From Facebook Regrets to Facebook Privacy Nudges

YANG WANG, PEDRO GIOVANNI LEON, XIAOXUAN CHEN, SARANGA KOMANDURI, GREGORY NORCIE, KEVIN SCOTT, ALESSANDRO ACQUISTI, LORRIE FAITH CRANOR & NORMAN SADEH*

TABLE OF CONTENTS

---

* Yang Wang is an Assistant Professor at the School of Information Studies, Syracuse University. Pedro Giovanni Leon is a Ph.D. candidate at the Department of Engineering and Public Policy, Carnegie Mellon University. Xiaoxuan Chen is an undergraduate student studying psychology at the University of Pittsburgh. Saranga Komanduri is a Ph.D. candidate at the School of Computer Science, Carnegie Mellon University. Gregory Norcie is a Ph.D. student at the School of Informatics and Computing, Indiana University. Kevin Scott is a Master's student at the Human Computer Interaction Institute, Carnegie Mellon University. Alessandro Acquisti is an Associate Professor of Information Technology and Public Policy at the Heinz College, Carnegie Mellon University. Lorrie Faith Cranor is an Associate Professor of Computer Science and Engineering and Public Policy, Carnegie Mellon University. Norman Sadeh is a Professor of Computer Science, Carnegie Mellon University. We would like to thank Rebecca Balebako, Eric Balebako, Eyal Peer, Jeffery Dyer, Abhishek Hindupur Devendraiah, Arvind Shrihari, and the members of the Privacy Nudge group at Carnegie Mellon University for invaluable research assistance. This Article is based upon work supported by the National Science Foundation under grants Nos. 0946825, DGE-0903659, and CNS-1012763 (*Nudging Users Towards Privacy*), IWT SBO Project on Security and Privacy for Online Social Networks (SPION), as well as by Google under a Focused Research Award on *Privacy Nudges*.

I. INTRODUCTION

As social networking sites (SNSs) gain in popularity, instances of regrets following online (over)sharing continue to be reported. In June 2010, a pierogi mascot for the Pittsburgh Pirates was fired because he posted disparaging comments about the team on his Facebook page.[1] More recently, a high school teacher was forced to resign because she posted a picture on Facebook in which she was holding a glass of wine and a mug of beer.[2] These incidents illustrate how, in addition to fostering socialization and interaction between friends and strangers, the ease and immediacy of communication that SNSs make possible can sometimes also negatively impact their users.

In this Article, we summarize empirical research that our team has conducted in the past few years, aimed at understanding what actions people regret having conducted in SNSs, and whether it is possible to help them avoid those regrets without diminishing the value users can extract from participating in these online communities. In particular, this Article is based on qualitative and quantitative studies investigating instances of regret on Facebook and alternatives to prevent it.[3]

---

[1] Christina Boyle, *Pittsburgh Pirate Pierogi Mascot Fired for Bashing Team on Facebook Page*, N.Y. DAILY NEWS, June 19, 2010, http://www.nydailynews.com/news/national/pittsburgh-pirate-pierogi-mascot-fired-bashing-team-facebook-page-article-1.180649.

[2] *Did the Internet Kill Privacy?* CBS NEWS (Feb. 6, 2011, 7:21 PM), http://www.cbsnews.com/8301-3445_162-7323148.html.

[3] Material in this Article was previously published in the following papers: Rebecca Balebako, Pedro G. Leon, Hazim Almuhimedi, Patrick Gage Kelley, Jonathan Mugan, Alessandro Acquisti, Lorrie Faith Cranor & Norman Sadeh, *Nudging Users Towards Privacy on Mobile Devices*, *in* PROCEEDINGS OF THE 2ND INTERNATIONAL WORKSHOP ON PERSUASION, NUDGE, INFLUENCE, & COERCION THROUGH MOBILE DEVICES 23, 23 (2011); Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon & Lorrie Faith Cranor, *"I Regretted the Minute I Pressed Share": A Qualitative Study of Regrets on Facebook*, *in* PROCEEDINGS OF THE 7TH SYMPOSIUM ON USABLE PRIVACY AND SECURITY (2011) [hereinafter Wang et al., *Regrets*]; Yang Wang, Pedro Giovanni Leon, Kevin Scott, Xiaoxuan Chen, Alessandro Acquisti & Lorrie Faith Cranor, *Privacy Nudges for Social Media: An Exploratory Facebook Study*, *in* PROCEEDINGS OF THE 22ND INTERNATIONAL WORLD WIDE WEB CONFERENCE 763, 763 (2013) [hereinafter Wang et al., *Privacy Nudges*].

With more than a billion users, Facebook has become the world's largest SNS.[4] While well-evolved norms guide socialization and self-disclosure in the offline world, in the online world it can be more difficult to identify one's audience, control the scope of one's actions, and predict others' reactions to them. As a consequence, Facebook users might not always anticipate the negative consequences of their online activities and might end up engaging in actions that they later regret.

Because they are common experiences that people can recognize and describe, we use regrets as an analytic lens to investigate users' negative experiences with Facebook. In the regret studies summarized in this paper, we asked our participants about things that they posted on Facebook and then regretted.[5] Since one of our goals was to understand how Facebook users think about regret, we used the word "regret" without defining it, and left the interpretation to our participants. In doing so, we sought to give voice to participants' own ways of understanding regrets and related concerns. After analyzing our participants' responses, we can summarize regret as a feeling of sadness, repentance, or disappointment over one's own actions and their actual or potential consequences.

While regrets in the real world have been studied extensively,[6] little work has investigated regrets in online contexts. Our work takes a first step into examining people's regrets in social media in general, and Facebook in particular. We identify different kinds of regrets, analyze their causes and consequences, and examine users' existing coping mechanisms.

To help individuals avoid regrettable online disclosures, we employed lessons from behavioral decision research and research on soft paternalism to design mechanisms that "nudge" users to consider the content and context of their online disclosures before posting them.[7]

Specifically, we describe the application of soft paternalistic interventions (or libertarian paternalism)[8] to mitigate the effects of behavioral and cognitive biases on information disclosure decisions. Using Facebook as an application domain, we explored the possibility of nudging users to make better (that is, less likely to be regretted) decisions about disclosing information in social media.

Following an iterative design–evaluate process, we designed a privacy nudge on Facebook based on results from pilot tests of previous designs. The nudging mechanism provides visual cues about the audience of a post and includes time delays before a post is published. We tested the nudge design in a

---

[4] *Top Sites*, ALEXA.COM, http://www.alexa.com/topsites (last visited July 10, 2013) (showing that Facebook has the highest traffic among all SNS sites in the United States).

[5] *See* Wang et al., *Regrets*, *supra* note 3, at 3.

[6] *See, e.g.*, Neal J. Roese & Amy Summerville, *What We Regret Most . . . and Why*, 31 PERSONALITY & SOC. PSYCHOL. BULL. 1273, 1273 (2005) (providing a meta-analysis of studies of regrets in the real world).

[7] *See* Wang et al., *Privacy Nudges*, *supra* note 3, at 763.

[8] *See, e.g.*, Richard H. Thaler & Cass R. Sunstein, *Libertarian Paternalism*, 93 AM. ECON. REV. 175, 175 (2003).

three-week field trial with twenty-one Facebook users. Quantitative analysis of our system logs does not show any statistically significant effect of the nudge on participants' posting behavior. However, a careful participant-level analysis triangulating participants' behavioral data with exit survey results reveals that the nudge did have a positive effect on some participants but not on others. This result suggests that privacy nudges have the potential to prevent unintended disclosure for some people. We discuss limitations of the current nudge design and future directions for improvement as well as implications for public policy.

II. FACEBOOK REGRETS[9]

Privacy researchers in the fields of information systems (IS), computer-mediated communication (CMC), and human-computer interaction (HCI) have studied users' privacy attitudes and use of privacy settings in the context of SNSs.[10] Less investigated is the issue of which disclosures and activities users may actually regret. We chose to directly investigate regrets on SNSs and their causes, with the ultimate goal of designing countermeasures to help users avoid them.[11]

A. *Study Methodology*

When we started this research, there was already some heated debate about Facebook privacy issues. The *New York Times* published a blog post that solicited readers to submit their privacy questions to Facebook.[12] This was a good place for us to start understanding Facebook users' opinions on this topic. We first analyzed reader comments on this blog post and then developed a survey to probe whether the concerns expressed in those comments were typical of American Facebook users. After analyzing the results from that survey, we conducted semi-structured interviews to ask in-depth questions about users' experiences on SNSs.

While the interviews capture the most memorable experiences of the interviewees, we also wanted users' daily, often mundane Facebook

---

[9] Material in this section was previously published in Wang et al., *Regrets*, *supra* note 3.

[10] *See, e.g.*, Alessandro Acquisti & Ralph Gross, *Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook*, *in* PROCEEDINGS OF THE 6TH INTERNATIONAL CONFERENCE ON PRIVACY ENHANCING TECHNOLOGY 36, 36 (2006); Bernhard Debatin et al., *Facebook and Online Privacy: Attitudes, Behaviors, and Unintended Consequences*, 15 J. COMPUTER-MEDIATED COMM. 83, 83 (2009); *see also* Adam N. Joinson, *"Looking at," "Looking up" or "Keeping up with" People? Motives and Uses of Facebook*, *in* PROCEEDINGS OF THE SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS 1027, 1028 (2008).

[11] *See* Wang et al., *Regrets*, *supra* note 3, at 2.

[12] Jenna Wortham, *Ask Facebook Your Privacy Questions*, N.Y. TIMES BITS (May 6, 2010, 3:21 PM), http://bits.blogs.nytimes.com/2010/05/06/ask-facebook-your-privacy-questions.

experiences, which they might forget or take for granted. We also hoped to explore how regrets might affect users' subsequent behavior on Facebook. For these reasons, we designed a diary study and invited the interviewees to log their daily Facebook experiences for a month. These studies raised additional questions about regrets on Facebook, and we conducted another online survey to gain further insights.

While Facebook's user population is quite diverse, the majority of prior research was conducted with college students.[13] Our research seeks to gain a more comprehensive understanding of the SNS user population by studying American Facebook users from a wider range of ages and occupations. We recruited survey participants using the Mechanical Turk crowd sourcing site and recruited interviewees from the Pittsburgh Craigslist website. We report on two surveys in this paper, and refer to them as "survey1" and "survey2." Our studies were approved as minimal risk studies by Carnegie Mellon University's Institutional Review Board (IRB).

In survey1, the interview study, and the diary study, we did not focus solely on users who had regrets on Facebook. The studies were designed to gain a better understanding of Facebook users' privacy-related experiences and behavior on Facebook. In this paper we only focus on the responses to the question: "Have you ever posted something on a social network and then regretted doing it? If so, what happened?" For survey2, however, we asked people to take our survey only if they had posted something on Facebook and later regretted it.

Survey2 contained thirty-four questions. We began by asking survey participants: "Have you ever regretted posting something (status updates, pictures, likes, comments, locations, etc.) on Facebook? For example, have you ever posted something that you felt bad about later or wished you hadn't posted?" We then asked how many times they regretted posting on Facebook in the last twelve months. In order to help participants recall specific details about their regrets, we asked them to think about the one posting that they regret the most and then answer our survey questions with respect to that post. We then asked the participants several multiple-choice and open-ended questions to learn about their post, specifically: why the post was made, what happened after the post, when the regret occurred, the reason(s) they regretted the post, how much they regretted the post, and what they did in response to the regret. We also asked about the participants' moods when they posted the regrettable content (e.g., very happy or sad) and whether they were under the influence of drugs or alcohol.

---

[13] *See, e.g.*, Acquisti & Gross, *supra* note 10, at 36; Ralph Gross & Alessandro Acquisti, *Information Revelation and Privacy in Online Social Networks*, *in* PROCEEDINGS OF THE 2005 ACM WORKSHOP ON PRIVACY IN THE ELECTRONIC SOCIETY 71, 71 (2005); Nicole B. Ellison et al., *The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites*, 12 J. COMPUTER-MEDIATED COMM. 1143, 1148 (2007).

Both survey1 and survey2 were hosted on SurveyGizmo, an online survey hosting service. Survey1 and survey2 were deployed on Mechanical Turk, each for about one week in March and May 2011, respectively. We paid each respondent $0.50.

B. *Results*

The results that we report below include data mostly from survey2 and some data from the interviews and user diaries, as well as answers to several regret-related, open-ended questions in survey1. As with the interview data, we coded the free responses from the two surveys and categorized them post-hoc to produce a list of common themes.

Our initial study was a three-part study consisting of survey1, the interviews, and the diary study. For these initial studies, we recruited Facebook users regardless of whether they had any regrets. Some of our studies gathered data on both the regrets of our study participants (first-party) and the regrets of friends of our study participants (third-party). We had a total of 340 participants from these initial studies, including 321 survey1 respondents and 19 participants in the interview/diary study. We found that 66 out of 321 survey1 respondents (21%) and 11 out of 19 (58%) interview/diary participants reported having first-party regrets. For the remainder of the paper we discuss only those participants who reported first-party regrets.

For survey1, there were 117 male respondents (36.4%) and 204 female respondents (63.6%). The average age of survey1 respondents was thirty-one years old (σ=11.0). For survey2, we had 492 valid responses. There were 216 male respondents (43.9%) and 276 female respondents (56.1%). The average age of survey2 respondents was twenty-eight years old (σ=8.6).

To protect the privacy of our research participants and to differentiate between studies, we use anonymous identifiers. The eleven participants in the interview and diary studies are denoted with P#. For instance, we use P1 to represent the first interviewee (and diary participant). Survey respondents are not identified by number. Instead, we specify which survey the data is from when we report it, e.g., "a survey1 respondent said . . . ."

1. *What Do People Regret Posting?*

In this section, we focus on participants' responses to questions of the form: "Have you posted something on Facebook and then regretted doing it? If so, what happened?" We see that regrettable postings revolve around sensitive topics (e.g., alcohol consumption, sex, politics, religion) and content with strong sentiment (e.g., arguments and criticism).

Our participants reported several types of sensitive content that they regretted posting. We loosely categorize that content here. In some cases, e.g., illegal drug use, merely posting this content is enough to cause regret. In other cases, sensitive content can be part of a deeper cause of regret. For example, we

find that profanity can sometimes be offensive on its own, or it can be used to insult others.

Many participants regretted posts about drinking. One survey2 respondent said, "I posted photos from a party that got a bit out of hand, and the photos were not very flattering. What bothered me was that I realized I posted them and my profile was public and other people could see them." He then explained why he posted them: ". . . out of habit; after an event with friends most of us post the photos." This quote suggests that the culture and norms of a person's social circle play a role in one's decision to post. In this case, most of the participant's friends post event photos.

If such posts are the norm, why did this participant regret it? He said, "I realized they weren't something I wanted other people to see that didn't know me, because they'd get the wrong idea." This highlights the issues of unintended audience (in this case, people who did not know him) and impression management. He felt uncomfortable because these photos might lead to a particular impression that violates how he wants himself to be perceived by others. He also said, "one person asked me to remove the tag of their photo." These posts can also violate others' self-representations.

Some regrettable posts mentioned illegal drugs. One survey2 respondent said, "I regretted posting a picture of me smoking marijuana at a party. People in my family seen it and other people I didn't want seeing it." He posted it because "I thought it was cool at the time. I had an I didn't care attitude." He regretted posting the picture because it embarrassed others: "Certain people around me give me a sense of disapproval when I was around them. My mom for example told me it was embarrassing for her."

Posting sexual content was another common regret. One survey2 respondent said, "I accidentally posted a video of my husband and I having sex . . . . I didn't mean to post it, I had accidentally clicked on the video of my daughter taking her first steps and on that video and they both uploaded together . . . . I didn't know I had posted it until the day after, when I logged on again, and saw all the comments from all of our friends and family, and my husbands coworkers (he's in the army)." She regretted posting "because it was a personal video between my husband and I." In this case, the posting was an accident, and not a result of failing to foresee consequences.

People can specify their religious or political beliefs in their Facebook profiles. However, posts that express these beliefs can cause debates, offend people, and damage relationships. One survey2 respondent said, "[I posted] my beliefs about religion. Because my name was also tied to my business, people who disagreed with my beliefs about religion took action against my business . . . . My business was given bad online reviews."

Postings with profanity or obscenity can be a cause of regret. One survey2 respondent said, "I said something along the lines of Hey Bob at ST, stop treating us women like trash . . . fuck you!" The profanity is often a result of the users' mood at the time when they posted the content. In this case, the

respondent explained, "I posted it because I was very angry. He is a customer at my place of business and hates women . . . . I was only venting my frustration."

Sometimes people share their personal issues to gain support, but it is tricky to balance how much to share and how much to keep private. A survey2 respondent said, "I posted that I was no longer single and I was dating this guy in my class . . . . I was happy and excited about myself . . . People read it and told my parents and they did not approve." This shows that people sometimes post things when they are in an extremely positive mood that they later regret. On the other hand, sometimes family issues are brought up when in a negative mood. One survey1 respondent wrote, "I did post something about a fight with my husband once and regretted it after he saw it and was offended that I was airing our 'dirty laundry' for everyone to see."

Participants reported that they regretted posting strongly negative or offensive comments as well as engaging in arguments on Facebook. People often post negative content because they are in a bad mood, and we heard many accounts of regret due to angry posts. One survey2 respondent said, "[I] posted a negative comment to a man I care about . . . emotions high with frustrations lashing out at him when I should instead be more in control . . . . I regret hurting him especially in writing when I can't change it later. No back button or undo. It hurts to hurt him so I regret doing it."

Our participants also reported regrets caused by posting about their work or company in a negative way. One survey2 respondent said, "When I badmouthed my job due to disciplinary I was on for b.s. stuff. My managers are my friends on facebook and ended up ugly at work." He then explained, "I was mad . . . I said it out of anger and not thinking."

2. *Why Do People Make Regrettable Posts?*

In this section, we consider the reasons why Facebook users make regrettable posts. We first describe the intended purposes of the posts, and then we explore why they turned out to be problematic and led to regret.

In many instances, users report that they had no specific purpose for posting. In others, they explain the reason behind their posts in order to explain their regrets. We categorize and explain commonly reported reasons here.

Some people reported wanting to be perceived as interesting or unique; however, when the content or behavior described in the post was controversial, this caused regret. One survey2 respondent said, "I posted a photo of me smoking hooka and got in trouble with it from my employer . . . at the time I thought it was cool. I lost my job because of it. My boss talked to me about it and told me they did not want that image in the company."

Trying to be funny is another source of regret when what was thought to be funny turns out to be offensive. One survey2 respondent wrote, "My post was about the Border Patrol not doing their job. I was trying to make an interesting event sound funny. One of my friend's husbands is an agent and [my friend] was very offended."

Users in a highly emotional state often vent their feelings on Facebook. A survey2 respondent wrote,

> I posted something about my feelings about an argument I had with a friend. I didn't mention her by name but it was fairly obvious to those who knew about the argument who I was referring to. I felt the need to vent and get the situation off of my chest. Also, I'm sure a small part of me wanted her to read it and feel bad.

Users want to express their frustration in a public forum, though they sometimes regret doing so.

Sometimes regrettable posts are made with the best of intentions. One survey2 respondent said,

> I posted something about a friend who had gained a lot of weight recently. I hadn't seen her in a long time and I just thought my friend was pregnant at the time I posted it. I was congratulating her on her upcoming pregnancy. So I asked if she was pregnant and she told me no, she had gained a lot of weight. I felt horrible.

Another survey2 respondent wanted to provide useful information but then was misunderstood. He said,

> [I] made a location check in at a club with some friends . . . to let a friend we were waiting for know we arrived. The boyfriend of one of my friends I was with thought she was cheating on him with me and they started to argue. He called me and started to yell that I was stealing his girl. He then broke up with my friend, his girlfriend.

When posting on Facebook becomes habitual, people rarely think about why they post things. The following survey2 respondent's story is telling: "I was so addicted to facebook! It's like an involuntary action. You feel something and you express that in facebook."

Some users also did not think about the potential consequences of their postings. One survey2 respondent reported posting a photo of his underage friend getting drunk and tagging him in it: "I didn't think his parents would see it, and I didn't think about any of the consequences at the time."

Users often regret things they posted while in a highly emotional state, or while under the influence of alcohol or drugs. One survey2 respondent said, "a few occasions if I was emotional or had too much to drink I wrote some things that were personal that I later took down." "Hot" states can lead to a lack of concern for consequences. One survey2 respondent said, "I told them that they are nothing but a desperate loser. I knew the post would hurt her feelings, and I would probably regret it; however, at that time I just didn't care." This respondent actually considered the possible consequences and foresaw his own regret, but posted anyway.

3. *How Do Posts Become Regrets?*

In this section we examine various errors that can lead to regret. They often stem from unforeseen or ignored consequences, but they can also be caused by a misunderstanding of SNSs and usability issues.

Users often do not remember or know who might see their Facebook content. In some cases, they were only concerned about their Facebook friends. For example, one survey1 participant said, "I once posted how frustrated I was with an interview and I regretted the minute I pressed 'share' because I suddenly realized some former employers were friends to me on Facebook."

In other cases, they regretted because people beyond their Facebook friends were involved. A survey2 respondent told us:

> It was a picture of me and my girlfriend together in front of a Waterfall kissing, nothing obscene or disturbing. I posted it because she wanted to see all the pictures we took from our trip to the waterfalls. I regret posting it because relatives saw the pictures on facebook and started commenting on it. When I thought on restricting the image it was too late because a lot of people had posted on it and the harm was already done. It became some sort of gossip in the small town I live in, especially because I hadn't told anyone, not even my parents that I had a girlfriend. So the first thing they see is me kissing my new girlfriend, and it is not a good idea coming from a catholic conservative family to let your relatives see this online. They always assume the worst.

We also heard several reports in which users' SNS content ended up in the hands of judges and prosecutors. P7 told us that he and his wife were undergoing a divorce and their fight spread into Facebook:

> My wife didn't pay spousal support . . . she posted on her Facebook that she got a job from somewhere. I took a screen shot of that post and gave it to the court and judge can use it as evidence. She was mad and blocked me on Facebook . . . . My daughter called me and suggested me to change my privacy setting to 'friends only,' and I did it.

Relatively new Facebook users tend to have problems understanding the Facebook platform, and experienced users can still be caught by surprise. For instance, one survey2 participant did not realize that it was possible for a friend's friend on Facebook to see what he posts: "I stated something about daughter's boyfriend which was observed by him through a mutual friends facebook wall."

Some users do not understand that their identities can be tied to their actions. For instance, a survey2 participant did not anticipate that the negative comments he posted on his company's fan page would be associated with him.

Facebook usability problems contribute to some user regrets. In one case we described earlier, the user accidentally posted a sexual video of hers: "I didn't know I had posted it until the day after." Facebook could better prevent users

from making these types of mistakes if it provided clear feedback on content being posted. In another case, a user said that when he posted things from his phone, he could not delete them. This user expected the same functionality from Facebook on every platform.

C. *Discussion*

We have seen from our data that users have many reasons for making posts on SNSs. For instance, a user might post things because they hope to be perceived as cool or funny. In other words, users sometimes try to present themselves in a way that matches how they want to be perceived by other people. In his influential book *The Presentation of Self in Everyday Life* sociologist Erving Goffman explains that we "perform," producing different images of ourselves depending on context, similar to the way actors perform in the theater.[14] For example, we may look or behave quite differently in a business meeting than at dinner with a close friend. This performative aspect of our lives was later conceptualized as "impression management."[15] This conceptual framework has been used to explain both offline and online behavior. In the domain of SNSs, for instance, boyd and Heer suggest that users' profiles on SNSs are dynamic performances of their online identities.[16]

Impression management theory can be used to understand the problem of unintended audience. The "wrong" self-presentation is perceived by the unintended audience. For instance, one participant explained his use of swear words: "It's inappropriate for my family. ok for friends, but not family or church friends." His comment expresses a desire to convey a different impression to each group.

Sometimes unintended audience becomes an issue when posts are taken out of their original context.[17] Philosopher Helen Nissenbaum has introduced an analytical construct called "contextual integrity."[18] She noted that "contextual integrity ties adequate protection for privacy to norms of specific contexts, demanding that information gathering and dissemination be appropriate to that

---

[14] ERVING GOFFMAN, THE PRESENTATION OF SELF IN EVERYDAY LIFE 254 (1959).

[15] *See* BARRY R. SCHLENKER, IMPRESSION MANAGEMENT: THE SELF-CONCEPT, SOCIAL IDENTITY, AND INTERPERSONAL RELATIONS 33–41 (Lawrence S. Wrightsman et al. eds., 1980).

[16] danah boyd & Jeffrey Heer, *Profiles as Conversation: Networked Identity Performance on Friendster*, *in* 3 PROCEEDINGS OF THE 39TH ANNUAL HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES 1, 5 (2006).

[17] danah michele boyd, Taken Out of Context: American Teen Sociality in Networked Publics 38 (Fall 2008) (unpublished Ph.D. dissertation, University of California, Berkeley), *available at* http://ssrn.com/abstract=1344756. *See generally* Woodrow Hartzog, *Social Data*, 74 OHIO ST. L.J. 995 (2013).

[18] *See* Helen Nissenbaum, *Privacy as Contextual Integrity*, 79 WASH. L. REV. 119, 119 (2004) (positing a "new construct, 'contextual integrity,' as an alternative benchmark for privacy, to capture the nature of challenges posed by information technologies").

context and obey the governing norms of distribution within it."[19] A teacher holding alcohol in a school or public context may conflict with its social norms, whereas the same person holding alcohol in a bar during her vacation seems reasonable with the social norms of that circumstance.

The problem is that sites like Facebook are becoming what danah boyd calls "networked publics"[20]—public places on the Internet, where different conflicting contexts and social norms coexist. We observed that some users posted troublesome content, like drinking pictures, because most of their friends post this kind of content. Thus, posting pictures of oneself drinking became the accepted norm of those users' small social circles, but this norm clashes with norms of other contexts. For example, this personal context could clash with the professional context if a user "friends" coworkers.

Even if a posting was only seen by its intended audience, it could still backfire because users cannot always foresee how others might perceive their postings. Users may not have enough information at the time of posting or they may underestimate the consequences of their posts.

We observed many incidents where people posted things when they were in an overly emotional mood ("hot" state) and later regretted their posts. For instance, one survey2 respondent said, "It was, 'I'm so fucking pissed right now.' I was overwhelmingly angry at something that had happened, and needed some sort of outlet. At the time, Facebook made sense, for some reason." We also found that when people were overly happy or excited, they could also post things they later regretted. We observed one example where a girl posted that she was excited about dating a new boyfriend, but her parents saw the post and disapproved of this relationship.

In social science literature, researchers have shown that being emotional may cause people to behave irrationally. Behavioral-economist George Loewenstein showed that visceral influences overwhelm logical thinking and contribute to people being "out of control."[21] Another survey2 respondent's experience was a telling example: ". . . emotions high with frustrations lashing out at him when I should instead be more in control."

To help individuals avoid regrettable online disclosures, we employed lessons from behavioral decision research and research on soft paternalism to design mechanisms that "nudge" users to consider the content and context of their online disclosures before posting them.

---

[19] *Id.*

[20] *See* boyd, *supra* note 17, at 15. *See generally* Tal Z. Zarsky & Norberto Nuno Gomes de Andrade, *Regulating Electronic Identity Intermediaries: The "Soft eID" Conundrum*, 74 OHIO ST. L.J. 1335 (2013).

[21] George Loewenstein, *Out of Control: Visceral Influences on Behavior*, 65 ORGANIZATIONAL BEHAV. & HUM. DECISION PROCESSES 272, 275–76 (1996).

III. FACEBOOK PRIVACY NUDGES[22]

For several decades, social scientists have pointed to the role of heuristics and cognitive or behavioral biases (such as bounded rationality and hyperbolic discounting) in affecting economic decision making.[23] Some of those biases and heuristics are likely to also affect online disclosure habits, explaining why making the "right" privacy decision—a decision an individual will not later regret—is difficult online,[24] and why regrettable disclosures may be common. Indeed, privacy blunders in social media offer vivid examples of the hurdles faced by users. Services such as Facebook facilitate the seamless, rapid broadcasting of intimate disclosures to audiences of both friends and strangers, often using interfaces fraught with complex settings. A considerable proportion of users of social media end up sharing online information and feelings that they later regret disclosing. As we discussed, those disclosures sometimes carry substantial consequences, such as losing a relationship or a job.[25]

In the field of behavioral economics, researchers have proposed soft (or asymmetric or libertarian) paternalistic interventions that nudge (instead of force) individuals toward certain behaviors.[26] Thaler and Sunstein popularized the idea of nudging as a form of soft paternalism to help people overcome cognitive or behavioral biases in decision making.[27] They define a nudge as "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives."[28] For instance, a radar speed sign that displays the driver's current driving speed (e.g., 85 mph) does not force her to slow down when the speed limit is 60 mph, but rather nudges her to slow down. Inspired by our work on Facebook regrets[29] and by the literature on behavioral decision research, our work explores a novel approach to help people protect their privacy in social media.

The application of soft paternalistic techniques to online privacy (and security) problems may help users make better online decisions and avoid regrets.

---

[22] Material in this section was previously published in Wang et al., *Privacy Nudges*, *supra* note 3. © 2013 International World Wide Web Conference Committee.

[23] *See* Herbert A. Simon, *A Behavioral Model of Rational Choice*, 69 Q. J. ECON. 99, 99, 114 (1955); *see also* David Laibson, *Golden Eggs and Hyperbolic Discounting*, 112 Q. J. ECON. 443, 445–46 (1997).

[24] *See, e.g.*, Alessandro Acquisti, *Privacy in Electronic Commerce and the Economics of Immediate Gratification*, *in* PROCEEDINGS OF THE 5TH ACM CONFERENCE ON ELECTRONIC COMMERCE 21, 24 (2004).

[25] *See, e.g.*, Wang et al., *Regrets*, *supra* note 3, at 4–6.

[26] *See, e.g.*, Thaler & Sunstein, *supra* note 8.

[27] *See* RICHARD H. THALER & CASS R. SUNSTEIN, NUDGE: IMPROVING DECISIONS ABOUT HEALTH, WEALTH, AND HAPPINESS 5–6 (2008).

[28] *Id.* at 6.

[29] *See, e.g.*, Wang et al., *Regrets*, *supra* note 3; Wang et al., *Privacy Nudges*, *supra* note 3.

While there is a large body of research on human behavioral modification,[30] so far little attention has focused on behavioral modifications related to online disclosures, particularly in social media.[31]

There has been some previous work attempting to apply nudging to computer security. For instance, Brustoloni and Villamarín-Salomón developed security dialogs in which users were held accountable for their decisions to open email attachments.[32] Those who took unjustified risks could be "subject to a variety of sanctions, such as being unable to use the application for increasing periods of time."[33] A user study found that these dialogs resulted in significantly fewer unjustified risks.[34]

We describe the application of soft paternalistic interventions to mitigate the effects of behavioral and cognitive biases on information disclosure decisions. We designed and evaluated three mechanisms that nudge users to consider more carefully the content and context of their disclosures on Facebook. One nudging mechanism provides visual cues about the audience of a post; a second one includes time delays before a post is published; a third one gives users feedback about their posts. We also developed a platform that enables us to deploy nudges and test them with Facebook users "in the wild."

A. *Nudge Designs*

Inspired by the literature on cognitive and behavioral biases in decision making, past research on online information disclosures, and the concept of soft paternalism, we designed three types of privacy nudges. The general ideas behind the design of our nudges can be applied to various services or domains that involve information disclosure, such as Twitter or FourSquare.

Our prior research has found that Facebook users often do not think about who is in their audience and do not have a clear idea of who can see their posts. They also struggle to remember all of their Facebook friends, and often do not understand their privacy settings entirely. As a consequence, Facebook users often post content that can be viewed by unintended audiences; in many cases, this leads to regret.[35] In an attempt to address such regret, we implemented a

---

[30] *See, e.g.*, RAYMOND G. MILTENBERGER, BEHAVIOR MODIFICATION: PRINCIPLES AND PROCEDURES 10–11 (Marianne Taflinger et al. eds., 2001).

[31] *See* Alessandro Acquisti, *Nudging Privacy: The Behavioral Economics of Personal Information*, 7 IEEE SECURITY & PRIVACY, Nov.–Dec. 2009, at 82, 84; *see also* Balebako et al., *supra* note 3.

[32] José Carlos Brustoloni & Ricardo Villamarín-Salomón, *Improving Security Decisions with Polymorphic and Audited Dialogs*, *in* PROCEEDINGS OF THE 3RD SYMPOSIUM ON USABLE PRIVACY AND SECURITY 76, 84 (2007). *See generally* Claudia Diaz, Omer Tene & Seda Gürses, *Hero or Villain: The Data Controller in Privacy Law and Technologies*, 74 OHIO ST. L.J. 923 (2013).

[33] Brustoloni & Villamarín-Salomón, *supra* note 32, at 77.

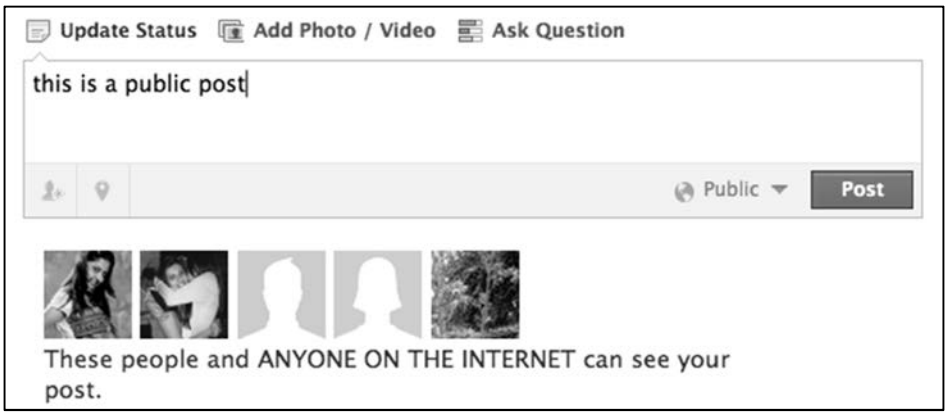
[34] *Id.* at 84.

[35] *See* Wang et al., *Regrets*, *supra* note 3, at 7.

nudge designed to lead users to consider the audience for their posts while they are composing them. We refer to this nudge as the "profile picture nudge."

Our profile picture nudge attempts to encourage users to pay attention to their audience by displaying five profile pictures, randomly selected from the pool of people who could view the post being created. These profile pictures serve as visual cues to remind users of the potential audience for their post. As shown in Figure 1, the profile pictures are displayed as a user starts typing in the "post" text box. The nudge also displays a notice to the user based on the user's current sharing setting. For example, if the post is to be visible to friends of friends, the notice states, "These people, your friends, AND FRIENDS OF YOUR FRIENDS can see your post."

Figure 1: *Profile Picture Nudge. A notice about the potential audience for the post and five profile pictures randomly selected from the set of people who will be able to see it are shown under the text box.*



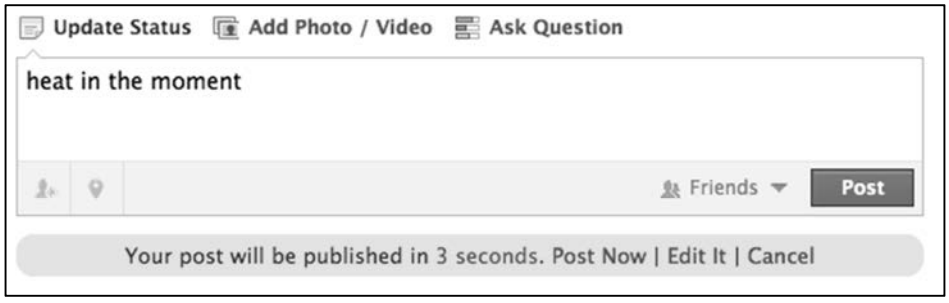
Acquisti has discussed how individuals may trade their personal information for immediate gratification.[36] Prior research on regrettable behavior on social media has also found that people often create regrettable posts "in the spur of the moment."[37] To encourage users to reflect on their posts, we designed a timer nudge that inserts a short time delay before a post is actually posted. We refer to this nudge as the "timer nudge."

Figure 2 shows a screenshot of the timer nudge interface after the user clicks the "Post" button. When a user starts typing a status update or comment, a message with a yellow background appears stating, "You will have 10 seconds to cancel after you post the update." After the user clicks the "Post" button, the user is given the option to "Cancel" or "Edit" the post during a ten-second countdown before the post gets published on Facebook. There is also an option to circumvent the timer by clicking a "Post Now" button.

---

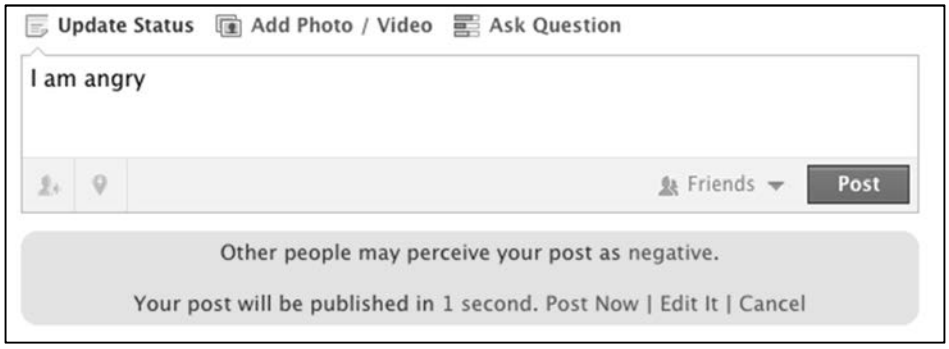
[36] *See* Acquisti, *supra* note 24, at 26–27.
[37] *See* Wang et al., *Regrets*, *supra* note 3, at 5.

Figure 2: *Timer Nudge. Countdown appears after the user clicks "Post."*



Our research has found that regrettable posts on Facebook often contain negativity, profanity, or sensitive topics like alcohol and sex.[38] Our third nudge sought to provide users with immediate feedback on the content of their posts. We designed a sentiment nudge that combines a countdown timer with a notice regarding the content of the post, as shown in Figure 3. After the user clicks "Post," the timer and a notice highlighted with a yellow background will appear below the text box. We refer to this nudge as the "sentiment nudge."

Figure 3: *Sentiment Nudge. Different sentiment notices are shown depending on the overall sentiment of the post content.*



For the sentiment nudge, we used an open-source sentiment-analysis module to analyze the content of each post.[39] This module uses AFINN-111—a list of 2,477 English words and phrases manually rated as negative or positive, on a scale between –5 (negative or very negative) and 5 (positive or very positive).[40] For each post, any words in the wordlist are scored, creating a

[38] *See id.* at 4–6.

[39] *See SentiMental—Putting the Mental in Sentimental*, GITHUB, https://github.com/thinkroth/Sentimental (last visited September 1, 2013).

[40] Lars Kai Hansen et al., *Good Friends, Bad News—Affect and Virality in Twitter*, 185 COMM. COMPUTER & INFO. SCI. 34, 39 (2011); Finn Årup Nielsen, *A New ANEW:*

weighted sum for the entire post. A text message corresponding to this sum is shown to the user. For example, a slightly negative weighted sum would lead to the message, "Other people may perceive your post as *negative*."

B. *Study Methodology*

To investigate how nudges would be perceived by active Facebook users and could impact their disclosures on Facebook, we conducted an exploratory field study with twenty-one participants, complemented with survey questionnaires and follow-up interviews. Participants remotely downloaded and installed a Chrome browser plug-in and a Facebook application, which they used over a period of three weeks. The study took place in Pittsburgh, Pa. and Syracuse, N.Y. during the summer of 2012. It was approved by the Carnegie Mellon University (CMU) IRB.

We sought active Facebook users who were also native English speakers. Since our plug-in was designed for the Chrome browser, we recruited participants who primarily used that web browser to access Facebook. Participants were recruited using Craigslist, flyers, email distribution lists, and a CMU research recruitment system. Participants were given a $10 Amazon gift card for each week they remained in the study, either three or four weeks, plus a $10 bonus for participating through the end of the study period and completing the final survey. Each participant who conducted an optional interview received an additional $10 Amazon gift card.

Recruitment material directed prospective participants to a screening survey. We invited via email fifty-one prospective participants, thirty-one of whom agreed to the online consent form and installed the Chrome plug-in and the Facebook application. Once participants had installed the plug-in, we verified that their self-reported Facebook usage was similar to their actual usage. We dropped one participant who in the screening survey self-reported posting several times a day but had only three posts recorded in the last thirty days. Two participants quit the study due to technical difficulties, and three more were dropped half-way through the study for not having answered the midterm survey. Four more participants never saw the profile picture nudge during the treatment period. We present results from the twenty-one participants who completed the field study and thirteen of them who participated in a follow-up interview.

Using a round-robin scheme, participants were randomly assigned to one of the three nudging interfaces: Profile Picture, Timer, and Sentiment. Study participants were required to install our plug-in and Facebook application, which allowed us to monitor participants' behavior on Facebook and to enable

---

Evaluation of a Word List for Sentiment Analysis in Microblogs, *in* PROCEEDINGS OF THE ESWC2011 WORKSHOP ON "MAKING SENSE OF MICROPOSTS": BIG THINGS COME IN SMALL PACKAGES 93, 94 (2011).

or disable the corresponding nudge treatment for each participant. The field study comprised two main stages.

During the first stage, the control stage, data collection took place without changes to the Facebook user interface. At the end of this stage, a midterm survey was administered to better understand the context in which each participant was making his or her posts and to identify external factors that could have affected participants' posting behaviors during the control period.

During the second stage, the treatment stage, in addition to data collection, each participant was shown one of the three nudges. On average, participants remained in the control and treatment conditions for eleven and twelve days, respectively. The specific time individual participants remained in the study depended on their response time to our midterm survey and the nudge they were assigned. In particular, participants in the profile picture nudge condition remained in the study for a longer time since we experienced technical difficulties showing profile pictures for posts with custom privacy settings and for comments on posts originally made using custom settings. Leaving the participants more time in the study allowed us to resolve some of these issues and increased the chance that users would use a different setting (e.g., friends only) for some of their posts, allowing them to see the profile pictures.

At the end of the field study, we administered a final survey that collected participants' opinions on the nudge they were shown. We further asked whether they were interested in participating in a follow-up interview. We extended this invitation to all participants who expressed interest, except the four participants in the profile picture treatment who, due to technical difficulties, never saw the profile pictures during the study.

The purpose of the follow-up interviews was to understand participants' attitudes and perceptions about, as well as experiences with the nudges. We asked participants about their main motivations for using Facebook, knowledge of Facebook privacy settings, first impressions with the nudge interface, and perceived benefits and drawbacks of that nudge. We then showed them three posts or comments they had made and asked them about the contexts of those posts and whether the nudge had affected their posting decision in any way. Towards the end of the interview, we asked them to log into their Facebook accounts using their own laptops or a lab computer with the Chrome plug-in installed. We reactivated the nudge they had seen during the field study and collected their ideas for design improvements while seeing the nudge on their Facebook page. Towards the end of the interview, we showed them a different nudge from the one they had used during the field study and collected their opinions about that other nudge.[41] We interviewed thirteen participants, and each interview took about thirty to forty-five minutes.

---

[41] Participants in the profile picture treatment were shown the sentiment nudge and participants in the sentiment or timer treatments were shown the profile picture nudge.

C. *Analysis*

We analyzed participants' responses to Likert questions, behavioral data collected using the Chrome plug-in, and interview data to explore the impact of our three nudges.

The final survey included Likert questions that queried participants' opinions about the usefulness of the nudges, their willingness to use these nudges, and their level of comfort with the nudges. The purpose of these analyses was not to compare statistically the results across the three nudge treatments, but to show a quantitative summary of opinions about these treatments.

We used the data collected with the Chrome plug-in to investigate whether there was any evidence of changes in Facebook usage before and after the participants started seeing the nudges. The metrics that we used to investigate behavioral changes included: number of changes in online privacy settings, number of canceled or edited posts, post frequency, and topic sensitivity. We focused on sensitive topics that our previous research identified as problematic on Facebook.[42] Given the number of factors other than our nudges that could have affected participants' behaviors during the study period, we do not claim any causality but only show instances that could have signaled an impact of a nudge on users' behavior. Similarly, given the exploratory nature of our study, the small sample size, and the uncontrolled environment of the study, we did not attempt to perform any statistical tests. If we had a larger sample size, we could have analyzed the results using a number of statistical techniques based on the distribution of the collected metrics. For example, we could use t-tests or Wilcoxon Rank Sum tests to perform both between- (across treatments) and within-subjects (control versus treatment) comparisons using the collected metrics as dependent variables.

Finally, we performed a qualitative analysis of the interview data. We developed a codebook of the comments that participants made during the follow-up interviews. We then grouped these comments into thematic strands, including perceived benefits and drawbacks, context in which the nudges could have a positive effect on users, and opportunities for design improvement. We report comments that were common among participants, as well as those that were unique. We illustrate these comments with a number of interview quotations.

D. *Results*

In this section we first describe our participants' demographics and overall posting behavior. Then we discuss participants' first impressions of the nudges, which were collected at the beginning of each interview. After that, we use system logs and interview data to describe the impact of these nudges on

---

[42] *See* Wang et al., *Regrets*, *supra* note 3, at 4–6.

participants' posting decisions. We further discuss the participants' perceptions of the benefits and drawbacks of these nudges. Finally, we discuss the results of the survey administered at the end of the field study.

Seven of our twenty-one participants were undergraduate students, five were graduate students, two were unemployed, and seven were employed in a variety of occupations. They included thirteen females and eight males between the ages of eighteen and forty-eight (mean age twenty-four). We use a combination of a letter and a number to refer to each participant. The letter represents the initial for the nudge treatment, and the number refers to the sequence within each treatment group. For instance, T-1, P-1, S-1 denote the first participant in the timer, picture, and sentiment nudge group, respectively.

During the three-week study period, our Chrome browser plug-ins stored a total of 1,209 posts (353 status updates and 856 comments) made by the twenty-one participants. On average, each participant made about two posts per day. For participants in the sentiment nudge, the number of nudge appearances include both positive ("Other people may perceive your post as [positive */ very positive*]) and negative ("Other people may perceive your post as [negative */ very negative*]) messages. The sentiment warning did not appear if the post was considered *neutral* by the sentiment analysis algorithm.

1. *Participants' First Impressions of Nudges*

During the interviews, we asked participants, "What was your impression when you first noticed the new interface on your Facebook page?"

Three of four interviewees in the timer nudge treatment commented that they thought the delay was a new feature introduced by Facebook, although they wondered why Facebook would want to introduce the new feature. T-2 explained that the first time she saw it she was annoyed by the time delay: "Why would it make me wait?" Later, she noticed that "Post Now," "Edit," and "Cancel" were clickable options and started to like the features because they allowed her to review her posts before making them public. When we switched participants to the timer and sentiment nudges, we experienced a few technical difficulties that prevented some of the participants' posts from being posted. T-4, who experienced this problem, also expressed a negative feeling. "The application was eating my posts," he said. Nevertheless, this participant later explained that once the problem was fixed, the timer nudge prevented him from posting trivial statements such as "hahaha," which he perceived as a benefit from the timer.

P-1 and P-4 wondered whether the profile pictures were a new Facebook feature or part of the user study. Another participant, P-2, thought it was a new Facebook feature that would allow her to tag people easily, but she soon realized that was not the case. She was surprised when she read that her post could be seen by such a large number of people. "It reminded me that I should probably clean up my friends list," she said.

S-3 immediately associated the sentiment nudge with the study. Both S-1 and S-3 wondered how the sentiment of their posts had been determined when they saw the "Other people may perceive" warning message. However, while S-1 expressed that "it made me think," S-3 mentioned she completely disregarded it. S-3 further explained, "I was like why would it think it's negative? Oh whatever, post now." She further elaborated that she did not like the warnings because "I'm giving a legitimate statement or opinion on something or I'm being sarcastic and my friends know that." This participant's comment highlights an important challenge of content or sentiment analysis: it should consider or understand the context around a post, not only the content of the post itself.

2. *Impact on Posting Behavior*

We logged participants' posting behavior on Facebook and their interactions with the nudges during the study. We analyzed participants' posting behavior during both the control and treatment periods. We found evidence of changes in posting behavior for some of our participants, and we combined those results with the interview data to better understand whether those behavioral changes could be associated with the nudges. We use concrete instances to illustrate the kinds of impacts that each nudge had on some participants' posting behavior.

Both P-2 and P-3 reported the profile picture nudge made them think about their privacy settings and the content of their posts. P-3 reported having changed the privacy settings of one post because she saw a picture of a person she did not recognize. When looking at her behavioral data in our system logs, we noticed that during the treatment period, she changed her privacy settings from "Friends" to "Friends except acquaintances" when she posted: "Survived one of the craziest, most exhausting days ever!" Based on the stored typing history of this post, we also found that the post was edited from the original, "Definitely just had one of the craziest/most exhausting days ever."

P-2 reported that she ended up canceling "a couple of posts" because of the profile picture nudge. She explained that she once canceled a negative post: "There wasn't any swear words or anything but it was a snide remark and then one of the pictures that popped up was one of the people I work with. It is probably not the best idea." She volunteered that she is often careless when posting on Facebook and the nudge "made me change, it did make me think." She added that she could probably benefit from the sentiment nudge as well, especially if she could configure a dictionary of curse words she normally uses. In contrast, although P-5 recognized that the profile picture nudge creates awareness about the audience of one's posts and encourages people to be more cautious, she did not believe that the nudge had a significant impact on her posting decisions. P-1 and P-4 both volunteered that they were ignoring the profile pictures for most of the study. P-4 explained, "I only make my posts available to friends," and he claimed he knew which people he had placed on

his friends list. He added, "If I were using different lists, [the profile pictures] would be very useful."

T-3 mentioned that the timer was "at times annoying and at times handy." He explained that it was annoying when he "knew exactly what I wanted to say" but had to wait for the timer to expire or hit "Post Now," which required extra time and effort. He also said it was handy because sometimes he edited his post to "make it a bit more publicly acceptable when it was a venting post" or to fix typos. He also said he canceled posts rather than wait for the timer "if I didn't need to say it." He further volunteered that he posted less often due to the time delay. However, we did not observe a change in the frequency of his posts during the study period.

T-4 reported that the timer made him think about the utility of his posts, explaining that he canceled several posts because the timer made him realize it was not really necessary to post them. Indeed, our collected data about him show that, on average, he reduced his posting activities in the treatment period by more than seven posts per day. In addition, while he did not post sensitive content during the treatment period, there were ten instances of sensitive content during the control period. He also edited a few of his posts in the treatment period. For example, one of his comments was, "Wow." Upon reviewing the typing history we stored for this comment, we found that he typed, "God damn. That's so cool man," and then deleted this sentence from the comment.

Both T-1 and T-2 agreed that the edit option was very convenient. They were using the time delay to review their posts, and they started liking the nudge after having used it for several days. T-1 reported caring about what she writes on Facebook and paying attention to grammar and spelling. She volunteered that she clicked "Edit" several times to improve the wording of her posts. Similarly, T-2 mentioned she used the "Edit" option a few times. For example, once when she posted a link to a movie cover, she edited out "this is the movie" because she felt it was redundant.

S-2 said the nudge reminded her that she was in the study, but that most of the time the sentiment meter was very sensitive or missing the context. Regardless, she remembered that the first time she saw the negative sentiment warning was when posting "damn the Steelers rock," and she decided to use the word "dang" instead. She further explained that she usually does not swear and she does not want to be perceived as a negative person.

Both S-1 and S-2 said that the nudges made them "stop and think" and review and edit their posts. S-3 volunteered that she only paid attention the first few times she saw the warning, ignoring it afterwards; she said she edited a few of her posts because of typos during the timer countdown. She also remembered canceling a post: "It was a link to a funny story. I just realized other friends had already posted it so I canceled the post." Her collected data further shows that her post frequency was reduced on average by almost four posts per day. We also found fewer (seven) instances of sensitive content during the treatment period than during the control period (thirteen). In contrast, S-4 commented that

each time he saw the sentiment warning he was given a positive score, which he thought was nice since "I do not want to be perceived as a jerk," but it did not have any effects on his posting habits. He further explained that as he is usually careful with what he posts, the sentiment nudge was not particularly useful to him. Behavioral data collected through the plug-in aligns with his claims because no sensitive content was found nor were changes in posting habits detected.

S-7 became annoyed when he saw the negative sentiment warning. He posted, "Also, apparently if I cuss on facebook I now get a warning that some people may find my post negative. As if I give a fuck." In another post that he ended up canceling, he claimed, "Now I just want to post a shit ton of bad words and see how facebook reacts to each one." These remarks show the potential negative effects of a sentiment warning and the importance of considering the form, style, and tone of the feedback given to users.

3. *Perceived Benefits and Drawbacks*

We asked participants, "Do you see any benefits from a Facebook interface like the one you tested?" Four out of seven interviewees in the timer or sentiment nudge mentioned the opportunity to stop and think as a benefit. Two of those participants also mentioned that it could deter people from posting trivial things. T-4 explained that the timer nudge helped him to post "better quality versus quantity." The same participant added that the timer nudge could prevent people from posting "politically incorrect statements." T-1 and T-2 also mentioned the timer nudge could be useful to correct typos. Three out of four interviewees who tested the profile picture nudge mentioned that it could be useful to remind those users who use customized groups to select the right group for each post. P-1 further mentioned that it could help to remember who is in each group. Moreover, P-3 mentioned that it was useful for creating awareness about who can see her posts, and P-2 thought it was a good reminder to clean up her friends list and to be cautious about what to post.

Apart from encouraging users to stop and think because of the time delay, the sentiment nudge was not perceived as being as useful as the other two nudges. Overall, users believed that the sentiment algorithm was taking isolated words and missing the context. However, S-3 recognized that it could be useful for people posting while in an emotional state. Towards the end of the interview, when the sentiment nudge was shown and explained to T-1, she disliked it because "sometimes people post things that might sound negative, but they need others' empathy and support." P-3 also thought the sentiment meter was not very useful for her; she added that the algorithm could "misinterpret sarcastic comments." However, she said it could be useful for people who had problems controlling their emotions. She mentioned children with autism as an example of those who could benefit from the sentiment nudge. P-4 also commented that the timer could help to cool people down when they engaged in a heated exchange of posts.

The downsides mentioned by our interviewees were mainly associated with performance issues such as Facebook page lag, posts not getting through or delayed posting. Nevertheless, participants appreciated the benefit of the nudges. In the words of P-2, "[There were] some technical things but the concept of having something there to remind you was fine."

4. *Exit Survey Opinions*

In the final survey, we used both open-ended and Likert questions to collect participants' opinions about the nudges they were shown. From the responses to the open-ended questions, we noticed that participants' opinions were significantly affected by some of the performance issues they experienced with the nudges. This distracted their attention from the actual functionalities of the nudges. In particular, due to technical difficulties that arose from changes rolled out by Facebook, the timer and sentiment nudges temporarily prevented posts from showing up.

Nevertheless, some of the participants valued the options offered by the timer nudge. In particular, when answering the survey question about whether our Facebook application was helpful in any way, T-3 typed, "[I h]ad time to think about what I posted and whether or not I really wanted to be represented in that way." T-7 further reported that the option to cancel "was interesting." Similarly, S-1 also believed the time delay was particularly useful; she said, "I liked the time available to cancel or edit a post."

As discussed earlier, we were unable to show profile pictures for every post that participants made. As a result, participants in that treatment were not exposed to the nudge as often as participants in the other two treatments. This issue probably prevented them from giving a completely informed opinion. For example, even when the system logs allowed us to determine that the pictures had showed up several times on some participants' Facebook pages, these participants forgot having seen them.

Towards the end of the final survey, we asked participants to rank their opinions about the likelihood of using the nudge application in their daily Facebook usage, and the likelihood of recommending it to a friend. We also asked about their perceived level of usefulness and comfort with it during the study period.

Overall, participants had a positive perception of the timer nudge. They were both willing to use it and believed it could be useful. In contrast, opinions of the sentiment and profile pictures nudges were mixed. Participants perceived benefits from the sentiment nudge, but the benefits mainly stemmed from the time delay and the opportunity it provided to stop and think. Participants mostly did not like the sentiment warnings, which we will discuss in detail in the next section.

Opinions captured from Likert questions about the profile picture nudge did not show a particular positive or negative trend. We attribute this result to the fact that participants in this treatment only saw the profile pictures a few times,

making it difficult for them to make an informed judgment about the nudge. However, as we discussed in the previous subsections, participants expressed a more positive opinion of the profile picture nudge during the interviews.

E. *Discussion*

The objective of our nudging approach was to help prevent users from posting things that they would later regret. Consistent with the tenets of soft paternalism, our nudging approach did not limit participants' ability to post on Facebook. Instead, it encouraged the participants to reflect on their posts and their audience.

1. *Stop and Think*

Our timer nudge was designed to encourage users to stop and think, so as to avoid regrettable, "spur of the moment" posts. We observed that this nudge was often successful in helping users reconsider their posts. It had an additional benefit of helping users catch typos and minor errors in their posts. Some participants rephrased or even canceled their posts during the timer delay. However, this benefit comes at the cost of delaying every post participants made. Although we did provide a "Post Now" button, some participants wished it were more salient. Increasing the saliency of this button might lead users to get into the habit of clicking it without thinking, which would undermine the effectiveness of the nudge. Further research on time delay nudges might explore adjusting the duration of the delay, allowing users to customize this duration, or varying the delay automatically based on factors such as number of words in a post. Research might also consider other mechanisms that might nudge users to stop and think without imposing a delay.

2. *Content Feedback*

Our sentiment nudge was designed to help make users more aware of how others might perceive their posts, since past research has found that posts that are perceived as very negative or contain sensitive topics are among those most regretted.[43] However, participants who received sentiment warnings did not find them useful. Participants seeing only positive scores believed the feedback was needless since they were already being careful with their posts. Participants who saw negative scores often disliked the negative feedback because it did not account for the post's context; in addition, they tended to dislike the feeling of being judged. Other difficulties with our sentiment nudge implementation were its inability to identify sarcasm and its inability to distinguish potentially damaging negativity in posts from more benign expressions of negativity. However, a number of participants agreed that a similar nudge could be useful

[43] *See* Wang et al., *Regrets*, *supra* note 3, at 4–6.

for younger, less mature Facebook users. Further work might focus on improving the feedback algorithm, by allowing users to customize it based on their past posts and typical vocabulary or providing a list of words they would like to avoid posting.

3. *Pay Attention to the Audience*

Our picture nudge was designed to remind Facebook users of who can see their posts, as prior research has found that users often forget who their Facebook friends are or have trouble understanding their privacy settings.[44] This feature was positively received by participants and seemed to have improved some participants' behavior. Showing profile pictures of people who might see a given post encouraged users to be more aware of and more cautious about their posts. For example, one participant adjusted her privacy settings in response to the nudge, and another reconsidered the size of her friend list. These anecdotes suggest that this nudge can assist users with making better privacy decisions at least in some situations. This nudge might be further improved by refining the number of pictures, the algorithm for selecting pictures, and the proximity of the pictures to the posts; or by providing additional cues about the audience.

4. *Study Limitations*

Conducting our investigation as a field study provided the advantage of users interacting with our nudges in a natural environment. However, it also introduced difficulties, such as external factors influencing participants' posting behavior. Further, while we were able to observe posts made using our Chrome plug-in and Facebook application, we were unable to analyze posts the participants may have made using other browsers. We also experienced technical difficulties when Facebook implemented changes to its interface.

Our recruitment was affected by biases. Our plug-in was designed for users of the Chrome web browser, and participants were informed that their Facebook activities would be monitored. Therefore, our sample might be biased towards users with fewer privacy concerns and with browser preferences different from that of the general population of Facebook users.

Measuring the effectiveness of our nudges in preventing regret is challenging because only a small fraction of the posts made by users lead to regret, and arguably even fewer lead to the short-term regret we could detect in this study. Instead, we could measure when a participant modified his or her post in response to a nudge. In addition, it is often difficult to measure the effect of a nudge; users may not react to them in a noticeable way or the reaction might be gradual.

---

[44] *See id.* at 7.

Some of our participants reported that they began to ignore our nudges after several days. Future work might investigate this habituation effect and how to mitigate it—for example, by varying the presence or content of the warning messages. Nudges could also be designed to appear only when a warning is needed (e.g., a post contains controversial topics), rather than appear for every post. However, determining when to display a warning is in itself a challenging research question. Alternatively, a more interactive system, similar to ELIZA, could be used to make nudges more engaging.[45] For instance, the system could provide feedback such as, "Do you think people will respond well to your post?" or, "You sound upset. Would you like to rephrase your post?"

Despite these limitations, this study provides interesting preliminary results and directions for future work. With further refinements, our experimental platform will be useful for conducting large-scale, longitudinal field trials, testing a variety of nudges.

5. *Ethical Considerations of Nudging*

Next, we discuss the ethics of nudging. Nudging can be seen as affecting users' agency. However, we argue that any system or design is inevitably not neutral. Designers build values into their systems with certain intended uses. Nudging is "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives."[46] As such, a nudging approach does not force people to do things, but rather stewards them toward a direction that the designer believes is good for them. We took the nudging approach because we recognize that people sometimes have difficulty making "rational" information-disclosure decisions, and we seek to help users with those difficulties.

6. *Implications for Public Policy*

While our preliminary results of the nudge showed limitations of the current design, the general nudge approach seemed promising. We advocate that SNS service providers, such as Facebook, consider adopting this nudging approach as part of their overall strategy to help their users avoid any regrettable experiences on their platforms for a number of reasons. First, users can have various cognitive and behavioral biases that hamper their rational decision making and lead to regrettable disclosures or postings. Second, platforms such as Facebook have complex privacy settings that users could misunderstand and that could become a source of regrets, as we have discussed previously. Third, the nudging approach is relatively lightweight and can be incorporated into the

---

[45] *See* Joseph Weizenbaum, *ELIZA—a Computer Program for the Study of Natural Language Communication Between Man and Machine*, 9 COMM. ACM 36, 36 (1966) ("ELIZA is a [computer] program . . . which makes certain kinds of natural language conversation between man and computer possible.").

[46] *See* THALER & SUNSTEIN, *supra* note 27, at 6.

existing Facebook user interface. Fourth, legal scholar Daniel Solove insightfully pointed out that the privacy self-management approach (i.e., user notice and control) is insufficient and suggested that the nudging approach, a middle ground between self-management and paternalistic policies, is a promising complement to existing privacy protection mechanisms.[47] Lastly, services such as Facebook would not want to become known as a site where people post things they regret. By including simple nudges they could demonstrate that they care about this issue and want to help their users.

There are still many open questions for public policy that are worth further research. For instance, should the law require or forbid social media systems to provide certain types of nudges to their users? What kinds of nudges should they provide? And can users disable the nudges?

IV. CONCLUSION

As the Internet has become an increasingly powerful medium for information sharing, a considerable proportion of users have shared information and feelings that they later regret disclosing. Our study of Facebook regrets showed that people have various cognitive and behavioral biases that affect their decision making and they make posts that they later regret. These regretted disclosures sometimes carry substantial consequences, such as loss of a relationship or a job.

Drawing on behavioral and decision research, we designed three privacy nudges that attempt to nudge users to think carefully before posting. While our field trial of the nudges was exploratory, our results suggested that privacy nudges can potentially be a powerful mechanism to help some people avoid unintended disclosures. Although we provide a Facebook case study, this idea of privacy nudges can be extended to similar services such as Twitter, or to other types of services such as e-commerce, location sharing, and smart phone applications. Finally, we advocate the privacy nudging approach to researchers, service providers, and policy-makers to explore the rich design space of nudging to help protect people's privacy.

---

[47] *See* Daniel J. Solove, *Introduction: Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1880, 1901 (2013).