

# Two Truths and a Lie: Exploring Soft Moderation of COVID-19 Misinformation with Amazon Alexa

Donald L. Gover  
 DePaul University  
 Chicago, Illinois, USA  
 dgover@mail.depaul.edu

Filipo Sharevski  
 DePaul University  
 Chicago, Illinois, USA  
 fsharevs@cdm.depaul.edu

## ABSTRACT

In this paper, we analyzed the perceived accuracy of COVID-19 vaccine Tweets when they were spoken back by a third-party Amazon Alexa skill. We mimicked the soft moderation that Twitter applies to COVID-19 misinformation content in both forms of warning covers and warning tags to investigate whether the third-party skill could affect how and when users heed these warnings. The results from a 304-participant study suggest that the spoken back warning covers may not work as intended, even when converted from text to speech. We controlled for COVID-19 vaccination hesitancy and political leanings and found that the vaccination hesitant Alexa users ignored any type of warning as long as the Tweets align with their personal beliefs. The politically independent users trusted Alexa less than their politically-laden counterparts and that helped them accurately perceiving truthful COVID-19 information. We discuss soft moderation adaptations for voice assistants to achieve the intended effect of curbing COVID-19 misinformation.

## KEYWORDS

Alexa, Misinformation, COVID-19, Twitter Soft Moderation

### ACM Reference Format:

Donald L. Gover and Filipo Sharevski. 2021. Two Truths and a Lie: Exploring Soft Moderation of COVID-19 Misinformation with Amazon Alexa. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnn>

## 1 INTRODUCTION

Misinformation and rumors about viruses are not a new thing. What's new is the size of the target audience and the speed with which such damaging information spreads online. It took weeks and a publication in foreign media outlets for the KGB-initiated "Operation Infektion," spreading the rumour that HIV/AIDS was a misfired American biological weapon in the 1980s, to gain traction and manipulate the public opinion, at least outside of the U.S. [7]. Fast forward 40 years, one can choose an aspect of the novel COVID-19 virus misinformation (e.g. origin, vaccines' adverse effects, efficacy) as well as audience (e.g. Twitter, Facebook, Parler)

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*Conference'17, July 2017, Washington, DC, USA*  
 © 2021 Association for Computing Machinery.  
 ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00  
<https://doi.org/10.1145/nnnnnnnn.nnnnnnn>

and manipulate the public opinion in a matter of seconds both in the U.S. and everywhere in the world [18, 25].

While there was no structure of countering the old-style vaccination rumors, a coordinated soft moderation of misinformation was rather quickly put in place during the COVID-19 pandemic on the mainstream social media platforms [29]. Such an effort was warranted because the real world implications of (mis)information and unverified rumors have a direct impact on public health, especially of mass immunization. According to Twitter, who opted to use warning tags and covers as misinformation labels, the supposed aim of the soft moderation is to reduce misleading or harmful information that could "incite people to action and cause widespread panic, social unrest or large-scale disorder" [29].

Consequently, the academic community set forth to investigate whether this seemingly noble aim was in fact achieved among the social media users and the general public. In an early study, the soft moderation labels were found to have a counter effect since the warnings convinced people to believe the misinformation even more than if the labels were not there [11]. A study exploring this so-called "soft moderation" implemented by Twitter found that the platforms' content with warning labels generated more action than content without said labels [43]. The soft moderated Tweets were more likely to be distributed than a "valid" Tweet through discourse, not always because the soft moderated Tweets contained misinformation but because they contained a response to mock or disclaim an original author or a valid Tweet. Interestingly, a mere 1% of the Tweets gathered for the study were labeled with a COVID-19 warning and a number of these few Tweets were found to be mislabeled simply because they contained the words "oxygen" and "frequency". Another study, in this context, found that some users did not trust the soft moderation intervention and felt that Twitter itself was biased and mislabeling content [16]. Even without the warning labels, a varying degree of users' perceptions and beliefs regarding vaccines in general and about COVID-19 vaccines in particular plays a role in individuals' reaction to (mis)information Tweets. A study on vaccine misinformation spreading and acceptance on social media platforms showed the effectiveness of "influential users" within like minded vaccine-fearing followers to be very high [12].

Mainstream social media platforms like Twitter allow for visual discernment of the information including the warning labels, formation of so-called "influencer" accounts, and direct communication of the engagement with the content metrics such as number of replies, re-Tweets, likes, and shares. While all of these factors certainly affect the receptivity of any COVID-19 vaccine information, little attention is devoted to exploring how people respond to soft moderated COVID-19 vaccine Tweets when these are delivered through a voice assistant like Amazon Alexa. Unlike the traditional

Twitter interface, Alexa is the sole authority or “influencer” that delivers the Tweet when prompted possibly without disclosing the details of any labeling or engagement metrics (depending on the configuration for retrieval and presentation of one’s twitter diet). Users usually trust Alexa and worry only about Alexa intruding into their privacy, but not about the validity of the information delivered by Alexa [21]. Since the configuration of how one’s twitter diet is spoken-back by Alexa can be customized and implemented by various third-party applications, called “skills” we set to explore how users will respond when such a third-party skill is used to deliver soft moderated COVID-19 vaccine Twitter content. Studies in the past have shown that third-party skills could be dangerous in that can silently rephrase information from any source to mislead a user and induce misperception about a polarizing topic such as vaccination, free speech, or government actions [31–33].

## 2 SOFT MODERATION BY ALEXA

Malicious skills for Alexa, unsurprisingly, exist on the Amazon Skills Store [40]. Some of them pray on malicious voice commands (songs, phonemes, inaudible noises) to be interpreted as legitimate by Alexa and invoke a skill that snoops on the user or attempts to steal their credentials [8, 20, 42]. Some of them try to lure the user to spell their password directly to Alexa in the voice variant of phishing or “vishing” [30]. Some of them don’t bother with the invocation logic or users’ credentials, but instead, manipulate the content that is spoken back to the users [34]. This new type of malicious skills, in a man-in-the-middle fashion, rewords a legitimate content in such way that the user makes an improbable interpretation of a set of true facts. The goal of this skill is to induce misperception, that is, to distort the reality of a targeted user in a similar way information operations attempt to do with disseminating alternative narratives on social media [37]. The authors of this proof-of-concept malicious skill, manipulated regulatory news to present the government in two different lights as a pro-worker and as a pro-business. A user study was conducted that validated the potential of the malicious skill to induce misperception about the government’s proclivity towards workers and the businesses.

The possibility for manipulating spoken back content by the voice assistant prompted us to think how such a skill could be weaponized if the content comes from Twitter. Twitter is one the main “battlefields” for information operations with numerous alternative narratives surrounding the COVID-19 pandemic. Recognized that these alternative dangerously pollute the public discourse on the platform and started to actively label them as “COVID-19 misinformation” [29]. The labels come in two forms: (1) a warning cover with a verbose text hiding the text and allowing the user to skip a COVID-19 misinformation Tweet; and (2) a warning tag underneath a content providing a link for users to “get the facts about COVID-19.” Twitter is not just providing content but also provides context on its truthfulness to the user with such label. It has been found that similar labels on social media reduce the perceived accuracy of an alternative narrative, e.g. it help users recognize an attempt for inducing misperception [11].

Circling back to the possibility for inducing misperception by manipulating content by a malicious skill, we are also interested in the ability of this skill to suppress or deliberately insert such labels

in order to enhance or attenuate the misperception-inducing effect of a selected Twitter content. In other words, we wanted to know how the user will perceive a spoken back Twitter content by the malicious skill in the following scenarios: (1) Alexa utters the warning label text and speaks back the content of a misinformation/valid Tweet; (2) Alexa speaks back the content of a misinformation/valid Tweet and then utters the warning tag; and (3) Alexa suppresses the warning label text/tag and only speaks back the content of a misinformation/valid Tweet;

## 3 ALEXA TWEETER READER SKILL

The thrust of our research was not to fully implement a malicious Alexa skill that reads Tweets but rather to test the effect should one be developed (we elaborate the ethics of our research in the Discussion section), therefore, we produced a proof-of-concept that closely resembled such a skill. The “Twitter Reader” skill could automatically interact with both the user and Twitter adjusting the interaction depending on the “blueprint” selected for the user interaction [5]. The Amazon Alexa skill ecosystem has a broad selection of “blueprints” or prepackaged skills from which one can easily deploy voice activated applications to the Alexa with minimal effort. Initially for this study we started using the Flash Briefing skill which allowed us to present to a user a predefined set of Tweets. This however was limited as having to statically define the Tweets. The requirements for dynamic content retrieved from Twitter, possible presentation of multimedia files, and conditional responses quickly outgrew the Flash Briefing skill and a more sophisticated platform was required. We migrated our implementation to VoiceFlow where we developed the Twitter Reader skill flow shown in Figure 1 [2].

To invoke the Twitter Reader skill the user utters the skill’s invocation name Alexa, open the Twitter Reader. Alexa response with the welcome message Welcome to the Alexa Official Twitter Reader and ask the name of the user. This is done to allow for reading Tweets from multiple accounts in one household, for example. Once Alexa connects the name to the configured Twitter account, Alexa prompts the user whether they want to hear the latest Tweet in their Tweeter stream. Assuming a confirmation answer from the user, Alexa retrieves the Tweet together with any warning labels attached to it. If a COVID-19 misinformation warning label is appended at the end of the Tweet, Alexa first converts the Tweet’s text into speech, clearing special characters like “@” for account links and “#” for hashtags, and responds, for example A second nytimes article quotes doctors who say the mRNA technology used in COVID vaccines may cause immune thrombocytopenia, a blood disorder that last month led to the death of a Florida doctor after getting the Pfizer vaccine (this Tweet is also shown in Figure 2a). Once done reading the Tweet, Alexa appends the warning tag, removing the exclamation mark favicon, and utters Get the information about COVID nineteen.

If a warning cover precedes the Tweet, Alexa first converts the text of the warning to speech and utters it back to the user. In the case of COVID-19 misinformation, this cover reads: This Tweet violated the Twitter Rules about spreading misleading and potentially harmful information related to COVID-19. However, Twitter has determined that it may be in the

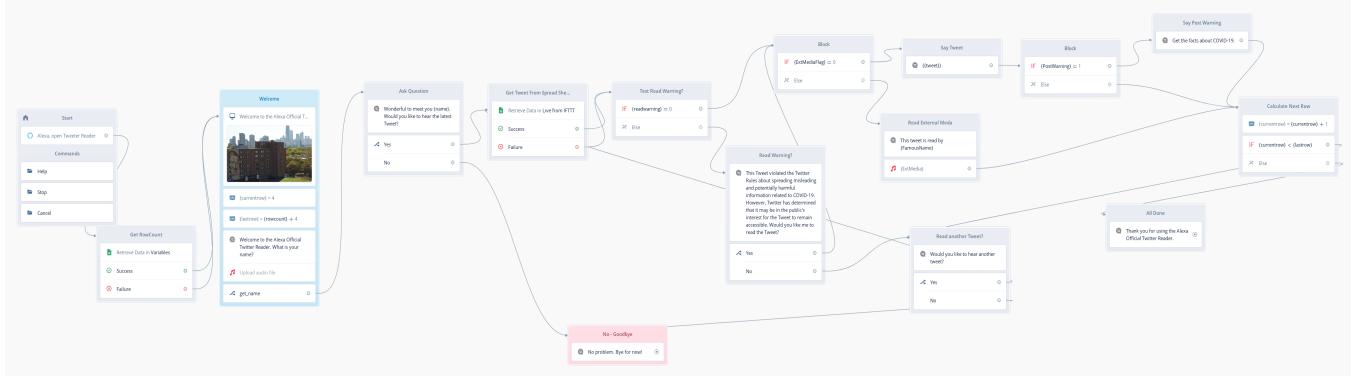


Figure 1: The Alexa Twitter Reader Skill (logical flow)

public's interest for the Tweet to remain accessible. Would you like me to read the Tweet?. If the user says Yes, Alexa proceeds and converts the Tweet's text into speech as in the scenario shown in Figure 3. If the user says No, then Alexa responds Would you like to hear another Tweet?. If the user again declines to hear another Tweet, Alexa simply responds Thank you for using the Alexa Official Twitter Reader. Otherwise, Alexa proceeds and selects the next Tweet in the stream.

## 4 RESEARCH STUDY

### 4.1 Misperceptions: Accuracy of Tweets

To test the misperception inducing-effect we investigated whether (mis)information Tweets about the COVID-19 vaccine efficacy, spoken back by Alexa and in the presence or absence of warning labels affect individuals' perceptions of the Tweet's accuracy with the following set of hypotheses:

- H1: The utterance of a warning tag following a Tweet containing *misleading* information about COVID-19 vaccines by Alexa will not reduce the perceived accuracy of the spoken back Tweet's content relative to a no warning tag condition.
- H2: The utterance of a warning cover before a Tweet containing *misleading* information about COVID-19 vaccines by Alexa will not reduce the perceived accuracy of the spoken back Tweet's content relative to a no warning cover condition.
- H3: The utterance of a warning tag following a Tweet containing *valid* information about COVID-19 vaccines by Alexa will not reduce the perceived accuracy of the spoken back Tweet's content relative to a no warning tag condition.
- H4: The utterance of a warning cover before a Tweet containing *valid* information about COVID-19 vaccines by Alexa will not reduce the perceived accuracy of the spoken back Tweet's content relative to a no warning cover condition.

To test the first hypothesis we utilized the Tweets containing *misleading information* shown in Figure 2a and Figure 2b. The Tweet in Figure 2a has a warning tag indicating that the above content is misinformation. The Tweet promulgates COVID-19 rumour about a rare adverse effect that was linked to the SARS-CoV-2 virus, not

the vaccine, at the time of writing [9]. To remove bias due to the "influencer" effect, the Tweet comes from a verified account "The-Vaccinator" (which we made up). An alteration of the same Tweet is shown in Figure 2b without the accompanying warning tag. To test the second hypothesis we utilized the Tweets shown in Figure 2b and Figure 3 (which includes a warning cover instead of a warning tag). In all the cases, the content of the Tweet and the warning covers/tags were converted from text-to-speech by the malicious Alexa skill as described above.

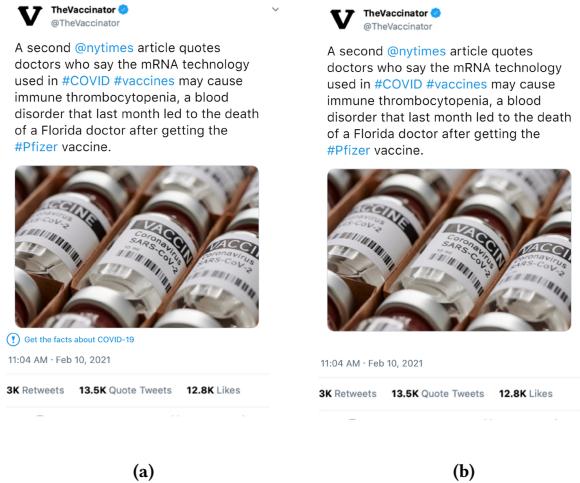


Figure 2: A Misleading Tweet: (a) With a Warning Tag; (b) Without a Warning Tag for Misleading Information.

To test the third hypothesis we utilized the Tweets containing *verified information* shown in Figure 4a and Figure 4b. The original Tweet content cites the CDC's announcement about proceeding with the second dose of the COVID-19 vaccine in case an individual has a serious reaction from the first dose, altered to include a warning tag in Figure 4b [13]. To control for bias, the Tweet comes from a verified account "TheVirusMonitor" instead of the CDC account. To test the fourth hypothesis we utilized the Tweets containing *verified information* shown in Figure 4a and Figure 5. We retained

**Figure 3: A Warning Cover Preceding a Misleading Tweet**

Figure 4a for the comparison of the conditions and altered the labeling in the Figure 4b to include a warning cover instead of a warning tag. Similarly, the text from the twitter content and the labels was converted as a text-to-speech by Alexa.

**Figure 4: A Verified Tweet: (a) Without a Warning Tag; (b) With a Warning Tag for Misleading Information.**

## 4.2 Misperceptions: Hesitancy and Political Leanings

We also controlled for hesitancy and political leanings, following the moderating effect of these variables reported in [43], [27], to examine the perceived accuracy of spoken-back Tweets with COVID-19 vaccine information in presence/absence of uttering warning labels. We used the same Tweets as shown in Figures 2-5 together to test the following hypotheses:

- H5a: The COVID-19 vaccine personal hesitancy will not affect the perception of accuracy of a spoken back Tweet with *misleading*

**Figure 5: A Warning Cover Preceding a Verified Tweet**

information about COVID-19 in any condition (with a warning a preceding warning cover, with a following warning tag, or without any warning)

- H5b: The COVID-19 vaccine personal hesitancy will not affect the perception of accuracy of a spoken back Tweet with *valid* information about COVID-19 in any condition (with a warning a preceding warning cover, with a following warning tag, or without any warning)

To test the association between one's political leanings and the perceived accuracy of the Tweets from Figures 2-5 we asked:

- RQ1: Is there a difference in the perceived accuracy of spoken-back COVID-19 misleading/verified Tweets with warning labels (tags or covers) between Republican, Democrat, and Independent users?

## 4.3 Sampling and Instrumentation

We set to sample a population using Amazon Mechanical Turk that is 18 years or above old, is a Twitter user and an Amazon Alexa user, and has encountered at least one Tweet into their feed that relates to COVID-19 vaccines. Because we were not allowed to physically invite the participants in our lab, we recorded the interaction between a user prompting Alexa to open the Twitter Reader skill and read the Tweets, which was offered as a recording to each participant. The study was approved by the Institutional Review Board (IRB) as a non-full disclosure experiment. Consequently, participants were initially told that they are being asked to gauge the effectiveness or usability of the Alexa skill as an experimental COVID-19 Tweeter Reader. After participation, each participant was debriefed and offered small compensation.

We crafted the content of the Tweets to be of relevance to the participants such that they meaningfully engage with the Tweet's content (i.e., their responses are not arbitrary). We assumed participants understood the Amazon Alexa interfaces and were aware of the COVID-19 pandemic in general. However, we acknowledge that the level of interest regarding the COVID-19 vaccines could vary among the individual participants, affecting the extent to which

their responses reflect their opinions. To assess the perceived accuracy, we used the questionnaire from [11], adapted to the content presented in Figures 2-5. The questionnaire assesses the perceived accuracy of each of the Tweets on a 4-point Likert scale (not at all accurate, not very accurate, somewhat accurate, very accurate).

To assess participants' subjective attitudes and beliefs regarding the COVID-19 vaccine, we used a modified questionnaire from [6]. To assess the subjective attitudes we asked if the participants will receive a COVID-19 vaccine (Yes/No/I Don't Know); and (c) politically speaking, do they consider as (Republican/Democrat/Independent). The overall questionnaire was anonymous with no personally identifiable data collected from the participants. We utilized a 2x3 experimental design where participants will be randomized into one of six groups: (1) spoken back misleading Tweet with a follow up warning tag; (2) spoken back misleading Tweet without a warning tag suppressed by Alexa; (3) spoken back misleading Tweet preceded by an utterance of a warning cover; (4) spoken back verified Tweet; (5) spoken back verified Tweet with a follow up warning tag; and (6) verified Tweet preceded by an utterance of a warning cover.

## 5 RESULTS

We conducted an online survey ( $N = 304$ ) in January and February 2021. There were 195 (64.1%) males and 103 (33.9%) females, with 6 participants (2.0%) identifying as non-cis. The age brackets in the sample were distributed as follows: 26 (8.6%) [18 - 24], 112 (36.8%) [25 - 34], 106 (34.9%) [35 - 44], 35 (11.5%) [45 - 54], 18 (5.9%) [55 - 64], 6 (2.0%) [65 - 74], and 1 (.3%) [75 - 84]. Our sample was younger-leaning and slightly skewed towards male Twitter and Alexa users. The sample was also slightly Democrat-leaning with 51 (16.8%) Republicans, 173 (56.9%) Democrats, and 80 (26.3%) Independent.

### 5.1 Misperceptions: Accuracy of Tweets

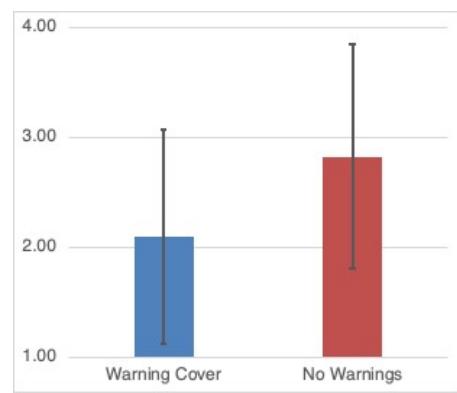
We first hypothesized that the utterance of a warning tag following a Tweet containing *misleading* information (Figure 2a) about COVID-19 vaccines by Alexa will not reduce the perceived accuracy of the spoken back Tweet's content relative to a no warning tag condition (Figure 2b). We have to confirm this hypothesis since we haven't found any statistically significant result in the perceived accuracy ( $U = 1288.5, p = .917, (\alpha = 0.05)$ ). In practical terms, this means that the follow up warning tag doesn't work as intended - both groups perceived the Tweet as "somewhat accurate" on average. Alexa users either ignored the warning tag utterance at the end or perhaps the warning tag saying Get the facts about COVID nineteen is ambiguous and logically doesn't actually say that the particular Tweet is indeed misinformation.

Granted, tagging the content is primarily intended for visual inspection and includes a warning exclamation favicon and a link to the COVID-19 verified information, which is not available directly to the users. We did allow for the participants to repeat the interaction with Alexa and change their reported perception of accuracy, but that didn't seem to matter. Perhaps, for a future use, Alexa could proceed and turn the warning tag into a question, prompting the user with This Tweet was labeled as COVID nineteen misinformation. Would you like to hear the facts about the COVID nineteen?. A case for such an adaptation also brings

the test result for the third hypothesis, which was also didn't yield a significant result ( $U = 1177, p = .490, (\alpha = 0.05)$ ). While it seems that the Alexa users correctly ignored the tag because the Twitter Reader skill inserted the warning tag utterance after it spoke back the Tweet as shown in Figure 4b, we don't know if this is because the Alexa users considered the warning tag at all.

We also hypothesized that the utterance of a warning cover before a Tweet containing *misleading* information about COVID-19 vaccines by Alexa (Figure 3) will not reduce the perceived accuracy of the spoken back Tweet's content relative to a no warning cover condition (Figure 3a). Here too, we didn't find a statistically significant result in the perceived accuracy ( $U = 1280, p = .889, (\alpha = 0.05)$ ). This made us speculate if the perhaps the content of the Tweet was more so perceived as potentially an unverified rumor rather than a misinformation altogether (at least at the time of the study there was no official repudiation of the link between the COVID-19 vaccine and the death caused by the immune thrombocytopenia). What made us think of the nuanced choice of misinformation and COVID-19 rumors is the rejection of the fourth hypothesis where the utterance of the warning cover before Alexa spoke back the information from the valid Tweet shown on Figure 5 ( $U = 1601, p = .002^*, (\alpha = 0.05)$ , Cohen's  $d = 0.619$  medium).

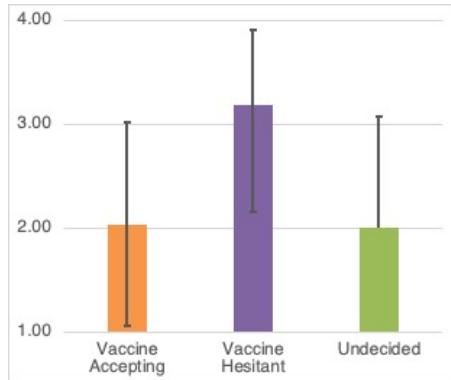
The cover was sufficiently potent for the users in the test group to perceive the following Tweet as "not very accurate" while the ones in the control condition as "somewhat accurate" as shown in Figure 6. This result suggest that the warning cover is a potent way of swaying users about potential misinformation, even if the following Tweet containing in fact verified information released by the CDC. This result, considering the possibility of developing a purely malicious Twitter Reader skill, could inject misinformation warning covers to a Tweet of choice and meddle with the public opinion on the COVID-19 vaccine effectiveness. A similar effort, though with actual Tweets and not warning labels, already surfaced on Twitter promoting a homegrown Russian vaccine and undercutting the vaccines from the rivals such as AstraZeneca [14].



**Figure 6: A difference in perceived accuracy between the group exposed to a warning cover utterance (Figure 5) and the group exposed only to the spoken back Tweet (Figure 4a)**

## 5.2 Misperceptions: Hesitancy

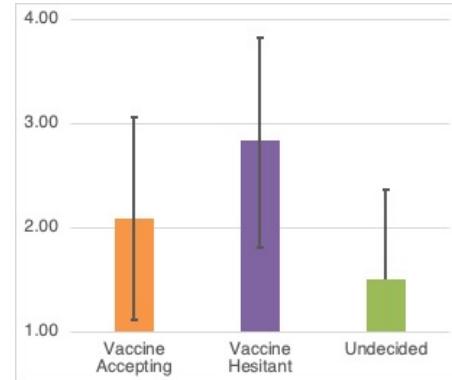
Hesitancy of vaccination has been shown to be a discriminative lens through which people critically discern online information, which also includes COVID-19 Tweets on the COVID-19 topic [6]. Therefore, we wanted to test and see if the COVID-19 vaccine hesitancy will affect the perception of accuracy of a spoken back Tweet with *misleading* information about COVID-19 in any condition (with a warning a preceding warning cover, with a following warning tag, or without any warning). When controlling for hesitancy, we found a significant difference only in the warning cover condition (Figure 3). The vaccine hesitant participants were more likely to perceive the spoken back misleading Tweet as “somewhat accurate”, while the vaccine accepting and undecided as “inaccurate” as shown in Figure 7 ( $\chi^2(2) = 10.058, p = .007^*, (\alpha = 0.05)$ ). It seems that the vaccine hesitant participants were alerted by the warning cover to contextually access the grim outlook linking the COVID-19 vaccine with death with their sceptic outlook of the COVID-19 vaccination, otherwise we would have observed a similar result for the the warning tag condition.



**Figure 7:** Vaccine hesitancy and perceived accuracy of the misleading Tweet in Figure 3.

That this is a plausible explanation of the obvious misperception suggest the similar test of the verified Tweet condition, where we also found a significant result only for the condition where a warning cover was uttered before the verified Twitter is spoken back by Alexa (Figure 5). Again, the vaccine hesitant participants were more likely to perceive the spoken back Tweet, this time containing verified information from the CDC, as “somewhat accurate”, while the vaccine accepting and undecided as “inaccurate” as shown in Figure 8 ( $\chi^2(2) = 6.432, p = .040^*, (\alpha = 0.05)$ ). We previously noted that the warning cover, spoken back by Alexa, seems sufficiently potent to add credibility to Alexa as the “best friend forever” [28]. After all, the best friends don’t lie and we trust them when they warn us that what they have found on Twitter might be a disinformation if they say so. But if this Twitter content aligns with our deeply held beliefs that the COVID-19 vaccines are bad and we for sure won’t get vaccinated, we might disregard what Twitter has told Alexa to convey to us, and instead believe the content. After all, we might have heard that many users feel Twitter itself is biased and mislabeling content [16], so who knows, maybe Alexa

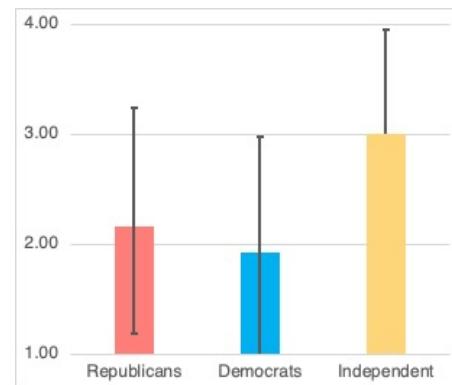
stumbled upon a mislabeled Tweet with an ominous COVID-19 misinformation warning cover.



**Figure 8:** Vaccine hesitancy and perceived accuracy of the verified Tweet in Figure 5.

## 5.3 Misperceptions: Political Leanings

The COVID-19 pandemic was not “immune” from politicization, expectedly, as a highly polarizing topic. Curious to test the early evidence that perceptions are modulated by one’s political leanings [10, 23, 43], we controlled for participant’s political identification to compare the perceived accuracy of the spoken back content by Alexa with and without warning covers/tags. We only found a statistical difference in for the condition where Alexa utters a warning cover before the valid Tweet (Figure 5). While the republican and democrat users took the Alexa “warning” and reported that this content is “not so accurate” the independent ones dismissed this warning and perceived this content as “somewhat accurate” as shown in Figure 9 ( $\chi^2(2) = 6.171, p = .046^*, (\alpha = 0.05)$ ).



**Figure 9:** Vaccine hesitancy and perceived accuracy of the verified Tweet in Figure 5.

This result contrast the divided receptivity to allegedly misinformation content between the republicans (further drawn to believe it in presence of warnings) and democrats (repelled by the warnings) previously reported. But such a comparison might not hold relevant

in our case first because the warnings and content in the previous studies focused on political social media content and second because the information interface was a screen, not a voice assistant. It might be expected for politically-laden users to assume the authority and trust Alexa as telling the truth so it comes to no surprise that they heeded the warning cover and payed little attention to the following Tweet content. The higher level of critical discernment for the independent might be the varied exposure to information to both left, center, and right leaning media and Twitter content in general [27].

## 6 DISCUSSION

Exploring the soft moderation when communicated by Alexa instead of a label on a screen, we found that warning covers when converted from text to speech and uttered by Alexa work as intended in grabbing the attention of the Alexa users. All would have been good and promising, if it wasn't for the malicious skill inserting these covers as utterances before Alexa spoke back a Tweet content that was true and validated by the CDC. The users in our study trusted Alexa and it's seemingly benign third-party skill "Twitter Reader" to tell the truth without critically discerning the Tweets spoken back by Alexa. We explored what might cause such an effect and found a difference in perception based on one's political leanings - the users considering themselves independent were the ones actually trying to make sense of the Tweets regardless of what Alexa (e.g. Twitter) had to say about the content. Alexa in our study wasn't able to dispel any biases that were rooted in personal beliefs [22], [39]. One's personal hesitancy from COVID-19 vaccination sufficed for biased perception of the information from Alexa despite any labeling as long as the Tweets echoed their sceptic outlook of the whole COVID-19 vaccination business [23, 26].

### 6.1 Ethical Implications

While we set to investigate the effect of a third-party skill that arbitrary manipulates the soft moderation covers and tags attached by Twitter to dubious COVID-19 vaccine content and debriefed the participants at the end of the study, the results could have several ethical implications nonetheless. We exposed the participants to moderated content that could potentially affect participants' stance on COVID-19 vaccination and the pandemic. The exposure might not sway participants on their vaccine hesitancy in the long run, but could make the participants reconsider their approach of obtaining the vaccine for themselves or their families. The exposure to Twitter content spoken back by Alexa as a novel interface could also affect the participants' stance on communicating important topics with their voice assistants. A recent study found that personification of Alexa is associated with increased levels of satisfaction, so learning that Alexa could be secretly controlled by a malicious third-party skill to insert or drop Twitter misinformation warnings might affect the user satisfaction with Alexa [28].

That a portion of the participants were able to critically discern the verified content even in presence of the maliciously inserted Alexa's warnings is reassuring and proves that users keep a critical mindset despite the proclivity for anthropomorphism towards Alexa [41]. However, the ease of crafting a malicious skill that could suppress or add warning utterances, could have unintended

consequences. With the evidence of nation-states disseminating misinformation, it is possible that they could resort to malware for voice assistants to avoid both the soft and hard moderation on social media platforms like Twitter [38]. Sure, this could be far from the realm of possibility, even if the capabilities exist, but for such a sensitive topic as COVID-19 vaccination, meddling with spoken-back content from Alexa could give an edge to a vaccine competitor in the global race for development and procurement of a COVID-19 solution. We certainly condemn such ideas and such a misuse of our research results and therefore only presented a very limited proof-of-concept flow for such a malicious skill. For example, evidence for such a misinformation campaign has already surfaced on Twitter, promoting homegrown Russian vaccines and undercutting rivals [14]. One could also point the malicious third-party skill to pull content from Parler instead of Twitter and package the disinformation vector as a skill that reads the "real" truth about the vaccine and avoid soft moderation altogether. Again, we condemn development of such a capability outside of academic research, even if it doesn't pose an imminent problem for the defense community [36].

### 6.2 Future Research

We acknowledge that there is further research to be done into investigating the manipulation of soft moderation and communication of Twitter content through voice assistants. Visually assessing warning labels might have a different effect when they are spoken by Alexa so it will be informative to see how a variation of the warning text could affect the perceived accuracy of content to which they are attached. Our results suggest that the follow up warning tag utterance has no effect at all due, we suspected, the ambiguous language. If perhaps Alexa speaks with a different tone and is more direct, saying This Tweet seems like is spreading misinformation. Do you want me to check the COVID-19 facts on the official CDC website?, one might hypothesize that the intended effect of soft moderation be better achieved. Certainly, the proposed adaptation entails extensive usable security research to determine the optimal way of delivering soft moderation warnings, especially with the option for Alexa to express emotions (e.g. "disappointed" and low volume in uttering the warning covers or "disappointed" and high volume in uttering the said tag [15]). This is yet another step in future research, and we plan to broadly explore the domain of voice-based security warnings.

Along this line and the politicization of COVID-19 debate, an interesting experiment we plan to conduct is to go beyond the emotions of Alexa but instead test a voice application that reads the Tweets in the voice of a famous political person. Siri, Apple's voice assistant allows for customization with accents (e.g. Irish) so one would expect that the voice assistants in future could be configured to speak with the voice of our beloved celebrity or authoritative figures [35]. A neural network that could synthesize speech directly from inputted text exists and with a moderate training models can be created to generate the speech from almost any individual [1]. It would be interesting to see how users will respond to same Tweets and warning covers read by John Oliver, Hillary Clinton, Donald Trump, or Barack Obama (one could already test these voices on the Vocodes website [3]).

### 6.3 Combating Malicious Skills

The misperception-inducing logic is enabled by customizing, in a relatively easy way, a blueprint template and registering a seemingly benign skill. As discussed in [32], a thorough certification process could uncover the malicious logic and remove the skill from the Amazon Skills Store. Another solution is monitoring for suspicious skills' behaviour with a tool like the *SkillExplorer* proposed in [17]. Again, a skill can evade both certification and exploring by claiming that the rephrasing aims to communicate important COVID-19 vaccine information in an assistive way, for example, to non-native English speakers [19].

As an additional layer of protection, feedback from users post-release could help close this gap. Twitter similarly hopes to identify and address misinformation through the use of pre-selected user "fact checkers", piloted in its Birdwatch program [24]. Amazon could similarly crowd-source its Alexa skill moderation and allow users with a high "helpfulness" score, as in Birdwatch, to identify potentially malicious or misinforming skills for further review and removal. Though allowing users to flag skills may be helpful in eliminating misinformation, this crowd-sourced soft moderation could be exploited by malicious users to flag legitimate skills or hijacked by partisan users if a skill's content has been highly politicized.

In the context of malicious third party skills, instead of manipulating warning labels, a skill might be directed to an RSS feed that steadily promotes rumours and unverified COVID-19 vaccine information. One might not need to make a Twitter reader, but a Parler reader, to access a wealth of unverified claims about COVID-19 vaccination and supply Alexa with "Parler COVID-19 briefings". For one, a widely shared information tidbit by "influencers" about the COVID-19 vaccine on Parler is that it contains HIV [4]. We experimented with mostly rumours about COVID-19 in our study, but exposing participants to such blatant misinformation, spoken by a trusted intermediary (Alexa, and not Alex Jones), could possibly uncover important dynamics in the relationship or personification of Alexa as a "Best Friend Forever" [28].

### 6.4 Scope Limitations

The current study has important limitations. First, we limited our questions to couple of Tweets that were relevant to the state of the pandemic and mass immunization during the period of January–February 2021, which could be perceived with a different level of accuracy after a certain period of time. Overall, the findings may be specific to the effect the malicious skill has only on COVID-19 mass immunization and may not be generalizable to other topics. Second, though our survey asked participants whether they intend to receive a COVID-19 vaccine, we did not ask participants who answered in the affirmative how soon they intended to get vaccinated. An affirmative intention to vaccinate does not indicate an intention to vaccinate immediately and unconditionally, and therefore, the results cannot be interpreted as such. We likewise did not ask why participants who answered in the negative why they did not intend to get vaccinated or whether any factors could change this. A negative intention to vaccinate does not indicate an intention to never receive the COVID-19 vaccine, and likewise, these results should not be interpreted as such.

Third, regular Alexa or voice assistant participants in general may be desensitized to the spoken back information, which may have affected their perception of the COVID-19 vaccine irrespective of the rephrasing. Our experiment was limited to Alexa as a voice assistant of choice and the Twitter as a COVID-19 vaccination information. We were limited to evaluating the effects of COVID-19 vaccination only in the U.S., and this information might not be relevant for places where other vaccines (the AstraZeneca, Sinopharm or Gamaleya vaccines) are used. We were limited to the choice of soft moderation labels present at the time of the study that could change the perception of the spoken-back information if reworded later by Twitter. Finally, although we tried to sample a representative set of participants for our study using Amazon Mechanical Turk, the outcomes might have been different if we used another platform, or another type of sampling. Also, a larger sample size, one that was gender and politically balanced, could have provided a more nuanced view of Alexa as a "Tweeter Reader", but we had limited funding for this study.

## 7 CONCLUSION

In this study, we explored whether a third-party Alexa skill could successfully affect the perceived accuracy of COVID-19 vaccine Tweets and induce misperceptions in users by simply manipulating the soft moderation applied by Twitter. Additionally, we examined whether participants' vaccine hesitant sentiment and political leanings had any effect on the perception of accuracy of the Tweets spoken back by Alexa. Our findings suggest that users were most likely to be misled on Twitter information when Alexa utters a warning speech before it delivers the Tweet's content, regardless of its validity. Participants' perception of the accuracy of both misleading and verified COVID-19 Tweets appear warped by participants' personal biases - participants judged Alexa's accuracy by how closely the response aligned with their own sceptic beliefs on the subject. We found a significant difference in perceived accuracy only for the apolitical Alexa users, which appear not to contextualize any misinformation labeling before they actually inspect a Tweet. All of our findings might be an effect of the lack of interaction or response users have when interacting with a voice assistant that they do not lack in online discourse on social media. Given the ease with which a user lacking developer experience can craft and share a third-party skill via the Amazon Skills Store, we believe it necessary to augment existing practices to catch malicious and misinforming skills like the one we showcased in this study. Likewise, we believe that the adaptation of the soft moderation with better verbal misinformation warnings may help break the confirmation bias feedback loop that reinforces listeners' biased vaccine outlooks.

## REFERENCES

- [1] 2021. <https://github.com/NVIDIA/tacotron2>
- [2] 2021. Visual Skill Building. <https://www.voiceflow.com/>
- [3] 2021. Voice Encoder. <https://vo.codes/#use>
- [4] Max Aliapoulos, Emmi Bevensee, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Savvas Zannettou. 2021. An Early Look at the Parler Online Social Network. arXiv:2101.03820v2 [cs.SI]
- [5] Amazon. 2019. Amazon Alexa | Skill Blueprints. <https://blueprints.amazon.com/>.
- [6] Luigi Roberto Biasio, Guglielmo Bonaccorsi, Chiara Lorini, and Sergio Pecorelli. 2020. Assessing COVID-19 vaccine literacy: A preliminary online survey. *Human*

- Vaccines & Immunotherapy* 0, 0 (2020), 1–9. <https://doi.org/10.1080/21645515.2020.1829315>
- [7] Thomas Boghardt. 2009. Operation INFESTATION: Soviet Bloc Intelligence and Its AIDS Disinformation Campaign. <https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/vols3no4/soviet-bloc-intelligence-and-its-aids.html>
  - [8] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. 2016. Hidden Voice Commands. In *25th USENIX Security Symposium (USENIX Security 16)*. USENIX Association, Austin, TX, 513–530. <https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/carlini>
  - [9] Bill Chappell. 2021. Instagram Bars Robert F. Kennedy Jr. For Spreading Vaccine Misinformation. <https://www.npr.org/sections/coronavirus-live-updates/2021/02/11/966902737/instagram-bars-robert-f-kennedy-jr-for-spreading-vaccine-misinformation>.
  - [10] Dino P Christensen, Sarah E Kreps, and Douglas L Kriner. 2020. Contemporary Presidency: Going Public in an Era of Social Media: Tweets, Corrections, and Public Opinion. *Presidential Studies Quarterly* (2020).
  - [11] Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. 2019. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior* (2019), 1–23.
  - [12] Jieyu D. Featherstone, George A. Barnett, Jeanette B. Ruiz, Yurong Zhuang, and Benjamin J. Millam. 2020. Exploring childhood anti-vaccine and pro-vaccine communities on twitter – a perspective from influential users. *Online Social Networks and Media* 20 (2020), 100105. <https://doi.org/10.1016/j.osnem.2020.100105>
  - [13] Centers for Disease Control and Prevention. 2021. COVID-19 Vaccines and Allergic Reactions. <https://www.cdc.gov/coronavirus/2019-ncov/vaccines/safety/allergic-reaction.html>.
  - [14] Sheera Frenkel, Maria Abi-Habib, and Julian E. Barnes. 2021. Russian Campaign Promotes Homegrown Vaccine and Undercuts Rivals. <https://www.nytimes.com/2021/02/05/technology/russia-covid-vaccine-disinformation.html>.
  - [15] Gao, Catherine. 2019. Use New Alexa Emotions and Speaking Styles to Create a More Natural and Intuitive Voice Experience. <https://developer.amazon.com/en-US/blogs/alexa/alexa-skills-kit/2019/11/new-alexa-emotions-and-speaking-styles>.
  - [16] Christine Geeng, Tiona Francisco, Jevin West, and Franziska Roesner. 2020. Social Media COVID-19 Misinformation Interventions Viewed Positively, But Have Limited Impact. arXiv:2012.11055 [cs.CY]
  - [17] Zhixiu Guo, Zijin Lin, Pan Li, and Kai Chen. 2020. SkillExplorer: Understanding the Behavior of Skills in Large Scale. In *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association, 2649–2666. <https://www.usenix.org/conference/usenixsecurity20/presentation/guo>
  - [18] Peter Jachim, Filipo Sharevski, and Paige Treebridge. 2020. TrollHunter [Evader]: Automated Detection [Evasion] of Twitter Trolls During the COVID-19 Pandemic. In *New Security Paradigms Workshop 2020* (Online, USA) (*NSPW '20*). Association for Computing Machinery, New York, NY, USA, 59–75. <https://doi.org/10.1145/3442167.3442169>
  - [19] Yeongjin Jang, Chengyu Song, Simon P. Chung, Tielei Wang, and Wenke Lee. 2014. A11Y Attacks: Exploiting Accessibility in Operating Systems. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (Scottsdale, Arizona, USA) (*CCS '14*). ACM, New York, NY, USA, 103–115. <https://doi.org/10.1145/2660267.2660295>
  - [20] Deepak Kumar, Riccardo Paccagnella, Paul Murley, Eric Hennenfent, Joshua Mason, Adam Bates, and Michael Bailey. 2018. Skill Squatting Attacks on Amazon Alexa. In *27th USENIX Security Symposium (USENIX Security 18)*. USENIX Association, Baltimore, MD, 33–47. <https://www.usenix.org/conference/usenixsecurity18/presentation/kumar>
  - [21] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, Are You Listening?: Privacy Perceptions, Concerns and Privacy-seeking Behaviors with Smart Speakers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 102 (Nov. 2018), 31 pages. <https://doi.org/10.1145/3274371>
  - [22] Amanda R. Mercadante and Anandi V. Law. 2020. Will they, or Won't they? Examining patients' vaccine intention for flu and COVID-19 using the Health Belief Model. *Research in Social and Administrative Pharmacy* (2020). <https://doi.org/10.1016/j.sapharm.2020.12.012>
  - [23] Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32, 2 (2010), 303–330.
  - [24] Barbara Ortutay. 2021. Twitter launches crowd-sourced fact-checking project. *Associated Press - AP News* (2021). <https://apnews.com/article/twitter-launch-crowd-sourced-fact-check-589809d4c9a7eceda1ea8293b0a14af2>
  - [25] Emma Peironi, Peter Jachim, Nathaniel Jachim, and Filipo Sharevski. 2021. Parler-monium: A Data-Driven UX Design Evaluation of the Parler Platform. In *Critical Thinking in the Age of Misinformation CHI 2021*.
  - [26] Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. 2020. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science* (2020).
  - [27] Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general* 147, 12 (2018), 1865.
  - [28] Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. 2017. "Alexa is My New BFF": Social Roles, User Satisfaction, and Personification of the Amazon Echo. In *CHI Extended Abstracts* (Denver, Colorado, USA) (*CHI EA '17*). Association for Computing Machinery, New York, NY, USA, 2853–2859. <https://doi.org/10.1145/3027063.3053246>
  - [29] Yoel Roth and Nick Pickles. 2020. Updating our approach to misleading information. [https://blog.twitter.com/en\\_us/topics/product/2020/updating-our-approach-to-misleading-information.html](https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html)
  - [30] Security Research Labs. 2019. Smart Spies: Alexa and Google Home expose users to vishing and eavesdropping. <https://srlabs.de/bites/smart-spies/>.
  - [31] Filipo Sharevski, Peter Jachim, and Kevin Florek. 2020. To Tweet or Not to Tweet: Covertly Manipulating a Twitter Debate on Vaccines Using Malware-Induced Misperceptions. In *Proceedings of the 15th International Conference on Availability, Reliability and Security (Virtual Event, Ireland) (ARES '20)*. Association for Computing Machinery, New York, NY, USA, Article 75, 12 pages. <https://doi.org/10.1145/3407023.3407025>
  - [32] Filipo Sharevski, Peter Jachim, Paige Treebridge, Audrey Li, Adam Babin, and Christopher Adadevoh. 2021. Meet Malexa, Alexa's malicious twin: Malware-induced misperception through intelligent voice assistants. *International Journal of Human-Computer Studies* 149 (2021), 102604. <https://doi.org/10.1016/j.ijhcs.2021.102604>
  - [33] Filipo Sharevski, Paige Treebridge, Peter Jachim, Audrey Li, Adam Babin, and Jessica Westbrook. 2020. Beyond Trolling: Malware-Induced Misperception Attacks on Polarized Facebook Discourse. arXiv:2002.03885 [cs.HC]
  - [34] Filipo Sharevski, Paige Treebridge, Peter Jachim, Audrey Li, Adam Babin, and Jessica Westbrook. 2020. Meet Malexa, Alexa's Malicious Twin: Malware-Induced Misperception Through Intelligent Voice Assistants. arXiv:2002.03466 [cs.CR]
  - [35] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvriannakis, and Y. Wu. 2018. Natural TTS Synthesis by Conditioning WaveNet on MEL Spectrogram Predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4779–4783. <https://doi.org/10.1109/ICASSP.2018.8461368>
  - [36] Toby Shevlane and Allan Dafoe. 2020. The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse?. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) (*AIES '20*). Association for Computing Machinery, New York, NY, USA, 173–179. <https://doi.org/10.1145/3375627.3375815>
  - [37] Kate Starbird. 2017. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter.
  - [38] Kurt Thomas, Chris Grier, and Vern Paxson. 2012. Adapting social spam infrastructure for political censorship. In *5th {USENIX} Workshop on Large-Scale Exploits and Emergent Threats ({LEET} 12)*.
  - [39] Emily Thorson. 2016. Belief echoes: The persistent effects of corrected misinformation. *Political Communication* 33, 3 (2016), 460–480.
  - [40] Tavish Vaidya, Yuankai Zhang, Micah Sherr, and Clay Shields. 2015. Cocaine Noodles: Exploiting the Gap between Human and Machine Speech Recognition. In *9th USENIX Workshop on Offensive Technologies (WOOT 15)*. USENIX Association, Washington, D.C. <https://www.usenix.org/conference/woot15/workshop-program/presentation/vaidya>
  - [41] Katja Wagner and Hanna Schramm-Klein. 2019. Alexa, are you human? Investigating anthropomorphism of digital voice assistants—A qualitative approach. *Robot Interactions and Interfaces* (2019).
  - [42] Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, XiaoFeng Wang, and Carl A. Gunter. 2018. Commandersong: A Systematic Approach for Practical Adversarial Voice Recognition. In *Proceedings of the 27th USENIX Conference on Security Symposium* (Baltimore, MD, USA) (*SEC '18*). USENIX Association, USA, 49–64.
  - [43] Savvas Zannettou. 2021. "I Won the Election!": An Empirical Analysis of Soft Moderation Interventions on Twitter. *arXiv* 2101.07183v1 (18 January 2021). <https://arxiv.org/pdf/2101.07183.pdf>.