Designing User Interface Elements to Improve the Quality and Civility of Discourse in Online Commenting Behaviors

Joseph Seering
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
jseering@andrew.cmu.edu

Tianmi Fang
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
tianmif@andrew.cmu.edu

Luca Damasco
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
ldamasco@andrew.cmu.edu

Mianhong 'Cherie' Chen
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
mianhonc@andrew.cmu.edu

Likang Sun
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
likangs@andrew.cmu.edu

Geoff Kaufman
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
gfk@andrew.cmu.edu

ABSTRACT

Ensuring high-quality, civil social interactions remains a vexing challenge in many online spaces. In the present work, we introduce a novel approach to address this problem: using psychologically "embedded" CAPTCHAs containing stimuli intended to prime positive emotions and mindsets. An exploratory randomized experiment (N = 454 Mechanical Turk workers) tested the impact of eight new CAPTCHA designs implemented on a simulated, politically charged comment thread. Results revealed that the two interventions that were the most successful at activating positive affect also significantly increased the positivity of tone and analytical complexity of argumentation in participants' responses. A focused follow-up experiment (N = 120 Mechanical Turk workers) revealed that exposure to CAPTCHAs featuring image sets previously validated to evoke low-arousal positive emotions significantly increased the positivity of sentiment and the levels of complexity and social connectedness in participants' posts. We offer several explanations for these results and discuss the practical and ethical implications of designing interfaces to influence discourse in online forums.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**;

KEYWORDS

Persuasive design; User interfaces, CAPTCHAs, Commenting, Online communities

ACM Reference Format:
Joseph Seering, Tianmi Fang, Luca Damasco, Mianhong 'Cherie' Chen, Likang Sun, and Geoff Kaufman. 2019. Designing User Interface Elements to Improve the Quality and Civility of Discourse in Online Commenting Behaviors. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3290605.3300836

1 INTRODUCTION

The proliferation of harmful behaviors in online spaces is by no means a new problem. As illustrated by the examples provided by Dibbell's classic article, "A Rape in Cyberspace" [20], through Donath's work on identity deception [22], to Phillips's work on Facebook trolls [43], evidence abounds that harassment and socio-political polarization are becoming increasingly evident dynamics on the social web. As seen in examples ranging from aggressive, targeted posts on news articles or YouTube videos to harassment campaigns organized from within subreddits [28], few sites have managed to reign in comment sections once they get out of control. For this reason, a number of sites intended for the discussion of news media, and personal interests, including IMDB, Popular Science, and even NPR, have closed their comment sections on articles and discussion forums in recent years. The inability of sites to host civil, thoughtful discussions online is a major social problem and a sobering testament

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300836

to the deterioration of the quality of discourse in online commenting.

Most interventions intended to address these behaviors tend to rely on one of two paradigms of *reactive* intervention: detection at scale that aims for the difficult-to-attain ideal of correctly classifying every problematic comment (assuming that a mechanistic definition for misbehaviors can even be reached) [39]; or social action taken by other members of a community or users of a site in response to bad actors [34, 49], noting specifically that social identity factors can have substantive impacts on behavior [48]. Mostly absent from existing work is the presence of *proactive* interventions: how can we encourage users not to misbehave in the first place — and to discourage bad behaviors before the damage is done? With few notable exceptions (such as Kriplean and colleagues' work on ConsiderIt, a public deliberation forum designed invite users to contribute to balanced discussions and to consider multiple perspectives on focal issues [29]), empirical research in this space has not uncovered promising methods for promoting more positive, productive discourse.

In order to attempt to address this problem, we used a novel design process with interaction designers, software engineers, and psychologists working side-by-side to create a set of interface design interventions, with an accompanying realistic, ecologically-valid simulation of a news article and comment section that we could manipulate freely to study in controlled, randomized experiments. Specifically, we created a range of novel CAPTCHAs, brief "tests" used widely online to differentiate humans from bots and to screen out the latter. Our CAPTCHAs contained embedded imagery or activities that, based on prior work, we predicted would evoke positive mindsets or goals — causing users to feel more self-aware, empathetic, compassionate, or simply to heighten their level of positive affect — prior to submitting a comment to a politically charged message thread. We performed two randomized, controlled experiments to test the efficacy and impact of these interventions - first, an exploratory study testing several distinct CAPTCHA interventions side-by-side, and a second, more focused follow-up study utilizing image sets previously validated to evoke particular levels of positive or negative affect.

This work is an initial venture into the challenging and relatively unexplored research space of *implicitly persuasive UI design*. The two studies reported provide initial evidence for the potential of this approach to directly impact various aspects of user commenting behavior — including the complexity of argumentation (Study 1) and the positivity of sentiment and level of social connectedness (Study 2) in response to a politically charged post and comment thread. In addition, we offer our code base for other researchers to use in their own work. More broadly, this work illustrates how carefully designed, theoretically grounded interface interventions can serve as a complement to a variety of other approaches to socio-technical change, ranging from automated detection of anti-social behavior [39] to tools that facilitate social support and solidarity [34] to movements that focus on achieving international change [32]. Our approach to interface design is guided by a premise that has been thoroughly articulated by many scholars across numerous fields [25] — that there is no such thing as a "neutral" design; we either spend the resources to understand and guide the impact of our design choices or we allow them to produce results we don't choose or understand. At the same time, we acknowledge that the use of implicit persuasive techniques introduces a host of ethical questions and considerations, as we discuss at length later in this paper.

2 RELATED WORK

We briefly explore here three different bodies of literature related to our focal goal. First, we discuss previous and current approaches to addressing anti-social behaviors online. Second, we summarize current literature on persuasive design in HCI. Third, we present a diverse set of psychological theories that form the foundations for the design of our CAPTCHA interventions.

Moderation and community management

The research studying, proposing, and creating approaches to addressing anti-social behaviors is as diverse as the types of behaviors it studies. Many researchers have taken computational approaches to the detection of misbehavior in text. Cheng, Danescu-Niculescu-Mizil, and Leskovec identified message characteristics that can be used to "flag" users early on before they can misbehave [14]. A number of papers have tackled the detection of hate speech and other negative behaviors after they have been posted, on sites from Yahoo [39] to Twitter [19] to Tumblr [8]. Chandrasekharan et al., offered a "Bag of Communities" approach that allows detection of potentially unwanted content in new contexts without the need for training data from that context by using a "mixed bag" of training data from other communities [10].

Other research focuses on the ability of other users to convince, pressure, or otherwise cause offenders to behave properly or not misbehave in the first place. While early research found that negative feedback was not effective in deterring users who create low-quality content [12], subsequent work has found situations in which it can be effective. For example, Munger found that higher status users, particularly within a user's identity group, can influence behavior change via light rebukes [36]. Seering, Kraut, and Dabbish revealed that higher status users can impact other users' behaviors simply by modeling positive behaviors, but also that bans can be very effective at deterring misbehavior among onlookers [47]. From a systems perspective, Mahar, Zhang, and Karger

[34] developed a "friendsourced" approach to moderation, whereby a user's personal network can help moderate their inbox and potentially mitigate the harmful impact of having to review harassment-laden content personally.

A third thread of research focuses on site or platform-level roles in addressing misbehavior and malicious content [25]. Massanari [35] details the impact of both Reddit's rating algorithms and its governance decisions in creating spaces for toxic content. Chandrakhesaran et al., [9] found a Reddit decision to ban two of these toxic spaces to be successful on a number of metrics, suggesting that it ultimately reduced the volume of toxic content on the site. Pater et al., take a policy approach, finding that sites' policies on harassment are generally vague, inconsistent, or even nonexistent [40]. However, platform-level shifts are particularly complex to implement and evaluate, and as such relying on them exclusively is unlikely to lead to fruitful change.

A notable missing thread in the literature is the impact of the design of specific user interface elements. While much work, including some of the above-cited research, has studied the impact of the presence of some major affordances or features, none of this work has explored the impact of "micro-interventions": small tweaks in user interfaces designed to shift users' behavior to be more positive.

Persuasive Design in HCI

The last two decades of HCI research have seen a dramatic rise in the number of technological interventions designed with the intention to persuade users to change their way of thinking or behaving. This growing body of work on "persuasive technologies" has been largely guided by the influential and widely utilized behavioral model of persuasive design developed by Fogg [23], and not surprisingly, have prioritized behavior change in domains such as health and fitness and sustainability as their focal goal [15, 18, 45, 52]. Because these interventions focus on changing behaviors through deliberative goal pursuit, they primarily employ models and methods of explicit persuasion, specifically those pertaining to persuasive communication and overt forms of social influence (e.g., the Elaboration Likelihood Model of Persuasion [42]) and social influence (e.g., foundational research on compliance and persuasive social norms: [16, 46]).

In contrast, comparatively little work in HCI has leveraged the literature from social psychology on *implicit* methods of persuasion: techniques that rely on inducing mindsets, goals, emotions, or traits in less overt ways in order to facilitate attitude and behavior change in the absence of users' conscious deliberation or intention [24]. Foremost among these techniques is *priming*: the incidental activation of mental constructs by stimuli in the present situational context, which, as prior literature has demonstrated, can produce automatic effects on a host of outcomes, including perceptions, judgments, motivations, and behaviors [3, 11, 21]. Among the advantages of implicit persuasive techniques like priming is that they are not reliant on users' motivation or awareness to change their attitudes or behaviors and they can be introduced subtly and unobtrusively in contexts to shape or shift responses. On the other hand, to be effective at changing attitudes or behaviors, priming techniques require that a target not become cognizant of their persuasive influence [2]; for this reason, priming stimuli (most commonly words or images) are either presented subliminally or introduced within tasks (such as perceptual or linguistic tasks) that are carefully separated from the context of impact. In addition, the short-lived nature of priming effects demands that the introduction of a priming stimulus be followed closely in time by the intended persuasive outcome. We note that the combination of these characteristics makes implicit persuasion techniques a potential tool for subtle manipulation of users online. We discuss this danger at length later in this paper.

Despite their ubiquity in the social psychological literature, priming has yet to make a significant entrance in the domain of persuasive design in HCI. One notable exception is the deployment of a priming manipulation by Lewis and colleagues [31]: these researchers found that displaying a photo of a smiling infant (versus a neutral stimulus, such as a hammer) alongside a web form soliciting ideas for novel uses of everyday objects activated positive affect and increased the number of creative exemplars generated by participants. In a similar vein, Riot Games received notable press coverage of their implementation of priming within their popular title *League of Legends*: specifically, a set of informal studies showed that altering the font color of game tips presented on loading screens (e.g., using blue text instead of white in a message about cooperation) reduced subsequent rates of hostility among players.[1]

Psychological Theories Informing Intervention Design

We drew inspiration from these two examples as compelling illustrations of how priming can lend itself to strategic integration within a user interface to effect positive changes in cognition and behavior. With the goals of reducing aggression and promoting greater civility in mind, we turned to the social psychological literature to identify four constructs that had potential for promoting positive interpersonal mindsets and behaviors.

Positive Affect. Affective priming - the induction of positive moods or emotional states - is one of the most widely employed of all priming applications, and prior work has successfully used exposure to happy faces [37] and smiley

[1] http://www.punchingsnakes.com/?p=771

emoticons [17] to activate underlying positive affect (even in the absence of consciously felt changes to emotion: [56]. In the same vein, the present work employed smiley faces (within a drawing task in Study 1) to prime positive affect. Study 2 utilized standardized image sets previously validated to trigger positive (or negative) affect.

Self-Awareness. Prior work has revealed a consistent link between the priming of self-focus and subsequent self-control and adherence to personal standards for conduct (e.g., [6, 7]). Study 1 employed a CAPTCHA embedding a commonly utilized technique for activating self-awareness: a sentence- unscrambling task that embeds self-related personal pronouns ("I" or "my": [1, 55]).

Empathy and Perspective-taking. Priming attentiveness and recognition of others' emotional states has been shown to be an important means of promoting empathic responses [5]. In the Study 1, we built a version of a validated instrument designed to gauge individuals' recognition of emotions - the Reading the Mind in the Eyes task [4] - directly into a CAPTCHA interface as a means of inducing empathy and perspective-taking.

Altruism and Pro-Social Goals. Prior research has repeatedly shown that priming with prosocial stimuli, such as helping-related words or imagery such as superheroes, can increase altruistic behaviors ([33, 38]). Taking the lead from these investigations, Study 1 aimed to use imagery to prime the goal of caretaking via virtually "feeding" a sad animal.

3 DESIGN AND DEVELOPMENT PROCESS

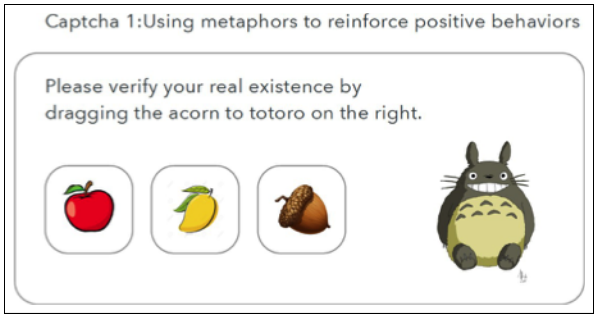
Our team for this project was composed equally of researchers with a social psychological background and interaction designers with fluency in using prototyping and programming tools. We began with a session for rapid ideation, generating more than 30 possible concepts for interventions to reduce aggression, ranging from using soothing audio cues to actively deleting a user's text if they typed a taboo word to requiring users to look at pictures of other users' faces to try to understand their feelings before they could respond.

This process allowed us to identify the idea with the most promise for our purposes: the creation of variants of traditional CAPTCHAs. CAPTCHAs have long been used as a security measure to require users to prove they are human [53, 57], and in some cases variants of CAPTCHAs have allowed collection of data to solve computational problems by tagging content [53]. Given recent advances in CAPTCHA security, particularly through the Google checkbox reCAPTCHA method, it is no longer strictly necessary for the *content* of a CAPTCHA to test a user's humanity [51].

We saw an opportunity in this — to use CAPTCHAs to expose users to specific content that induces mindsets or goals to nudge their interaction behaviors in a positive direction. This approach holds practical appeal as an intervention - compared to overhauls of moderation features, it requires relatively little redesign for sites to implement.

CAPTCHA design

During our initial ideation and prototyping phase, we developed ten initial concepts for CAPTCHAs based on the four social psychological principles described above. Figure 1 shows an early version of one of our mock-ups that was eventually used in the first of the two reported studies. In this process, a wide variety of other ideas were developed (e.g., asking users to write about their day to clicking checkboxes with positive attributes they feel describe them) and assessed for their feasibility of implementation within typical CAPTCHA formats.



Captcha 1:Using metaphors to reinforce positive behaviors

Please verify your real existence by dragging the acorn to totoro on the right.

Figure 1: Drag-and-drop CAPTCHA prototype

Over the course of our design process, we performed informal user testing with 30 volunteer participants in several waves. In our first round of testing, we showed these mock-ups to potential users and asked them to replicate the motions they would use and to comment on the process and the aesthetic of the mock-up. These tests revealed that users had a relatively low tolerance for time-consuming and/or complex activities, including checkbox tasks or tasks requiring the generation of lengthy textual content. We also found that users were even more cautious about exposing personal information, opinions, and personal values online than we had presumed. For example, users found tasks eliciting their personal values to be highly aversive, to the point that they reported a desire to leave the hypothetical site once encountered. Some users felt uncomfortable even when asked simply to write a sentence about their day (to activate self-awareness).

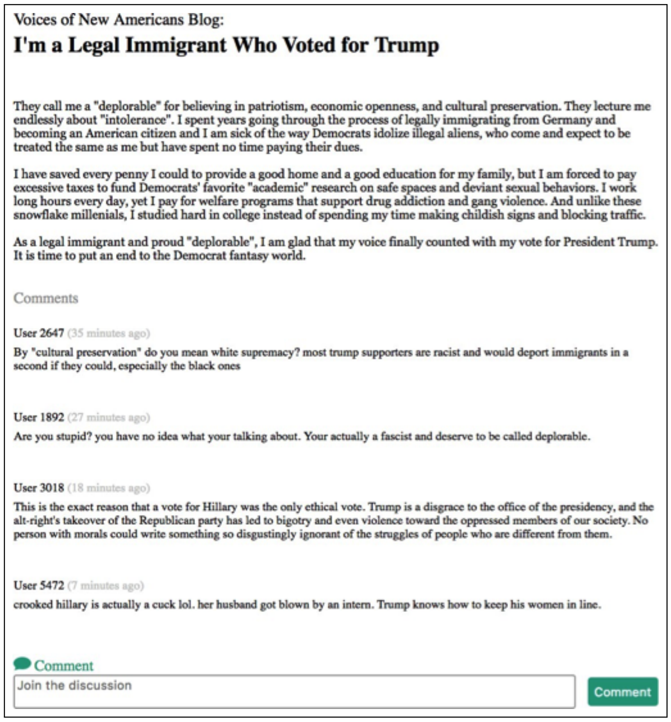
As a result, we narrowed our ideas to only those that could be completed quickly and relatively impersonally. These included variants of the three shown above plus an additional CAPTCHA that was ideated late in the process that required users to unscramble a set of words, inspired by similar tasks used in the self-awareness priming experiments mentioned above ([1, 55]). For this reason, we opted to conduct an exploratory initial study featuring a range of different CAPTCHAs utilizing different psychological mechanisms.

Political Forum and Comment Thread Design

The second part of our design process, which happened in parallel with the prototyping and refinement of CAPTCHAs in preparation for Study 1, focused on creating a meaningful environment in which to embed the CAPTCHAs. We elected to match the paradigm from Cheng et al. [13], using a simulated comment thread to capture the impact of our interventions on user behavior. We felt that it was realistic to introduce a CAPTCHA into the flow of comment posting, reasoning that users would not find it highly unusual to be asked to prove their humanity before being allowed to compose their reply. In our comment thread we wanted to offer users a relatively complete experience modeled on existing forums like Reddit and comment sections on news sites like the *Washington Post.*

While Cheng et al. showed that users could be induced to misbehave when put in a bad mood or when exposed to offensive content, we wanted to see if we could reverse this process. In order to test our interventions' impact on commenting behaviors, we needed a baseline level of "bad behaviors" to work with — that is, a stimulus that would, in the absence of any intervention, elicit a relatively high level of negative or aggressive response: a (fictitious) blog post with a strong conservative viewpoint on the topic of immigration. To balance the ideology of the post, we skewed our comments to be fairly strongly liberal, matching language from existing debate threads. Pre-testing revealed that the post elicited heated comments at a level and degree similar to the numbers achieved by Cheng et al [13]. Moreover, users reported that the post itself was high in face validity and credibility, many reporting that it triggered strong emotional responses akin to those they experienced when encountering such content in the wild. Figure 2 shows the final blog post used in both Studies 1 and 2.

One of the most difficult design decisions we made was determining at what point in a user's experience to activate the CAPTCHA intervention. We considered whether to place it on a page prior to exposure to the blog and comments, to activate it after the post was read but before comments were made visible, or to activate it upon the attempt to begin typing. If we waited until too late in the process, users' emotional reactions may have already solidified, but if we placed the CAPTCHAs too early, their impact might have faded by the time users began thinking of their comment. We decided, due to the expected short-term duration of the effects, that it was best to implement the CAPTCHA directly before users would begin typing their comments.



Voices of New Americans Blog:

I'm a Legal Immigrant Who Voted for Trump

They call me a "deplorable" for believing in patriotism, economic openness, and cultural preservation. They lecture me endlessly about "intolerance". I spent years going through the process of legally immigrating from Germany and becoming an American citizen and I am sick of the way Democrats idolize illegal aliens, who come and expect to be treated the same as me but have spent no time paying their dues.

I have saved every penny I could to provide a good home and a good education for my family, but I am forced to pay excessive taxes to fund Democrats' favorite "academic" research on safe spaces and deviant sexual behaviors. I work long hours every day, yet I pay for welfare programs that support drug addiction and gang violence. And unlike these snowflake millenials, I studied hard in college instead of spending my time making childish signs and blocking traffic.

As a legal immigrant and proud "deplorable", I am glad that my voice finally counted with my vote for President Trump. It is time to put an end to the Democrat fantasy world.

Comments

User 2647 (35 minutes ago)

By "cultural preservation" do you mean white supremacy? most trump supporters are racist and would deport immigrants in a second if they could, especially the black ones

User 1892 (27 minutes ago)

Are you stupid? you have no idea what your talking about. Your actually a fascist and deserve to be called deplorable.

User 3018 (18 minutes ago)

This is the exact reason that a vote for Hillary was the only ethical vote. Trump is a disgrace to the office of the presidency, and the alt-right's takeover of the Republican party has led to bigotry and even violence toward the oppressed members of our society. No person with morals could write something so disgustingly ignorant of the struggles of people who are different from them.

User 5472 (7 minutes ago)

crooked hillary is actually a cuck lol. her husband got blown by an intern. Trump knows how to keep his women in line.

Figure 2: Final version of comment thread page

Comment Forum Technical Development. The final version of the forum utilized in both studies consisted of a blog post, several pre-seeded comments, secondary interaction options for each of those comments, and an input where users could narrate and submit their own comments. Secondary interaction options included reporting, upvoting and downvoting, and the ability to leave a text reply (these additional interaction features were not the focus of the present research and, thus, will not be discussed further). Front end development was completed entirely using Javascript, HTML and CSS. The webpage itself was hosted with Github Pages, allowing for quick iteration.

We developed our interventions in two ways. We created the drawing interventions as standalone HTML5 apps utilizing Javascript, HTML and CSS as well as jQuery v3.3.1. The remaining interventions were developed using the Wick Editor[2], a free and open source interactive multimedia creation engine which exports HTML5 canvas applications. In order to record data, our team launched several Google Apps utilizing the Google Sheets API. When users committed a meaningful action on the page, including landing on a new page, upvoting, downvoting, reporting, replying, commenting, and starting or completing an intervention, the action was posted to the Google Sheet using a basic ajax command. Information such as the DOM element interacted with on

[2] http://wickeditor.com/

the page, the action committed, a timestamp, comment or reply text if applicable, or interaction with a CAPTCHA was collected.

Users were assigned a persistent random string of letters and numbers as an ID upon beginning the task, with which all of their actions were associated. No workers' IDs or identifying information beyond their age, gender, and political leaning, which they provided at the end of the task, were associated with their data.

This design has substantial utility as a framework for future comment-related experiments. The text of the main post can be edited freely, as can the number and text of comments. Voting, replies, and reporting, can be disabled or enabled freely. A wide variety of types of interactive interventions can easily be designed on Wick and inserted into the tool by uploading Wick's output. We encourage other researchers to use our code base in future experiments and modify it as they see fit to develop novel interface interventions and test additional empirical questions.[3]

4 STUDY 1: AN EXPLORATORY INVESTIGATION OF FOUR CAPTCHA INTERVENTIONS

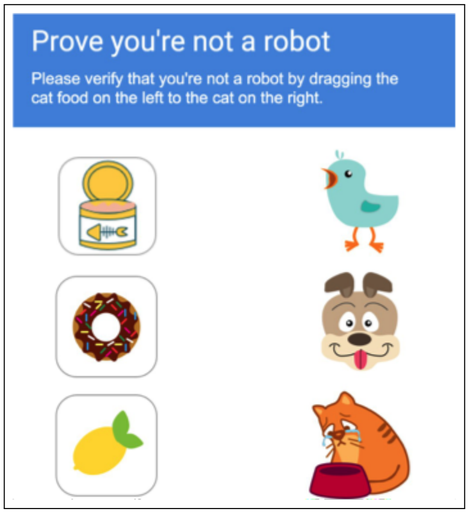
Overview

In Study 1, we tested the impact of four different CAPTCHA types, each designed to prime a specific positive emotional state in users: (1) a drag-and-drop "feeding" CAPTCHA intended to prime a helpful, altruistic mindset (see Figure 3); (2) a "face drawing" CAPTCHA intended to prime positive affect (see Figure 4); (3) a "sentence unscramble" CAPTCHA that utilized a first-person pronoun to activate self-awareness (see Figure 5); and (4) an "eye-reading" CAPTCHA intended to prime perspective-taking through emotion recognition (see Figure 7). For each "positive" version of these CAPTCHAs, we created corresponding "neutral" versions: (1) a neutral "feeding" CAPTCHA in which users dragged a fish to a box; (2) a neutral "so-so" face drawing CAPTCHA; (3) a neutral "eye-reading" CAPTCHA (utilizing the same eye images as the positive version but instructing users to select all eyes not looking at the camera); and (4) a neutral sentence unscramble CAPTCHA (replacing "I" with the gender-ambiguous name Chris). The study also utilized two control groups: a "standard CAPTCHA" condition (in which users were asked to select which images from an array of nine included houses) and a "no-CAPTCHA" condition (in which users were not given a human verification task of any kind prior to posting their comment). To measure the impact of each of the CAPTCHA conditions on user behaviors, we evaluated users' responses to the comment thread on a number of dimensions, including levels of complexity and aggression (rated by human coders) and computationally rated linguistic markers such as sociability and sentiment.

Participants

Four hundred fifty-four participants (M_{age}=38, 199 female, 245 male, 10 other) who met the requirement of being United States residents (to ensure that political content of the experimental stimulus would be relevant) were recruited from Amazon Mechanical Turk and were compensated $1 (which, given that the mean completion time for the entire study was 6 minutes, is equivalent to a pay rate of $10/hour). All participants were randomly assigned to one of the ten experimental conditions, consisting of the eight experimental CAPTCHAs, the standard CAPTCHA control, and the no-CAPTCHA control.

All procedures were approved by our university's Institutional Review Board. After completing a consent form, participants were provided with a brief description of the details of the task and were then directed to the comment thread page. On this page, they first read the blog post (see Figure 2), read through comments from other "users," and clicked the comment box to start writing their comment. Participants in all conditions, with the exception of the no-CAPTCHA control condition, were then shown the version of the CAPTCHA intervention they had been randomly assigned to receive.



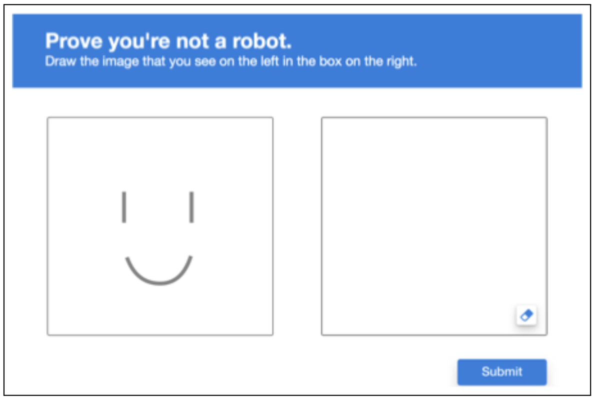
Prove you're not a robot

Please verify that you're not a robot by dragging the cat food on the left to the cat on the right.

Figure 3: Drag-and-drop CAPTCHA, "Positive" version

It is worth noting that we did not elect to require participants to solve the CAPTCHA *correctly* in this study; once they "completed" the CAPTCHA (successfully or not), they were directed back to the comment box to type their response. Nonetheless, of the five conditions where participants could

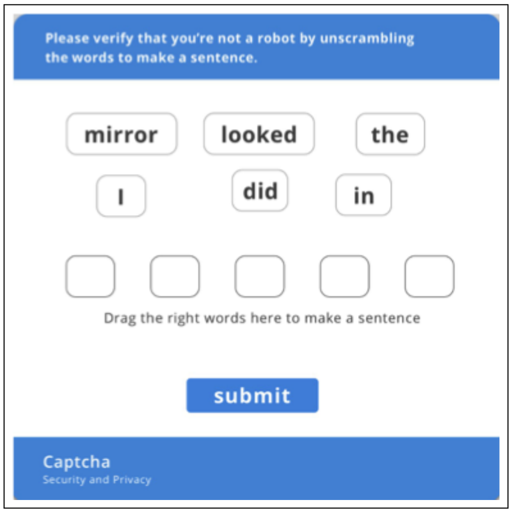
[3] A persistent version of this tool can be found at https://scomp-research.github.io/commentAnalysisAllConditions/pages/exercise.html



Prove you're not a robot.

Draw the image that you see on the left in the box on the right.

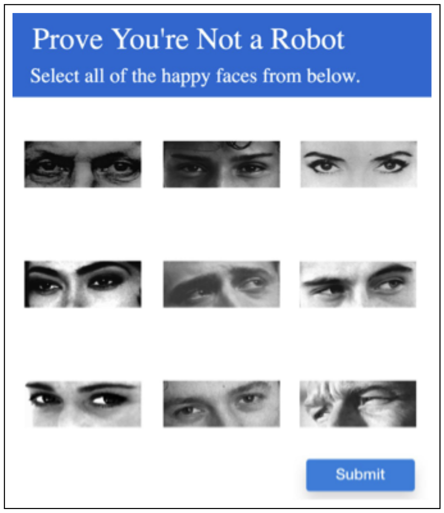
Figure 4: Drawing CAPTCHA, “Positive” version



Please verify that you're not a robot by unscrambling the words to make a sentence.

Drag the right words here to make a sentence

Figure 5: Sentence unscramble CAPTCHA, “Positive” version



Prove You're Not a Robot

Select all of the happy faces from below.

Figure 6: Eye-reading CAPTCHA: “Positive” version

possibly do the CAPTCHA incorrectly, the two drawing, two eye-reading, two sentence-unscramble, and standard CAPTCHAs, almost all did the CAPTCHAs correctly. For example, only three participants of the 96 combined participants in the Drawing conditions drew something other than the face they were shown. The error rate was highest for the eye-reading emotion identification task, but this was expected. Even in a laboratory setting, the average score on these faces was approximately 75% [4]. In our results, we found an average accuracy score of approximately 67%, unsurprising for fast-paced work. However, in subsequent analyses, correct vs. incorrect completion of a CAPTCHA did not have a significant impact on any of our analyses of the subsequent comments in any condition.

After completing the CAPTCHA, participants were directed to write their comment. Comments averaged 44 words, equivalent to three sentences. The shortest comment was one word, (“farts”) and the longest was 379 words. Participants took an average of two minutes and twenty-five seconds to write their comment, with a minimum time of one second and a maximum time of twenty-one minutes. Time taken to complete a comment was moderately strongly positively correlated with its length (r = 0.63). After submitting their comment, participants were given the option to interact with the other comments via voting, replying, or reporting, but they were told in the instructions that this was optional (as discussed earlier, these options were interaction were intended to increase the realism of the forum and comment thread and were not the focus of the present study).

Finally, participants completed a brief demographics survey. They were asked to rate their political leaning on a 7-point Likert scale, from “extremely liberal” to “extremely conservative”; to give their age; and to share their gender, with options “Male”, “Female”, and “Other” with a text entry option. The mean political leaning reported by participants was 3.42, between “moderate” and “slightly liberal.” While gender, age, and political leaning, all significantly affected several of the comment analyses to be reported below, there were no significant interactions between demographics and the CAPTCHA condition, so we do not discuss demographic results further. Following completion of all tasks, participants were presented with a debrief explaining the purpose of the study, what deception was used, and why deception was necessary. We discuss issues related to disclosure of deception at the end of this paper.

Results

To analyze the quality and sentiment expressed by participants in their comments, we used the Linguistic Inquiry and Word Count (LIWC) software package [41] to analyze participants' comments. We focused our analyses on three key dimensions that we deemed to be the most relevant for measuring the impact of the CAPTCHA interventions on commenting behaviors: *tone* (the level of positive versus negative sentiment expressed), *analytical complexity* (the level of logical, high-level thinking versus personal, low-level thinking expressed), and *social* (the level of connectedness versus separation expressed).

An example of a comment that the LIWC ranked among the lowest for sentiment was the following: *"God bless you for immigrating the right way. The demonrats [sic] have sold their soul to the illegal invaders and will pay dearly for it at the polls in November. Glad to see you wear "deplorable" like the badge of honor it is. Pay no attention to the crying snowflakes that call you names then run to their safe spaces because the truth "triggers" them. We're taking our country back and nothing can stop us!"*

And an example of a comment that the LIWC rated highly for both analytical complexity and social connectedness was the following: *"I think it's great that you immigrated the legal and proper way, but on the other side of the coin, I don't think life in Germany is the same as life in Mexico. I understand the factors related to our economy and social services and how people (citizens) would be upset. But I also think it's important to think like a human rather than an American at times, just to try and understand the other side of the issue. Many of these mexican immigrants are trying to escape extreme poverty and oppression..."*

For each of these three key dimensions, we conducted pairwise t-tests comparing the mean for each of the eight CAPTCHA conditions (positive and neutral) to the two control conditions (the no-CAPTCHA and standard CAPTCHA conditions), with a Bonferroni correction to control for familywise error (adjusted p-values are provided below). See Table 1 for a complete set of means and standard deviations by condition for the three LIWC dimensions analyzed.

Tone. Results revealed two significant pairwise comparisons for ratings of comment tone (with higher scores indicating more positive sentiment). First, the average tone rating for comments in the positive "drawing" CAPTCHA condition (in which participants drew a happy face) was significantly higher than the average tone rating in the no-CAPTCHA control condition, t(88) = 3.50, p_{adj} = .011, d = .74. Second, the average tone rating in the neutral "drag-and-drop" condition (in which participants clicked and dragged a fish to a box) was also significantly higher than the average tone rating in the no-CAPTCHA condition, t(88) = 3.07, p_{adj} = .032, d = .65. No other significant comparisons emerged.

Analytical. A similar pattern of results emerged for LIWC ratings of the analytical complexity of participants' comments. First, the average analytical rating for comments in the positive "drawing" CAPTCHA condition was significantly higher than the average analytical rating in the standard CAPTCHA control condition, t(88) = 3.29, p_{adj} = .022, d = .69. Second, the average analytical rating in the neutral "drag-and-drop" condition was also significantly higher than the average analytical rating in the no-CAPTCHA condition, t(88) = 3.13, p_{adj} = .038, d = .66. No other significant comparisons emerged.

Condition	Tone	Analytical	Social
No-CAPTCHA	44.9 (27.3)	40.6 (20.6)	11.4 (8.6)
Standard	49.6 (30.1)	38.9 (14.5)	11.8 (6.9)
Drag-Pos	49.0 (30.8)	37.0 (22.2)	12.6 (5.9)
Drag-Neutral	63.4 (29.9)	48.7 (15.2)	14.5 (7.7)
Draw-Pos	66.6 (31.3)	49.9 (17.1)	14.6 (9.5)
Draw-Neutral	49.5 (27.3)	41.7 (23.2)	12.7 (7.2)
Scramble-Pos	59.4 (29.9)	43.7 (23.0)	14.4 (10.0)
Scramble-Neutral	51.5 (30.2)	42.0 (30.2)	13.2 (7.8)
Eyes-Pos	51.6 (31.1)	41.8 (18.8)	13.8 (6.9)
Eyes-Neutral	46.7 (26.1)	42.3 (21.6)	13.9 (7.1)

Table 1: Means and Standard Deviations (in Parentheses) for LIWC Tone, Analytical, and Social Ratings by Condition

Social. Although the same pattern resulted once again for LIWC ratings of social connectedness — with participants in the neutral drag-and-drop and positive face-drawing CAPTCHA conditions exhibiting the highest mean levels of social connectedness in their comments (See Table 1), none of the pairwise comparisons achieved statistical significance when applying the Bonferroni correction to correct for familywise error.

Discussion. In sum, across two of the three linguistic dimensions — tone and analytical complexity — the two CAPTCHAs that produced significant effects on user comments were the "positive" face-drawing task and, unexpectedly, the "neutral" drag-and-drop task. The fact that the latter CAPTCHA was so effective was a surprising finding and one that does not lend itself to an obvious interpretation. One could conjecture, for instance, that the absurdity of simulating the act of placing a live fish in a cardboard box was humorous to participants. Alternatively, perhaps the act of placing the fish in a box might have unintentionally primed a mindset of caregiving or empathy; moreover, perhaps dragging the fish to the cat might have focused users on the perspective of the fish rather than the cat, rendering this capture decidedly less "positive" than originally intended.

We reasoned that, regardless of the underlying causes, these two CAPTCHAs may have been the two to elicit the highest rates of positive affect in participants. The results of a separate validation study supported this view: we recruited an additional 360 Mechanical Turk participants and randomly assigned them to receive one of the eight experimental CAPTCHAs included in Study 1 or to a no-CAPTCHA control condition and measured participants' emotional states using self-report items from the positive affect/negative affect (PANAS-1) scale [54] following exposure. Results indicated that, in line with the pattern of results from Study 1, that participants who received the "neutral" drag-and-drop CAPTCHA reported feeling significantly more "compassionate" compared to participants in the control condition. Second, participants in the "positive" drawing condition reported feeling significantly more "pleasant" compared to participants in the no-CAPTCHA control. No other differences in positive affect emerged, again paralleling the lack of results for the other CAPTCHA conditions in Study 1.

At first blush, the results for analytical complexity are not as straightforward; however, there is a theoretical basis for explaining the potential role played by positive affect in increasing the complexity of the arguments expressed by participants. For example, Isen et al. [27] found that positive affect increases the number of cognitive elements available for processing, broadens attention, and increases the diversity of cognitive elements and degree of cognitive flexibility that individuals bring to bear to a judgment and decision-making task. Moreover, Isen [26] showed that positive affect increases self-regulation; specifically, that:

> .... the cautious or avoidant responses of people in positive affect who are faced with negative material, may simply reflect sensible choices where they appear to be acceptable and appropriate. This is because, where it is clear that the negative material needs to be addressed, or where it is in the person's long-term interest to do so, people in positive affect states do engage the materials, and when they do, they demonstrate greater elaboration and coping.

In this way, the heightened positive affect triggered by the neutral dragging CAPTCHA and positive face-drawing CAPTCHA may have placed participants in a mindset characterized by the highest levels of abstraction, which, combined with their positivity of tone, produced commenting behavior that was focused more on appealing to logic than directly confronting or derogating the original poster or other commenters in the simulated forum.

In sum, Study 1 provided preliminary evidence that two of the experimental CAPTCHAs that appeared to most successful at eliciting positive affect also had the greatest effect on influencing the positivity and complexity of users' comments. Thus, at least in the context of CAPTCHA design, it appears that this intervention approach may lend itself better to the activation of positive emotions compared to the activation of perspective-taking and heightened self-awareness (as evidenced by the relative ineffectiveness of these CAPTCHAs in the present study). One reason for this could be the fact that, as discussed earlier, emotional perspective-taking and self-consciousness may be states that are more difficult or complex to prime with a short (ten-second) activity, particularly given that the tasks these CAPTCHAs were based on utilized many more items to achieve the desired effects on individuals' mindsets.

Thus, we opted to further explore the link between "affective priming" and user commenting behavior by conducting a follow-up study in which we utilized standardized image sets that had been previously validated to elicit particular affective responses in individuals, as a means of replicating the findings from Study 1 using more deliberate and systematic approach to selecting stimuli to embed within the CAPTCHA intervention.

5 STUDY 2: A CONFIRMATORY INVESTIGATION OF THE IMPACT OF AFFECTIVE PRIMING VIA CAPTCHAS

Overview

Study 2 was a conceptual replication of Study 1 focusing on the priming of positive (versus negative) affective states using CAPTCHAs including image sets validated to elicit particular types of emotional response (in terms of both valence and arousal). These images were taken from the Open Affective Standardized Image Set (OASIS: [30]), an open-source online repository of color images with accompanying normative ratings of emotional valence and arousal, the two key components included in the emotional circumplex model of affective response [44]. One key difference in Study 2 is that we were unable to include a valid "neutral" image condition. This reflected a limitation of the OASIS dataset itself; though we were careful in the positive and negative image conditions to balance the number of images with and without human faces, there was not a sufficient number of images within the neutral range of arousal/valence continuum to achieve this balance. Thus, while we acknowledge that an ideal experimental design would have included a neutral-image control condition, we opted to avoid the potential threat to internal validity that would have resulted from including a neutral condition composed of images that were qualitatively different in important respects from those used in the experimental conditions.

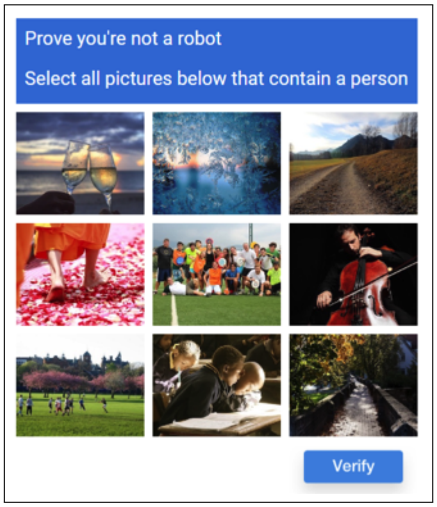
Participants

One hundred twenty participants (M_{age}=34, 57 female, 58 male, 5 non-binary/other) were recruited from Amazon Mechanical Turk to take part in the study and were compensated

$1. All participants were randomly assigned to one of the four experimental conditions (corresponding to the CAPTCHAs that included image sets validated to elicit either low-arousal positive affect, high-arousal positive affect, low-arousal negative affect, or high-arousal negative affect: see Figures 7-10) or to a no-CAPTCHA control condition.

Procedure

All procedures were approved by our university's Institutional Review Board. With the exception of the specific CAPTCHAs used, all procedures in Study 2 were identical to those described above for Study 1. It is important to note that, while participants were not explicitly told that they would encounter images that would evoke positive or negative emotional states, they were informed that they could cease participation at any time without penalty if they encountered any stimuli that they found to be aversive. In addition, as part of the standard consent procedure, participants were provided with contact information for the investigator if they needed to seek additional support in the case of any negative responses to the study materials and procedures.



Prove you're not a robot

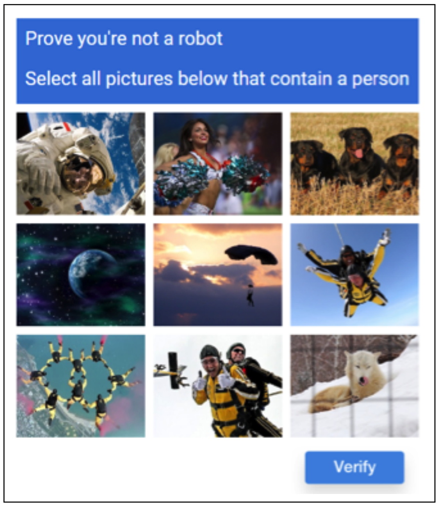
Select all pictures below that contain a person

Figure 7: Low Arousal, Positive Valence CAPTCHA

Results

As in Study 1, we utilized the LIWC analyses of tone, analytical complexity, and social connectedness to measure the impact of CAPTCHA condition on commenting behaviors. We conducted pairwise comparisons of each of the four CAPTCHA conditions to the no-CAPTCHA control, with a Bonferroni correction to adjust for familywise error.

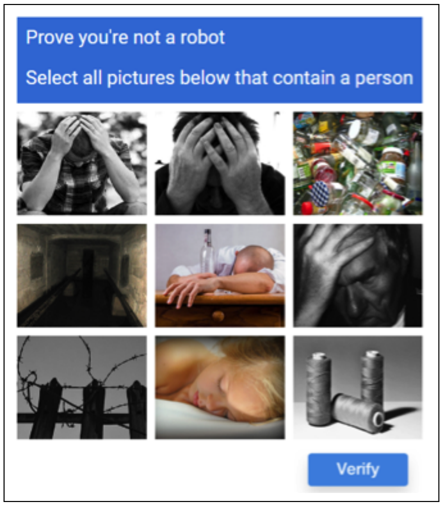
Tone. Results revealed that, relative to the no-CAPTCHA control condition, participants in the low arousal/positive valence CAPTCHA condition exhibited significantly higher levels of positive tone: t(48) = 2.22, p_{adj} = .030, *d* = .63. No other significant comparisons emerged.



Prove you're not a robot

Select all pictures below that contain a person

Figure 8: High Arousal, Positive Valence CAPTCHA



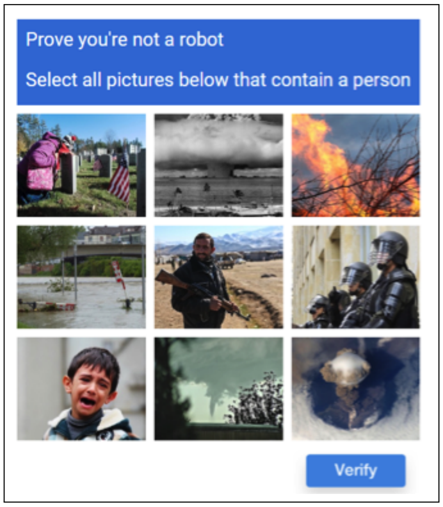
Prove you're not a robot

Select all pictures below that contain a person

Figure 9: Low Arousal, Negative Valence CAPTCHA

Condition	Tone	Analytical	Social
No-CAPTCHA	50.9 (18.6)	39.7 (24.1)	13.9 (7.0)
LowArousal-Pos	61.7 (15.7)	51.8 (20.1)	19.6 (10.0)
HighArousal-Pos	54.9 (19.0)	43.8 (19.5)	17.2 (7.8)
LowArousal-Neg	45.9 (20.8)	38.7 (22.2)	12.4 (5.7)
HighArousal-Neg	52.1 (21.8)	37.5 (21.8)	14.4 (5.9)

Table 2: Means and Standard Deviations (in Parentheses) for LIWC Tone, Analytical, and Social Ratings by Condition



Prove you're not a robot

Select all pictures below that contain a person

Figure 10: High Arousal, Negative Valence CAPTCHA

Analytical. A similar pattern of results emerged for ratings of the analytical complexity: participants in the low arousal/positive valence CAPTCHA condition exhibited significantly higher levels of analytical complexity relative to the no-CAPTCHA control condition: t(48) = 1.99, p_{adj} = .049, d = .55. No other significant comparisons emerged.

Social. Once again, the sole significant pairwise comparison revealed that participants in the low arousal/positive valence CAPTCHA condition exhibited significantly higher levels of social connection in their comments compared to participants in the no-CAPTCHA control condition: t(48) = 2.33, p_{adj} = .023, d = .66.

Discussion

In sum, across all three linguistic indicators, participants in the *low arousal, positive valence* CAPTCHA condition exhibited commenting behaviors that aligned with the valence of emotion elicited by their assigned image set: that is, more positive sentiment, complexity, and sociability in their responses relative to participants in the no-CAPTCHA control condition. These findings attest to the successful implementation of validated image sets to prime positive affect. It is particularly noteworthy that the benefits of positive affective priming were found only for the low arousal, positive stimulus set. One possible explanation for this result is that these participants in the *high arousal*, positive valence condition at least partially (mis)-attributed any increase in arousal that they experienced to the context of the heated blog post and comment thread, which may have made the negative tone of the forum more salient and served to dampen the positive affect triggered by the image set [[50].

6 GENERAL DISCUSSION

Across two randomized, controlled experiments, results revealed the promise of interventions utilizing affective priming to trigger positive emotional states and, consequently, enhance the complexity, positivity, and interpersonal connectedness of user posts. Moreover, the effect sizes for these significant results, which ranged from .50 to .70 (representing medium to medium-large effects), are on par with those reported in the most well-cited examples from the psychological literature (e.g., [3]) and higher than those in prior studies investigating interventions to address issues of incivility in online spaces (e.g., [36]).

The demonstrated success of the CAPCTHA interventions we implemented in the current work show their potential as a seamless, easily implemented technique for mitigating bad behavior. At the same time, it is equally important to acknowledge the pitfalls and ethical issues that arise in employing priming as an implicit method for influencing users in a stealthy, covert fashion. Because the efficacy of priming effects rests on participants not recognizing an explicit connection between the priming activity and the subsequent outcome that is intended to be influenced by the prime, some degree of deception is required for any intervention utilizing priming as a focal strategy. For this reason, users may be cognizant of the behaviors they are exhibiting (e.g., their use of more complex reasoning or positive sentiment) but not necessarily the situational primes that triggered the mindsets that produced them. While it is the case that interface design already inherently involves the use of priming (which textual and visual elements, such as fonts, images, and content placement intended to evoke desired responses from users), researchers and practitioners who are justifiably reluctant to use any level of deception in their design of interface may wish to consider forms of implicit persuasion that require less subterfuge to be successful. In fact, while our attempts in the present work to implicitly influence affect were not successful, persuasive attempts using the three other psychological constructs included in Study 1 — the heightening of self-awareness, the inducement of empathy and perspective-taking, and the encouragement of altruism — are effective at changing behavior even if users are aware of the mindsets induced by the interventions that employ them.

Another key ethical concern raised by the present work is the possibility that priming can be used for unscrupulous purposes. While this is already happening with great frequency in the deployment of online political propaganda, targeted marketing campaigns, spambots, and the like. — one can argue that our work provides even more specific guidelines for designing effective priming methods to achieve desired persuasive ends. We share these concerns but, at the same time, point out that there is a limit to the degree to which priming itself can prompt significant changes in consequential behaviors (such as voting preferences). Indeed, while priming can temporarily change emotional states, behaviors, and goals, it is not an effective means of inducing individuals to engage in counterattitudinal behaviors or to endorse beliefs or actions that are not already within their latitude of acceptance. Indeed, in the present work, we did not observe a change in political attitude among participants but, rather, a change in the tone and content of the communication of their beliefs. Moreover, we take comfort in the fact that, as observed in Study 2, exposing participants to negative affective imagery did not cause them to be any more aggressive or vitriolic in their responses compared to participants in the no-image control condition.

We conclude here by presenting three broad questions for future discussion and posit (perhaps controversially) that one of these three questions has already been answered, at least in part. First we must consider what forms of implicit persuasion are acceptable; what types of designs are permissible in commercial contexts? Political contexts? Health and wellness contexts? Second, we must consider what level of disclosure we think is ethically necessary. Should organizations be required (legally or otherwise) to disclose that they are using implicit persuasion? Should they be required to disclose specifically how they are doing so? And third, should anyone be using implicit persuasion at all? We suggest that this last question already has an answer; implicit persuasion is core to all designs and is inseparable from them. ‘Choosing not to design through implicit persuasion’ is equivalent to not understanding how a design will implicitly persuade.

7 CONCLUSION

In this work we used a collaborative design process, bringing together interaction designers and psychologists, to create and test a variety of CAPTCHA-based priming interventions intended to prime positive affective states and an accompanying comment thread tool designed for conducting experiments on commenting behaviors. In the process, we have also created what we believe to be an ecologically valid platform for other researchers to deploy and evaluate their own approaches to mitigating aggression and encouraging greater civility in online forums. Indeed, across both studies, which involved close to six hundred participants, only three questioned whether or not the blog post was real. As the results revealed, many participants engaged deeply with the (fake) authors of the original blog post and their fellow commenters, engaging with them to an extent that we believe would not have occurred if they had felt that they were not addressing real people.

We hope that the present research provides a first step in establishing new lines of inquiry on ‘micro-interventions’ in user interfaces. This is a new and challenging research area, and, looking ahead, the landscape is ripe for identifying and investigating any number of “entry points” in interface design for the embedding of implicit persuasive stimuli, including priming, that could directly influence the quality of tone of interactions in online communities.

ACKNOWLEDGEMENTS

We would like to thank Brooke Sachs for supporting work with the OASIS image dataset. We would also like to thank members of the eHeart Lab for feedback on research design throughout the process.

REFERENCES

[1] Hugo JEM Alberts, Carolien Martijn, and Nanne K De Vries. 2011. Fighting self-control failure: Overcoming ego depletion by increasing self-awareness. *Journal of Experimental Social Psychology* 47, 1 (2011), 58–62.

[2] John A Bargh. 2016. Awareness of the prime versus awareness of its influence: implications for the real-world scope of unconscious higher mental processes. *Current Opinion in Psychology* 12 (2016), 49–52. Issue December 2016.

[3] John A Bargh, Mark Chen, and Lara Burrows. 1996. Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology* 71, 2 (1996), 230.

[4] Simon Baron-Cohen, Sally Wheelwright, Jacqueline Hill, Yogini Raste, and Ian Plumb. 2001. The “Reading the Mind in the Eyes” test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry* 42, 2 (2001), 241–251.

[5] C Daniel Batson, Shannon Early, and Giovanni Salvarani. 1997. Perspective taking: Imagining how another feels versus imaging how you would feel. *Personality and Social Psychology Bulletin* 23, 7 (1997), 751–758.

[6] Arthur L Beaman, Bonnel Klentz, Edward Diener, and Soren Svanum. 1979. Self-awareness and transgression in children: Two field studies. *Journal of Personality and Social Psychology* 37, 10 (1979), 1835.

[7] Charles S Carver and Michael F Scheier. 1981. The self-attention-induced feedback loop and social facilitation. *Journal of Experimental Social Psychology* 17, 6 (1981), 545–568.

[8] Stevie Chancellor, Yannis Kalantidis, Jessica A Pater, Munmun De Choudhury, and David A Shamma. 2017. Multimodal classification of moderated online pro-eating disorder content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3213–3226.

[9] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can’t stay here: The efficacy of Reddit’s 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 31.

[10] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The bag of communities: identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3175–3187.

[11] Tanya L Chartrand and John A Bargh. 1996. Automatic activation of impression formation and memorization goals: Nonconscious goal priming reproduces effects of explicit task instructions. *Journal of Personality and Social Psychology* 71, 3 (1996), 464.

[12] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. 2014. Can cascades be predicted?. In *Proceedings of WWW 2014*. IW3C2.

[13] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 1217–1230.

[14] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial Behavior in Online Discussion Communities.. In *Proceedings of ICWSM*. AAAI, 61–70.

[15] Meng-Chieh Chiu, Shih-Ping Chang, Yu-Chen Chang, Hao-Hua Chu, Cheryl Chia-Hui Chen, Fei-Hsiu Hsiao, and Ju-Chun Ko. 2009. Playful bottle: A mobile social persuasion system to motivate healthy water intake. In *Proceedings of the 11th International Conference on Ubiquitous Computing*. ACM, 185–194.

[16] Robert B Cialdini. 1987. *Influence*. Vol. 3. A. Michel Port Harcourt.

[17] Montserrat Comesaña, Ana Paula Soares, Manuel Perea, Ana P Piñeiro, Isabel Fraga, and Ana Pinheiro. 2013. ERP correlates of masked affective priming with emoticons. *Computers in Human Behavior* 29, 3 (2013), 588–595.

[18] Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, et al. 2008. Activity sensing in the wild: A field trial of ubifit garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1797–1806.

[19] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International AAAI Conference on Web and Social Media*. AAAI.

[20] Julian Dibbell. 1994. A rape in cyberspace or how an evil clown, a Haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society. *Annual Survey of American Law* (1994), 471.

[21] AP Dijksterhuis and John A Bargh. 2001. The perception-behavior expressway: Automatic effects of social perception on social behavior. In *Advances in Experimental Social Psychology*. Vol. 33. Elsevier, 1–40.

[22] Judith S Donath. 2002. Identity and deception in the virtual community. In *Communities in cyberspace*. Routledge, 37–68.

[23] Brian J Fogg. 2002. Persuasive technology: Using computers to change what we think and do. *Ubiquity* 2002, December (2002), 5.

[24] Bertram Gawronski and Galen V Bodenhausen. 2006. Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. *Psychological Bulletin* 132, 5 (2006), 692.

[25] Tarleton Gillespie. 2010. The politics of 'platforms'. *New Media & Society* 12, 3 (2010), 347–364.

[26] Alice M Isen. 2000. Some perspectives on positive affect and self-regulation. *Psychological Inquiry* 11, 3 (2000), 184–187.

[27] Alice M Isen, Kimberly A Daubman, and Gary P Nowicki. 1987. Positive affect facilitates creative problem solving. *Journal of Personality and Social Psychology* 52, 6 (1987), 1122.

[28] S. Jhaver, L. Chan, and A. Bruckman. 2018. The border between controversial speech and harassment on Kotaku in Action. *First Monday* 23, 2 (2018). https://doi.org/10.5210/fm.v23i2.8232

[29] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012. Supporting Reflective Public Thought with Considerit. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 265–274. https://doi.org/10.1145/2145204.2145249

[30] Benedek Kurdi, Shayn Lozano, and Mahzarin R Banaji. 2017. Introducing the open affective standardized image set (OASIS). *Behavior Research Methods* 49, 2 (2017), 457–470.

[31] Sheena Lewis, Mira Dontcheva, and Elizabeth Gerber. 2011. Affective computational priming and creativity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 735–744.

[32] Peng Liu, Xianghua Ding, and Ning Gu. 2016. "Helping others makes me happy": Social interaction and integration of people with disabilities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 1596–1608.

[33] C Neil Macrae and Lucy Johnston. 1998. Help, I need somebody: Automatic action and inaction. *Social Cognition* 16, 4 (1998), 400–417.

[34] Kaitlin Mahar, Amy X Zhang, and David Karger. 2018. Squadbox: A tool to combat email harassment using friendsourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 586:1–13.

[35] Adrienne Massanari. 2017. #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society* 19, 3 (2017), 329–346.

[36] Kevin Munger. 2017. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior* 39, 3 (2017), 629–649.

[37] Sheila T Murphy and Robert B Zajonc. 1993. Affect, cognition, and awareness: affective priming with optimal and suboptimal stimulus exposures. *Journal of Personality and Social Psychology* 64, 5 (1993), 723.

[38] Leif D Nelson and Michael I Norton. 2005. From student to superhero: Situational primes shape future helping. *Journal of Experimental Social Psychology* 41, 4 (2005), 423–430.

[39] C. Nobata and J. Tetreault. 2016. Abusive language detection in online user content. In *ACM WWW Conference*. ACM Press.

[40] Jessica A Pater, Moon K Kim, Elizabeth D Mynatt, and Casey Fiesler. 2016. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th International Conference on Supporting Group Work*. ACM, 369–374.

[41] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71 (2001).

[42] Richard E Petty and John T Cacioppo. 1986. The elaboration likelihood model of persuasion. In *Communication and persuasion*. Springer, 1–24.

[43] Whitney Phillips. 2011. LOLing at tragedy: Facebook trolls, memorial pages and resistance to grief online. *First Monday* 16, 12 (2011).

[44] Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology* 17, 3 (2005), 715–734.

[45] Stephen Purpura, Victoria Schwanda, Kaiton Williams, William Stubler, and Phoebe Sengers. 2011. Fit4life: the design of a persuasive technology promoting healthy behavior and ideal weight. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 423–432.

[46] P Wesley Schultz, Jessica M Nolan, Robert B Cialdini, Noah J Goldstein, and Vladas Griskevicius. 2007. The constructive, destructive, and reconstructive power of social norms. *Psychological Science* 18, 5 (2007), 429–434.

[47] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 111–125. https://doi.org/10.1145/2998181.2998277

[48] Joseph Seering, Felicia Ng, Zheng Yao, and Geoff Kaufman. 2018. Applications of Social Identity Theory to Research and Design in Computer-Supported Cooperative Work. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 201 (Nov. 2018), 34 pages. https://doi.org/10.1145/3274771

[49] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator Engagement and Community Development in the Age of Algorithms. *New Media & Society* (2019).

[50] Robert C Sinclair, Curt Hoffman, Melvin M Mark, Leonard L Martin, and Tracie L Pickering. 1994. Construct accessibility and the misattribution of arousal: Schachter and Singer revisited. *Psychological Science* 5, 1 (1994), 15–19.

[51] Suphannee Sivakorn, Jason Polakis, and Angelos D Keromytis. 2016. I'm not a human: Breaking the Google reCAPTCHA. *Black Hat* (2016).

[52] Anja Thieme, Rob Comber, Julia Miebach, Jack Weeden, Nicole Kraemer, Shaun Lawson, and Patrick Olivier. 2012. We've bin watching you: designing for reflection and social persuasion to promote sustainable lifestyles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2337–2346.

[53] Luis Von Ahn. 2008. Human computation. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*. IEEE Computer Society, 1–2.

[54] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology* 54, 6 (1988), 1063.

[55] Carina J Wiekens and Diederik A Stapel. 2008. I versus we: The effects of self-construal level on diversity. *Social Cognition* 26, 3 (2008), 368–377.

[56] Piotr Winkielman, Kent C Berridge, and Julia L Wilbarger. 2005. Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value. *Personality and Social Psychology Bulletin* 31, 1 (2005), 121–135.

[57] Jeff Yan and Ahmad Salah El Ahmad. 2008. Usability of CAPTCHAs or usability issues in CAPTCHA design. In *Proceedings of the 4th Symposium on Usable Privacy and Security*. ACM, 44–52.