

# Understanding Blind People's Experiences with Computer-Generated Captions of Social Media Images

Haley MacLeod<sup>1</sup>, Cynthia L. Bennett<sup>2</sup>, Meredith Ringel Morris<sup>3</sup>, Edward Cutrell<sup>3</sup>

<sup>1</sup>School of Informatics & Computing  
Indiana University  
Bloomington, IN, USA  
hemacleo@indiana.edu

<sup>2</sup>Human Centered Design & Engineering  
University of Washington  
Seattle, WA, USA  
benec3@uw.edu

<sup>3</sup>Microsoft Research  
Redmond, WA, USA  
{merrie, cutrell}@microsoft.com

## ABSTRACT

Research advancements allow computational systems to automatically caption social media images. Often, these captions are evaluated with sighted humans using the image as a reference. Here, we explore how blind and visually impaired people experience these captions in two studies about social media images. Using a contextual inquiry approach (n=6 blind/visually impaired), we found that blind people place a lot of trust in automatically generated captions, filling in details to resolve differences between an image's context and an incongruent caption. We built on this in-person study with a second, larger online experiment (n=100 blind/visually impaired) to investigate the role of phrasing in encouraging trust or skepticism in captions. We found that captions emphasizing the probability of error, rather than correctness, encouraged people to attribute incongruence to an incorrect caption, rather than missing details. Where existing research has focused on encouraging trust in intelligent systems, we conclude by challenging this assumption and consider the benefits of encouraging appropriate skepticism.

## Author Keywords

Automatic image captioning; alt text; accessibility; blindness; social media; Twitter.

## ACM Classification Keywords

K.4.2. Social issues: Assistive technologies for persons with disabilities.

## INTRODUCTION

Thanks to advances in AI, computational systems are now capable of automatically generating captions describing objects, people, and scenery in images (e.g., [9,12,27]). While these systems vary in accuracy, they are prominent enough that we are beginning to see them integrated into social media platforms (e.g., [33]). One group that stands to benefit from these advancements are blind and visually

impaired people (BVIP), who have expressed frustration with increasingly visual content on social media [22,29,32]. Knowing the content of images in social contexts allows BVIPs to more fully participate in social conversations, get more information from news, and enjoy entertainment or humor [22]. Automatic captioning tools have the potential to empower BVIPs to know more about these images without having to rely on human-authored alt text (which is often missing [4,13,26]) or asking a sighted person (which can be time consuming or burdensome [3,5,6]).

Because AI systems used for image captioning were not designed with BVIP use as a primary scenario, they have been evaluated by measuring the similarity between a machine's output and that of a sighted human [16,23]. Sometimes this is done through user studies where sighted people are asked to rate the quality of a caption for a given image [25,27] or choose the best from a series of captions [9]. This does not take into account the experiences of BVIPs, who cannot use the visual image as a reference point and would be experiencing these captions outside of a controlled experimental context.

In this paper, we explore how blind and visually impaired people experience automatically generated captions on social media. We focus specifically on Twitter, since it is heavily used by BVIPs [5] to access a range of types of content (news, humor, etc.) across different relationships (personal, professional, strangers) [22]. Using a contextual inquiry approach, we find that BVIPs place a great deal of trust in these captions, often filling in details to resolve differences between a tweet's text and an incongruent caption (where the image caption does not seem to match the content or context of the tweet). We build on these findings by conducting an online experiment to explore this phenomenon on a larger scale and investigate the role of caption phrasing in encouraging trust or skepticism. Our findings suggest that captions worded in a way that emphasize the probability of error, rather than correctness, encourage BVIPs to attribute incongruence to an incorrect caption rather than to missing details. The specific contributions of this work are:

- 1) A description of blind people's experiences with automatically generated image captions.
- 2) An evaluation of the role of phrasing on a blind person's trust in a caption.
- 3) Recommendations for automated captioning systems and further areas of research.

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 the author(s) 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).

CHI 2017, May 06 - 11, 2017, Denver, CO, USA

Copyright is held by the owner/author(s). Publication rights licensed to ACM.

ACM 978-1-4503-4655-9/17/05...\$15.00

DOI: <http://dx.doi.org/10.1145/3025453.3025814>

## RELATED WORK

### Social Media & Visual Impairments

Social media use continues to be popular, and this is also true among BVIPs [5,22,29,32]. Specifically, previous work has explored BVIP experiences on Facebook and Twitter. Wu and Adamic [32] found many similarities between blind and sighted Facebook users, but noted that while visually impaired users post more status updates, they post and engage with fewer photos. Voykinska et al. [29] explored this phenomenon further, focusing on how blind people interact with visual content on Facebook. They describe blind users' interest in participating in common Facebook activities such as posting a photo or commenting on a friend's photo, alongside a fear of getting something wrong and a frustration with the lack of useful contextual information. The authors outline strategies people use to interpret visual content, finding that author-generated descriptions, when available, were considered the most useful. More recently, Facebook released an automatic image captioning tool that provides a list of objects and scenes identified in a photo [33].

In Brady et al.'s [5] study of how BVIPs use social media sites, they found that blind people are especially active on Twitter because of its simple, text-based interface. In June 2011, Twitter added capabilities for photo sharing [11] and Morris et al. [22] described the impact this change had on BVIPs. Their study highlighted challenges with the increasing prevalence of images. Since this time, Twitter added the ability to include alt text descriptions of images with a tweet [19], but this setting must be enabled explicitly.

### Automatic Image Captioning

While existing research has examined ways of generating better captions manually [21], an interesting new approach to addressing the lack of descriptions for social media images is to generate these alt text captions automatically. Research in this area has generally been focused on the necessary artificial intelligence advancements (particularly in computer vision and natural language generation) to be able to accomplish such a task with a reasonable level of accuracy [10,17,23,25,27,34]. Recent successes have resulted in the deployment of such tools in a variety of applications, including as part of social platforms [33].

These solutions are typically evaluated using standardized metrics measuring the similarity between a machine's output and that of a sighted human [16,23]. These metrics help compare various algorithms when they are run on common datasets. Researchers also often conduct user studies, asking sighted individuals to rate the quality of a caption for a given image [25,27], or asking them to choose the best of a series of captions [9] to assess the quality difference between human-authored captions and machine-generated captions. This makes sense given that most work on automated captioning is not focused on generating alt text for BVIPs, but rather is motivated by scenarios like providing metadata to improve image search. The evaluation criteria for caption

quality and the cost/benefit tradeoffs for different types of errors are different than they would be if designed with accessibility as the primary scenario. The fact that such systems are now being repurposed for accessibility purposes (as in [33]) requires a reexamination of their fundamental assumptions, such as what makes a good caption or what the relative risks are in the precision/recall tradeoff.

In this work, we build on previous research on BVIP experiences with visual content on social media and focus specifically on automatically generated captions. We explore how these captions are experienced *in situ* by blind users who cannot see an image and therefore do not have the same frame of reference for the quality of the caption as the sighted people that have historically evaluated these algorithms. We contribute to understanding how to adapt existing caption generation tools to accessibility scenarios.

### Framing Effects

Previous research has shown that BVIPs appreciate hearing captions provided as complete sentences, rather than as a list of keywords [33]. This requires not only advanced computer vision and deep learning techniques but also advanced techniques for converting those results to natural language. As such, some implementations (such as on Facebook [33]) choose to provide a list of keywords to better prioritize accuracy. Other implementations do integrate the captions into full sentences (such as in Microsoft's Cognitive Services Computer Vision API [36]), but little is known about the implications of the way in which these captions are worded.

We turn to research from psychology on framing effects [28] to explore different ways of wording captions to give people a more accurate sense of how reliable they are. Work on framing effects discusses how people react to information when it is framed differently, and the impact this has on decision making. Commonly this is done as a study of positive versus negative framing. For example, Marteau [20] compared differences in a decision to undergo surgery between people presented with either the phrasing "X% chance of dying" (negative framing) or "Y% chance of surviving" (positive framing), finding that people tended to be more risk-averse in the negative framing scenario and more risk-tolerant in the positive framing scenario.

Framing effects are also sometimes studied by comparing natural language and numeric risk descriptors. Young et al. [35] compared participants' perceptions of risk when presented with semantic descriptions (i.e., plain English) and with numeric values, finding that people tend to overestimate the likelihood of low probability events when provided with natural language descriptions.

Framing effects have been studied in many different contexts, and have been used to inform the design of persuasive technology within the HCI community [7,8,14,18]. In our work, we explore the application of framing to image captioning as a way of communicating confidence in the accuracy of a caption.

## CONTEXTUAL INQUIRY ON EXPERIENCES WITH IMAGE CAPTIONS IN THE TWITTER FEED

In the first phase of our research, we used contextual observations and interviews with a small number of blind Twitter users (n=6) to better understand their current experiences navigating images in their Twitter feed. We also introduce captioned images into the Twitter feed to gauge responses to the automatically generated captions.

### Method

#### Recruitment & Participants

We targeted Twitter users for our study, so we recruited participants primarily through Twitter and additionally through snowball sampling. We sent tweets from our personal and organizational accounts using hashtags relevant to our target demographic, such as *#blind* and *#ally* (a general hashtag related to accessibility). We purchased advertising on Twitter to promote our tweets, targeting our ads to Twitter users in our geographic area who had previously used hashtags relating to accessibility or visual impairments, and had interests similar to accounts relating to blindness (e.g. advocacy organizations or services used by people who are visually impaired).

To be eligible for our in-person study, participants had to be at least 18 years old and use Twitter on an iOS device using the VoiceOver screen reader. This VoiceOver screen reader converts on-screen information to aural speech. In the case of a mobile Twitter app, VoiceOver will read out loud the tweet author, the tweet text, descriptions of any emojis, any alt text associated with an image, and the number of likes or retweets. Participants were instructed to bring their personal iOS device to the study, so they could easily use their preferred settings (such as a comfortable speed for VoiceOver). Each participant received a \$150 VISA gift card as gratuity. Information about the six participants who took part in this study can be found in Table 1.

#### Procedure

To ground our understanding of blind people's experiences on Twitter, we began by asking participants to open their preferred Twitter application and use it as normal. We asked them to think out loud, describing what they were doing. We used example tweets and images from their personal feeds to ground conversations around their experience with images in the feed and with Twitter as a whole.

#	Gender	Age	Visual Impairment	App
P1	F	57	Blind since birth	Twitterrific
P2	M	34	Blind since birth	Twitterrific
P3	F	56	Blind since birth	Twitterrific
P4	M	24	Blind since birth	Twitterrific
P5	M	35	Minimal light perception since birth	Twitter
P6	F	52	Minimal light perception for past 7 years.	Twitter

Table 1.: Contextual Inquiry Participant Demographics.

We then directed participants to a specific Twitter account we created. This account consisted of 14 tweets, each with an image and an alt text caption. The tweets in this account were collected from hashtags that were trending during the period leading up to the study (e.g., *#ThrowbackThursday*, *#MyUnOlympicEventWouldBe*, *#AlwaysMakesMeSmile*, *#WhenIWasYoung*) as well as from popular Twitter accounts (e.g., Ellen DeGeneres, BuzzFeed, The New York Times). We selected tweets to represent a range of topics including memes, humor, news, politics, and celebrities.

We used the Microsoft Cognitive Services API's [37] to automatically generate captions for the images. These captions are full sentence descriptions of objects and known celebrities in the image. We also used Optical Character Recognition (OCR) to transcribe text in images that were text-heavy. An important note is that these captions ranged substantially in their accuracy: some captions were mostly accurate and complete (6 captions), some were correct but were missing some important information (4 captions), and some were completely wrong (4 captions). The Microsoft Cognitive Services Computer Vision API [36] provides a confidence measure of the caption results, and Microsoft's CaptionBot.ai tool [38] integrates these into the full sentence descriptions, typically in the form "I'm not really sure but I think it's..." or "I'm 98% sure that's..."

We asked participants to search for this Twitter account and go through these tweets, again thinking out loud. We told participants that the images' alt text came from a computer algorithm, and explained that phrasing like "I'm not really sure but I think..." or "I'm 98% sure that is a picture of..." would give them a sense of how confident the algorithm was in the caption it had generated. We warned them explicitly that captions would sometimes be wrong.

After participants reviewed these tweets, we asked questions about their experience with the captions and what they found useful. We then asked how much they trusted the accuracy of the captions and whether there were any captions they thought were wrong or incomplete. We revealed some of the tweets that had entirely incorrect captions and provided more accurate descriptions of the images. We additionally asked how they would behave had these tweets shown up in their regular news feed, and whether they preferred more information (even if it was sometimes wrong) or less information (but have it be more reliable).

#### Analysis

We recorded audio and video of our observations and interview sessions with participants. Each study session took approximately one hour. We iteratively analyzed these sessions using open and axial coding

#### Findings

Participants tended to trust the captions, commenting "*I felt pretty confident about them being accurate*" (P2). P1 explained, "*I have to trust them because I don't have any form of reference...Unless somebody sighted were saying 'I*



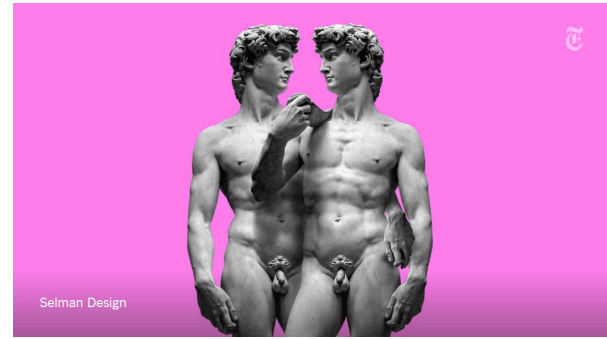
**Figure 1:** Image tweeted by Hillary Clinton used as a probe. The tweet text reads, *"Some on the other side may say our best days are behind us. Let's prove them wrong."* The computer generated caption says, *"I am not really confident, but I think it's a man is doing a trick on a skateboard at night."* The image is actually a black and white photo of Hillary Clinton walking onto a stage with a crowd of people in the background.

*don't know why they said that, that's not what it is at all... You know if somebody were sitting beside me and said 'That's not a dog, that's a something else'...then I would have to say 'well gee, somebody wrote that incorrectly'"*

Some participants expressed skepticism when asked about trust in captions in general, but this was not reflected in their behavior during the think-aloud protocol. For example, P3 claimed that she was only 50% trusting of the captions and said that she normally confirms the content of an image with her daughter before taking an action. She explains, *"if I'm going to retweet something, I want to know that I'm retweeting what I think it is... I definitely wouldn't retweet that without somebody looking at it first. And after all these years my daughter is getting to the point where's she's like 'moooooommmmm' 'Oh just tell me! Minimum words! I know you can do it in 145 characters!' and I'll go 'okay, I'm going to retweet that to people I know'. So I don't retweet a photo unless I truly know."* Yet, when it came to specific examples from our set of tweets (in this case, the Hillary Clinton tweet from Figure 1) she expressed, *"I probably would have just retweeted it thinking it was a photo of a skateboarder."* This could have potential social consequences for her: *"I would have been so in trouble by my friends...People would be like, 'I thought we don't talk about politicals...', 'I thought we keep this clean....', 'Are you supporting her?' I'd be all like 'calm down, calm down I thought it was this.'"*

#### Resolving Dissonance

Participants described assessing the accuracy of the caption based on how well it matched their expectations given the tweet text. P4 explained that his trust varied depending on this mismatch: *"When it made sense, I tended to trust it. When it didn't make sense or I couldn't see how it related, I was like 'where does this come from?' Maybe I don't trust it."* Yet when participants were reviewing the tweets (after they had been warned captions *could* be wrong, but before they had been told which ones *were* wrong), they resolved this dissonance not by suggesting that the caption was wrong, but by describing scenarios or explanations that would



**Figure 2:** Image tweeted by the New York Times used as a probe. The tweet text reads, *"Why isn't Italy kinder to gays?"* The computer generated caption says, *"I'm not really confident, but I think it's a person holding a tennis racquet and she seems neutral face."* The image is actually of two copies of the statue of David digitally altered so one has his arm around the other and they are looking at each other.

connect an unexpected image to the tweet text. In one example, P2 reviewed the tweet in Figure 2 and resolved this dissonance with the explanation, *"Okay, so it's a tennis player. Who, if I recognized tennis players, maybe she's gay."* Similarly, P3 explained the incongruence in Hillary Clinton's tweet in Figure 1 by filling in a detail about it being an older man on a skateboard: *"If you say, there's an older man on a skateboard at night that would make more sense. That's the way I take 'our best days are behind us' because I'm an old lady...which is not true because I've seen people in their 90s do amazing things."*

P5 didn't have a resolution for the dissonance, but instead thought the answer might be found in the linked article and was curious to find out more: *"Like huh. Why is a tennis racket involved in this story? ... like the way pictures are eye catching to people, it would engage a different side of my perception. Saying, oh tennis racket. I don't play tennis, but maybe that's even more interesting than I thought. Because, why is there a tennis racket involved in this story?"*

#### Awareness of Disclaimers

Participants showed signs of at least being aware of the disclaimers and confidence warnings, sometimes stopping to comment on them. P1 stopped after a tweet to remark *"That's funny..."* explaining, *"Just their comment, 'I think it's a tree but I'm not sure.' It's like, okay, whatever."* P3 was similarly amused by the captions, mimicking *"He THINKS it's a plate of pasta and broccoli? You don't know????? How bad is that picture...?"* P5 told us that the confidence levels were the least useful of all the information that had been presented to him, stating that *"Me personally I know, and I think most people would understand, that it is a computer algorithm."*

#### Expertise on Computer Vision

Participants worked hard to explain the failings of the captioning algorithm or described how they imagined it worked. Several participants brought up some background they had in software development or some familiarity with

technology that they credited for this understanding. As P4 described, *“I don’t see how a table looks like another guy at all. Or a tree looks like an infographic. I mean it’s got text it should be able to recognize that. I’m a software developer for a living so I’d say my opinions are kind of educated.”*

Even participants who did not mention any expertise in technology worked to understand how a caption might be wrong, often either personifying the algorithm or referring to a programmer that was responsible for the captions. P3 reviewed an image that the algorithm had labelled as pasta with broccoli and asked, *“Is [the image] fuzzy? Is it not clear enough to know the difference between a green and a tan colored food?”* After we described the food as linguini with spinach, she concluded, *“No wonder he was confused it might be broccoli. Greenness kind of confuses people who don’t know vegetables. It’s green right?”*

When describing their understanding of the limitations of such algorithms, some participants still underestimated how wrong a caption could possibly be. Before receiving the accurate description of the image in Figure 2, P5 stated, *“I would understand if it turned out to be a beach ball or something whatever looks like a tennis racquet. Or a badminton paddle or something. A racquetball racquet. I would understand if it turned out to be different.”*

#### Risk Tolerance

Participants varied in their willingness to encounter wrong captions from time to time (in exchange for more information) as opposed to having limited information (that was consistently correct). P2 preferred some risk, feeling that *“most of the time it’s going to be fairly accurate”* whereas P3 expressed that *“even if it was less information but it was what it was, I would have more confidence in retweeting it.”*

P4 was trusting of the captions before hearing descriptions of the images but quickly stated, *“It doesn’t really give me confidence if it can be THAT wrong”* after he heard accurate descriptions of some images. And yet when asked about the tradeoff between correctness and detail, he still said *“I think I’d rather have more information that could be wrong. I can make that decision whether I want to trust it or not.”*

#### Summary

Overall, our observations suggested that participants believed many incorrect captions, despite their self-reported double-checking strategies and abilities to correctly assess the accuracy of a caption. Rather than suspect a caption was wrong, they filled in details to resolve incongruences between a tweet’s text and its caption.

#### Limitations

It can be challenging to recruit BVIPs for in-person studies, and our sample size from this first study is accordingly limited. However, these exploratory contextual inquiry sessions with BVIPs revealed a number of challenges with automatically generated captions. Even when we explicitly, deliberately primed participants (telling them the captions came from algorithms, warning them the captions might be

wrong, explaining how to interpret the confidence disclaimers) issues of trust rose to the surface. Being aware of these issues made it possible for us to construct a second study (our online experiment) to explore this phenomenon on a larger scale online.

Additionally, we are conscious that behaviors and perceptions communicated in this lab study, particularly those about behaviors in response to captions and anticipated social consequences, may not be reflective of *in situ* interactions. This is a challenging area to study in the wild, because so few images on Twitter currently have any image descriptions, and to our knowledge none are automatically generated. This may be an interesting area of future study as captioning systems become more widely used.

Finally, although we did not explicitly collect data about education or employment from participants in this study, several participants referred to a background in software engineering or some relevant technical field during our discussions with them. These were people who seemed to have high numeracy skills, and we were unsure whether responses to disclaimers like *“I’m 98% sure that’s...”* might be different for different levels of education or numeracy. It’s also possible that people with strong technical backgrounds might understand captioning systems differently from people without that same experience. In our second study, we explicitly asked about education level and technical proficiency of participants and found these items were not predictive of trust in captions.

#### EXPERIMENT ON THE IMPACT OF CAPTION PHRASING ON TRUST AND SKEPTICISM

Building from our interview results, we conducted an online experiment to explore themes of trust in incongruent captions on a larger scale. We additionally explored ways of phrasing the captions to encourage a higher level of distrust in captions that intuitively seem off. This experiment was designed to answer three main research questions:

- 1) To what extent are our findings from the contextual inquiry study consistent on a larger scale?
- 2) Is research on positive and negative framing effects relevant in constructing captions to direct trust?
- 3) Is there a difference between a caption phrased using numeric values and a caption phrased using natural language with respect to directing trust?

#### Method

##### Recruitment & Respondents

As in our contextual interviews, we recruited through Twitter using the *#blind* and *#ally* hashtags. We again purchased advertising on Twitter, targeting users in the United States who had previously used hashtags relating to accessibility or visual impairments with interests similar to accounts relating to blindness. Participants had to be at least 18 years old, living in the U.S., with a visual impairment, and use Twitter. Each respondent received a \$5 Amazon gift card as gratuity.

	Positive Framing	Negative Framing
<b>Natural Language Framing</b>	"I'm really confident that's a cat sitting on a counter next to a window."	"There's a small chance I could be wrong, but I think that's a cat sitting on a counter next to a window."
<b>Numeric Framing</b>	"There's an 80% chance that's a cat sitting on a counter next to a window."	"There's a 20% chance I'm wrong, but I think that's a cat sitting on a counter next to a window"

Table 2: Four different framings of a caption.

We received 100 valid responses to our experiment. Participants ranged in age from 20 – 66 ( $M=33.8$ ). 30% identified as female and 69% as male. They had been experiencing their visual impairment for 1 – 66 years ( $M=12$ ) and 21% had experienced their visual impairment since birth. Respondents had tweeted a median of 3,918 tweets (mean 19,653). They had a median of 270 followers (mean 707) and were following a median of 379 accounts (mean 1,812). Participants had varied educational and professional backgrounds; 38% had completed college, and 44% had a career involving technology in some capacity.

#### Procedure

We designed an online experiment using the SurveyGizmo tool, and pilot-tested the experimental questionnaire with a range of screen readers to ensure accessibility. There were four parts to this instrument:

- 1) **Background questions:** Description and length of visual impairment, age, gender, education, skill level and background in technology, frequency of Twitter use, Twitter client(s) used, and Twitter handle (used to look up number of tweets posted, number of followers, and number of accounts followed).
- 2) **Tweets and understanding:** We provided ten tweets (in random order). As in the first study, we warned participants that the captions were generated by a computer algorithm and may not always be accurate. For each tweet, participants heard the tweet author, the tweet text, and a computer-generated caption describing the tweet's associated image attachment. After each tweet, we asked participants to rate on Likert scales the extent to which the caption improved their *understanding of the image* and their *understanding of the tweet as a whole*. Drawing from research on Bayesian Truth Serum [24] (a method for eliciting truthful subjective data where objective truth is unknowable), we also asked the extent to which the caption would *be helpful to other visually impaired people*.
- 3) **Overall trust:** After having reviewed all ten tweets, respondents rated how much they *trusted the captions overall* and whether they thought *other visually impaired people could trust the captions*. We also asked them to rate the *intelligence of the computer algorithm* that had generated the captions.

Confidence	Positive Framing	Negative Framing
>85%	I'm absolutely sure	There's no way I'm wrong
61% - 85%	I'm really confident	There's a small chance I could be wrong
26% - 60%	I'm pretty sure	I'm not completely sure
≤25%	I'm only sort of confident	I have absolutely no idea but my best guess is

Table 3: Captions in four natural language bins.

- 4) **Request for further information:** We asked respondents which of the images they would like more information on and why/why not. Options for why/why not included being *interested/disinterested* in the tweet, *trusting/distrusting* the caption, finding the caption to be *sufficiently/insufficiently detailed*, or a write in option.
- 5) **Optional debriefing:** At the end of the questionnaire, respondents had the option to click a link that took them to accurate, human-authored descriptions of the images.

Respondents were randomly assigned to four possible study conditions. We varied *framing* [whether the confidence measure accompanying each caption was framed *positively* (i.e., how confident the algorithm was that the caption was correct) or *negatively* (i.e., how confident the algorithm was that the caption was incorrect)], and we varied *abstraction* [whether the confidence measure was presented as an exact *numeric* value (i.e., a percentage amount) or in a *natural language* frame (i.e., using words)], resulting in four possible variants (2 framings x 2 abstractions), as shown in Table 2. In the natural language conditions, the percentage values were grouped into four bins (Table 3). The captions and confidence values were taken from Microsoft's Cognitive Services Computer Vision API [36]. The caption text (the description of the image that follows the confidence warning) was consistent across study conditions.

The tweets in the study were taken from popular hashtags or Twitter accounts, across a range of topics. We selected images that would allow for a range of *congruence* with the tweet (i.e., how well the caption matches expectations based on the tweet), with a focus on the extreme cases; we included four tweets with *high congruence*, four tweets with *low congruence*, and two tweets in the middle.

We also selected tweets with images that would allow for a range of *confidence* (as reported by the Cognitive Services API), again focusing on the extreme cases; we included four tweets with low reported accuracy (*low confidence*), four with high reported accuracy (*high confidence*), and two in the middle (one on each side of 50%).

Note that these two taxonomies (confidence and congruence) overlap partially, but not entirely, and that the API's reported confidence level does not always match what a sighted human might assign as an accuracy score. The tweets used in this study, as well as our classifications of congruence and confidence can be found in Table 4.







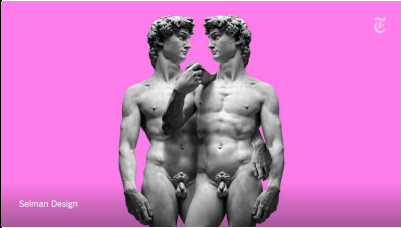
	Tweet author and text	Confidence Disclaimers	Caption Text	Image	Human-Authored Description
High Congruence	ALensOnAmerica. Share your best wildlife photo from a national park and get feedback from @chamiltonjames. #YourShot.	There's a 25% chance that's...	...a bear is swimming in the water.		The image shows two otters floating on their backs in the water. One is sleeping and the other has its mouth open as though it's smiling.
		There's a 75% chance I'm wrong, but I think that's...			
		I'm only sort of confident that's...			
		I have absolutely no idea, but my best guess is that it's...			
High Congruence	The Bloggess. I feel like I'm intruding in my own kitchen. "I'll just see myself out. I didn't know this room was reserved."	There's an 80% chance that's...	...a cat sitting on a counter next to a window.		There are three cats sitting on the counters in a kitchen staring somewhat menacingly at the camera.
		There's a 20% chance I'm wrong, but I think that's...			
		I'm really confident that's...			
		There's a small chance I could be wrong, but I think that's...			
High Congruence	Buzzfeed. 19 of the most disrespectful sandwiches ever made.	There's a 52% chance that's...	...a close up of a sandwich.		The image is two photos of open faced sandwiches. The first is an English muffin with a tiny piece of ham on one half and a tiny fried egg on the other half. The second is two slices of bread and one tiny piece of cheese on one half.
		There's a 48% chance I'm wrong, but I think that's...			
		I'm pretty sure that's...			
		I'm not completely sure, but I think that's...			
High Congruence	Whole Foods Market. Ready for #summer entertaining? Plan w/ special diets in mind. #glutenfree #paleo #vegan	There's a 94% chance that's...	...a plate of food with broccoli.		The image shows a plate of pasta with a green pesto sauce and fresh basil leaves mixed in.
		There's a 6% chance I'm wrong, but I think that's...			
		I'm absolutely sure that's...			
		There's no way I'm wrong about this. It's...			
Medium Congruence	Fizzington Womack. For the last time, quit pointing that thing at me. I am not a Pikachu! #PokemonGo	There's a 99% chance it's ...	...a cat sitting on a couch.		It's a cat sitting on a couch.
		There's a 1% chance I'm wrong, but I think it's a...			
		I'm absolutely sure it's...			
		There's no way I'm wrong about this. It's...			
Medium Congruence	Harvard Business Review. 3 psychological principles to help us interpret recent political events — and be better at predicting future outcomes.	There's a 99% chance that's...	... a table topped with plates of food and Donald Trump.		The image shows a dinner table with plates and cutlery all set out. Near the middle of the table there's an image of Donald Trump's head. It sort of looks like it was printed on paper and cut out and placed on the table.
		There's a 1% chance I'm wrong, but I think it's...			
		I'm absolutely sure it's...			
		There's no way I'm wrong about this. It's...			

Table 4. Tweets used in the online experiment (continued on next page).

Table 4: Tweets used in the online experiment (continued from previous page).

Low Congruence	Hillary Clinton. Some on the other side may say our best days are behind us. Let's prove them wrong.	There's a 23% chance that's...	...a man doing a trick on a skateboard at night.		The image shows Hilary Clinton walking onto a stage in front of a large crowd of people.
		There's a 67% chance I'm wrong, but I think that's...			
		I'm only sort of confident that's...			
		I have absolutely no idea but my best guess is that it's...			
	The New York Times. Why isn't Italy kinder to gays?	There's a 10% chance that's...	...a person holding a tennis racquet.		The image shows two nude, male statues (both are the Statue of David) next to each other over a bright pink solid background.
		There's a 90% chance I'm wrong, but I think that's...			
		I'm only sort of confident that's...			
		I have absolutely no idea, but my best guess is that it's...			
	Seattle Police Department. Bank robber arrested after abandoning pants.	There's a 5% chance that's...	...a person on a surf board in a skate park.		The image shows text from a poem by Shel Silverstein called Something Missing. The poem goes like this: I remember I put on my socks, I remember I put on my shoes. I remember I put on my tie That was painted In beautiful purples and blues. I remember I put on my coat, To look perfectly grand at the dance, Yet I feel there is something I may have forgot— What is it? What is it? ...
		There's a 95% chance I'm wrong, but I think that's...			
		I'm only sort of confident that's...			
		I have absolutely no idea, but my best guess is that it's...			
	ESPN. A massive gator interrupted a group's golf round ... in Florida, of course.	There's a 98% chance that's...	.....a horse standing on top of a lush green field.		The image shows a golf course with an alligator walking across it.
		There's a 2% chance I'm wrong, but I think that's...			
		I'm absolutely sure that's...			
		There's no way I'm wrong about this. That's...			

### Analysis

In our analysis, we primarily considered two independent variables, each with two levels: 1) *framing* of the caption (positive vs. negative) and 2) the *abstraction* of the caption (numeric vs. natural language). Additionally, we sometimes considered 3) the *congruence* of the caption (high vs. low) or 4) the *confidence* reported in the caption (high vs. low). When we consider congruence, we exclude tweets with medium congruence as these are cases that did not fit cleanly into either the high or the low category (or could arguably fit into either). We similarly exclude the tweets with medium confidence when looking for differences by confidence level.

We focused our analysis on a number of different dependent variables. From part two of the experimental instrument we collected three Likert scale ratings for **understanding** (“This caption improved my understanding of the image,” “This caption improved my understanding of the tweet as a whole,” “I think other visually impaired people would find this

caption helpful”). We will refer to these three questions of **understanding** as U1, U2, and U3 throughout our findings. These questions were five-point scales answered for each tweet.

From part three, we collected three Likert scale ratings relating to **overall trust** (“Overall, I trust these captions,” “I think other visually impaired people could reasonably trust these captions,” and “How intelligent do you think the computer algorithm that came up with these captions is?”). We will refer to these three five-point **overall trust** scales as OT1, OT2, and OT3.

From part four, we collected data for whether or not a given caption was *sufficiently detailed* and whether or not participants had *trust that the caption is correct*. We refer to these three-point scales as D for **detailed** and T for **trust** throughout the findings. As in part two, these questions were answered for each tweet.

## FINDINGS

To explore the findings from our initial study (i.e., the contextual inquiry) on a larger scale we explored the extent to which a participant's self-reported trust in the captions overall aligned with how useful they found each specific caption (and whether this changed depending on how the caption was phrased). Implicit in this comparison are two assumptions: 1) at least one part of finding a caption to be useful in situ is believing it to be true, and 2) believing a caption to be useful may influence a decision about an image (such as whether to retweet it or not).

### Overall Trust in Captions vs. Perceived Utility of Captions

For the questions about overall trust in the captions (OT1 – OT3) we found that respondents were fairly trusting of captions overall; Likert scale responses were slightly skewed towards the upper part of the five-point scale (OT1 M=3.9, SD=0.8; OT2 M=3.7, SD=0.9; OT3 M=3.8, SD=0.8). We conducted Mann-Whitney's U tests to compare differences in trust overall (OT1 – OT3) and found no significant differences between study groups, which suggests that caption phrasing did not make a difference to one's assessment of the captioning technology's trustworthiness overall; participants reported that, as a whole, they found the captioning system to be trustworthy.

For responses to questions about the usefulness of specific captions (U1 – U3) we found that respondents found the captions to be helpful overall. As with trust, Likert scale responses were slightly skewed towards the upper part of the scale (U1 M=3.7, SD=0.9; U2 M=3.6, SD=0.9; U3 M=3.7, SD=0.9). We similarly conducted Mann-Whitney's U tests for U1 – U3 and found no significant differences between groups for these questions, suggesting that phrasing did not make a difference to one's assessment of usefulness overall.

### Differences in Trust by Caption Type

Next, we were interested in whether there were differences in trust for *different kinds* of captions. In particular, we were interested in cases of extreme (in)congruence. Tweets with *high congruence* were cases where the image caption was closely related to the tweet's context (i.e., the tweet text and author). Tweets with *low congruence* were cases where the caption was unexpected or weird, given the tweet text. These assignments were made manually by the research team and informed by our observations in the contextual interviews. Our goal by altering phrasing of tweets was to encourage BVIPs to respond to these cases of low congruence with more skepticism, considering the possibility the caption might be wrong rather than assuming it was correct and building unsupported narratives to fill in missing details.

We were also interested in cases where the caption was either *highly confident* or *highly uncertain*. In some cases, the confidence estimations were incorrect (e.g., "There's a 98% chance that's a horse standing on top of a lush green field" when there is actually an alligator in the image, not a horse), but we are nonetheless interested in understanding the extent

to which these confidence estimations play a role in one's trust or distrust of a given caption.

**Congruence.** We conducted a Wilcoxon Signed-rank test for each of the dependent variables (D, T, U1, U2, U3). We found significant differences between congruence levels for trust and all three understanding questions (Table 5). There was no significant difference in detail. (People found tweets with high congruence to be more trustworthy (M=2.0 vs. 1.9, three-point scale for T) and more helpful (3.9 vs. 3.6 for U1, 3.7 vs 3.5 for U2, 3.8 vs. 3.6 for U3, five-point scales).

T	Z=-2.59	p<0.01	r=0.13
U1	Z=-4.62	p<0.01	r=0.23
U2	Z=-2.88	p<0.01	r=0.14
U3	Z=-2.73	p<0.01	r=0.14

**Table 5: Results of Wilcoxon Signed-rank test showing significant differences in congruence for trust, understanding.**

We also examined *subsets* of tweets (only high congruence tweets and only low congruence tweets) to consider whether there were significant differences between conditions (i.e., simple effects for positive vs. negative framing or for numeric vs. natural language abstraction for the two levels of congruency). We conducted a Mann-Whitney's U test for each of the dependent variables in the two subsets of the data.

When we looked only at tweets with low congruence, we saw significant differences in trust (mean ranks of positive and negative framing were 214 and 191, respectively; respectively; U=17194, Z=-3.37, p<0.01, r=0.17 for T); people who received negatively framed captions trusted incongruent captions *significantly less often* than people who received positively framed captions (M=1.8 vs. 2.0 for T, three-point scale). ***This would suggest that people faced with negative framing are more distrusting of incongruent captions.*** There was no difference in trust when congruence was high.

We also saw significant differences in understanding. Understanding of the image (U1) was significantly different both when the tweets were highly congruent (mean ranks of positive and negative framing were 185 and 211, respectively; U=16870, Z=-2.36, p<0.02, r=0.12 for U1) and highly incongruent (mean ranks of positive and negative framing were 220 and 187, respectively; U=16076, Z=-3.19, p<<0.01, r=0.16 for U1). People who received negatively framed captions found the captions with high congruence to be more helpful to understanding the image (M=4.0 vs. 3.8 for U1, five-point scale) and the captions with low congruence to be less helpful (M= 3.4 vs. 3.8 for U1, five-point scale). This is ideal, as our goal is to encourage BVIPs to rely less on a caption that might be incorrect. People who received positively framed captions found captions with low congruence to be more helpful. This suggests that people who received positively phrased captions felt these mismatched captions added to their understanding of the image. Given that all four of our incongruent captions were at least partially wrong, it could be dangerous for an

individual to make a decision off the provided caption alone. *Negatively phrased captions better encouraged a lower level of reliance on incongruent captions, and a higher level of reliance on congruent captions.*

**Confidence.** We again conducted a Wilcoxon Signed-rank test for each of the dependent variables (D, T, U1, U2, U3) and only found a significant effect of confidence on trust ( $Z=-2.26$ ,  $p<0.03$ ,  $r=0.11$ ). People trusted tweets with high reported confidence *significantly more* than tweets with low reported confidence ( $M=2.0$  vs.  $1.96$ ). This makes intuitive sense, although we note again that the reported confidence and the actual accuracy of a caption may not always be well matched.

We also examined subsets of tweets by confidence, using Mann-Whitney's U tests. For tweets with low confidence, we found a significant difference in detail depending on the framing of the caption (mean ranks of positive and negative framing were 183 and 212, respectively;  $U=16633$ ,  $Z=-2.86$ ,  $p<0.01$ ,  $r=0.14$  for D). When tweets reported low confidence, people found negatively framed captions significantly more detailed than positively framed captions ( $M=2.0$  vs.  $1.8$ ).

## DISCUSSION

In our contextual inquiry, we saw that BVIPs are trusting of automatically-generated captions, even when they don't make sense. They filled in details or built unsupported narratives to resolve these differences, rather than suspect that captions might be wrong, even when we warned them the captions were authored by a fallible computer algorithm. We conducted an online experiment where we learned that a) our findings from the contextual inquiry study are consistent across a larger sample size and b) negatively framed captions are more appropriate for encouraging appropriate skepticism in uncertain captions (i.e., where congruence and/or confidence are low).

Thus, we recommend that automatically generated captions be phrased in a way that reinforces the possibility the caption might be wrong (negative framing). As there were no significant differences in results between the natural language and numeric groups, we leave decisions of abstraction up to the designers.

Negatively framed captions help BVIPs rely more on their intuition about a caption, rather than unquestioningly trusting a caption and making decisions based off misinformation. This may encourage people to seek a second opinion on the content of an image before taking an action (e.g., retweeting) and risking social embarrassment if they are wrong. While getting a second opinion for *every* image might be burdensome or inconvenient, the results of our experimental online study suggest that negatively framed captions help encourage skepticism most in situations where that second opinion is most important; where congruence and/or confidence are low and the risk of incorrectness is high.

There is a broader implication here, as well; while there is a great deal of current research on modelling uncertainty in intelligent systems, there is very little on *communicating* that uncertainty. In fact, there is even research on how to encourage people to have *more* trust in these intelligent systems (e.g., [1,2,15,30]). The problem here is that even the best intelligent systems or agents do make mistakes, and when there is complete trust in these systems there can also be consequences, be that social embarrassment from retweeting an image because of false information or even safety risks in the case of self-driving cars. While we, of course, encourage researchers to continue working toward improving the accuracy of these systems, we also stress the importance of being able to effectively communicate uncertainty to users so they can more appropriately decide how to behave or how much trust to place in a system. In our work, we took an initial step toward investigating ways of engendering skepticism in situations of uncertainty, recognizing that there are many other contexts where a similar investigation would be valuable.

## CONCLUSION

We conducted a small contextual inquiry where we learned that blind and visually impaired people are very trusting of even incorrect AI-generated captions, filling in details to reconcile incongruencies rather than suspecting the caption may be wrong. They described being skilled at detecting incorrect captions and as being consistent about double-checking, but this was not reflected in their behaviour. We then conducted an online experiment to validate these findings on a larger scale. Through this study, we additionally learned that negatively framed captions are best suited to encouraging distrust in incongruent or low confidence captions.

This work provided the first evaluation of full sentence algorithmically generated image captions with blind and visually impaired people, and our results have implications towards how to better adapt existing captioning systems for accessibility scenarios, such as describing social media images. This was also an initial step toward investigating ways of encouraging distrust in cases of uncertainty by intelligent systems, and we encourage researchers to pursue these questions of effective communication of that uncertainty in other domains as well.

## REFERENCES

1. Stavros Antifakos, Nicky Kern, Bernt Schiele, and Adrian Schwaninger. 2005. Towards Improving Trust in Context-aware Systems by Displaying System Confidence. In *Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices & Services* (MobileHCI '05), 9–14. <https://doi.org/10.1145/1085777.1085780>
2. Johannes Beller, Matthias Heesen, and Mark Vollrath. 2013. Improving the driver–automation interaction an approach using automation uncertainty. *Human Factors*:

- The Journal of the Human Factors and Ergonomics Society* 55, 6: 1130–1141.
3. Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, and Tom Yeh. 2010. VizWiz: Nearly Real-time Answers to Visual Questions. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology* (UIST '10), 333–342. <https://doi.org/10.1145/1866029.1866080>
  4. Jeffrey P. Bigham, Ryan S. Kaminsky, Richard E. Ladner, Oscar M. Danielsson, and Gordon L. Hempton. 2006. WebInSight:: Making Web Images Accessible. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility* (Assets '06), 181–188. <https://doi.org/10.1145/1168987.1169018>
  5. Erin L. Brady, Yu Zhong, Meredith Ringel Morris, and Jeffrey P. Bigham. 2013. Investigating the Appropriateness of Social Network Question Asking As a Resource for Blind Users. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (CSCW '13), 1225–1236. <https://doi.org/10.1145/2441776.2441915>
  6. Michele A. Burton, Erin Brady, Robin Brewer, Callie Neylan, Jeffrey P. Bigham, and Amy Hurst. 2012. Crowdsourcing Subjective Fashion Advice Using VizWiz: Challenges and Opportunities. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility* (ASSETS '12), 135–142. <https://doi.org/10.1145/2384916.2384941>
  7. Eun Kyoung Choe, Jaeyeon Jung, Bongshin Lee, and Kristie Fisher. 2013. Nudging People Away from Privacy-Invasive Mobile Apps through Visual Framing. In *Human-Computer Interaction – INTERACT 2013: 14th IFIP TC 13 International Conference, Cape Town, South Africa, September 2-6, 2013, Proceedings, Part III*, Paula Kotzé, Gary Marsden, Gitte Lindgaard, Janet Wesson and Marco Winckler (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 74–91. Retrieved from [http://dx.doi.org/10.1007/978-3-642-40477-1\\_5](http://dx.doi.org/10.1007/978-3-642-40477-1_5)
  8. Eun Kyoung Choe, Bongshin Lee, Sean Munson, Wanda Pratt, and Julie A. Kientz. 2013. Persuasive Performance Feedback: The Effect of Framing on Self-Efficacy. *AMIA Annual Symposium Proceedings* 2013: 825–833.
  9. Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language Models for Image Captioning: The Quirks and What Works. In *ACL – Association for Computational Linguistics*, 100.
  10. Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. 2015. Long-term recurrent convolutional networks for visual recognition and description. 2625–2634. <https://doi.org/10.1109/CVPR.2015.7298878>
  11. Jack Dorsey. 2011. search+photos. *Twitter Blogs*. Retrieved August 12, 2016 from <https://blog.twitter.com/2011/searchphotos>
  12. Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, 1473–1482. <https://doi.org/10.1109/CVPR.2015.7298754>
  13. Morten Goodwin, Deniz Susar, Annika Nietzio, Mikael Snaprud, and Christian S. Jensen. 2011. Global Web Accessibility Analysis of National Government Portals and Ministry Web Sites. *Journal of Information Technology & Politics* 8, 1: 41–67. <https://doi.org/10.1080/19331681.2010.508011>
  14. Jan Hartmann, Antonella De Angeli, and Alistair Sutcliffe. 2008. Framing the User Experience: Information Biases on Website Quality Judgement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '08), 855–864. <https://doi.org/10.1145/1357054.1357190>
  15. Tove Helldin, Göran Falkman, Maria Riveiro, and Staffan Davidsson. 2013. Presenting System Uncertainty in Automotive UIs for Supporting Trust Calibration in Autonomous Driving. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (AutomotiveUI '13), 210–217. <https://doi.org/10.1145/2516540.2516554>
  16. Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, and others. 2016. Visual Storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
  17. Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, 3128–3137. <https://doi.org/10.1109/CVPR.2015.7298932>
  18. Young-Ho Kim, Jae Ho Jeon, Eun Kyoung Choe, Bongshin Lee, KwonHyun Kim, and Jinwook Seo. 2016. TimeAware: Leveraging Framing Effects to Enhance Personal Productivity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 272–283. <https://doi.org/10.1145/2858036.2858428>
  19. Todd Kloots. 2016. Accessible images for everyone. *Twitter Blogs*. Retrieved August 12, 2016 from <https://blog.twitter.com/2016/accessible-images-for-everyone>
  20. Theresa M. Marteau. 1989. Framing of information: Its influence upon decisions of doctors and patients. *British*

- Journal of Social Psychology* 28, 1: 89–94.  
<https://doi.org/10.1111/j.2044-8309.1989.tb00849.x>
21. Valerie S. Morash, Yue-Ting Siu, Joshua A. Miele, Lucia Hasty, and Steven Landau. 2015. Guiding Novice Web Workers in Making Image Descriptions Using Templates. *ACM Transactions on Accessible Computing* 7, 4: 1–21. <https://doi.org/10.1145/2764916>
  22. Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P. Bigham, and Shaun K. Kane. 2016. “With Most of It Being Pictures Now, I Rarely Use It”: Understanding Twitter’s Evolving Accessibility to Blind Users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI ’16), 5506–5516.  
<https://doi.org/10.1145/2858036.2858116>
  23. Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Larry Zitnick, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating Natural Questions About an Image. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.
  24. D. Prelec. 2004. A Bayesian Truth Serum for Subjective Data. *Science* 306, 5695: 462–466.  
<https://doi.org/10.1126/science.1102081>
  25. Krishnan Ramnath, Simon Baker, Lucy Vanderwende, Motaz El-Saban, Sudipta N. Sinha, Anitha Kannan, Noran Hassan, Michel Galley, Yi Yang, Deva Ramanan, Alessandro Bergamo, and Lorenzo Torresani. 2014. AutoCaption: Automatic caption generation for personal photos. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, 1050–1057.  
<https://doi.org/10.1109/WACV.2014.6835988>
  26. Y. Shi. 2006. E-Government Web Site Accessibility in Australia and China: A Longitudinal Study. *Social Science Computer Review* 24, 3: 378–385.  
<https://doi.org/10.1177/0894439305283707>
  27. Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, and Chris Sienkiewicz. 2016. Rich Image Captioning in the Wild. *Proceedings of CVPR 2016*.
  28. Amos Tversky and Daniel Kahneman. 1985. The Framing of Decisions and the Psychology of Choice. In *Environmental Impact Assessment, Technology Assessment, and Risk Analysis: Contributions from the Psychological and Decision Sciences*, Vincent T. Covello, Jeryl L. Mumpower, Pieter J. M. Stallen and V. R. R. Uppuluri (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 107–129. Retrieved from [http://dx.doi.org/10.1007/978-3-642-70634-9\\_6](http://dx.doi.org/10.1007/978-3-642-70634-9_6)
  29. Violeta Voykinska, Shiri Azenkot, Shaomei Wu, and Gilly Leshed. 2016. How Blind People Interact with Visual Content on Social Networking Services. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (CSCW ’16), 1584–1595.  
<https://doi.org/10.1145/2818048.2820013>
  30. Michael Wagner and Philip Koopman. 2015. A Philosophy for Developing Trust in Self-driving Cars. In *Road Vehicle Automation 2*, Gereon Meyer and Sven Beiker (eds.). Springer International Publishing, Cham, 163–171. Retrieved from [http://dx.doi.org/10.1007/978-3-319-19078-5\\_14](http://dx.doi.org/10.1007/978-3-319-19078-5_14)
  31. Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI ’11), 143–146.  
<https://doi.org/10.1145/1978942.1978963>
  32. Shaomei Wu and Lada A. Adamic. 2014. Visually Impaired Users on an Online Social Network. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI ’14), 3133–3142.  
<https://doi.org/10.1145/2556288.2557415>
  33. Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Jill Schiller. 2017. Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service. In *Proceedings of the 20th ACM Conference on Computer Supported Cooperative Work and Social Computing*. <https://doi.org/10.1145/2998181.2998364>
  34. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of The 32nd International Conference on Machine Learning*, 2048–2057.
  35. Sean Young and Daniel M. Oppenheimer. 2009. Effect of communication strategy on personal risk perception and treatment adherence intentions. *Psychology, Health & Medicine* 14, 4: 430–442.  
<https://doi.org/10.1080/13548500902890103>
  36. Microsoft Cognitive Services - Computer Vision API. Retrieved August 24, 2016 from <https://www.microsoft.com/cognitive-services/en-us/computer-vision-api>
  37. Microsoft Cognitive Services. Retrieved August 24, 2016 from <https://www.microsoft.com/cognitive-services>
  38. CaptionBot - For pictures worth the thousand words. Retrieved August 24, 2016 from <https://www.captionbot.ai/>