomsbury Publishing USA, 2011.

[246] Bill Kovach and Tom Rosenstiel. *The elements of journalism: What newspeople should know and the public should expect*. Three Rivers Press (CA), 2014.

[247] Geza Kovacs, Zhengxuan Wu, and Michael S Bernstein. Not now, ask later: Users weaken their behavior change regimen over time, but expect to re-strengthen it imminently. *arXiv preprint arXiv:2101.11743*, 2021.

[248] Steven Kull, Clay Ramsay, and Evan Lewis. Misperceptions, the media, and the iraq war. *Political science quarterly*, 118(4):569–598, 2003.

[249] Srijan Kumar, Robert West, and Jure Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602, 2016.

[250] Jim A. Kuypers. *Press Bias and Politics: How the Media Frame Controversial Issues*. Greenwood Publishing Group, 2002. ISBN 978-0-275-97758-0. Google-Books-ID: GHIQimmDvbcC.

[251] Jim A. Kuypers. *Issues of Bias in the News Media*, page 127–154. Peter Lang, 2013. ISBN 1-4331-2094-1.

[252] MIT Media Lab. Overview ‹ flipfeed — mit media lab. https://www.media.mit.edu/projects/flipfeed/overview/, 2017. (Accessed on 08/31/2021).

[253] KK Lamberty and Janet L Kolodner. Camera talk: Making the camera a partial participant. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 839–848, New York, NY, USA, 2005. Association for Computing Machinery.

[254] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

[255] Issie Lapowski. Newsguard wants to fight fake news with humans, not algorithms, 2018.

[256] Joseph D Lasica. Transparency begets trust in the ever-expanding blogosphere. *Online Journalism Review*, 12, 2004.

[257] Neal Lathia, Stephen Hailes, Licia Capra, and Xavier Amatriain. Temporal diversity in recommender systems. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 210–217, New York, NY, USA, 2010. Association for Computing Machinery.

[258] Paul F Lazarsfeld, Bernard Berelson, and Hazel Gaudet. The people's choice. In *The people's choice*. Columbia University Press, 1968.

[259] David G Lebow and Dale W Lick. Hylighter: An effective interactive annotation innovation for distance education. In *20th Annual Conference on Distance Teaching and Learning*, pages 1–5, 2005.

[260] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*, 2017.

[261] Matthew L Lee and Anind K Dey. Reflecting on pills and phone use: supporting awareness of functional abilities for older adults. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2095–2104, New York, NY, USA, 2011. Association for Computing Machinery.

[262] Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. Mining behavioral economics to design persuasive technology for healthy choices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 325–334. ACM, 2011.

[263] Birthe A Lehmann, Gretchen B Chapman, Frits ME Franssen, Gerjo Kok, and Robert AC Ruiter. Changing the default to promote influenza vaccination among health care workers. *Vaccine*, 34(11):1389–1392, 2016.

[264] Ian Li, Anind Dey, and Jodi Forlizzi. Grafitter: leveraging social media for self reflection. *XRDS: Crossroads, The ACM Magazine for Students*, 16(2):12–13, 2009.

[265] Ian Li, Anind Dey, and Jodi Forlizzi. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 557–566, New York, NY, USA, 2010. Association for Computing Machinery.

[266] Ian Li, Anind K Dey, and Jodi Forlizzi. Understanding my data, myself: supporting self-reflection with ubicomp technologies. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 405–414, New York, NY, USA, 2011. Association for Computing Machinery.

[267] Juanzi Li, Jun Li, and Jie Tang. A flexible topic-driven framework for news exploration. In *Proceedings of KDD*, volume 2007, 2007.

[268] Matthew D Lieberman. Social cognitive neuroscience: a review of core processes. *Annu. Rev. Psychol.*, 58:259–289, 2007.

[269] Yehiel Limor and Rafi Mann. Journalism: Reporting, writing and editing. *Tel Aviv: Open University Press (in Hebrew)*, 1997.

[270] James J Lin, Lena Mamykina, Silvia Lindtner, Gregory Delajoux, and Henry B Strub. Fish'n'steps: Encouraging physical activity with an interactive computer game. In *International conference on ubiquitous computing*, pages 261–278, Berlin, Heidelberg, 2006. Springer, Springer Berlin Heidelberg.

[271] Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. Which side are you on?: identifying perspectives at the document and sentence levels. In *Proceedings of the tenth conference on computational natural language learning*, pages 109–116. Association for Computational Linguistics, 2006.

[272] Jordan Litman, Tiffany Hutchins, and Ryan Russon. Epistemic curiosity, feeling-of-knowing, and exploratory behaviour. *Cognition & Emotion*, 19(4):559–582, 2005.

[273] Elizabeth Lopatto. In its latest confusing decision, twitter reinstates the new york post - the verge. https://www.theverge.com/2020/10/30/21542801/twitter-lifts-ny-post-ban-policy-changes, 2020. (Accessed on 04/05/2021).

[274] Charles G Lord, Mark R Lepper, and Elizabeth Preston. Considering the opposite: a corrective strategy for social judgment. *Journal of personality and social psychology*, 47(6):1231, 1984.

[275] Tamari Lukava, Dafne Zuleima Morgado Ramirez, and Giulia Barbareschi. Two sides of the same coin: accessibility practices and neurodivergent users' experience of extended reality. *Journal of Enabling Technologies*, 16(2):75–90, 2022.

[276] Tessa Lyons. Replacing disputed flags with related articles - about facebook. https://about.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/, 2017. (Accessed on 09/11/2020).

[277] Tessa Lyons. Hard questions: How is facebook's fact-checking program working? - about facebook. https://about.fb.com/news/2018/06/hard-questions-fact-checking/, 2018. (Accessed on 09/11/2020).

[278] Diane M Mackie and Sarah Queller. The impact of group membership on persuasion: Revisiting "who says what to whom with what effect?". *Attitudes, behavior, and social context: The role of norms and group membership*, pages 135–155, 2000.

[279] Rakoen Maertens, Jon Roozenbeek, Melisa Basol, and Sander van der Linden. Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, 2020.

[280] Jonathan RA Maier and Georges M Fadel. Affordance-based methods for design. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 37017, pages 785–794, 2003.

[281] Scott R Maier. Setting the record straight: When the press errs, do corrections follow? *Journalism Practice*, 1(1):33–43, 2007.

[282] Adam Maksl, Seth Ashley, and Stephanie Craft. Measuring news media literacy. *Journal of Media Literacy Education*, 6(3):29–45, 2015.

[283] Sylvain Malacria, Joey Scarr, Andy Cockburn, Carl Gutwin, and Tovi Grossman. Skillometers: Reflective widgets that motivate and help users to improve performance. In *Proceedings of the 26th annual ACM symposium on User interface software and*

technology, pages 321–330, New York, NY, USA, 2013. Association for Computing Machinery.

[284] Ross A Malaga. Worst practices in search engine optimization. *Communications of the ACM*, 51(12):147–150, 2008.

[285] Giada Marino and Laura Iannelli. Seven years of studying the associations between political polarization and problematic information: a literature review. *Frontiers in Sociology*, 8:91, 2023.

[286] Zlatina Marinova, Jochen Spangenberg, Denis Teyssou, Symeon Papadopoulos, Nikos Sarris, Alexandre Alaphilippe, and Kalina Bontcheva. Weverify: Wider and enhanced verification for you project overview and tools. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–4. IEEE, 2020.

[287] Renee Martin-Kratzer and Esther Thorson. Use of anonymous sources declines in us newspapers. *Newspaper Research Journal*, 28(2):56–70, 2007.

[288] Nafiseh Masoudi, Georges M Fadel, Christopher C Pagano, and Maria Vittoria Elena. A review of affordances and affordance-based design to address usability. In *Proceedings of the Design Society: International Conference on Engineering Design*, volume 1, pages 1353–1362. Cambridge University Press, 2019.

[289] Kelly McBride and Tom Rosenstiel. *The new ethics of journalism: Principles for the 21st century*. CQ Press, 2013.

[290] Robert D McChesney. *The problem of the media: US communication politics in the twenty-first century*. NYU Press, 2004.

[291] Maxwell E McCombs and Donald L Shaw. The agenda-setting function of mass media. *Public opinion quarterly*, 36(2):176–187, 1972.

[292] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[293] Joanna McGrenere and Wayne Ho. Affordances: Clarifying and evolving a concept. In *Graphics interface*, volume 2000, pages 179–186, 2000.

[294] Sean M McNee, John Riedl, and Joseph A Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 1097–1101, New York, NY, USA, 2006. Association for Computing Machinery.

[295] Mediabiasfactcheck. About - media bias/fact check. https://mediabiasfactcheck.com/about/, 2019. (Accessed on 09/18/2019).

[296] Ania Medrek. NEWS BY ASSOCIATION: Designing a way out of the echo chamber. (April), 2018.

[297] Melvin Mencher and Wendy P Shilton. *News reporting and writing*. Brown & Benchmark Publishers, 1997.

[298] Solomon Messing and Sean J Westwood. Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication research*, 41(8):1042–1063, 2014.

[299] Miriam J Metzger. Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, 58(13):2078–2091, 2007.

[300] Miriam J Metzger, Andrew J Flanagin, and Ryan B Medders. Social and heuristic

approaches to credibility evaluation online. *Journal of communication*, 60(3):413–439, 2010.

[301] Miriam J Metzger, Ethan H Hartsell, and Andrew J Flanagin. Cognitive dissonance or credibility? a comparison of two theoretical explanations for selective exposure to partisan news. *Communication Research*, page 0093650215613136, 2015.

[302] Hans K Meyer, Doreen Marchionni, and Esther Thorson. The journalist behind the news: credibility of straight, collaborative, opinionated, and blogged "new". *American Behavioral Scientist*, 54(2):100–119, 2010.

[303] Philip Meyer. Defining and measuring credibility of newspapers: Developing an index. *Journalism quarterly*, 65(3):567–574, 1988.

[304] Susan Michie, Michelle Richardson, Marie Johnston, Charles Abraham, Jill Francis, Wendy Hardeman, Martin P Eccles, James Cane, and Caroline E Wood. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Annals of behavioral medicine*, 46(1):81–95, 2013.

[305] Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley, and Katerina Eva Matsa. Political polarization & media habits. *Pew Research Center*, 21, 2014.

[306] Amy Mitchell, Jeffrey Gottfried, Michael Barthel, and Nami Sumida. Can americans tell factual from opinion statements in the news? | pew research center. https://www.journalism.org/2018/06/18/distinguishing-between-factual-and-opinion-statements-in-the-news/, 2018. (Accessed on 01/23/2020).

[307] Tanushree Mitra and Eric Gilbert. Credbank: A large-scale social media corpus with associated credibility annotations. In *Proc. ICWSM'15*, 2015.

[308] Tanushree Mitra, Clayton J Hutto, and Eric Gilbert. Comparing person-and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1345–1354, 2015.

[309] Tanushree Mitra, Graham P Wright, and Eric Gilbert. A parsimonious language model of social media credibility across disparate events. In *Proc. CSCW*, pages 126–145. ACM, 2017.

[310] Ine Mols, Elise van den Hoven, and Berry Eggen. Informing design for reflection: An overview of current everyday practices. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, NordiCHI '16, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450347631. doi: 10.1145/2971485.2971494. URL https://doi.org/10.1145/2971485.2971494.

[311] Niv Mor and Zvi Reich. From "trust me" to "show me" journalism: Can documentcloud help to restore the deteriorating credibility of news? *Journalism Practice*, 12(9):1091–1108, 2018.

[312] Adam Mosseri. Addressing hoaxes and fake news - about facebook. https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/, 2016. (Accessed on 09/11/2020).

[313] Sendhil Mullainathan and Andrei Shleifer. Media bias, 2002.

[314] Sean A Munson and Paul Resnick. Presenting diverse political opinions: how and

how much. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1457–1466, 2010.

[315] Sean A Munson, Hasan Cavusoglu, Larry Frisch, and Sidney Fels. Sociotechnical challenges and progress in using social media for health. *Journal of medical Internet research*, 15(10):e226, 2013.

[316] Rosanna Nagtegaal, Lars Tummers, Mirko Noordegraaf, and Victor Bekkers. Designing to debias: Measuring and reducing public managers' anchoring bias. *Public Administration Review*, 80(4):565–576, 2020.

[317] Joseph Napolitan. *The election game and how to win it*. Doubleday, 1972.

[318] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

[319] Elmie Nekmat. Nudge effect of fact-check alerts: Source influence and media skepticism on sharing of news misinformation in social media. *Social Media+ Society*, 6(1): 2056305119897322, 2020.

[320] Neil Nemeth and Craig Sanders. Number of corrections increase at two national newspapers. *Newspaper Research Journal*, 30(3):90–104, 2009.

[321] First Draft News. Research on crosscheck journalists and readers suggests positive impact for project. https://firstdraftnews.org/latest/crosscheck-qualitative-research/, 2017. (Accessed on 01/22/2020).

[322] NJ Spotlight News. Funders | nj spotlight news. https://www.njspotlight.com/about/funders/, 2020. (Accessed on 10/11/2020).

[323] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

[324] Jasmin Niess and Paweł W Woźniak. Supporting meaningful personal fitness: The tracker goal evolution model. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, New York, NY, USA, 2018. Association for Computing Machinery.

[325] Matthew C Nisbet, Dietram A Scheufele, James Shanahan, Patricia Moy, Dominique Brossard, and Bruce V Lewenstein. Knowledge, reservations, or promise? a media effects model for public perceptions of science and technology. *Communication Research*, 29(5):584–608, 2002.

[326] Brian Keith Norambuena, Michael Horning, and Tanushree Mitra. Evaluating the inverted pyramid structure through automatic 5w1h extraction and summarization. In *Proc. of the 2020 Computation+ Journalism Symposium. Computation+ Journalism*, pages 1–7, 2020.

[327] Dan Norman. Signifiers, not affordances. 15(6):1–23, 2016.

[328] Don Norman. Affordances and design. *Unpublished article, available online at: http://www. jnd. org/dn. mss/affordances-and-design. html*, 2004.

[329] Don Norman. *The design of everyday things: Revised and expanded edition*. Basic books, 2013.

[330] Donald A Norman. *The psychology of everyday things.* Basic books, 1988.

[331] Donald A Norman. Affordance, conventions, and design. *interactions*, 6(3):38–43, 1999.

[332] Elena Novak, Rim Razzouk, and Tristan E. Johnson. The educational use of social annotation tools in higher education: A literature review. *Internet and Higher Education*, 15(1):39–49, 2012. ISSN 10967516. doi: 10.1016/j.iheduc.2011.09.002.

[333] Alexander Nussbaumer, Milos Kravcik, and Dietrich Albert. Supporting self-reflection in personal learning environments through user feedback. In *UMAP Workshops*, 2012.

[334] Adewale Obadimu, Esther Mead, Muhammad Nihal Hussain, and Nitin Agarwal. Identifying toxicity within youtube video comment. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 214–223. Springer, 2019.

[335] Onora O'neill. *A question of trust: The BBC Reith Lectures 2002*. Cambridge University Press, 2002.

[336] Ruchi Ookalkar, Kolli Vishal Reddy, and Eric Gilbert. Pop: Bursting news filter bubbles on twiter through diverse exposure. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, pages 18–21, 2019. doi: 10.1145/3311957.3359513.

[337] Ruchi Ookalkar, Kolli Vishal Reddy, and Eric Gilbert. Pop: Bursting news filter bubbles on twitter through diverse exposure. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, pages 18–22, 2019.

[338] OpenSources. Opensources. http://opensources.co, 2018.

[339] Cian O'Mahony, Maryanne Brassil, Gillian Murphy, and Conor Linehan. The efficacy of interventions in reducing belief in conspiracy theories: A systematic review. *Plos one*, 18(4):e0280902, 2023.

[340] Nirzar Pangarkar. How we designed page previews for wikipedia and what could be done with them in the future – wikimedia blog. https://blog.wikimedia.org/2018/04/18/how-we-designed-page-previews-for-wikipedia/, 2018. (Accessed on 10/15/2019).

[341] Eli Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.

[342] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. Newscube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 443–452, 2009.

[343] Souneil Park, Minsam Ko, Jungwoo Kim, Ho-jin Choi, and Junehwa Song. NewsCube2 . 0 : An Exploratory Design of a Social News Website for Media Bias Mitigation. *Workshop on Social Recommender Systems*, pages 1–5, 2011. URL https://pdfs.semanticscholar.org/b87b/f0986b2e9fe34a22ed0c19cfd32ed06857d0.pdf.

[344] Souneil Park, Kyung Soon Lee, and Junehwa Song. Contrasting opposing views of news articles on contentious issues. *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1: 340–349, 2011.

[345] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. A computational framework for media bias mitigation. *ACM Transactions on Interactive Intelligent Systems*, 2(2), 2012. ISSN 21606463. doi: 10.1145/2209310.2209311.

[346] Sara Parker and Derek Ruths. Is hate speech detection the solution the world wants? *Proceedings of the National Academy of Sciences*, 120(10):e2209384120, 2023.

[347] John W Payne, James R Bettman, and Eric J Johnson. *The adaptive decision maker*. Cambridge university press, 1993.

[348] Jessica Paynter, Sarah Luskin-Saxby, Deb Keen, Kathryn Fordyce, Grace Frost, Christine Imms, Scott Miller, David Trembath, Madonna Tucker, and Ullrich Ecker. Evaluation of a template for countering misinformation—real-world autism treatment myth debunking. *PloS one*, 14(1), 2019.

[349] Cornelia Pechmann. Predicting when two-sided ads will be more effective than one-sided ads: The role of correlational and correspondent inferences. *Journal of Marketing Research*, 29(4):441–453, 1992.

[350] Gordon Pennycook and David G Rand. Assessing the effect of "disputed" warnings and source salience on perceptions of fake news accuracy. 2017.

[351] Gordon Pennycook and David G Rand. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7):2521–2526, 2019.

[352] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David Rand. Understanding and reducing the spread of misinformation online. *Unpublished manuscript: https://psyarxiv. com/3n9u8*, 2019.

[353] Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, 2020.

[354] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, 31(7):770–780, 2020.

[355] Richard E Petty and John T Cacioppo. The elaboration likelihood model of persuasion. In *Communication and persuasion*, pages 1–24. Springer, 1986.

[356] Charlie Pinder, Jo Vermeulen, Russell Beale, and Robert Hendley. Subliminal priming of nonconscious goals on smartphones. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, pages 825–830, 2015.

[357] Val Pipps, Heather Walter, Kathleen Endres, and Patrick Tabatcher. Information recall of internet news: Does design make a difference? a pilot study. *Journal of Magazine Media*, 11(1):1–20, 2009.

[358] Patrick Lee Plaisance. Transparency: An assessment of the kantian roots of a key element in media ethics practice. *Journal of Mass Media Ethics*, 22(2-3):187–207, 2007.

[359] Patrick Lee Plaisance and Joan A Deppa. Perceptions and manifestations of autonomy, transparency and harm among us newspaper journalists. *Journalism & Communication Monographs*, 10(4):327–386, 2009.

[360] Bernd Ploderer, Wolfgang Reitberger, Harri Oinas-Kukkonen, and Julia van Gemert-Pijnen. Social interaction and reflection for behaviour change, 2014.

[361] Sara Pluviano, Caroline Watt, and Sergio Della Sala. Misinformation lingers in memory: failure of three pro-vaccination strategies. *PLoS One*, 12(7):e0181640, 2017.

[362] Horst Po¨ ttker. News and its communicative quality: The inverted pyramid—when and why did it appear? *Journalism Studies*, 4(4):501–511, 2003.

[363] Odette Pollar. *Surviving information overload: how to find, filter, and focus on what's important*. Thomson Crisp Learning, 2003.

[364] Colin Porlezza, Scott R Maier, and Stephan Russ-Mohl. News accuracy in switzerland and italy: a transatlantic comparison with the us press. *Journalism Practice*, 6(4): 530–546, 2012.

[365] Jon Porter. Twitter announces new api pricing, posing a challenge for small developers - the verge. https://www.theverge.com/2023/3/30/23662832/twitter-api-tiers-free-bot-novelty-accounts-basic-enterprice-monthly-price, 2023. (Accessed on 06/10/2023).

[366] Associated Press. Ap news values and principles. 2014.

[367] THE MEDIA INSIGHT PROJECT. Americans and the news media: What they do — and don't — understand about each other - american press institute. https://www.americanpressinstitute.org/publications/reports/survey-research/americans-and-the-news-media/, 2018. (Accessed on 09/07/2020).

[368] The Media Insight Project. What americans know, and don't, about how journalism works - american press institute. https://www.americanpressinstitute.org/publications/reports/survey-research/what-americans-know-about-journalism/, 2018. (Accessed on 10/02/2020).

[369] The Trust Project. Frontpage. https://thetrustproject.org/, 2015. (Accessed on 10/14/2019).

[370] The Trust Project. Collaborator materials. https://thetrustproject.org/collaborator-materials/, 2019. (Accessed on 01/23/2020).

[371] ProPublica. Supporters — propublica. https://www.propublica.org/supporters, 2017. (Accessed on 10/11/2020).

[372] ProPublica. This is what propublica is now covering — propublica. https://www.propublica.org/article/this-is-what-propublica-is-now-covering, 2017. (Accessed on 07/10/2020).

[373] Hoyt Purvis. Anonymous sources: More or less and why and where? *Southwestern Mass Communication Journal*, 30(2), 2015.

[374] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599. Association for Computational Linguistics, 2011.

[375] Muck Rack. Muck rack for journalists and public relations. https://muckrack.com/, 2009. (Accessed on 03/25/2021).

[376] Muck Rack. Muck rack | for journalists. https://muckrack.com/journalists, 2019. (Accessed on 09/14/2019).

[377] Danielle E Ramo and Judith J Prochaska. Broad reach and targeted recruitment using facebook for an online survey of young adult substance use. *Journal of medical Internet research*, 14(1):e28, 2012.

[378] Arthur S Reber. Implicit learning and tacit knowledge. *Journal of experimental psychology: General*, 118(3):219, 1989.

[379] Marc-Andre Reinhard and Siegfried L Sporer. Verbal and nonverbal behaviour as a basis for credibility attribution: The impact of task involvement and cognitive capacity. *Journal of Experimental Social Psychology*, 44(3):477–488, 2008.

[380] Paul Resnick, R Kelly Garrett, Travis Kriplean, Sean A Munson, and Natalie Jomini Stroud. Bursting your (filter) bubble: strategies for promoting diverse exposure. In

Proceedings of the 2013 conference on Computer supported cooperative work companion, pages 95–100, New York, NY, USA, 2013. Association for Computing Machinery.

[381] Matthias Revers. The twitterization of news making: Transparency and journalistic professionalism. *Journal of communication*, 64(5):806–826, 2014.

[382] Jens Riegelsberger. Interpersonal cues and consumer trust in e-commerce. In *CHI'03 Extended Abstracts on Human Factors in Computing Systems*, pages 674–675, 2003.

[383] Zacc Ritter. How much does the world trust journalists? https://news.gallup.com/opinion/gallup/272999/world-trust-journalists.aspx, 2019. (Accessed on 03/22/2021).

[384] Dimitris Rizopoulos. Glmmadaptive: generalized linear mixed models using adaptive gaussian quadrature. *R package version 0.5–1*, 2019.

[385] Marilyn Roberts, Wayne Wanta, and Tzong-Horng Dzwo. Agenda setting and issue salience online. *Communication research*, 29(4):452–465, 2002.

[386] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. Can the crowd identify misinformation objectively? the effects of judgment scale and assessor's background. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 439–448, 2020.

[387] Jon Roozenbeek and Sander van der Linden. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, 5(1):1–10, 2019.

[388] Guy Rosen. An update on our work to keep people informed and limit misinformation about covid-19 - about facebook. https://about.fb.com/news/2020/04/covid-19-misinfo-update/, 2020. (Accessed on 09/11/2020).

[389] Guy Rosen and Tessa Lyons. Remove, reduce, inform: New steps to manage problematic content - about facebook. https://about.fb.com/news/2019/04/remove-reduce-inform-new-steps/, 2019. (Accessed on 09/11/2020).

[390] Jay Rosen. Beyond objectivity. 1993.

[391] Matthew Rosenberg. How republican voters took qanon mainstream - the new york times. https://www.nytimes.com/2020/10/19/us/politics/qanon-trump-republicans.html, 10 2020. (Accessed on 11/08/2021).

[392] Mary Beth Rosson, John M Carroll, and Natalie Hill. *Usability engineering: scenario-based development of human-computer interaction*. Morgan Kaufmann, 2002.

[393] Yoel Roth and Ashita Achuthan. Building rules in public: Our approach to synthetic & manipulated media. https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html, 2020. (Accessed on 09/11/2020).

[394] Yoel Roth and Nick Pickles. Updating our approach to misleading information. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html, 2020. (Accessed on 09/11/2020).

[395] Twitter Safety. Strengthening our approach to deliberate attempts to mislead voters. https://blog.twitter.com/en_us/topics/company/2019/strengthening-our-approach-to-deliberate-attempts-to-mislead-vot.html, 2019. (Accessed on 09/11/2020).

[396] Twitter Safety. Expanding our policies to further protect the civic con-

versation. https://blog.twitter.com/en_us/topics/company/2020/civic-integrity-policy-update.html, 2020. (Accessed on 09/11/2020).

[397] Emily Saltz, Tommy Shane, Victoria Kwan, Claire Leibowicz, and Claire Wardle. It matters how platforms label manipulated media. here are 12 principles designers should follow. - the partnership on ai. https://www.partnershiponai.org/it-matters-how-platforms-label-manipulated-media-here-are-12-principles-designers- 2020. (Accessed on 07/01/2020).

[398] Corina Sas and Alan Dix. Designing for reflection on personal experience. *International Journal of Human-Computer Studies*, 69(5):281–282, 2011.

[399] Corina Sas and Irni Eliana Khairuddin. Design for trust: An exploration of the challenges and opportunities of bitcoin users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6499–6510, 2017.

[400] Roser Saurí and James Pustejovsky. Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43(3):227, 2009.

[401] Jorrit Schaap. Bubble Trouble – Venture Out of Your Filter Bubbles. pages 1–14, 2020.

[402] schema.org. Markup for news - schema.org. https://schema.org/docs/news.html, 2017. (Accessed on 01/24/2020).

[403] Dietram A Scheufele. Agenda-setting, priming, and framing revisited: Another look at cognitive effects of political communication. *Mass communication & society*, 3(2-3): 297–316, 2000.

[404] Dietram A Scheufele and David Tewksbury. Framing, agenda setting, and priming:

The evolution of three media effects models. *Journal of communication*, 57(1):9–20, 2007.

[405] Michael Schudson. *Discovering the news: A social history of American newspapers*. Basic Books, 1981.

[406] Paul Resnick Sean A. Munson, Stephanie Y. Lee. Encouraging reading of diverse political viewpoints with a browser widget. In *ICWSM*, 2013.

[407] Katharine Q Seelye. Times panel proposes steps to build credibility. *New York Times*, 9, 2005.

[408] Gwendolyn Seidman. Self-presentation and belonging on facebook: How personality influences social media use and motivations. *Personality and individual differences*, 54 (3):402–407, 2013.

[409] Evan Selinger and Kyle Whyte. Is there a right way to nudge? the practice and ethics of choice architecture. *Sociology Compass*, 5(10):923–935, 2011.

[410] Alina Selyukh, Maria Hollenhorst, and Katie Park. Disney-fox deal: Who controls digital media? conglomerates, brands in one chart|npr. https://www.npr.org/sections/alltechconsidered/2016/10/28/ 499495517/big-media-companies-and-their-many-brands-in-one-chart, 2016.

[411] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph Kaye. Reflective design. *Critical Computing - Between Sense and Sensibility - Proceedings of the 4th Decennial Aarhus Conference*, pages 49–58, 2005. doi: 10.1145/1094562.1094569.

[412] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9, 2018.

[413] Renee Shelby, Shalaleh Rismani, Kathryn Henne, Ajung Moon, Negar Rostamzadeh, Paul Nicholas, YILLA-AKBARI N'MAH, Jess Gallegos, Andrew Smart, and GURLEEN VIRK. Identifying sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. *arXiv preprint arXiv:2210.05791*, 2022.

[414] Richard M Shiffrin and Walter Schneider. Automatic and controlled processing revisited. 1984.

[415] Pamela J Shoemaker, Tsan-Kuo Chang, and Nancy Brendlinger. Deviance as a predictor of newsworthiness: Coverage of international events in the us media. *Annals of the International Communication Association*, 10(1):348–365, 1987.

[416] John Sides. Analysis | is the media biased toward clinton or trump? here is some actual hard data. *Washington Post*, Sep 2016. ISSN 0190-8286. URL https://www.washingtonpost.com/news/monkey-cage/wp/2016/09/20/is-the-media-biased-toward-clinton-or-trump-heres-some-actual-hard-data/.

[417] Suvi Silfverberg, Lassi A. Liikkanen, and Airi Lampinen. "I'll press play, but I won't listen". page 207, 2011. doi: 10.1145/1958824.1958855.

[418] Craig Silverman. *Regret the error: how media mistakes pollute the press and imperil free speech*. Sterling Publishing Company, Inc., 2009.

[419] Henry Silverman. Helping fact-checkers identify false claims faster - about facebook. https://about.fb.com/news/2019/12/helping-fact-checkers/, 2019. (Accessed on 09/11/2020).

[420] Slate. How people read online: Why you won't finish this article. https://slate.com/technology/2013/06/

how-people-read-online-why-you-wont-finish-this-article.html, 06 2013. (Accessed on 09/11/2022).

[421] Petr Slovak, Chris Frauenberger, and Geraldine Fitzpatrick. Reflective practicum: A framework of sensitising concepts to design for transformative reflection. *Conference on Human Factors in Computing Systems - Proceedings*, 2017-May:2696–2707, 2017. doi: 10.1145/3025453.3025516.

[422] Jeff Smith. Designing against misinformation - facebook design - medium. https://medium.com/facebook-design/designing-against-misinformation-e5846b3aa1e2, 2019. (Accessed on 09/18/2019).

[423] Jeff Smith, Alex Leavitt, and Grace Jackson. Designing new ways to give context to news stories | facebook newsroom. https://newsroom.fb.com/news/2018/04/inside-feed-article-context/, 2018.

[424] Oren Soffer. Algorithmic personalization and the two-step flow of communication. *Communication Theory*, 31(3):297–315, 2021.

[425] Sandeep Soni, Tanushree Mitra, Eric Gilbert, and Jacob Eisenstein. Modeling factuality judgments in social media text. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 415–420, 2014.

[426] Timo Spinde, Felix Hamborg, Karsten Donnay, Angelica Becerra, and Bela Gipp. Enabling news consumers to view and understand biased news coverage: a study on the perception and visualization of media bias. In *Proceedings of the ACM/IEEE joint conference on digital libraries in 2020*, pages 389–392, 2020.

[427] Steven A Stahl, Cynthia R Hynd, Bruce K Britton, Mary M McNish, and Dennis Bosquet. What happens when students read multiple source documents in history? *Reading Research Quarterly*, 31(4):430–456, 1996.

[428] Keith E Stanovich and Richard F West. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and brain sciences*, 23(5):645–665, 2000.

[429] Dominik A Stecula and Mark Pickup. Social media, cognitive reflection, and conspiracy beliefs. *Frontiers in Political Science*, 3:62, 2021.

[430] Anselm Strauss and Juliet Corbin. Grounded theory methodology. *Handbook of qualitative research*, 17:273–85, 1994.

[431] Natalie Jomini Stroud. Polarization and partisan selective exposure. *Journal of communication*, 60(3):556–576, 2010.

[432] Natalie Jomini Stroud. *Niche news: The politics of news choice*. Oxford University Press on Demand, 2011.

[433] Jay Sullivan. Introducing a forwarding limit on messenger - about facebook. https://about.fb.com/news/2020/09/introducing-a-forwarding-limit-on-messenger/, 2020. (Accessed on 09/11/2020).

[434] Róbert Sumi, Taha Yasseri, et al. Edit wars in wikipedia. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 724–727. IEEE, 2011.

[435] S Shyam Sundar. The main model: A heuristic approach to understanding technology effects on credibility. *Digital media, youth, and credibility*, 73100, 2008.

[436] S Shyam Sundar, Anne Oeldorf-Hirsch, and Qian Xu. The bandwagon effect of collaborative filtering technology. In *CHI'08 extended abstracts*, pages 3453–3458. ACM, 2008.

[437] S Shyam Sundar, Haiyan Jia, T Franklin Waddell, and Yan Huang. Toward a theory of interactive media effects (time). *The handbook of the psychology of communication technology*, pages 47–86, 2015.

[438] Cass R Sunstein. The law of group polarization. *University of Chicago Law School, John M. Olin Law & Economics Working Paper*, (91), 1999.

[439] Cass R Sunstein. The polarization of extremes. *The Chronicle of Higher Education*, 54(16):9, 2007.

[440] Cass R Sunstein. http://republic.com 2.0, 2009.

[441] Cass R Sunstein. *Why nudge?: The politics of libertarian paternalism*. Yale University Press, 2014.

[442] Cass R Sunstein. The ethics of nudging. *Yale J. on Reg.*, 32:413, 2015.

[443] Cass R Sunstein. Do people like nudges? 2016.

[444] Cass R Sunstein. Nudges that fail. *Behavioural Public Policy*, 1(1):4–25, 2017.

[445] Twitter Support. New labels for government and state-affiliated media accounts. https://blog.twitter.com/en_us/topics/product/2020/new-labels-for-government-and-state-affiliated-media-accounts.html, 2020. (Accessed on 09/11/2020).

[446] James Surowiecki. *The wisdom of crowds*. Anchor, 2005.

[447] Hanaa Tameez. Maybe greater transparency can increase trust in news — but readers have to find your transparency first » nieman journalism lab. https://www.niemanlab.org/2020/01/maybe-greater-transparency-can-increase-trust-in-news-but-readers-have-to-find-you 2020. (Accessed on 02/23/2021).

[448] Harsh Taneja and Katie Yaeger. Do people consume the news they trust? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing systems*, pages 1–10, 2019.

[449] David Tannenbaum, Craig R Fox, and Todd Rogers. On the misplaced politics of behavioural policy interventions. *Nature Human Behaviour*, 1(7):0130, 2017.

[450] David Tewksbury and Scott L Althaus. Differences in knowledge acquisition among readers of the paper and online versions of a national newspaper. *Journalism & Mass Communication Quarterly*, 77(3):457–479, 2000.

[451] Richard H Thaler. Nudge: Improving decisions about health, wealth, and happiness, 2008.

[452] Share the Facts. Share the facts. http://www.sharethefacts.org/, 2016. (Accessed on 01/22/2020).

[453] Jacob Thebault-Spieker, Stevie Chancellor, Michael Ann DeVito, Niloufar Salehi, Alex Leavitt, David Karger, and Katta Spiel. Do we fix it or burn it down? towards practicable critique at cscw. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '21, page 234–237, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384797. doi: 10.1145/3462204.3483281. URL https://doi.org/10.1145/3462204.3483281.

[454] Anja Thieme, Jayne Wallace, Paula Johnson, John McCarthy, Siân Lindley, Peter Wright, Patrick Olivier, and Thomas D Meyer. Design to promote mindfulness practice and sense of self for vulnerable women in secure hospital services. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2647–2656, New York, NY, USA, 2013. Association for Computing Machinery.

[455] Sreethu Thulasi. Understand why you're seeing certain ads and how you can adjust your ad experience - about facebook. https://about.fb.com/news/2019/07/understand-why-youre-seeing-ads/, 2019. (Accessed on 09/11/2020).

[456] NY Times. Timescast | march 22, 2010 - video - nytimes.com. https://www.nytimes.com/video/continuous/1247467418484/timescast.html, 2010. (Accessed on 09/29/2019).

[457] The New York Times. Product and design | the new york times company. https://www.nytco.com/careers/product-and-design/, 2019. (Accessed on 03/05/2021).

[458] Peter M. Todd and Gerd Gigerenzer. Précis of Simple heuristics that make us smart. *Behavioral and Brain Sciences*, 23(5):727–780, 2000. ISSN 0140525X. doi: 10.1017/S0140525X00003447.

[459] Newsroom Transparency Tracker. Discover the people, policies, and practices behind the news. https://www.newsroomtransparencytracker.com/, 2019. (Accessed on 01/24/2020).

[460] Shawn Tseng and BJ Fogg. Credibility and computing technology. *Communications of the ACM*, 42(5):39–44, 1999.

[461] Tiffany Tseng and Coram Bryant. Design, reflect, explore: encouraging children's reflections with mechanix. In *CHI'13 Extended Abstracts on Human Factors in Com-*

puting Systems, pages 619–624. Association for Computing Machinery, New York, NY, USA, 2013.

[462] Yariv Tsfati. Online news exposure and trust in the mainstream media: Exploring possible associations. *American Behavioral Scientist*, 54(1):22–42, 2010.

[463] Yariv Tsfati and Joseph N Cappella. Do people watch what they do not trust? exploring the association between news media skepticism and exposure. *Communication Research*, 30(5):504–529, 2003.

[464] Yariv Tsfati and Joseph N Cappella. Why do people watch news they do not trust? the need for cognition as a moderator in the association between news media skepticism and exposure. *Media psychology*, 7(3):251–271, 2005.

[465] Gaye Tuchman. Objectivity as strategic ritual: An examination of newsmen's notions of objectivity. *American Journal of sociology*, 77(4):660–679, 1972.

[466] Miroslav Tudjman and Nives Mikelic. Information science: Science about information, misinformation and disinformation. *Proceedings of Informing Science+ Information Technology Education*, 3:1513–1527, 2003.

[467] Twitter. Notices on twitter and what they mean. https://help.twitter.com/en/rules-and-policies/notices-on-twitter, 2019. (Accessed on 09/11/2020).

[468] Juliane Urban and Wolfgang Schweiger. News quality from the recipients' perspective: Investigating recipients' ability to judge the normative quality of news. *Journalism Studies*, 15(6):821–840, 2014.

[469] Nina Valkanova, Sergi Jorda, Martin Tomitsch, and Andrew Vande Moere. Reveal-it! the impact of a social visualization projection on public awareness and discourse.

In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3461–3470, New York, NY, USA, 2013. Association for Computing Machinery.

[470] Marshall Van Alstyne and Erik Brynjolfsson. Global village or cyber-balkans? modeling and measuring the integration of electronic communities. *Management science*, 51 (6):851–868, 2005.

[471] Kim J Vicente and Jens Rasmussen. Ecological interface design: Theoretical foundations. *IEEE Transactions on systems, man, and cybernetics*, 22(4):589–606, 1992.

[472] Anthony J Viera, Joanne M Garrett, et al. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363, 2005.

[473] Stella Vosniadou and William F Brewer. Theories of knowledge restructuring in development. *Review of educational research*, 57(1):51–67, 1987.

[474] Stella Vosniadou and Andrew Ortony. *Similarity and analogical reasoning*. Cambridge University Press, 1989.

[475] Emily Vraga, Melissa Tully, John E Kotcher, Anne-Bennett Smithson, and Melissa Broeckelman-Post. A multi-dimensional approach to measuring news media literacy. *Journal of Media Literacy Education*, 7(3):41–53, 2015.

[476] W3. Credible web community group. https://www.w3.org/community/credibility/, 2017. (Accessed on 01/30/2020).

[477] David Wagner, Gabriele Vollmar, and Heinz-Theo Wagner. The impact of information technology on knowledge creation: An affordance approach to social media. *Journal of Enterprise Information Management*, 2014.

[478] Jonathan Wai and Kaja Perina. Expertise in journalism: Factors shaping a cognitive and culturally elite profession. *Journal of Expertise*, 1(1):57–78, 2018.

[479] Canhui Wang, Min Zhang, Liyun Ru, and Shaoping Ma. An automatic online news topic keyphrase extraction system. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 214–219. IEEE, 2008.

[480] Wei Wang. Chinese news event 5w1h semantic elements extraction for event ontology population. In *Proceedings of the 21st International Conference on World Wide Web*, pages 197–202. ACM, 2012.

[481] Yang Wang, Pedro Giovanni Leon, Kevin Scott, Xiaoxuan Chen, Alessandro Acquisti, and Lorrie Faith Cranor. Privacy nudges for social media: an exploratory facebook study. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 763–770, 2013.

[482] Yang Wang, Pedro Giovanni Leon, Alessandro Acquisti, Lorrie Faith Cranor, Alain Forget, and Norman Sadeh. A field trial of privacy nudges for facebook. In *Proc. CHI*, pages 2367–2376. ACM, 2014.

[483] Yixue Wang and Siyu Yao. Study on intention-aware recommendation of youtube videos. 2020.

[484] Claire Wardle. Misinformation has created a new world disorder - scientific american. https://www.scientificamerican.com/article/misinformation-has-created-a-new-world-disorder/, 2019. (Accessed on 05/27/2020).

[485] C Kay Weaver, Judy Motion, and Juliet Roper. From propaganda to discourse (and back again): Truth, power, the public interest and public relations. *Public relations: Critical debates and contemporary practice*, 7:21, 2006.

[486] Matthew S Weber. Newspapers and the long-term implications of hyperlinking. *Journal of Computer-Mediated Communication*, 17(2):187–201, 2012.

[487] Karl E Weick. *Sensemaking in organizations*, volume 3. Sage, 1995.

[488] Karl E Weick, Kathleen M Sutcliffe, and David Obstfeld. Organizing and the process of sensemaking. *Organization science*, 16(4):409–421, 2005.

[489] David Manning White. The "gate keeper": A case study in the selection of news. *Journalism quarterly*, 27(4):383–390, 1950.

[490] Wikipedia. Help:infobox - wikipedia. https://en.wikipedia.org/wiki/Help:Infobox, 2006. (Accessed on 10/01/2019).

[491] Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H. Chi, and Jennifer Gillenwater. Practical diversified recommendations on youtube with determinantal point processes. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 2165–2173, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360142. doi: 10.1145/3269206.3272018. URL https://doi.org/10.1145/3269206.3272018.

[492] WisconsinWatch. Funding | wisconsinwatch.org. https://www.wisconsinwatch.org/about/funding/. (Accessed on 10/11/2020).

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

[494] Thomas Wood and Ethan Porter. The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, 41(1):135–163, 2019.

[495] Robin Worrall. Social media used to spread, create covid-19 falsehoods – harvard gazette. https://news.harvard.edu/gazette/story/2020/05/social-media-used-to-spread-create-covid-19-falsehoods/, 2020. (Accessed on 09/12/2020).

[496] Caiming Xiong, Victor Zhong, and Richard Socher. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604*, 2016.

[497] Yan Xu, Erika Shehan Poole, Andrew D Miller, Elsa Eiriksdottir, Dan Kestranek, Richard Catrambone, and Elizabeth D Mynatt. This is not a one-horse race: understanding player types in multiplayer pervasive health games for youth. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 843–852, New York, NY, USA, 2012. Association for Computing Machinery.

[498] Hsiao-chen You and Kuohsiang Chen. Applications of affordance and semantics in product design. *Design Studies*, 28(1):23–38, 2007.

[499] John R Zaller et al. *The nature and origins of mass opinion*. Cambridge university press, 1992.

[500] Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, et al. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference 2018*, pages 603–612, 2018.

[501] Pengyi Zhang and Dagobert Soergel. Cognitive mechanisms in sensemaking: A qualitative user study. *Journal of the Association for Information Science and Technology*, 71(2):158–171, 2020. ISSN 23301643. doi: 10.1002/asi.24221.

[502] Shanyang Zhao, Sherri Grasmuck, and Jason Martin. Identity construction on facebook: Digital empowerment in anchored relationships. *Computers in human behavior*, 24(5):1816–1836, 2008.

[503] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th international conference on world wide web*, pages 1395–1405, 2015.

[504] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515, 2010.

[505] Fabiana Zollo and Walter Quattrociocchi. Misinformation spreading on facebook. In *Complex spreading phenomena in social systems*, pages 177–196. Springer, 2018.

Appendices

Appendix A

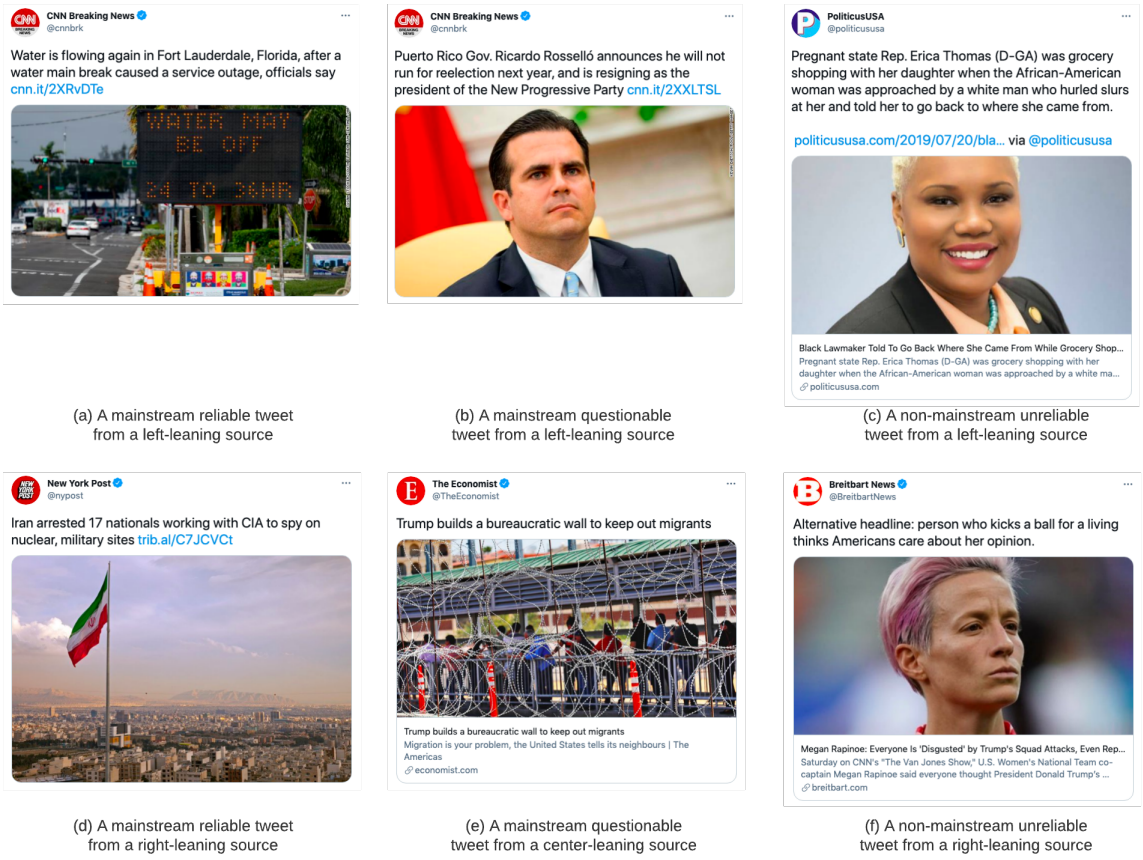
NudgeCred

A.1 Example Tweets Used in Study 1

Figure A.1 shows several tweets used in Study 1. These tweets were selected by finding the most popular tweets from the last 48 hours from various partisan sources. Notice the partisan nature of the items including immigration, racism and LGBTQ+ issues.

A.2 Study 2: Interview Questionnaire

- Could you tell me about your news reading on twitter? How often do you read news? and what type of news do you read?
- How often do you come across misinformation?
- Have you ever felt the need for any tools to improve news reading on twitter?
- Have you ever used any tools to improve news reading on twitter? What tool? How did it work?
- (Asking them to share their screen for twitter feed) How would you compare your Twitter use during the study to how you normally do?



(a) A mainstream reliable tweet from a left-leaning source

(b) A mainstream questionable tweet from a left-leaning source

(c) A non-mainstream unreliable tweet from a left-leaning source

(d) A mainstream reliable tweet from a right-leaning source

(e) A mainstream questionable tweet from a center-leaning source

(f) A non-mainstream unreliable tweet from a right-leaning source

Figure A.1: Sample tweets used in Study 1 without the interventions. The examples include reliable, questionable and unreliable tweets from left-/center-/right-leaning sources. Here, there is a mix of politically contentious (e.g., immigration, racism and LGBTQ+) and not so contentious issues (e.g., flood and national security).

- Could you tell me about a time when you paid more attention to a news on Twitter in the last 5 days? Why? What was it about?
- How satisfied were you with what you saw on Twitter in the last 5 days?
- What aspect of the intervention did you notice most? How did that impact you?
- When using the extension, did you think of anything it was missing? What more should it do?
- Can you think of any other application of this extension that you would like?
- During the study, did you have any issues with NudgeCred? Is there any aspect of the usability (eg., design, speed, accuracy) you thought could be improved?
- How did you feel about the plugin overall? What did you like about it? What did you dislike about it?
- Would you continue using it after this study?

Appendix B

NewsComp

B.1 Thinkaloud Interviews Questionnaires

- What are the viewpoints expressed in each article? How would you compare the viewpoints between the articles?
- How would you compare the numbers of actors reported in each article?
- How would you compare each article providing complete information about what happened/where/when/who was involved?
- How would you compare the analytical quality in each article? Do they provide information on causes, consequences, evaluations and claims of/from the event?
- How transparent are the authors for each article about their sources (e.g,name, function, circumstances of quote)?

	Imp. Recall (M1)			Imp. Precision (M2)			Conn. Recall(M3)			Conn. Precision(M4)		
	$\boldsymbol{\beta}$	**std.**	**err.**	$\boldsymbol{\beta}$	**std.**	**err.**	$\boldsymbol{\beta}$	**std.**	**err.**	$\boldsymbol{\beta}$	**std.**	**err.**
(Intercept)	0.32***		0.08	0.21***		0.04	0.34***		0.05	0.37***		0.05
CEK[Low]	0.35**		0.13	0.05		0.07	0.07		0.06	0.02		0.07
VML[Low]	0.18		0.12	-0.02		0.06	-0.04		0.06	0.02		0.07
	R^2=0.12			R^2=0.01			R^2=0.04			R^2=0.19		
N_{obs}=62										*p<0.05, **p<0.01, ***p<0.001		

Table B.1: Linear models of recall and precision for connection-making and importance detection with user characteristics as predictors.

	Qual(M5)		*Cred*(M6)	
	β	**std. err.**	β	**std. err.**
(Intercept)	0.67***	0.05	0.65***	0.05
Group[Treat.]	0.01	0.07	0.11	0.07
Article[Abortion(R)]	0.04	0.07	0.01	0.07
Article[Immigration(L)]	0.03	0.07	0.11	0.07
Article[Immigration(R)]	-0.01	0.05	0.04	0.05
Group [Treat.] * Article[Abortion(R)]	-0.07	0.09	-0.17*	0.08
Group [Treat.] * [Immigration(L)]	0.00	0.09	-0.18	0.09
Group [Treat.] * Article[Immigration(R)]	0.02	0.08	-0.12	0.08
	$R^2 = 0.43$		$R^2 = 0.48$	
$N_{user} = 109$, $N_{article} = 4$, $N_{obs} = 218$			*p<0.05, **p<0.01, ***p<0.001	

Table B.2: Mixed-effects regression on quality and credibility score using the interaction of experimental condition and articles.

- How would you compare the comprehensibility of the article pair (e.g., simplicity in terms/phasing, conciseness, coherence)?
- How would you compare impartiality in content presentation between the articles (balanced viewpoints and actors, article
- author personally evaluating/judging the reported situation)?
- How would you compare ethical standards (e.g., discriminating any party involved, neutral phrasing) between the reports?
- Whose perspective this article represents more than others? Is there any particular group/party/side that the article focus to represent compared to the other article?

B.2 Articles Used in the Deployment

- E_1L: Immigration (Left)
- E_1R: Immigration (Right)

- E_2L: Abortion (Left)
- E_2R: Abortion (Right)

B.3 Effect of User Characteristics

We modeled user characteristics to predict precision and recall in annotation tasks, shown in table B.1. We accounted for several factors in these models, including users' demographic characteristics (age, gender, education, and political affiliation) and news expertise metrics (CEK and VML).

B.4 RQ2: Mixed-Effects Models

Besides ANOVA, we also performed a series of mixed-effects regression model on users' quality and credibility perception using experimental variables, in Table B.2. Similar to ANOVA results, we found significant interaction effect on credibility only for high contrast article.

Appendix C

OtherTube

C.1 Distribution of Users Who Passed Eligibility Criteria and Signed Up

Total 318 users signed up for our study. Figure C.1 shows the demography of these users.

C.2 Need for Reflection and Insight Questionnaire

These questions have been taken from Halttu et. al. [183].

C.2.1 Need for Self-Reflection

- I am not really interested in analyzing my behavior. (R)
- It is important for me to evaluate the things that I do.
- I am very interested in examining what I think about.
- It is important to me to try to understand what my feelings mean.
- I have a definite need to understand the way that my mind works.
- It is important for me to be able to understand how my thoughts arise.

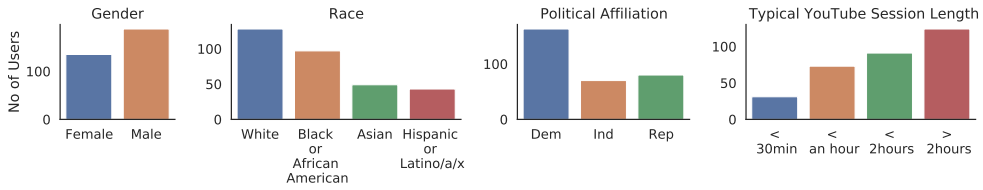


Figure C.1: Demography of the participants who passed screening criteria and signed up for the study.

C.2.2 Insight

- I usually have a very clear idea about why I've behaved in a certain way.
- My behavior often puzzles me. (R)
- Thinking about my thoughts makes me more confused. (R)
- Often I find it difficult to make sense of the way I feel about things.(R)
- I usually know why I feel the way I do.

C.3 Semi-Structured Interview Questions

- How do you typically use YouTube to find new content?
- How would you describe the benefits and limitations in the features provided by YouTube to find new content?
- Can you walk me through your recommendation feed? How did you use features [sharing video, setting persona, browsing OtherTube] in OtherTube?
- What were your impressions when you saw strangers' profiles and recommended videos?
- Could you describe any profile/recommended videos from the strangers during the study that stood out or were memorable to you? Why?

- How would you compare strangers' recommendations to your YouTube recommendation?
- Would you continue to use OtherTube after the study? If so, what would be the purpose/motivation? If not, why not?
- Among the features provided in OtherTube about choosing your persona, what was important for you to present about yourself to others?
- How would you describe the features offered by the tool to share/remove recommended videos from your feed?
- What concerns would you have regarding sharing your profile and recommendation?
- Is there anything else where the tool should be more transparent about?
- Other than using the plugin, was there anything that you had to do and you have not done regularly for this study?
- What did you like about the tool? What did you not like about the tool? Do you have any suggested changes on the tool for us to improve it?