Explaining Why Fake Photos are Fake: Does It Work?

MARGIE RUFFIN, University of Illinois at Urbana-Champaign, USA
GANG WANG, University of Illinois at Urbana-Champaign, USA
KIRILL LEVCHENKO, University of Illinois at Urbana-Champaign, USA

Today's disinformation campaigns may use deceptively altered photographs to promote a false narrative. In some cases, viewers may be unaware of the alteration and thus may more readily accept the promoted narrative. In this work, we consider whether this effect can be lessened by explaining to the viewer how an image has been manipulated. To explore this idea, we conduct a two-part study. We started with a survey (n = 113) to examine whether users are indeed bad at identifying manipulated images. Our result validated this conjecture as participants performed barely better than random guessing (60% accuracy). Then we explored our main hypothesis in a second survey (n = 543). We selected manipulated images circulated on the Internet that pictured political figures and opinion influencers. Participants were divided into three groups to view the original (unaltered) images, the manipulated images, and the manipulated images with *explanations*, respectively. Each image represents a single case study and is evaluated independently of the others. We find that simply highlighting and explaining the manipulation to users was not always effective. When it was effective, it did help to make users less agreeing with the intended messages behind the manipulation. However, surprisingly, the explanation also had an opposite (e.g., negative) effect on users' feeling/sentiment toward the subjects in the images. Based on these results, we discuss open-ended questions which could serve as the basis for future research in this area.

CCS Concepts: • **Networks** → **Social media networks**; • **Computing methodologies** → **Image manipulation**; • **Human-centered computing** → *Empirical studies in HCI*.

Additional Key Words and Phrases: human perception, psychological theory, disinformation explanation

ACM Reference Format:
Margie Ruffin, Gang Wang, and Kirill Levchenko. 2023. Explaining Why Fake Photos are Fake: Does It Work?. *Proc. ACM Hum.-Comput. Interact.* 7, GROUP, Article 8 (January 2023), 22 pages. https://doi.org/10.1145/3567558

1 INTRODUCTION

Social media has brought about a new era of citizen journalism by democratizing the ability to distribute content. But it has also eroded the editorial gatekeeping that was a hallmark of mainstream media, making those of us who get our news from social media more vulnerable to disinformation. Malicious actors exploit this editorial vacuum to spread false information and sow division [18, 40].

Among the more pernicious tools of this trade are photo and video manipulation, altering real photographs to produce convincing fakes that advance a false narrative [38]. What makes them particularly effective is that visual information is so much more compelling than text—a picture is worth a thousand words, after all—and can have an immediate impact on the viewer.

Authors' addresses: Margie Ruffin, University of Illinois at Urbana-Champaign, 1308 W Main St, Urbana, Illinois, USA, 61801, mruffin2@illinois.edu; Gang Wang, University of Illinois at Urbana-Champaign, 201 N Goodwin Ave, Urbana, Illinois, USA, 61801, gangw@illinois.edu; Kirill Levchenko, University of Illinois at Urbana-Champaign, 1308 W Main St, Urbana, Illinois, USA, 61801, klevchen@illinois.edu.

2573-0142/2023/1-ART8 $15.00
https://doi.org/10.1145/3567558

Another reason manipulated photographs are such a powerful influence tool is that we trust photographs. Barring clear evidence to the contrary, many of us still accept a photograph as a true representation of reality, so when we are presented with a well-crafted fake, we accept it as truth. In a recent study, Nightingale *et al.* [30] found that participants could not tell which photos were real and which were fake: participant accuracy was only 66% (compared to 50% for a random guess).

This suggests that if people are not good at telling real and fake photos apart, one way to lessen the effect of fake photos is to help people identify them. We ask a simple question in this paper: if we showed someone a fake photograph, *but explained how it was altered*, can we counteract the negative effect that the image manipulation aimed to achieve?

Before examining this question, however, we began by confirming our underlying assumptions, namely that people really are bad at telling real and fake images apart. To do so, we ran a study similar to the one carried out by Nightingale *et al.* to test participants' ability to identify manipulated photographs. The difference between our two studies is that we used manipulated photos that were actually circulated on the internet (whereas Nightingale *et al.* used lab-created fakes). We assume that people are bad at this task, not only because of the work from Nightingale *et al.*but also because of the sophistication of today's photo editing techniques [14, 22]. Image manipulation can appear in several ways. For the purpose of our study, we consider manipulation as changes made to images with the intention of significantly altering the perception of the subject of the image. Simple photo adjustments (e.g., adjusting the brightness or contrast, simple face touch-ups) are not considered. Our results confirm those of Nightingale *et al.*: participants were able to identify manipulated images with 60% accuracy.

The second assumption behind our question is that images do influence viewers' opinions. Although there is a strong body of work supporting this assumption [24, 25, 28, 35] , we wanted to determine whether actual fake images that circulated on the internet had the intended effect. In particular, we focus on intentionally manipulated images for the purpose of spreading disinformation and altering an opinion or invoking certain feelings from the viewers for political gain. We collected several fake photos from Reddit, Twitter, and the fact-checking site Snopes and then located the original photographs on which they were based, using reverse image search engines. This allowed us to test the differential effect of the manipulation of the viewer. Each image set and its results were evaluated independently as its own case study. In our survey, we showed one group of participants the original image and the other its manipulated derivative. We then asked participants two questions related to the images: first, whether they agreed with a statement that (in our judgment) the image alteration intended to convey. Second, we asked the participants' opinions about the person pictured in the photo. We had mixed results: manipulation swayed participants to be more agreeing with the intended statement in some cases but not in others (e.g., ineffective on well-known political figures). For well-known political figures, users' prior opinion towards them is a more consistently influencing factor.

To find answers to our research question, we showed the third group the manipulated image with an *explanation* of how it was derived from the original (see Figures 2–8 throughout the paper). Our results showed that the explanation was not always effective, which was dependent on the specific manipulation cases. When the explanation was effective, it did help to make participants less agreeing with the intended statement behind the manipulation. However, very surprisingly, the explanation also had a *negative effect* on their feeling/sentiment toward the pictured subjects. Overall, the results suggested that the impact of explanation was not simply positive or negative — it can even have the opposite impact on people's "thinking" and their "feeling". Based on the experimental results, we further discuss the open questions that demand further research explorations.

Collectively, our results highlight the need for better tools to help users identify and understand manipulated images in disinformation campaigns. Meanwhile, such tools should be carefully designed to reduce their own negative effect on users.

In summary, in this work, we present the results of our study that aim to answer the following three research questions (and our findings in parentheses):

- **R1:** Can viewers detect manipulated images? *(Poorly)*
- **R2:** Are viewers' opinions influenced by manipulated images? *(Sometimes)*
- **R3:** Does explaining how an image has been altered protect against the effects of image manipulation? *(Sometimes)*

To facilitate future research, all the images and questionnaires used in our study will be made publicly available to other researchers. For paper submissions, we hosted them under an anonymous link.[1]

This paper begins with a review of related work in Section 2. Section 3 presents the results of the first study (answering R1). Section 4 presents the results of the second study (answering R2 and R3). Section 5 discusses the implications of our results and the open research questions. Section 6 concludes the paper.

2 RELATED WORK

Image Manipulation and Detection. Numerous prior studies have shown that images influence viewers' memories, emotions, and opinions about themselves and others [24, 25, 28, 35]. As a result, images, especially manipulated ones, are often used in disinformation campaigns on social media sites to deceive users [15, 26, 34, 38]. Image manipulation involves transformation and/or alternation of the image to enhance the image or achieve deception [13, 14, 23, 30]. Researchers have studied technical methods (e.g., using deep neural networks) to detect digitally manipulated images by searching for various artifacts/anomalies [4, 5, 7–9, 14, 23]. Such detection methods are still far from perfect, and it is also challenging to automatically determine the (malicious) intent of image editing. Meanwhile, researchers find that people have difficulty detecting image manipulation [1, 22, 30]. Our first study is inspired by the work of Nightingale *et al.* [30], and our findings agree with theirs.

Combating Disinformation. There are three main practices used to combat misinformation today. The first is fact-checking, both by mainstream media outlets such as CNN,[2] The New York Times,[3] and the Washington Post,[4] as well as by independent sites such as Snopes.[5] In particular, several of the manipulated images we use in our study were found on Snopes. Our hypothesis (explaining image manipulation) adopts the spirit of fact-checking, that is, that the best way to fight lies is with the truth. Prior work on correcting misinformation suggests that misinformation can persist and continue to influence decision-making [37]. Our study specifically targets visual information and presents corrections alongside the manipulated image rather than at a later point.

Another practice, this one widely used by social media, is to remove offending content outright. A notable example of this practice is the permanent suspension of former US President Donald Trump from Twitter [36].

Finally, some social media platforms such as Facebook [27] display a warning and de-rank offending content. A recent study by Kaiser *et al.* [21] suggests that well-designed security warnings

[1] https://shorturl.at/qxDZ8.
[2] https://www.cnn.com/specials/politics/fact-check-politics
[3] https://www.nytimes.com/spotlight/fact-checks
[4] https://www.washingtonpost.com/news/fact-checker
[5] https://snopes.com

can counteract misinformation. The intervention studied in our paper—direct engagement with the content—is expected to share this advantage of immediacy provided by security warnings.

Continued Influence Effect. Efforts to debunk disinformation may not always be effective. Previous research has found that discredited information can continue to influence people, despite the explicit instruction to disregard it [19, 37]. One of the contributors to the continued influence is that disinformation is increasingly exposed to users during the process of "debunking," which is likely to make users more familiar with the disinformation— the so-called *familiarity backfire effect* [32]. The additional mention of the disinformation may activate key concepts in memory for subjects who see it and trigger counter-responses. There has also been work done to discredit the backfire effect. Authors in [39] show that after conducting their experiments, citizens heed factual information, even when such information challenges their ideological commitments. In our study, we will present the "correction" information alongside the manipulated images and study how they influence viewers.

Dual Process Theory. A piece of information (or disinformation) can influence people in various ways. Our study design is inspired by the dual-process theory, which describes a bi-system framework for the cognitive processes of human minds. System 1 operates automatically and quickly with little or no effort (handling subconscious emotion), and System 2 allocates attention to the effortful mental activities (handling conscious reasoning) [20]. Prior work has shown that susceptibility to fake news is driven more by lazy thinking than partisan bias [33]. In the meantime, increased deliberation facilitates accurate belief formation [2]. In our study, we design survey questions to capture people's subconscious feeling as well as their conscious reasoning.

In summary, it is not yet well understood whether we can effectively counteract the negative effect that image manipulation has in disinformation campaigns by showing how the image is altered. Our paper aims to fill in this gap.

3 DETECTING MANIPULATION

We start with the first survey to evaluate whether users can correctly identify manipulated images and spot the manipulated areas (our research question R1).

3.1 Methodology

Image Selection. We used manipulated images that were actually disseminated on the Internet. As shown in Figure 1, we found such images from social media platforms Reddit and Twitter, a fact-checking website Snopes, and public datasets shared by researchers [17, 31]. We chose manipulated images from these sources because they represent what social media users would see in the real world. When selecting images, we considered the following factors. First, the manipulation in the image needs to meet our definition. For the purpose of our studies, we regard manipulation as changes made to images with the intention of significantly altering the perception of the subject of the image, i.e., violating the integrity of the original image. For example, such manipulation may involve the addition or subtraction of an object (or a person), or major changes to a subject's face and/or body or other areas of the image. However, light image adjustments (on brightness or contrast) or simple face touch-ups (e.g., make-up filters, smoothing the skin) are not considered. Second, we selected manipulated images for which we can obtain their original images (either directly from the fact-checking site or using reverse image search on Google/TinEye). Third, we prioritized the selection of images of political figures and those that had been used in disinformation campaigns.

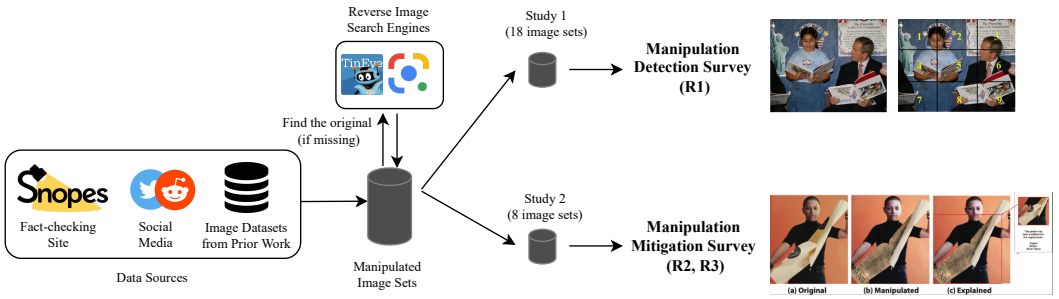


Fig. 1. **Workflow**—For our study, we first collected manipulated images that were actually disseminated on the Internet. We then obtained the original images of those manipulated images using reverse image search engines. The collected images were split into two groups for two different studies to explore research questions regarding manipulation detection (R1) and manipulation mitigation (R2 and R3).

As shown in Figure 1, that images used in the two surveys are non-overlapping. For this current survey (on manipulation detection), we selected 18 manipulated images and their original (un-manipulated) ones.

Survey Design. We asked participants to examine six images (three manipulated and three original) in a randomized order. To measure the *attentiveness* of the participant, we inserted one additional control image into a random position. The control image is obviously manipulated (see Figure 9 in the Appendix), under which we asked participants to select a specified answer out of the provided choices. We expect attentive participants will follow the instruction to select the correct answer. We run the study with three batches of images (18 experimental images and three control images). Each participant can only take the survey once and cannot join multiple batches.

When a participant viewed an image, they were first asked to determine whether the image was manipulated. They were given three choices: "Yes", "No", and "I'm not sure". We design the question to collect the binary determination from participants on an image (in order to measure their detection accuracy). If users cannot make a binary determination, they can select "I'm not sure". If the participant suspected manipulation i.e., selecting "Yes", they were asked to identify the manipulation region by selecting from a 3×3 grid overlayed on the image. The manipulated regions may cover multiple regions of the grid (i.e., a multi-choice question). If the participant did not think the image was manipulated, they were asked to select the option: "I don't think it is manipulated, or I'm not sure". After examining all the images, we collected basic demographic information of the participants (including age, gender, and educational background)[6].

Our methodology is inspired by a recent study [30]. The key difference is that our study used *real-world manipulated images*. In comparison, their study used images of daily objects/scenes, and the manipulation was inserted by the researchers.

Later in Section 3.2, we will compare our findings with those of [30]. In addition, to make sure the participants have correctly interpreted the meaning/definition of manipulation, we also run a quick validation study to confirm the validity of our results (details are in Appendix A).

Participant Recruitment. We recruited participants from Amazon Mechanical Turk after our study protocol was approved by IRB. A more detailed discussion of research ethics is presented in Section 5.

To get high-quality responses, we recruited workers who have completed at least 100 tasks with an approval rate greater than 95% and they should be located in the U.S. (to get English speakers).

[6]We include all the survey questions and images in a folder, hosted under an anonymous link: https://shorturl.at/qxDZ8

We recognize that the population of social media users is not solely made up by Americans and will further discuss this limitation later in Section 5.

In addition, we excluded inattentive participants who failed the attention check questions. We also implemented two other filters to remove unreliable results: (1) if participants completed the survey in an amount of time significantly lower than average, we manually inspected their answers to check validity. (2) We checked for obvious patterns in the answers (e.g., selecting the same answer for all questions) to remove invalid responses. Out of 128 total participants, we eventually obtained valid results from $n = 113$ participants: 76 identified as male, and 37 as female. Additional participant demographic information is presented in Table 1. The survey took a median of 3.23 minutes to complete, and each participant was compensated $1.34 for their time.

	Survey 1	Survey 2			
Attributes	**Total**	**Original**	**Manipulated**	**Explained**	**Total**
Gender					
Male	76	114	115	111	340
Female	37	58	71	68	197
Other	0	1	4	1	6
Age					
Under 18	0	0	0	0	0
18-24	9	17	16	4	37
25-34	16	83	83	69	235
35-44	7	46	57	62	165
45-54	62	15	6	20	41
55-64	19	9	17	18	44
65+	0	2	1	7	10
Not Disclosed	0	1	0	0	1
Education					
< High school	0	2	0	0	2
High school	5	12	8	15	35
Some college	54	12	9	9	30
Associate	30	9	15	10	34
Bachelor	13	114	118	117	349
Graduate	9	22	40	29	91
Not Disclosed	2	2	0	0	2
Total	**113**	**173**	**190**	**180**	**543**

Table 1. **Demographics**—We show the demographics information of participants of Survey 1 (Detecting Manipulation) and Survey 2 (Mitigating Manipulation).

3.2 Results

To answer our research question R1, we first examine how well participants identified manipulated images from the non-manipulated ones. We had 113 participants, and each of them examined 6 images (678 data points in total). Out of the 678 answers, only 40 (5.9%) were "I'm not sure". To be conservative, we regarded them as participants *not* able to recognize the manipulation (i.e., coded as "No"). Table 2 shows the confusion matrix for the identification results. We run a Chi-squared test to examine the correlation between participants' answers and the true labels of the images, which returns χ^2 $(1, n = 678) = 50.03$, $p < 0.001$. This means the participants' answers positively correlate with the true labels, i.e., participants have some ability to identify manipulated images.

However, when we further calculate the identification accuracy for each user, we obtain a mean accuracy of 60%, with a 95% confidence interval [56.5%, 64.5%]. This means the participants' overall identification accuracy is not high.

We further analyze how well participants located the manipulated regions after they correctly determined an image was manipulated. Recall that out of 678 responses, 241 of them were true positives (where the image was a true manipulated image, and the participant correctly determined the image was manipulated). For this analysis, we only consider this subset of responses because we want to capture the participants' accuracy of locating the manipulation region when there was indeed one. For this question, participants may select one or more regions out of the 3×3 grid. We regard the participant's answer as correct if the selected cells exactly match or are a subset of the true manipulated regions. This analysis returns a mean accuracy of 38%, with a 95% confidence interval of [32.4%, 43.8%]. This result indicates that participants cannot accurately identify the manipulated regions.

		Manipulated Image?	
		Yes	No
User Answers	Yes	241	150
	No	98	189

Table 2. Confusion matrix for participants' detection results of manipulated images.

In this survey, the participants spent on average 26.58 seconds (*SD* = 35 seconds) on each image, which is within a reasonable range. As a reference, a Facebook report shows that people on average spend 1.7 seconds on a piece of content on Facebook's news feed on mobile devices (2.5 seconds when using desktops) [12]. This suggests that users are unlikely to spend a long time processing the validity of the information in practice.

Our finding echos those of [30]. Recall this study used lab-created manipulation for their images, and their reported accuracy for manipulated image identification was 66%. In our survey, we focus on *real-world manipulated images* and find that participants exhibit some ability to distinguish manipulated images, but their accuracy is also not high (60%). These results collectively suggest that there is a need to build better tools to help users identify and understand image manipulations.

4 MITIGATING MANIPULATION

Our first survey showed that viewers are not good at identifying manipulated images, doing only slightly better than random (60% versus 50%) at this task. To determine whether viewers' opinions are influenced by manipulated images (research question R2), we ran a survey that compared the effect of seeing an original image versus a manipulated one.

To measure the differential effect of the two images, we asked survey participants two questions. The first, which we term the *message agreement* question, asks participants whether they agree with a specific statement that we believe the manipulated image intended to portray. The second question, which we term *subject sentiment* question, asks participants their opinion of the person who is shown in the image.

To answer our third research question (R3), that is, whether explaining how an image was manipulated would cancel the negative effects of image manipulation, we showed the third group of participants an image that explains how the original image was modified to produce the manipulated version and collected their responses to the questions above.

In the following, we describe our experiment design in more detail. In Section 4.2 we present the results.

4.1 Methodology

Image Selection. Similar to the first survey, we selected manipulated images that had been widely circulated online. While the first survey was focused on generic image manipulation, for this survey, we further focused on manipulated images that were involved in campaigns intentionally distributing false information for political means [11]. Following a similar image selection process, as shown in Figure 1, we identified eight sets of images. We chose these images in particular because they included well-known (political) subjects and were representative of disinformation campaigns on social media. For each pair of images (original and manipulated), we created a third image that explains how the original image was manipulated to produce the manipulated version.

Each of the image sets was evaluated as an independent case study. This was because each image depicted different subjects, and the corresponding manipulation had different contexts and purposes. In other words, we could not measure the impact of the image manipulation without setting up such contexts for users. Here, our goal was not to build a predictive model for the effect of image manipulation (or explanation) *across all images*. Instead, we treated each image set as a

Fig. 2. **The Squad**—Text in the explanation (c) reads, "*The appearance of clothing has been modified from the original photo. Orignal Section Shown Above.*"

case study to capture the nuanced intention of each manipulated image. Then based on the specific context, we *tailored* the questions for each image to measure how much users agreed with the *intended statement* behind each image and their *sentiment/feeling* toward the pictured subjects. We believe such an in-depth examination is needed to reveal the potentially complex impact of image manipulation and explanations. To tailor the questions and caption for each manipulated image, we carefully considered context information such as where the image was found and what messages were sent along with the image (based on information from Snopes and the corresponding social media sites). All three authors had to agree on the theme and specific wording for each question. We recognize that this method still has limitations in accurately capturing the intention of the actual image manipulators. We will further discuss this limitation later in Section 5.

Figure 2 shows one such set of three images used in the survey. The original shows six left-leaning democratic lawmakers commonly called "The Squad" in the media: Ilhan Omar, Alexandria Ocasio-Cortez, Rashida Tlaib, Ayanna Pressley, Jamaal Bowman, and Cori Bush. Figure 2(b) shows the manipulated image that was shared across the Internet in 2021. The clothes and some of the face masks of the lawmakers had been altered by adding stars that form the shape of a Nazi swastika. Figure 2(c) explains how the manipulated image was derived from the original.

Among the eight selected image sets, two sets depicted former US President Donald Trump. They were similar in nature and produced comparable results; we omitted one of these two sets from the results due to space constraints (presented in the Appendix instead). The remaining six image sets used in our analysis are shown in Figures 3 through 8 (shown alongside the text describing the results).

Groups. Survey participants were randomly assigned to one of three groups (conditions):

- **C1: Original.** Participants are shown original images, e.g. Figure 2(a).
- **C2: Manipulated.** Participants are shown manipulated versions of the images, e.g., Figure 2(b).
- **C3: Explained.** Participants are shown an image explaining how the manipulated image was produced, e.g., Figure 2(c).

One participant can only see one condition. We chose the between-subjects design to avoid the continued influence of seeing the same subject/image multiple times under different manipulation conditions.

Survey Design. Each participant was shown 4 images plus 1 control image (explained below) in a randomized order. Above each image was a caption identifying the person or people pictured. The same caption was shown for all three images in the set. For the set of images in Figure 2, the caption

read "Pictured below are six Democratic lawmakers." Below each image, we ask participants two questions described below.

Q1: Message agreement. The first question asks participants whether they agree with an image-specific statement that, in our judgment, the manipulated image intended to make. For example, the message agreement question accompanying Figure 2 asked, "Do you believe that Democratic lawmakers—Ilhan Omar of Minnesota, Alexandria Ocasio-Cortez of New York, Rashida Tlaib of Michigan, and Ayanna Pressley of Massachusetts, also known as the squad—hold extreme views?" While the manipulated image showed Nazi imagery, we assumed that the intent was not to imply that The Squad subscribed to Nazi ideology but that their views were extreme and formulated the message agreement question accordingly. The answer is recorded on a five-point Likert scale: strongly disagree, disagree, neither agree nor disagree, agree, strongly agree.

Q2: Subject sentiment. The second question asks their opinion of the subject(s) of interest in the image. This question asked, "From negative to positive, what is your opinion of the person(s) in this photo?" The answer is recorded on a five-point Likert scale: extremely negative, negative, neutral, positive, extremely positive. For images that depicted multiple subjects, we slightly rephrased the question to point out the person of interest. For example, Figure 8 showed two subjects, one hitting the other. For this question, we asked, "From negative to positive, what is your opinion of the person assaulting law enforcement in this photo?"

The message agreement question was designed to test the effect of the image on a concrete, specific statement, while the subject sentiment question was designed to determine whether the image induced a generally negative attitude toward the subject. A difference in the response to the two questions might arise for several reasons. First, viewers might perceive a different, but still negative, intended message that would affect their opinion of the subject. Second, even if viewers do not accept the intended message, the negative messaging might still taint their opinion of the subject. The dual-process theory of psychology suggests that this might happen if the questions invoke different thought processes in the viewer [20].

Q3: Prior opinion. Each image is shown on a separate page, and participants could not go back to change their prior answers. After participants complete the questions under all five images, they are asked for their prior opinion about the subjects of each of the non-control images shown to them (four in total). For example, participants who were shown one of the images in Figure 2 were asked, at the end of the survey, "Before taking this study, did you have a favorable, unfavorable, or neutral opinion of Democratic lawmakers?". Participants select from *favorable*, *neutral*, *unfavorable*, and *no opinion*. The prior-opinion question was asked after the survey participants had already seen and answered the *message agreement* and *sentiment* questions for all their assigned images. The reason was that we did not want to prime participants prior to answering these main questions. As existing literature suggests, once a person declares their feelings related to a specific subject, they may feel the need to commit to those feelings when questioned later as a way to confirm their existing beliefs [29]. We placed the prior-opinion questions in the latter part of the survey to avoid such undue influences.

After answering the questions above, we collected the participants' demographic information, including age, gender, and educational background.

Participant Recruitment. Recruitment, IRB review, and consent protocols were the same as for the first survey (Section 3.1). The quality control method was also the same (an example attention-check image is shown in Figure 10 in the Appendix). We recruited 543 participants for the survey: C1 $n = 173$, C2 $n = 190$, and C3 $n = 180$. In total, 340 identified as male, 197 as female, and 6 as other.

Additional participant demographic information is presented in Table 1. It took a median of 2.44 minutes to complete, and each participant was compensated $1.16 for their time.

4.2 Results

We modeled the outcome of the survey as a linear regression with the following factors:

- **Image type:** *original*, *manipulated*, or *manipulated with explanation*; coded as binary variables referenced to *manipulated*.
- **Prior opinion of subject:** *unfavorable*, *neutral*, *favorable*, or *none*; coded as binary variables referenced to *unfavorable*.
- **Viewer gender:** *male* or *female*; referenced to *male*. (Our survey also included *other* and *unspecified*, but because of the small number of such responses (6 out of 543), they were excluded from the regression analysis.)

The Squad				
	Message Agreement		Subject Sentiment	
Variable	β	p	β	p
Image (Reference = Manipulated)				
Original	0.12	0.53	-0.06	0.64
Explained	-0.09	0.60	-0.27	0.03*
Prior Opinion (Reference = Unfavorable)				
Favorable	-0.56	0.01**	1.22	<0.001***
Neutral	-0.45	0.07	1.00	<0.001***
None	-0.67	0.17	0.50	0.15
Gender (Reference = Male)				
Female	0.01	0.02*	-0.02	0.87

Table 3. **Linear Regression Model**—We show the Estimate β and the p value for each variable. For each image set, significance is denoted by *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$).

The above factors were coded as binary variables. The outcome variable was coded as an integer in the range -2 to 2. For the message agreement question, -2 represented *strongly disagree* and 2 *strong agree*, with remaining options mapped naturally between these two extremes. For subject sentiment, -2 represented *extremely negative* and 2 *extremely positive*, with remaining options mapped similarly.

Because the effect of image manipulation, if any, would heavily depend on the nature of the images themselves, we consider each of our seven images as separate studies and apply the same analysis to each. The remainder of this section discusses the results for each image.

The Squad. The images of the Squad are shown in Figure 2. The caption for these images read, "Pictured below are six Democratic lawmakers." The message agreement question asked, "Do you believe that Democratic lawmakers—Ilhan Omar of Minnesota, Alexandria Ocasio-Cortez of New York, Rashida Tlaib of Michigan, and Ayanna Pressley of Massachusetts, also known as the squad—hold extreme views?"

Table 3 shows the regression results.

Image type did not have a statistically significant effect on message agreement (Q1). Holding a favorable prior opinion had a statistically significant ($p < 0.01$) effect of over a half-step closer to *disagree* ($\beta = -0.56$) compared to holding an unfavorable prior opinion. Participant gender had a very small effect ($\beta = 0.01$, $p = 0.02$) towards agreement with the intended message.

On the subject sentiment question (Q2), the difference between original and manipulated was not statistically significant. However, explaining how the image was manipulated did have a statistically significant effect ($p = 0.03$). What was surprising to us was the *direction* of the effect: participants who saw the manipulated image with an explanation of how it was manipulated came out with a more *negative* opinion (by 0.27 Likert scale steps) of the Squad. We see this effect again with two other images; in Section 5, we discuss potential explanations for this phenomenon.

Participant sentiment toward the Squad was largely shaped by prior opinion. Viewers who stated they had a favorable prior opinion[7] also had a more favorable opinion after viewing the image compared to those who had an unfavorable prior opinion ($\beta = 1.22$, $p < 0.001$). Similarly,

[7]Note, however, that the prior opinion question was posed after participants completed the main part of the survey. Participants may be more likely to state a favorable prior opinion after expressing a favorable current opinion.

participants reporting a neutral prior opinion had a more favorable opinion after viewing the image compared to those who had an unfavorable prior opinion ($\beta = 1.00$, $p < 0.001$).

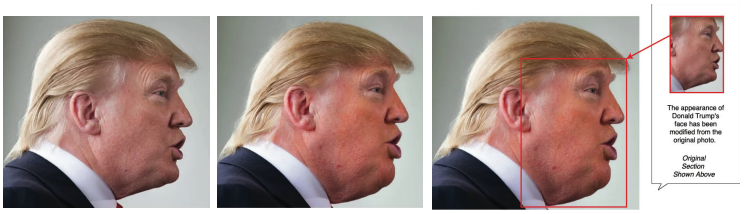


Fig. 3. **D. Trump-1**—Text in the explanation (c) reads, "*The appearance of Donald Trump's face has been modified from the original photo. Original Section Shown Above.*"

D. Trump. Figure 3 shows former U.S. president Donald Trump. In the manipulated image, his face is given a bloated and unattractive appearance. The caption for these images read, "A photo of former President of the United States Donald Trump." The message agreement question (Q1) asks, "Do you believe that former United States President Donald Trump is an unattractive person?". The regression results are shown in Table 4.

As with the Squad, we did not see a statistically significant effect in original versus manipulated and explained versus manipulated (Q1). Prior opinion had a stronger statistically significant effect on participant opinion: participants who reported a favorable prior opinion were 0.71 scale steps closer to disagreeing with the statement above compared to those holding an unfavorable prior opinion ($p < 0.001$), and similarly for neutral versus unfavorable prior opinion ($\beta = -0.73$, $p < 0.001$). Gender did not have a statistically significant effect.

D. Trump				
	Message Agreement		Subject Sentiment	
Variable	β	p	β	p
Image (Reference = Manipulated)				
Original	-0.04	0.85	0.17	0.30
Explained	-0.08	0.65	-0.08	0.62
Prior Opinion (Reference = Unfavorable)				
Favorable	-0.71	0.001^{***}	2.08	0.001^{***}
Neutral	-0.73	0.001^{***}	1.35	0.001^{***}
None	-0.63	0.18	1.35	0.001^{***}
Gender (Reference = Male)				
Female	0.05	0.76	0.07	0.60

Table 4. **Linear Regression Model**—We show the Estimate β and the p value for each variable. For each image set, significance is denoted by *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$).

On the subject sentiment question (Q2), image type and gender had no statistically significant effect. Sentiment was heavily influenced by prior opinion: over two scale steps more positive for favorable versus unfavorable prior opinion ($p < 0.001$), and 1.35 scale steps more positive for both neutral and none versus unfavorable ($p < 0.001$).

Fig. 4. **J. Biden**—Text in the explanation (c) reads, "*The appearance of Joseph Biden and Amy Parnes has been modified from the original photo. Original Section Shown Above.*"

J. Biden. Figure 4 shows current US president Joe Biden along with his wife Dr. Jill Biden, reporter Amie Parnes, and producer Chris Donovan. In the manipulated image, Joe Biden's hands are positioned on the breasts, rather than the waist, of Amie Parnes, and a bottle of bourbon whiskey was added in the bottom left corner of the image. The caption for these images read, "Pictured from the left is United States President Joe Biden with Amie Parnes, Chris Donovan, and Jill Biden." The intended message question (Q1) asks, "Do you believe United States President Joe Biden inappropriately touches women?"

J. Biden				
	Message Agreement		Subject Sentiment	
Variable	β	p	β	p
Image (Reference = Manipulated)				
Original	-0.20	0.28	0.03	0.86
Explained	-0.57	0.01**	-0.53	0.01**
Prior Opinion (Reference = Unfavorable)				
Favorable	-0.43	0.02*	1.30	0.001***
Neutral	-0.24	0.27	1.06	0.001***
None	-0.79	0.14	0.95	0.03*
Gender (Reference = Male)				
Female	-0.30	0.06	-0.03	0.80

Table 5. **Linear Regression Model**—We show the Estimate β and the p value for each variable. For each image set, significance is denoted by *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$).

As shown in Table 5, for the intended message question (Q1), there was no statistically significant effect of original versus manipulated image. However, showing the manipulated image explanation had a negating effect: viewers who saw the manipulated image with explanation were 0.57 scale steps closer to disagreeing with the statement above than those seeing the manipulated image ($p < 0.01$). We were surprised to find that the negating effect of seeing the explanation was stronger than seeing the original rather than manipulated image ($\beta = -0.20$, $p = 0.28$).

As with many of the images, prior opinion had a strong effect: participants who reported a favorable prior opinion were 0.43 steps closer to disagreeing with the statement compared to those holding a negative prior opinion ($p = 0.02$).

For the subject sentiment question (Q2), participants who saw the manipulated image with explanation were 0.53 steps more *negative* than those who saw the manipulated image ($p < 0.01$). (Original versus manipulated had no statistically significant effect.) This was surprising: we expected the effect of the explanation, which counteracted the effect of image manipulation for the first question, to have a similar counteracting effect for the second question. The effect, however, was the opposite. As with other images, stated prior opinion had a statistically significant effect.

(a) Original (b) Manipulated (c) Explained

Fig. 5. **J. Varney**—Text in the explanation (c) reads, "*The head of Jeffery Epstein has been added to the original photo. Original Section Shown Above.*"

J. Varney. The original image in Figure 5 shows actors Jim Varney and Robin Williams. The caption for these images read, "The person pictured on the left is actor Jim Varney." In the manipulated photo, Robin Williams is replaced with American financier and convicted sex offender Jeffery Epstein. Other public figures had come under fire for their association with Epstein, some facing accusations of similar offenses themselves. A photograph of Varney with Epstein

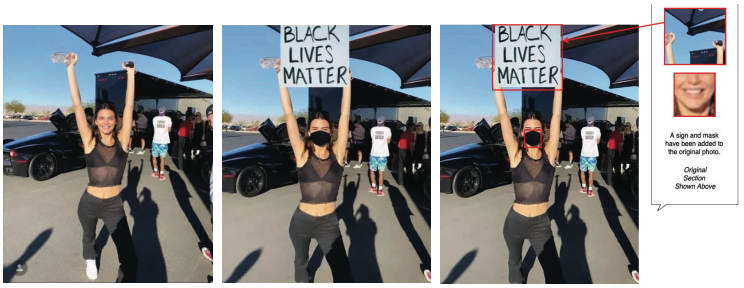
might implicate Varney in Epstein's offenses, so we formulated the message agreement question as, "Do you believe that actor Jim Varney may be guilty of the sex trafficking of minors?"

As shown in Table 6, on the message agreement question (Q1), we see a statistically significant effect of the manipulated image versus the original ($\beta = -0.44$, $p = 0.02$). This result is different from those of the Squad, Biden, and Trump (well-known political figures), where no significant effect was observed comparing manipulated with the original. In Varney's case, explaining the manipulation has a similar effect ($\beta = -0.58$, $p < 0.01$). That is, seeing an explanation of how the image was manipulated is comparable to the effect of seeing the original, non-manipulated image. There was no statistically significant effect for prior favorable versus prior unfavorable opinion nor neutral versus unfavorable. Participants who reported to have had no prior opinion of Varney were 0.79 scale steps more in disagreement with the statement above ($p = 0.02$).

J. Varney				
	Message Agreement		Subject Sentiment	
Variable	β	p	β	p
Image (Reference = Manipulated)				
Original	-0.44	0.02*	0.21	0.10
Explained	-0.58	0.01**	-0.40	0.01**
Prior Opinion (Reference = Unfavorable)				
Favorable	-0.20	0.41	0.76	0.001***
Neutral	-0.09	0.73	0.21	0.24
None	-0.79	0.02*	-0.40	0.07
Gender (Reference = Male)				
Female	0.13	0.41	0.09	0.40

Table 6. **Linear Regression Model**—We show the Estimate β and the p value for each variable. For each image set, significance is denoted by *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$).

On the subject sentiment question (Q2), viewers who saw the manipulated image with explanation were 0.4 scale steps more *negative* versus manipulated image. As with the Biden image, this is surprising: we expected the negating effect seen for message agreement to extend to subject sentiment also. As with other images, a favorable reported prior opinion resulted in a 0.76 scale step more positive subject sentiment versus unfavorable prior opinion ($p < 0.001$). No other factors were statistically significant.



(a) Original (b) Manipulated (c) Explained

Fig. 6. **K. Jenner**—Text in the explanation (c) reads, "*A sign and mask have been added to the original photo. Original Section Shown Above.*"

K. Jenner. Figure 6 shows an American model Kendal Jenner. In the original image, she appears holding a water bottle. In the manipulated image, the water bottle is gone, and she is holding a hand-lettered sign that says BLACK LIVES MATTER. The manipulated image also shows her wearing a black face mask. The caption for these images read, "A photo of 24-year-old model, Kendall Jenner." The message agreement question asks, "Do you believe that Kendall Jenner supports wearing protective face masks and supports the Black Lives Matter Movement?"

As shown in Table 7, for the message agreement question (Q1), there was a statistically significant 0.5 scale step shift toward disagreeing with the statement for participants who saw the original versus manipulated image ($p < 0.001$). (That is, seeing the manipulated image with the BLACK LIVES MATTER led viewers to believe Jenner supported the movement.) Again, this is similar to

Varney's case but different from well-known political figures, including Squad, Trump, and Biden. In Jenner's case, seeing the manipulated image with the explanation of the modification did not show a statistically significant effect versus the manipulated image by itself. A reported favorable prior rating had a statistically significant effect toward agreement versus an unfavorable prior opinion ($\beta = 0.48$, $p < 0.01$). No other factors were statistically significant.

This was the only image in our study where the intended message was not manifestly negative.[8] For images with negative messages, a prior favorable opinion of the subject (versus unfavorable) meant greater disagreement with the (negative) message, while for Jenner, a prior favorable opinion resulted in greater agreement with the message.

For the subject sentiment question (Q2), image type had no statistically significant effect. A stated favorable prior opinion had a positive effect, 1.15 scale steps toward favorable, compared to participants with a stated unfavorable prior opinion. Similarly, for participants with a stated neutral prior opinion versus unfavorable, the effect was 0.66 scale steps more favorable ($p < 0.001$). No other factors were statistically significant.

K. Jenner				
	Message Agreement		Subject Sentiment	
Variable	β	p	β	p
Image (Reference = Manipulated)				
Original	-0.50	0.001***	-0.02	0.90
Explained	-0.15	0.31	-0.21	0.10
Prior Opinion (Reference = Unfavorable)				
Favorable	0.48	0.01**	1.15	0.001***
Neutral	0.25	0.18	0.66	0.001***
None	0.01	0.98	0.38	0.14
Gender (Reference = Male)				
Female	-0.11	0.40	-0.05	0.69

Table 7. **Linear Regression Model**—We show the Estimate β and the p value for each variable. For each image set, significance is denoted by *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$).

(a) Original (b) Manipulated (c) Explained

Fig. 7. **Gun Activism**—Text in the explanation (c) reads, "*This section has been modified from the orignal photo. Original Section Shown Above.*"

Gun Activism. Figure 7 shows Emma González, a survivor of the Marjory Stoneman Douglas High School mass shooting. The original image shows González tearing paper shooting practice target in two. In the manipulated image, the target is replaced by the US constitution. The caption for these images read, "The person pictured in the photo is a survivor of the Marjory Stoneman Douglas High School mass shooting, Emma González campaigning against gun violence." We believe that the subject of this image is not necessarily Emma González herself, but her act of campaigning against gun violence. Therefore, we formulated the message agreement question as "Do you believe that those who campaign against guns do not respect the Constitution of the United States?" The regression results are shown in Table 8.

We found that seeing the manipulated image (versus the original) had a statistically significant effect on the message agreement question (Q1), with participants who saw the manipulated image being 0.62 scale steps closer to agreeing with the message above ($p < 0.01$). Seeing the explanation

[8]According to recent polling, support for the Black Lives Matter movement is split, with 44% of Americans supporting the movement and 43% opposing it [6].

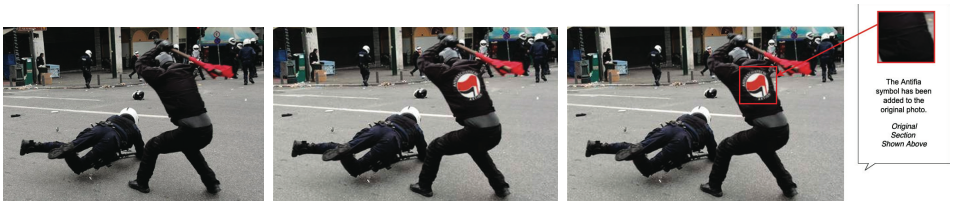
of how the image was manipulated had a similar effect as seeing the original ($\beta = -0.44$, $p = 0.02$). In this case, explaining what modifications were made to the image had the intended effect of counteracting the image manipulation.

The prior opinion question (Q3) for this image asked, "Before taking this study, did you have a favorable, unfavorable, or neutral opinion about gun control?" (We asked about "gun control" instead of "Emma González" to align with the intended messages.) Response to the prior opinion question did not have a statistically significant effect on the message agreement question (Q1).

For the subject sentiment question (Q2), we asked, "From negative to positive, what is your opinion of the person in this photo?" Participants who reported a favorable prior or neutral prior opinion (versus unfavorable) about gun control had, respectively, 0.58 and 0.24 scale steps more favorable opinions about González after viewing the photo (in both cases $p < 0.001$). No other factors had a statistically significant effect.

Gun Activism				
	Message Agreement		Subject Sentiment	
Variable	β	p	β	p
Image (Reference = Manipulated)				
Original	-0.62	0.01**	0.20	0.14
Explained	-0.44	0.02*	0.01	0.93
Prior Opinion (Reference = Unfavorable)				
Favorable	-0.23	0.24	0.58	0.001***
Neutral	-0.08	0.74	0.24	0.001***
None	-.0.54	0.55	0.84	0.20
Gender (Reference = Male)				
Female	-0.17	0.28	-0.07	0.56

Table 8. **Linear Regression Model**—We show the Estimate β and the p value for each variable. For each image set, significance is denoted by *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$).



(a) Original (b) Manipulated (c) Explained

Fig. 8. **Antifa**—Text in the explanation (c) reads, "*The Antifa symbol has been added to the original photo. Original Section Shown Above.*"

Antifa. Figure 8 shows what appears to be a person hitting a police officer who has fallen to the ground with a stick. In the manipulated image, an Antifa logo was added to the black jacket of the person assaulting the fallen officer. The caption for these images read, "A photo that shows a protester assaulting a law enforcement officer." The message agreement question asked, "Do you believe Antifa protesters are violent?" The prior opinion question asked, "Before taking this study, did you have a favorable, unfavorable, or neutral opinion about the Antifa Movement?" (We asked about the Antifa Movement to align with the intended messages.)

As shown in Table 9, image type had no statistically significant effect on the answer to the message agreement question (Q1). Favorable, neutral, and no prior opinion (versus unfavorable prior opinion) had a statistically significant effect on the answer. Participants who stated a favorable or neutral prior opinion were 0.87 more scale steps away from agreement with the statement that Antifa protesters are violent than those who stated an unfavorable prior opinion (in both cases $p < 0.001$). Participants who stated that they had no prior opinion about the Antifa movement were 1.03 more scale steps away from agreement with the message agreement question ($p < 0.01$).

Female respondents were 0.43 scale steps more in agreement with the statement that Antifa protesters are violent than men ($p < 0.01$). This was the only image where gender had a non-negligible statistically significant effect on the answer to the message agreement question.

For the subject sentiment question (Q2), we asked, "From negative to positive, what is your opinion of the person assaulting law enforcement in this photo?" Image type had no statistically significant effect on the answer. Participants who stated a favorable prior opinion of the Antifa movement responded 1.3 scale steps more favorably about the subject of the photo ("the person assaulting law enforcement") than those who stated an unfavorable prior opinion ($p < 0.001$). Participants who stated a neutral prior opinion had a more favorable opinion of the subject of the photo than those who stated an unfavorable prior opinion ($\beta = 0.53$, $p < 0.01$). This image set was the only case where gender also had a statistically significant effect on the answer to the subject sentiment question. Female participants had a less favorable opinion of the subject than male participants ($\beta = -0.3$, $p = 0.03$).

	Antifa			
	Message Agreement		Subject Sentiment	
Variable	β	p	β	p
Image (Reference = Manipulated)				
Original	-0.19	0.26	-0.20	0.22
Explained	-0.25	0.14	0.13	0.42
Prior Opinion (Reference = Unfavorable)				
Favorable	-0.87	0.001^{***}	1.30	0.001^{***}
Neutral	-0.87	0.001^{***}	0.53	0.01^{**}
None	-1.03	$<0.01^{**}$	-0.28	0.38
Gender (Reference = Male)				
Female	0.43	0.01^{**}	-0.30	0.03^{*}

Table 9. **Linear Regression Model**—We show the Estimate β and the p value for each variable. For each image set, significance is denoted by *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$).

5 DISCUSSION

In this section, we summarize our findings and discuss the implications for designs. We then discuss the open questions derived from these case studies and the limitation of this work.

5.1 Findings and Implications

Our first survey confirmed the results of Nightingale et al. that formed the premise for our work, namely that users are not effective at determining whether an image was manipulated or not. Our second survey tested whether explaining how an image was manipulated counteracted the messaging associated with the image. We found that in only three of the seven study images did explaining how an image was manipulated sway viewers' agreement with the factual message implied by the manipulated image (J. Biden, J. Varney, Gun Activism), and in two cases, showing the original (non-manipulated) image was more effective than showing our explanation. Moreover, in three cases, our explanation of manipulation had an unexpected greater negative effect on viewers' sentiment toward the subject than the manipulated image itself (The Squad, J. Biden, J. Varney).

Our results show that attempting to explain how an image was manipulated should be done with caution. In particular, showing a manipulated image, even in the context of an explanation, may still produce an intended negative effect of the manipulated image. This suggests that future work countering visual disinformation should avoid showing the manipulated image, for example, by providing a warning before showing a manipulated image, or emphasizing the correct, rather than incorrect, information by showing the original (non-manipulated) image with text stating that the original social media post contained a manipulated version of this image. Our results also show that users' prior opinion towards the pictured (political) subjects has a more consistent influence across different case studies. This suggests future explanation designs may account for people's political preferences or standings.

For social media platforms and users, we need better tools to trace and search for credible sources of a piece of information. Currently, such tools (e.g., reverse image search and fact-checking services) are mostly ad-hoc and are not well integrated with social media platforms. Moreover, designing such tools needs to carefully consider their influence on users. For example, when discrediting a piece of disinformation (e.g., a manipulated image), social media platforms should *avoid re-exposing users to the disinformation* as much as possible while focusing on the correction message and the

truthful materials. Also, it is important to prevent disinformation from reaching a large number of users in the first place, as discrediting disinformation afterward is a difficult task.

5.2 Connections to Existing Theory

Continued Influence Effect. As introduced in Section 2, prior works have mixed conclusions regarding the continued influence effect of disinformation. Some argue that disinformation can be increasingly exposed to users during the process of debunking, which makes discredited information continue to influence people [19, 32, 37]. Others find that users would heed factual information, even when such information challenges their ideological commitments [39]. In our study, we observed that participants tended to feel more negatively about the subjects in the image when they were presented with explanations of how the image had been manipulated. One possible reason is that the continued exposure of the manipulated images had introduced a negative influence. This leads to an open question to be explored for future work: *does the influence of manipulated images occur and persist during the process of "debunking" disinformation cause a less favorable opinion of subjects pictured in the image?*

In the current study, we presented the manipulated images to participants along with a small text box to explain how the image was manipulated. While the text box presented an instruction to discredit the image, it also highlighted the manipulated areas of the altered images. In a future study, researchers may explore different ways to present the explanations (e.g., without showing the complete manipulated images) for more desired outcomes.

Dual Process Theory. As introduced in Section 2, the dual-process theory describes a bi-system framework for the cognitive processes of human minds, including System 1 that handles subconscious emotion and System 2 that handles conscious reasoning [20]. Our results showed that after viewing the explanation, participants were generally less agreeing with the intended messages, as expected. A possible explanation is the question about their agreement on the intended messages (Q1) requires deliberate thinking and reasoning (System 2). Participants need to read the explanation and associate the information with the image to answer this question, which counteracts the manipulation effect. In comparison, when asked about the sentiment towards the pictured subject (Q2), it is possible that participants were relying more on System 1 as the question was related to their feelings [3, 19]. This leads to an open question: *can we counteract the influence of manipulated images more effectively by nudging users to perform more deliberate thinking and reasoning? If so, how to facilitate this process?* We will explore this question as part of future work.

5.3 Limitations

We note several limitations of our studies. *First*, participants recruited from Amazon Mechanical Turn (MTurk) may not be representative of the entire Internet population, and there is no guarantee of participant attention (even with attention checks). While these are known limitations of MTurk, recent studies also shed some confidence in the quality of MTurk results. For example, research shows replication experiments conducted on Mturk can obtain comparable results with those obtained from national samples [10], and MTurk workers are at least as attentive as the subject pool participants [16]. *Second*, the explanation method used in the survey is not necessarily the most effective design (which is not the main focus of this study). Future work may focus on the design aspects to explore ways to improve the effectiveness of the explanation while suppressing its negative influence. For example, one direction is to add interactive features to stimulate the cognitive reasoning of participants. *Third*, our study only covers a limited number of case studies with a focus on opinion influencing or political subjects in the United States. Future work may extend the study scope to cover more categories of manipulated images (e.g., those from outside of the U.S.), include

survey participants from different regions of the world, and consider more demographic factors such as political preference. *Fourth*, when creating the captions for the images, we have, with our best efforts, incorporated the context information regarding where the manipulated images were posted and the messages sent along with the images in the orignal campaign. We acknowledge this approach may still introduce biases (from the authors). *Finally*, in the second survey, we choose a between-subjects design in order to compare the results of the three image types/conditions. This design seeks to avoid the continued influence of exposing the same image to participants multiple times. A limitation of this approach is it does not measure how the same user's perception changes from viewing the original/manipulated image to viewing the explained image. Future work may explicitly study this "continued influence" on the same participants using a within-subject design.

5.4 Ethics

We have taken careful steps to ensure research ethics. First, we worked closely with our IRB and obtained their approval before running the surveys. Second, we did not collect any personally identifiable information (PII) from the participants during the study. Third, consent was given prior to the survey, and participants were also given the opportunity to withdraw their data after the study at any time. The study presents little to no risk compared to those encountered in people's everyday online activities. Meanwhile, the results from the study can benefit social media platforms and users, and the Internet community to build better tools to fight against disinformation. We believe the benefit outweighs the potential risk.

6 CONCLUSION

We report on the results of two surveys aimed to determine how well viewers can identify manipulated photographs, whether such photographs influence viewers' opinions, and whether explaining the manipulation to users would counteract the negative effect of image manipulation. We found that users were not good at identifying manipulated images (and were worse at locating the manipulated regions). Also, simply highlighting and explaining the manipulation to users was not always effective. When it was effective, it did help to make users less agreeing with the intended messages behind the manipulation. However, surprisingly, the highlighting and explanation led viewers to hold less favorable opinions about the subjects pictured. The results from our case studies inspire new questions for future research to study the continued influence of manipulated images during the debunking process and ideas for more effective interventions. While we need better tools to help users identify and understand manipulated images in disinformation campaigns, we argue that such tools must be carefully designed to avoid introducing their own negative effects on users.

ACKNOWLEDGEMENTS

This work was supported in part by NSF grants 2030521, and the Graduate Research Fellowship Program under Grant No 21-46756. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Who falls for online political manipulation?. In *The Web Conference (WWW)*.

[2] Bence Bago, David Gertler Rand, and Gordon Pennycook. 2019. Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology* (2019).

[3] Roy F. Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen C. Vohs. 2001. Bad is stronger than good. *SAGE Journals* (2001).

[4] Belhassen Bayar and Matthew C Stamm. 2016. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *ACM Information Hiding and Multimedia Security Workshop*.

[5] Belhassen Bayar and Matthew C. Stamm. 2018. Constrained Convolutional Neural Networks: A New Approach Towards General Purpose Image Manipulation Detection. *IEEE Transactions on Information Forensics and Security* 13, 11 (2018).

[6] Claretta Bellamy. 2021. Support for Black Lives Matter movement is declining, according to new poll. https://www.nbcnews.com/news/nbcblk/support-black-lives-matter-movement-declining-according-new-poll-rcna5746

[7] Guoyong Cai and Binbin Xia. 2015. Convolutional neural networks for multimedia sentiment analysis. In *Lecture Notes in Computer Science*. Springer Verlag.

[8] Víctor Campos, Brendan Jou, and Xavier Giró-i Nieto. 2017. From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction. *Image and Vision Computing* (2017). arXiv:1604.03489

[9] Andreu Casas and Nora Webb Williams. 2017. Computer Vision for Political Science Research : A Study of Online Protest Images. *New Faces in Political Methodology IX* (2017).

[10] Alexander Coppock. 2019. Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach. *Political Science Research and Methods* (2019).

[11] Joan Donovan. 2021. What is Media Manipulation? https://just-infras.illinois.edu/speaker-series/joan-donovan/

[12] Facebook. 2016. Capturing attention in feed: The science behind effective video creative. Facebook IQ. https://www.facebook.com/business/news/insights/capturing-attention-feed-video-creative

[13] Hany Farid. 2006. Digital doctoring: How to tell the real from the fake. *Significance* 3, 4 (2006), 162–166.

[14] Farid Hany. 2009. A Survey of Image Forgery Detection. *IEEE Signal Processing Magazine* (March 2009).

[15] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy. In *The Web Conference (WWW) Companion*.

[16] David J. Hauser and Norbert Schwarz. 2016. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods* (2016).

[17] Silvan Heller, Luca Rossetto, and Heiko Schuldt. 2018. The PS-Battles Dataset – an Image Collection for Image Manipulation Detection. *CoRR* abs/1804.04866 (2018).

[18] Pik-Mai Hui, Chengcheng Shao, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. 2018. The Hoaxy Misinformation and Fact-Checking Diffusion Network. In *International AAAI Conference on Web and Social Media (ICWSM)*.

[19] Hollyn Johnson and Colleen Seifert. 1994. Sources of the Continued Influence Effect: When Misinformation in Memory Affects Later Inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition* (1994).

[20] Daniel Kahneman. 2011. *Thinking, fast and slow*. Farrar, Straus and Giroux.

[21] Ben Kaiser, Jerry Wei, Eli Lucherini, Kevin Lee, J. Nathan Matias, and Jonathan Mayer. 2021. Adapting Security Warnings to Counter Online Disinformation. In *USENIX Security Symposium*.

[22] Mona Kasra, Cuihua Shen, and James F. O'Brien. 2018. Seeing is believing: How people fail to identify fake images on the web. In *Conference on Human Factors in Computing Systems (CHI)*.

[23] Eric Kee, James O'brien, and Hany Farid. 2013. Exposing photo manipulation with inconsistent shadows. *ACM Transactions on Graphics* (2013).

[24] Mariska Kleemans, Serena Daalmans, Ilana Carbaat, and Doeschka Anschütz. 2018. Picture Perfect: The Direct Effect of Manipulated Instagram Photos on Body Image in Adolescent Girls. *Media Psychology* (jan 2018).

[25] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. 1997. International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention* 1, 39-58 (1997), 3.

[26] Yiyi Li and Ying Xie. 2020. Is a Picture Worth a Thousand Words? An Empirical Study of Image Content and Social Media Engagement. *Journal of Marketing Research* (nov 2020).

[27] Meta Journalism Project. 2021. How Facebook's third-party fact-checking program works.

[28] Robert A. Nash, Kimberley A. Wade, and Rebecca J. Brewer. 2009. Why do doctored images distort memory? *Consciousness and Cognition* 18, 3 (2009), 773–780.

[29] Raymond S. Nickerson. 1998. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* 2, 2 (1998), 175–220.

[30] Sophie J. Nightingale, Kimberley A. Wade, and Derrick G. Watson. 2017. Can people identify original and manipulated photos of real-world scenes? *Cognitive Research: Principles and Implications* (dec 2017).

[31] Adam Novozamsky, Babak Mahdian, and Stanislav Saic. 2020. IMD2020: A Large-Scale Annotated Dataset Tailored for Detecting MockupManipulated Images. In *IEEE Winter Applications of Computer Vision Workshops (WACVW)*.

[32] Brendan Nyhan and Jason Reifler. 2010. When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior* 32, 2 (2010), 303–330.

[33] Gordon Pennycook and David G. Rand. 2018. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188 (2018), 39–50.

[34] Gustavo Resende, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. 2019. (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures. In *The World Wide Web Conference*. Association for Computing Machinery, 818–828. https://doi.org/10.1145/3308558.3313688

[35] Shawn W. Rosenberg and Patrick Mccafferty. 1987. The Image and the Vote: Manipulating Voters' Preferences. *Public Opinion Quarterly* 51, 1 (01 1987), 31–47.

[36] Twitter, Inc. 2021. Permanent suspension of @realDonaldTrump. https://blog.twitter.com/en_us/topics/company/2020/suspension.

[37] Nathan Walter and Riva Tukachinsky. 2020. A Meta-Analytic Examination of the Continued Influence of Misinformation in the Face of Correction: How Powerful Is It, Why Does It Happen, and How to Stop It? *Communication Research* 47, 2 (2020), 155–177.

[38] Yuping Wang, Fatemeh Tahmasbi, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, David Magerman, Savvas Zannettou, and Gianluca Stringhini. 2021. Understanding the Use of Fauxtography on Social Media. In *The International AAAI Conference on Web and Social Media (ICWSM)*.

[39] Thomas Wood and Ethan Porter. 2017. The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Forthcoming, Political Behavior* (Dec 2017).

[40] Savvas Zannettou, Michael Sirivianos, Tristan Caulfield, Gianluca Stringhini, Emiliano De Cristofaro, and Jeremy Blackburn. 2019. Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. In *The Web Conference (WWW)*.

A APPENDIX: VALIDATION STUDY FOR MANIPULATION DETECTION

In our study, we consider image manipulation as changes made to the image with the intention of significantly altering the perception of the subject of the image. In other words, the changes violate the integrity of the image at the semantic level. We followed this definition when selecting images for our study. Images with only minor adjustments (on brightness or contrast) or simple face touch-ups (e.g., make-up filters, smoothing the skin) were not included to avoid confusion among participants.

In our first survey (Section 3), we asked participants to identify the manipulated images and the manipulated regions in such images *without priming users* on what manipulated images look like. To make sure participants could correctly interpret the meaning of image manipulation, we performed a separate validation test. More specifically, we re-run the first survey, where we provided an explicit definition for the type of image manipulation considered in the study (as described above) and reminded participants about the definition under each of the displayed images. We made it clear that minor image adjustments (that do not alter the semantic meaning of the image) were not considered. Like the original experiment, participants can select from "Yes", "No", and "I'm not sure" (to eliminate the pressure of making a binary choice). We recruited 47 participants (31 male, 16 female) for this validation test. The new test returned similar conclusions (consistent with our survey in Section 3). The results show that participants are not good at identifying manipulation with an accuracy 53%, with a 95% confidence interval of [48%, 59.1%]. Their accuracy of locating the manipulated region in an image was still low, with an accuracy of 34%, with a 95% confidence interval of [26.4%, 43.2%]. This confirms that our survey results described in Section 3 are reliable.

Case Study	**Prior Opinion**			
	Favorable	Neutral	Unfavorable	No Opinion
The Squad	58.78%	24.43%	14.12%	2.67%
D. Trump-1	45.88%	22.35%	29.02%	2.75%
J. Biden	48.60%	21.68%	27.27%	2.45%
J. Varney	50.00%	29.01%	10.69%	10.31%
K. Jenner	44.92%	33.59%	15.23%	6.25%
Gun Activism	58.39%	18.98%	21.90%	0.73%
Antifa	34.27%	28.67%	32.17%	4.90%

Table 10. **Prior Opinions**—Participants' opinion of the pictured subjects prior to the study.

Received May 2022; revised August 2022; accepted September 2022

Image Subject	1. Message Agreement Question 2. Subject Sentiment Question 3. Prior Opinion Question
The Squad	Q1. Do you believe that Democratic lawmakers - Reps. Ilhan Omar of Minnesota, Alexandria Ocasio - Cortez of New York, Rashida Tlaib of Michigan, and Ayanna Pressley of Massachusetts, also known as the squad, hold extreme views? Q2. From negative to positive, what is your opinion of the person(s) in this photo? Q3. Before taking this study, did you have a favorable, unfavorable, or neutral opinion of Democratic lawmakers?
D. Trump -1	Q1. Do you believe that former United States President Donald Trump is an unattractive person? Q2. From negative to positive, what is your opinion of the person in this photo? Q3. Before taking this study, did you have a favorable, unfavorable, or neutral opinion of former United States President Donald Trump?
J. Varney	Q1. Do you believe that actor Jim Varney may be guilty of the sex trafficking of minors? Q2. From negative to positive, what is your opinion of the person on the left in this photo? Q3. Before taking this study, did you have a favorable, unfavorable, or neutral opinion of actor Jim Varney?
K. Jenner	Q1. Do you believe that Kendall Jenner supports wearing protective face masks and supports the Black Lives Matter Movement? Q2. From negative to positive, what is your opinion of the person in this photo? Q3. Before taking this study, did you have a favorable, unfavorable, or neutral opinion of Kendall Jenner?
Gun Activism	Q1. Do you believe that those who campaign against guns do not respect the Constitution of the United States? Q2. From negative to positive, what is your opinion of the person in this photo? Q3. Before taking this study, did you have a favorable, unfavorable, or neutral opinion about gun control?
Antifa	Q1. Do you believe Antifa protesters are violent? Q2. From negative to positive, what is your opinion of the person assaulting law enforcement in this photo? Q3. Before taking this study, did you have a favorable, unfavorable, or neutral opinion about the Antifa Movement?

Table 11. **Survey Questions**—For each image case, we list the corresponding Message Agreement Question (Q1), Subject Sentiment Question (Q2), and Prior Opinion Question (Q3).

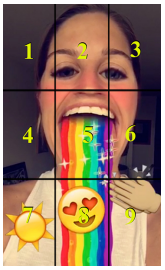


Fig. 9. **Control Image for Survey 1**—An example control image used in Survey 1 (Detecting Manipulation) for attention check. We show a 3 x 3 grid on top of the image. The participants are instructed to select region #6.

Fig. 10. **Control Image for Survey 2**—An example control image used in Survey 2 (Mitigating Manipulation) for attention check. The participants are instructed to pick a specified answer when answering the question.

Fig. 11. **D. Trump 2**—The additional image set of former United States President Donald Trump. This image set garnered similar results to the D. Trump case study (Figure 3), and thus was omitted from the main paper.