Engineering Information Disclosure: Norm Shaping Designs

Daphne Chang[1], **Erin L. Krupka**[1], **Eytan Adar**[1], **Alessandro Acquisti**[2]

[1]University of Michigan,
Ann Arbor, MI
{daphnec,ekrupka,eadar}@umich.edu

[2]Carnegie Mellon University
Pittsburgh, PA
acquisti@cmu.edu

ABSTRACT

Nudging behaviors through user interface design is a practice that is well-studied in HCI research. Corporations often use this knowledge to modify online interfaces to influence user information disclosure. In this paper, we experimentally test the impact of a norm-shaping design patterns on information divulging behavior. We show that (1) a set of images, biased toward more revealing figures, change subjects' personal views of appropriate information to share; (2) that shifts in perceptions significantly increases the probability that a subject divulges personal information; and (3) that these shift also increases the probability that the subject advises others to do so. Our main contribution is empirically identifying a key mechanism by which norm-shaping designs can change beliefs and subsequent disclosure behaviors.

Author Keywords

Social Norms; Privacy Calculus; Social Media

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

Interface design in social media systems can greatly influence when, how much, and why people disclose private information. For example, the addition of a user interface element that displays friends' birthdays (Figure 1) may lead a user to perceive that sharing birthday information is the norm and subsequently to share this data with the system. However, the mechanisms through which interfaces affect disclosure behavior are still poorly understood. In particular, we understand little about how interfaces signal social norms and, in doing so, encourage information disclosures. We broadly know that signals of norms can assist newcomers and both encourage and discourage pro- or anti-social behavior. We also broadly understand that decision contexts may be altered to nudge behavior towards target outcomes. However, we know much less about how norms, contexts, and interfaces combine to form social guideposts that shape users' beliefs about what kind, and how much, information is acceptable to reveal. Work in economics and psychology describes a social learning mechanism through which actors observe behavior that informs their beliefs about what is appropriate to do in a particular context and then change their own behavior to conform with those (newly updated) social expectations. In this paper, we design an experiment that tests a causal pathway from design choices, via changes in beliefs about what is appropriate, to subsequent behavior change.

Shaping behaviors through user interface design is a well-established principle in modern Human-Computer Interaction (HCI) practice. Interfaces often provide cues or affordances to help a user (or communities [21]) achieve some end-goal, and many designs that are successful are encoded as patterns (using the terminology of the software engineering literature). For example, a multi-layer pattern hides complexity so that the end-user is not overwhelmed [7], and a voting pattern uses a thumbs-up or down icon to encourage voting [14].

In contrast to these helpful patterns, there are other patterns that are not intended to benefit the user directly, but rather to serve the interest of another party–such as pushing users to provide personal information, or even "dark" patterns that nudge users to take security risks [13]. As an example, forced disclosure patterns require users to fill in information before gaining access to a service. The preference that system owners and builders hold for certain designs over others may be motivated by explicit corporate interests (e.g., advertising revenue or engagement) and behavior targets (e.g., design *A* drives more photo sharing than design *B*). These pressures might mean that incentives to maintain privacy conflict with systems' goals to reduce privacy-maintaining actions and may implicitly create an impression of information sharing norms [39].

Norm-shaping patterns need not be positive or negative. However, the fact that such patterns can be used both ways makes them particularly insidious, because users cannot infer what the designer's intent was or the downstream consequences of their behavior may be [15]. Policy makers and well-intentioned designers have no mechanism for assessing how their design choices shape norms. Our claim is that *design choices have the power to engineer personal information giveaway by changing beliefs and, subsequently, behavior*. The implications are not only that the user has revealed more information to the system than they may have intended, but that this opens up the downstream uses of this data in ways that the user never envisioned (e.g., [4]).

CHI'16, May 07-12, 2016, San Jose, CA, USA
 ISBN 978-1-4503-3362-7/16/05$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858346

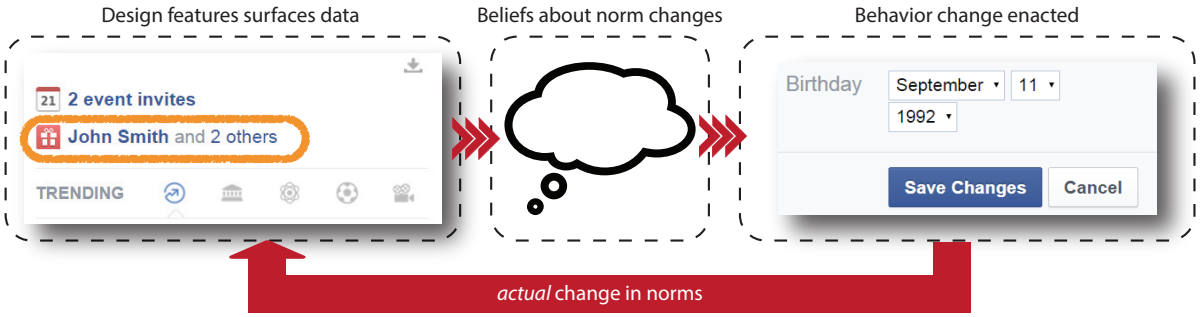

Figure 1. How designers influence information sharing.

In this paper we experimentally identify the impact of design on the perception of appropriate behavior and we test the impact of norm-shaping design patterns on subsequent information divulging behavior. We find that subjects' perception of what images they personally would find acceptable to post on a social media site can be altered when we manipulate the context to contain more "risky" or "explicit" images. Further, we find that a shift in perception affects subsequent information divulging behavior in a subsequent task as well as the advice given to others about revealing personal information.

Our main contribution is to experimentally unpack and identify a causal pathway by which design patterns can work to affect disclosure: they can modify perceptions of appropriate behavior which, in turn, impact subsequent behavior. We show a chain of influence, where designers may nudge users' beliefs about what is appropriate (in a situation), and subsequently alter users' behaviors in that situation. Our second contribution is to demonstrate that the shift in perceptions can leave a larger footprint on user behavior than one might think. This is because the shift in perceptions of appropriate behavior also impacts the advice that a user gives others. This gives rise to a cycle of nudging individual behavior through design patterns. The cycle begins with a design pattern that shapes *perceptions* of what is personally acceptable to do, which then nudges behavior which then shapes the norms of the community through altered behavior and through altered advice given to others (Figure 1). Further, we demonstrate that nudges in a one medium (images) affect norm perceptions for that medium (posting images on a social network website) and have a spillover to a second domain (revealing information to the experimenter) and advice given to another user (e.g., advice on what is appropriate to reveal). Our work leverages a novel experimental design, which combines methodologies and theory from experimental economics with HCI methods. Taken together, our findings have significant implications for security and privacy.

RELATED WORK

Much of the modeling of information disclosure decisions rests on the assumption that a user can at least tell you how much she values privacy. Various studies investigate how people navigate the trade-off between sharing private information and other instrumental goals such as better recommendations or financial remuneration [1]. A subset of these studies highlight the dichotomy between professed privacy attitudes and actual self-revelatory behavior [3, 5, 27, 31].

Some of the (numerous, and not mutually exclusive) explanations for the dichotomy reside in the hurdles that hamper individuals' privacy-sensitive decision making. In particular, this literature posits that the disparity between professed attitudes and behavior stems from either uncertainty due to incomplete information about one's preferences or from uncertainty about which social rules apply in that context [2]. Users may experience *preference uncertainty*–that is, they have a vague sense of, or they just don't know, how much they value privacy. Alternatively users know their preferences but are trying to figure out how to trade-off between their preferences for privacy and the social norms and expectations that others have of them to share information.

These hurdles make privacy *attitudes* appear inconsistent and/or easily malleable. The hurdles make disclosure *behavior* highly contextually sensitive and surprising or seemingly contradictory. Both explanations (uncertainty over preferences or uncertainty over social norms) imply that users may rely on social cues (such as the behavior of others) as they decide how much to disclose [34]. As an example, Acquisti et al. [6] find that disclosure behavior is comparative in nature: People's willingness to divulge sensitive information depends on judgments about others' readiness to divulge that information.

This evidence is suggestive of an interplay between others' behaviors, perceptions of social norms, and one's own behavior. And although sharing behavior, as well as the drivers of personal disclosures, have been investigated from a number of different disciplinary angles [25, 33], several new questions emerge. How critical are social and contextual cues to shaping the *perception* of norms as well as the nudging of individual behavior? Do context and social cues operate on behavior by changing beliefs about the norms? The implication of such relationships is a cycle of nudging individual behavior through design patterns which, in turn, affects the *perceptions* of norms, which then nudges behavior which then shapes the norms of the community.

The importance of norm-shaping in social media and computer-supported cooperative work has been recognized by a number of researchers (many summarized in [21]). As social media systems gain popularity, designers of social media interfaces

are creating a set of interface heuristics (e.g., having a "conversational" style question such as "What are you up to?" instead of "Type status here." to make users more comfortable) [14]. These design patterns have evolved as a result of conventional wisdom, aesthetic concerns, and ad-hoc experimentation, but rarely through principled studies and even more rarely in information disclosure contexts (though see [35, 36]). However, much of this work focuses on the impact of these signals on the end-outcome (behavior) and does not focus on why the behavior changed in response to the social signal. One pathway by which a social signal (such as observing what others are doing) can impact behavior is by changing beliefs about what is appropriate to do or the social norms.

A long tradition of work in psychology and, later in economics, shows that by observing others, people learn what behaviors are considered appropriate as well as expected. This work predicts a positive relationship between one's action and what one observes others doing [10, 12]. In psychology, the classic experiments showing this influence involve observing how an individual's judgment of the length of a line segment varies depending on the responses of others [8, 16]. Recent work in economics has predominantly demonstrated this relationship in public goods games [10, 17, 26].

Work in human-computer interaction and computer-supported cooperative work has sought to understand the drivers behind information disclosure such as "folk models" ([37]), privacy concerns (e.g., [29]) and preference models (e.g., [19, 20]). Considerable work in psych (e.g. [12, 16]), economics (e.g. [9, 23]) and applied design (e.g. [18, 30]) demonstrated that showing a user what others are doing (or think should be done), leads to conformity in action (or with respect to normative expectations). However, this work does not establish casual mechanisms by which this correlation is observed.

More recently, researchers have begun to unpack the pathway by which this influence occurs: observing others influences the actor's beliefs about what actions are appropriate which in turn affects behavior [22]. Both economics and psychology offer a rich empirical body of work demonstrating the impact of others' behavior on an actor's norm compliant behavior. However, the pathway from observing others, to changing one's normative beliefs, to changing one's behavior in the context of information disclosure is not as well-understood (though the fact that such a pathway *exists* has been demonstrated in [6]).

A great deal of existing research on privacy in the context of social media platforms such as Facebook relies on self-reports (e.g, [11, 28, 29]). While informative, it is difficult to directly assess how a design element may impact perceptions and behavior. In our research we use experiments to test the impact of design patterns on users' perceptions of information sharing norms. Specifically, we manipulate the types of information our target users see others providing and use that to test the pathway from observation to perceptions of social norms. To accomplish this we adapt instruments developed in previous work by Krupka and Weber [24] and embed them in an experiment. Our primary goal is to demonstrate that perceptions about norms that govern information disclosure can be manipulated and affect subsequent disclosure behavior.

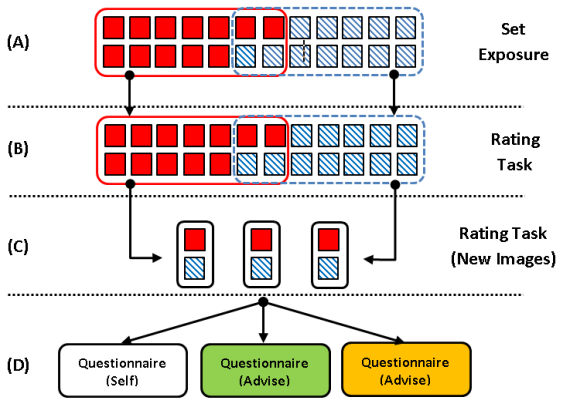

Figure 2. Subjects were randomly assigned into either the R condition and saw the R images (solid red squares) or the PG condition and saw the PG images (dashed blue squares). In addition, subjects also saw two R and two PG images (overlap in the middle). During *Set exposure* (A), participants saw their assigned images in a group. In the *Rating Task*, they rated each of these images individually (B). In the *New Image Rating Task* (C), they rated one of three possible sets of new R and PG images. Finally, subjects were randomly assigned to one of three possible questionnaire condition (D).

Specifically, we establish a causal pathway from beliefs about appropriate sharing, to disclosure, and to advice to others about disclosure.

EXPERIMENT

Overview of Design

We would like our experimental data to accomplish three goals. The first is to directly identify perceptions of norms associated with divulging information. Second, the data should allow us to test whether beliefs about norms can be nudged with norm-shaping design patterns, and third, whether subsequent behavior is affected. To accomplish these goals, we designed an experiment that tests how observing others' behaviors can influence (1) personal beliefs about information disclosure, (2) subsequent disclosure of information about oneself, and (3) subsequent advice to others about appropriate disclosure of information about themselves.

At a high level, our experiment tests whether people exposed to more (or less) provocative images display different perceptions of what is acceptable to share and subsequently change their information sharing behaviors. Our high level hypotheses are that more provocative imagery shapes norm perception and nudges subsequent behavior. We chose posting images (or "selfies") as our social media context because we were able to achieve a high degree of experimental control through systematic variation in the images shown to subjects (see Figure 3). More broadly though, designers have control over how and where widgets are shown; our experiment focuses on one such possible manipulation in one possible context. Though the experiment itself is highly controlled, it resembles what might actually be seen in real designs (we further explore generalizabiltiy in the discussion).

The experimental design consists of four steps. The first step exposes subjects to different image-exposure sets. They are randomly assigned to see images that are either more or less risky/provocative (in Figure 2A the solid red tiles depict more provocative images and the dashed blue tiles depict less provocative images). Throughout our analysis we will refer to the more provocative images as "R" and the less provocative images as "PG." In study 2, described below, we obtain independent ratings on each picture to be able to categorize them on appropriateness and on attractiveness.

In the second step, we ask subjects to rate how personally appropriate they find each of the images they received in the image-exposure set (Figure 2B). In the third step, we show them 2 new images that were not originally part of the image-exposure set and ask them to rate how personally appropriate they find the new images to be. The two new images contain one R (solid red tile in Figure 2C) and one PG image (dashed blue tile Figure 2C). We created three pairs of new R/PG images for this step which are matched to be similarly appropriate (ratings from study 2, described below, are used to do this matching). However, subjects are only shown *one*, randomly selected, new R/PG pair. We will refer to these images as "new images" in our analysis.

In the fourth step, we observe subjects' disclosure behaviors and advice-giving behavior through their responses to a questionnaire. Subjects are randomly assigned to respond to one of three possible questionnaires. Either they respond to a questionnaire about themselves or they are asked to be an adviser to another fictitious person who either has a more "vanilla" or a more provocative, "cinnamon," personality.[1]

The design of the main experiment consists of a 2 (initial image exposure set is R or PG) $\times$ 3 (new image pair #1, #2, or #3) $\times$ 3 (questionnaire regarding self or advice to vanilla or advice to cinnamon person) design. Further, we minimize self-image concerns, where a subject may assess the relative physical physique between image and him/herself, by giving our male subjects female selfies and vice versa. Table 1 summarizes the treatment conditions in the main experiment. A subject could only be assigned to one of these cells. In what follows we describe each step in detail and relate it back to our high-level hypotheses.

	Questionnaire Self	Questionnaire Advise Cinnamon-type	Questionnaire Advise Vanilla-type
R	R - Self	R - Advise (Cinnamon)	R - Advise (Vanilla)
PG	PG - Self	PG - Advise (Cinnamon)	PG - Advise (Vanilla)

Table 1. Participants were randomly assigned into one of these six conditions. Within each cell, they were randomly shown new image pair #1, #2, or #3. A participant was only assigned to one of the cells and did not participate in any of the other conditions. ***Note: the text in the cell describes the exposure set and questionnaire type a subject received.***

Step one: initial set exposure

To create the environment for step one, we generated a fake photo-driven social media site, ThisIs.Me (see Figure 3), and told subjects that we were introducing a new feature for the site. Our instructions read: "This plugin would scan pictures that you choose to post. For certain images, the plugin would ask you to wait for 10 minutes before deciding to post an image and then, after 10 minutes, it would ask you:'Do you want to post this?"' We informed subjects that their job was to teach the plugin what kinds of pictures the subject might later regret or wish that they had not posted. The instructions explained that "To help the plugin learn how to advise you, we will give you a set of 'training' images that others have posted [...] and ask you to rate how appropriate you think they are." Subjects used a scale from 1 "very inappropriate to post" to 6 "very appropriate to post" to rate how *personally* appropriate they felt it was to post to our hypothetical social network site.

After reading the instructions, subjects were exposed to the initial set of images all at once on one screen–we call this the initial set exposure. The initial set exposure mimics a strategic surfacing of images to a new user that might happen when they first log onto the site. The intent is to shape the user's perceptions of what others are posting and consider acceptable. The question we test in step two is whether initial exposure to an R or PG set can also change what the user *personally* believes is acceptable for him or herself to post on our social network site.

To manipulate the type of selfie posting behavior subjects saw in the initial set exposure, we collected two types of images from existing Instagram accounts. We collected one R (where the individual is mostly undressed) and one PG image (where the individual is mostly dressed) from the same Instagram account so that our R and PG sets contained comparable images. Pairs of images were selected so that the figure was roughly in the same pose and the picture was taken with the same camera angle. The initial R set contained a total of 14 images selected from Instagram accounts: 12 R images and also 2 "overlapping" PG images. The initial PG set also contained 14 images: 12 PG and also 2 "overlapping" R images.

We used the 2 PG and 2 R overlapping images to create a common group of 4 images that subjects in both the R and PG conditions saw. This is visually depicted in Figure 2A by the solid red and dashed blue outlines overlapping over the 4 central tiles. These overlapping images are used in the analysis because they allow us to test for how the initial set exposure affects subject perceptions of appropriateness on identical images.

Step two: Rating task

To test whether the initial exposure set impacts personal beliefs about what is appropriate to post, subjects rate each image from the initial set one at a time on subsequent screens (although the order was randomized at the subject level) (Figure 2B). Our participants used a Likert scale to rate how *personally* appropriate they felt it would be to post the image on the ThisIs.Me web site. The critical point here is that subjects were asked to tell us their *personal* opinions about how appropriate an image was to post. We purposefully did not ask them to tell us whether they thought *others* on the site would think the post was okay, but instead focused on how their personal views about posting the images were affected. Put another way, it would be very reasonable for an end-user

[1]We did not describe the fictitious person as vanilla or cinnamon to subjects, but adopt that language here for ease of description.

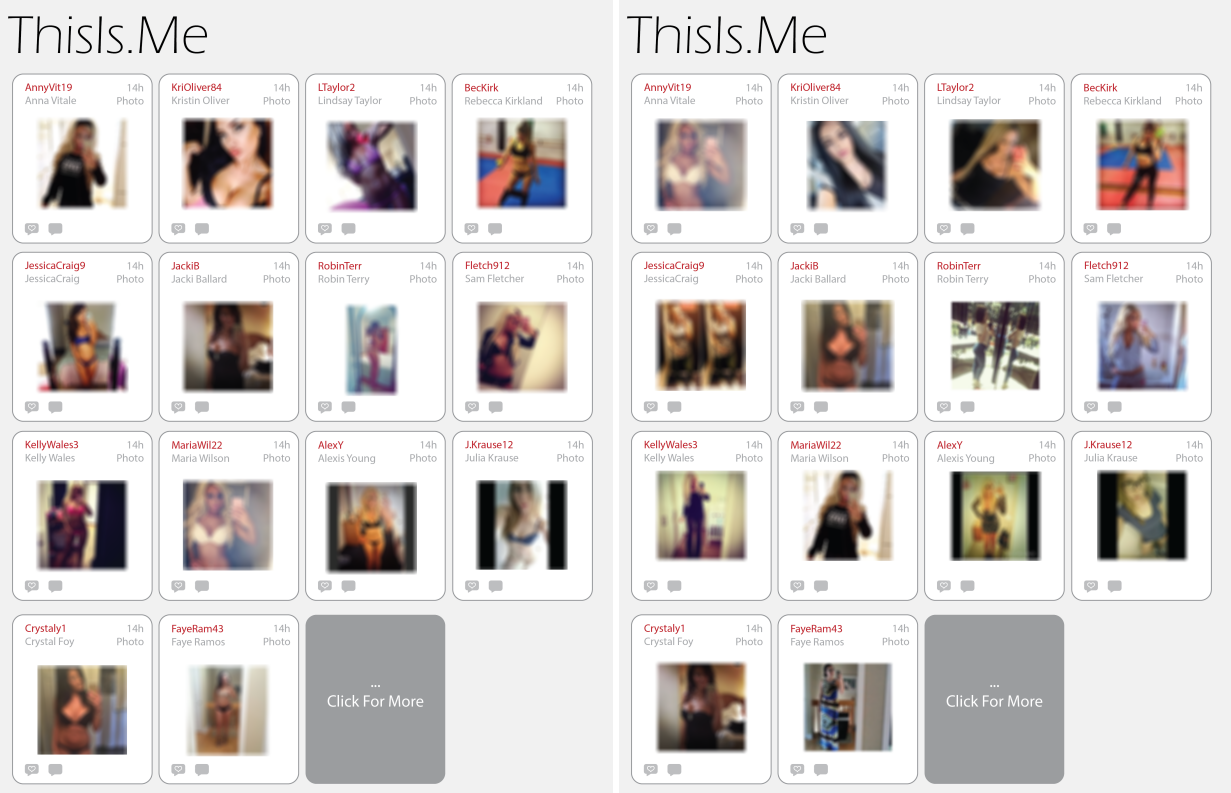

Figure 3. Examples of the ThisIs.Me page that subjects saw during *Set Exposure*. R individuals saw the image on the left while PG individuals saw the image on the right. The images in the experiment itself were not blurred.

to infer what others think is appropriate to do based on their posting behavior. What this design tests is how the behavior of others changes what the *subject* thinks is appropriate for *the subject herself* to do. With this data we can test our first hypothesis:

Hypothesis 1 *Individuals exposed to the R set will rate the pictures in the R set to be more personally appropriate than those exposed to the PG set.*

Once subjects finished rating the selfies from the initial set, they were randomly assigned to rate one of three pairs of new images (Figure 2C).[2] In this rating task, participants saw one new R selfie and one new PG selfie. We introduced these new images so that we can test the spill-over effects of exposure to the initial set onto new images.

Hypothesis 2 *Individuals exposed to the R set will rate new pictures to be more personally appropriate than those exposed to the PG set.*

Step three: Questionnaires

In step three we test whether subjects are more likely to divulge (or advise another to divulge) information in a subsequent task.

[2]To control for the possibility that the difference in attractiveness of the individual in the selfies may affect subjects' perception of appropriateness, we provided three different sets of new images, as opposed to just one. Each image was rated on attractiveness in study 2, described below, and all analyses control for this.

To test whether the effect of initial set exposure extends beyond changing beliefs about what is appropriate to do *within* the social media site, our subjects filled out a questionnaire (23 questions in total). The questionnaire immediately followed steps one and two, but we randomized subjects into one of three conditions.

In the self-questionnaire condition, subjects were told that for the ThisIs.Me site "...users may create a profile card that is visible to other members of the site. The profile card will have your username, your selfie and your answers to the series of questions in this section. If you prefer to not have an answer show up in your profile card, you may choose to skip that question by selecting the 'skip' option." The questionnaire (a complete copy is available in the supplementary materials) contained items that rang from less intrusive (e.g. "How often do you hold the door open for someone?") to very intrusive (e.g. "Did you ever have sex with someone who was too drunk to know what they were doing?"). These questions were scaled on intrusiveness in previous work by Acquisti et al. [6].

We also used the responses from the self-questionnaire to create two types of potential new users to the site. We selected one set of responses that were more conventional–we termed this our vanilla new user. In another case we selected a set of responses that were not conventional –we termed this our cinnamon new user. We chose these two types of respondents to mimic a scenario where a new user's answers are well outside of how most others are answering the questions.

In the advise-questionnaire conditions, subjects saw the new user's responses to the questionnaire items and subjects were asked to provide advice to the new user on whether to include this on the profile card. Subjects saw either the vanilla or the cinnamon new user's responses and were instructed: "A user of this site has filled out this questionnaire, but has not yet submitted and published their answers on their profile card. Given their potential answers, help them decide which questions they should choose to submit and publish and which they should choose to 'skip."'

Because subjects were randomized into the three questionnaire conditions (self, advise-vanilla and advise-cinnamon), we can test three hypotheses. We can use the self-questionnaire responses to test how initial set exposure affects the likelihood that a subject chooses not to divulge information about him/herself by skipping some questions. Second, we can test whether initial set exposure affects the advice a subject will give to a cinnamon or vanilla set of responses.

Hypothesis 3 *Individuals exposed to the initial R set will skip fewer questions in the self-questionnaire condition than those exposed to the initial PG set.*

Hypothesis 4 *Individuals exposed to the initial R set will advise a cinnamon type to skip fewer questions than individuals exposed to the initial PG set.*

Hypothesis 5 *Individuals exposed to the initial R set will advise a vanilla type to skip fewer questions than individuals exposed to the initial PG set.*

Study 2: Baseline ratings

Lastly, since the selfies may vary in appropriateness and attractiveness even within the R and PG groups, we ran a second study with different subjects to collect baseline appropriateness and attractiveness ratings for each of the images we used in the our main study. Subjects saw all of the R and PG images together and then rated the images. Thus, their appropriateness ratings were not made after being exposed to a biased set of R or PG images. All of our regressions control for the baseline appropriateness and attractiveness ratings. In our analysis we refer to these ratings as our "baseline appropriateness rating" and "attractiveness ratings." Together, we refer to them as our "controls."

RESULTS

To pilot the experiment we utilized a pool of students at a large academic institution (undergraduate and graduate students). To achieve greater analytical power and to generalize beyond this pool we utilized Amazon's Mechanical Turk infrastructure. A total of 387 Turk workers participated in our study. Of those, 305 participated in our main study (105 female and 200 male) while 82 (38 female and 44 male) participated in study 2. It took subjects an average of 6-7 minutes in both studies. We restricted participation to Turkers who have had at least 10,000 approved HITs and a HIT approval of at least 98%. Subjects were paid $0.50 for completing the survey. We note that our Turk-based results are consistent with our pilot experiment, giving us some confidence in the stability of the results across populations.

	(1) Images: Overlap R	(2) Images: Overlap R	(3) Images: Overlap PG	(4) Images: Overlap PG
Dependent variable: Appropriateness rating				
R set exposure	0.328**	0.299**	0.286***	0.295***
	(0.145)	(0.125)	(0.0917)	(0.0891)
Baseline appropriateness rating		0.940***		0.713***
		(0.335)		(0.195)
Attractiveness rating		0.518		-0.200
		(0.597)		(0.187)
Baseline rating differences		0.227		0.250***
		(0.248)		(0.0735)
Attractiveness rating differences		(Omitted)		(Omitted)
Male		(Omitted)		(Omitted)
		()		()
Constant	2.563***	-2.723	5.009***	1.563
	(0.0972)	(2.129)	(0.0634)	(1.653)
Observations	610	610	610	610
R-squared	0.015	0.265	0.021	0.162
Subjects	305	305	305	305

Robust standard errors in parentheses, clustered on ID.
*** p<0.01, ** p<0.05, * p<0.10

Table 2. OLS regressions comparing appropriateness ratings on the 4 overlapping images by initial R or PG set exposure.

Rating task

Initial set exposure affects image appropriateness ratings

To test the influence of the initial set exposure on personal appropriateness ratings, we regress the appropriateness ratings subjects provided on the 4 overlapping images depicted in Figure 2A, on a dummy variable for the initial set exposure (if "R" then "R set exposure" takes the value of 1 and 0 otherwise) and on controls. We find evidence consistent with Hypothesis 1 (see Table 2).

Looking at column (1) of Table 2, R-exposure set subjects rate the overlapping R images to be more appropriate to post than those exposed to the PG set ($p<0.05$). This effect does not diminish when we include controls for the baseline appropriateness ratings, the attractiveness ratings of the individuals depicted in the selfies, and the differences in baseline appropriateness ratings between the overlapping R and PG images (column (2)) ($p<0.05$). We find a similar effect of set exposure on the ratings of overlapping PG images. Set exposure to R increases personal appropriateness ratings on the PG images (column (3), $p<0.01$) and is not affected by controls (column (4), $p<0.01$).

We also wish to test the impact of the initial exposure set on how subjects rate the new images, Hypothesis 2. To do so, we regress appropriateness ratings for the new images on a dummy variable for initial set exposure to R ("R set exposure") and find evidence that supports Hypothesis 2. We report the coefficients of this regression in (Table 3).[3]

We find that those initially exposed to the R set rate the new selfie to be more appropriate to post than those exposed to the

[3]We note that solving for differences between the treatments using a generalized linear model is mathematically equivalent to ANOVA (ANOVA being a particular case of the linear regression model with factor levels represented by dummy variables–e.g. 1 for the R group and 0 for PG). Thus, the analysis and conclusions drawn from regression and ANOVA are equivalent here.

PG set. This is true when rating the new R image (column (1), $p<0.01$) and the new PG image (column (3), $p<0.01$). These results are also not sensitive to including controls for the selfie and a dummy for the participant's gender (columns (2) and (4)), ($p<0.01$).

	(1) Image: New R	(2) Image: New R	(3) Image: New PG	(4) Image: New PG
Dependent variable: Appropriateness ratings				
R set exposure	0.523***	0.513***	0.290***	0.292***
	(0.149)	(0.147)	(0.0937)	(0.0895)
Baseline appropriateness rating		(Omitted)		1.322***
				(0.427)
Attractiveness rating		0.204		0.313
		(0.730)		(0.642)
Baseline rating differences		0.307		(Omitted)
		(0.718)		
Attractiveness differences		-0.407		0.363
		(1.457)		(0.890)
Male		(Omitted)		0.192
				(1.010)
Constant	2.082***	0.632	5.335***	-3.614
	(0.0975)	(2.919)	(0.0716)	(3.924)
Observations	305	305	305	305
R-squared	0.039	0.082	0.030	0.127
Subjects	305	305	305	305

Robust standard errors in parentheses, clustered on ID.
*** p<0.01, ** p<0.05, * p<0.10

Table 3. OLS regressions comparing appropriateness ratings on the 2 new images by initial R or PG set exposure.

Questionnaire

Initial set exposure affects information disclosure behavior

Since the questions in our questionnaire vary by intrusiveness, we expect to see more individuals skipping the more intrusive questions relative to the less intrusive questions. However, we also hypothesize that those participants who were initially exposed to the R set would skip fewer questions (Hypothesis 3) relative to those initially exposed to the PG images.

Figure 4 graphs the frequency of skips in the self-questionnaire by initial set exposure (R or PG). The red (bottom) line plots the skip frequency of subjects who were initially exposed to the R set and were answering the self-questionnaire. The blue (top) line plots the skip rate for those respondents who were initially exposed to the PG set.[4] Consistent with Hypothesis 3, for all questions, the skip rate for those exposed to the R set is lower than those exposed to the PG set.

To formally test whether the differences observed in Figure 4 are significant, we perform a logistic regression. We regress a dummy variable (1 if the question is skipped and 0 otherwise) on a dummy variable for whether the individual was initially exposed to the R set ("R set exposure"). Consistent with Hypothesis 3, individuals initially exposed to the R set are significantly less likely to skip questions (average discrete effect is 6.75% at $p<0.10$).

Initial set exposure affects advice about disclosure

Finally, we test the influence of the initial exposure set on advice-giving behavior. Recall, that subjects who received the advice-questionnaire, saw a fictitious vanilla or cinnamon subject's responses and were asked to advise whether each

[4]The graphs plots the skipping frequency from least frequently skipped to most frequently skipped question on the *x*-axis.

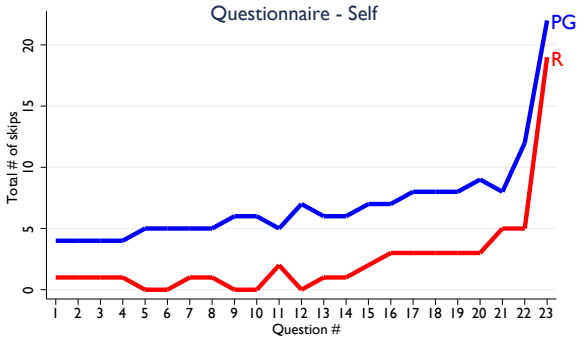

Figure 4. Frequency of skipped questions in the self-questionnaire condition by initial R (red line) and PG (blue line) set exposure.

response should be skipped or posted as part of the profile. Figure 5 graphs the total number of advised skips by question (*x*-axis) and by cinnamon (left panel) or vanilla condition (right panel). The red line plots the total advised skips made by subjects who were initially exposed to the R set and the blue line plots the total advised skips made by subjects initially exposed to the PG set.

Looking just at the advice given to the cinnamon type (left panel), we see that initial exposure to the R set causes subjects to advise the cinnamon type user to skip fewer questions than initial exposure to the PG set. This difference is not visually apparent in the advice given to the vanilla type individual (right panel).

To formally test whether these differences are statistically significant, we regress a skip-dummy (which takes the value of 1 if the advice was to skip and 0 otherwise) on whether the adviser was initially exposed to the R (dummy is equal to 1) or PG set (dummy takes value of 0). Consistent with Hypothesis 4, exposure to the R set makes individuals less likely to advise the cinnamon type user to skip questions than exposure to the PG set (average discrete effect: 6.34%, $p<0.10$).

By contrast, we find no such differences when advice is given to the vanilla type of user. Initial exposure to the R or PG set, does not significantly affect the skipping advice given to a vanilla type user ($p=0.406$). Hypothesis 5 is not supported.

These results suggest that the effects of set exposure spill over into other areas of information disclosure. Further, they suggest that the effects of set exposure on advice about information disclosure impact advising when it is, in some sense, very desirable that they do so. If the goal is to get interesting and information rich responses posted, then set exposure is an effective tool.

DISCUSSION

Given the ubiquity of social media systems there is a fundamental need to understand and mitigate against manipulative designs that lead to over-sharing. Users of these systems, many of whom are part of sensitive populations, are not always aware of the potential consequences of sharing their personal information (e.g., SSN inference, social phishing attacks). Further, as social media systems gain popularity, de-

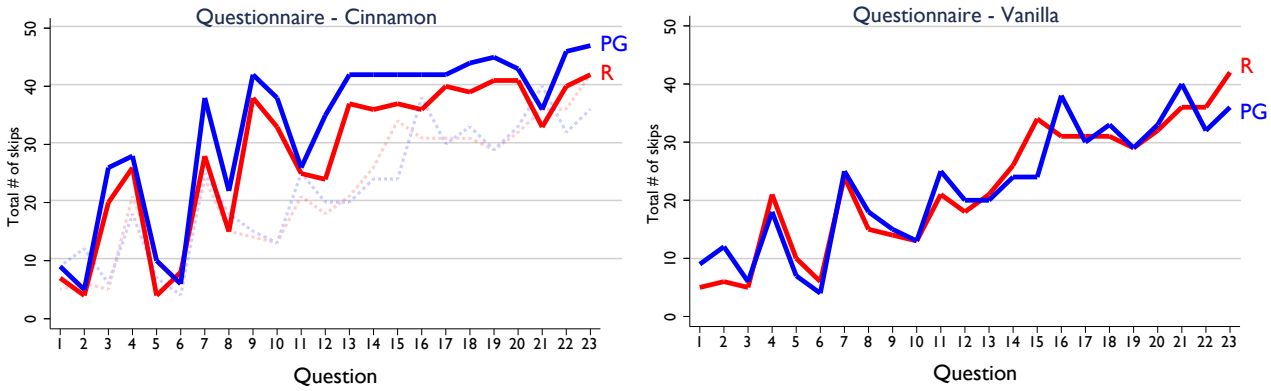

Figure 5. Total number of *advised* skips broken out by whether the adviser was initially expose to the R set of images or the PG set. The left panel plots the advice given to a cinnamon individual and the right panel plots the advice given to a vanilla individual. The lighter lines on the left panel are the superimposed lines from the right (vanilla) panel. This is done to ease comparison.

signers of social media interfaces are creating a set of interface heuristics [14].

While in some situations these heuristics emerge to satisfy certain design aesthetics, in many cases they are a result of conventional wisdom, theory, and scientific literature about motivating continued use of these systems and may favor designs that increase sharing of private or information rich data. For example, consider the following UI components (some existing, some hypothetical):

1. A widget that indicates that 3 of your friends have a birthday today. This element shows you that your friends have shared their birthdays and may change your belief about the prescriptive norm of sharing birthday information. A change in beliefs about how acceptable or frequent others' sharing is, may increase your own propensity to share.
2. A quiz element indicating "90% of your friends took the salary quiz, find out how you rank." This element explicitly indicates how many of your friends took the survey, again encouraging a change in your belief about norms around sharing financial information that may result in your sharing such information.
3. A feed element displaying a recent photo album posted by a friend with 3 representative images. Such an element surfaces a friend's behavior and encourages the belief that posting photo albums is a norm. More subtly, as we have shown, the strategic choice of representative images can also distort an end-user's models of the type of pictures that should be shared. For example, the album may contain 90% pictures of sunsets, by which a random selection of representative images would likely show 3 sunset images. However, algorithms could surface non-representative images that contain flesh tones, have comments, or tagged friends. The strategic selection of such images can change behavior (posting more swimsuit pictures, less sunset pictures) by changing perceptions of how acceptable certain images are.
4. A site ad for an application that "helps you track your sexual contacts." This ad's request for highly sensitive information is intentionally targeted to be far outside the norms of typical requests; it acts to erode taboos and tells users something about the norms (i.e., what is acceptable to talk about). Through the mere act of asking, designers give users the impression that acceptable topics for sharing encompass more provocative topics. Further, such a request could be used as part of a two-steps-forward-one-back strategy that explicitly asks end-users to share highly sensitive information so that other less sensitive, related requests–tracking your dating history–are perceived as more acceptable to share. Through this strategy designers can affect people's perceptions of norms regarding the acceptability of sharing the less sensitive information (after all, it is not sexual contacts!) and increasing the likelihood they do so.

Our results can be used to both safeguard and educate users, designers, and legislators. When automated, instruments can be developed to test new interface changes and provide early warning to users and those developing legal frameworks to better monitor, understand, and counteract the inappropriate use of these patterns and to enhance security.

Our research impacts the broader discourse and development of tools for the study of user interfaces as embodiments of social norms and other aspects of the culture and organization that the interface represents. Further, the combination of survey research and experimental economics, that allows us to explore differences between preferences and behavior and to explore issues of interface design, is a model that may be exploited in analogous research.

However, there are limitations to our findings that are important to mention here. Because of our experience with the more controlled pilot pool, we believe that the MTurk population is suitable for this experiment. Further, we purposefully chose a situation (posting pictures/questionnaire) that captures common online activities (from dating sites to Snapchat to Pinterest) where this type of information is routinely divulged. However, we believe it is also worth expanding the experiment to work within existing social media frameworks to further test for ecological validity. Though we believe that our experiments offer important advantages for studying the effects of UI designs, and our questions in particular, they also suffer from

known drawbacks (such as limited generalizability and the relative simplicity of choice environments in the "lab"), which make the use of multiple methods attractive. For this reason, future work will rely on surveys to gather norm-shaping design patterns that are found in the wild and on qualitative coding of design patterns and common "widgets" that are present in social media systems (making use of established patterns) (e.g., [14]) to identify common patterns used by designers.

Though participants were revealing their answers to us (the experimenters), a potential limitation is that the data were obtained from a hypothetical situation. Participants in the rating tasks were asked to evaluate how appropriate they personally felt the photos were to "post"-elicited ratings were for the participants' view of what is (in)appropriate to share on the hypothetical social network site. Though participants did not experience social consequences from revealing information, a user on an social network site would be engaging in the same type of thinking when deciding whether to reveal personal information. Further, studies have suggested that for some decision tasks, participants respond similarly when faced with real and hypothetical consequences (cf. [38]).

CONCLUSIONS

The design of social media interfaces greatly shapes the extent and timing of people's decision to reveal private information. The mechanisms through which interfaces affect disclosure behavior are still poorly understood. Previous research demonstrates how signals of social norms are useful for online social communities to assist newcomers, and can prevent certain behaviors and encourage others [21]. There is also extensive literature on how decision contexts may be altered to nudge behavior towards some target outcome [5, 32].

What we show is that signals embedded in design form social guideposts that shape users' beliefs about what kind, and how much, information is acceptable to reveal and ultimately translate into greater information disclosure. We experimentally unpack the causal pathway by which design patterns affect disclosure: they can modify perceptions of appropriate behavior which, in turn, impact subsequent behavior. We then demonstrate that this shift in perceptions can leave a larger footprint on user behavior than typically anticipated. This is because the shift in perceptions of appropriate behavior also impacts the advice that a user gives others.

Taken together, we identify a powerful cycle of behavior nudging and norm shaping through design patterns. The cycle begins with a design pattern that shapes *perceptions* of what is personally acceptable to do. These perceptions then nudge users to change behavior (in our case, share more about themselves by skipping fewer questions). An increase in sharing of private and information rich data then feeds back into the system and does shape the norms of the community. It does so in two very powerful ways. First, through altered behavior by individual users. Second, it changes the actual norms because those "nudged" users also change what advice they give other newcomers. Taken together, these findings have significant implications for security and privacy.

Social media systems touch the lives of hundreds of millions of end-users daily, and even minor design changes contribute to changes in behavior regarding the sharing of privacy-sensitive information among these users. By failing to account for the norm-setting implications of design choices, or only understanding them through vague intuitions, designers are poorly equipped to model the impact of their choices and the long term consequences to their end-users and sites. End-users and policy makers are similarly blind. In this paper we describe a novel approach to understanding the impact of social media user interfaces on social norms. Our study allows us to test how individuals trade-off the gains from information sharing against norms. As critically, we are also able to isolate a causal pathway from UI choices designers make to information sharing behavior.

ACKNOWLEDGEMENTS

We thank our anonymous reviewers for their feedback. This material is based upon work supported by the National Science Foundation under Grant No. 1537483/1537143.

REFERENCES

1. Alessandro Acquisti. 2004. Privacy in Electronic Commerce and the Economics of Immediate Gratification. In *Proceedings of the 5th ACM Conference on Electronic Commerce (EC '04)*. ACM, New York, NY, USA, 21–29. DOI:http://dx.doi.org/10.1145/988772.988777

2. Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2015. Privacy and human behavior in the age of information. *Science* 347, 6221 (2015), 509–514. DOI:http://dx.doi.org/10.1126/science.aaa1465

3. Alessandro Acquisti and Ralph Gross. 2006. Imagined communities: Awareness, information sharing, and privacy on the Facebook. In *Privacy Enhancing Technologies*. Springer, 36–58.

4. Alessandro Acquisti and Ralph Gross. 2009. Predicting Social Security numbers from public data. *Proceedings of the National Academy of Sciences* 106, 27 (2009), 10975–10980.

5. Alessandro Acquisti and Jens Grossklags. 2005. Privacy and rationality in individual decision making. *IEEE Security & Privacy* 1 (2005), 26–33.

6. Alessandro Acquisti, Leslie K. John, and George Loewenstein. 2012. The impact of relative standards on the propensity to disclose. *Journal of Marketing Research* 49, 2 (2012), 160–174.

7. Eytan Adar, Desney S. Tan, and Jaime Teevan. 2013. Benevolent Deception in Human Computer Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1863–1872. DOI: http://dx.doi.org/10.1145/2470654.2466246

8. Solomon E. Asch. 1956. Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs: General and Applied* 70, 9 (1956), 1–70.

9. Oriana Bandiera, Iwan Barankay, and Imran Rasul. 2005. Social Preferences and the Response to Incentives: Evidence from Personnel Data. *The Quarterly Journal of Economics* 120, 3 (2005), 917–962. DOI: http://dx.doi.org/10.1093/qje/120.3.917
10. Nicholas Bardsley and Rupert Sausgruber. 2005. Conformity and reciprocity in public good provision. *Journal of Economic Psychology* 26, 5 (2005), 664–681.
11. danah boyd and Eszter Hargittai. 2010. Facebook privacy settings: Who cares? *First Monday* 15, 8 (2010). http://firstmonday.org/ojs/index.php/fm/article/view/3086
12. Robert B. Cialdini, Raymond R. Reno, and Carl A. Kallgren. 1990. A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology* 58, 6 (1990), 1015.
13. Gregory Conti and Edward Sobiesk. 2010. Malicious Interface Design: Exploiting the User. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. ACM, New York, NY, USA, 271–280. DOI: http://dx.doi.org/10.1145/1772690.1772719
14. Christian Crumlish and Erin Malone. 2009. *Designing social interfaces: Principles, patterns, and practices for improving the user experience*. " O'Reilly Media, Inc.".
15. Bernhard Debatin, Jennette P. Lovejoy, Ann-Kathrin Horn, and Brittany N. Hughes. 2009. Facebook and Online Privacy: Attitudes, Behaviors, and Unintended Consequences. *Journal of Computer-Mediated Communication* 15, 1 (2009), 83–108. DOI: http://dx.doi.org/10.1111/j.1083-6101.2009.01494.x
16. Morton Deutsch and Harold B. Gerard. 1955. A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology* 51, 3 (1955), 629.
17. Simon Gächter and Elke Renner. 2003. *Leading by example in the presence of free rider incentives*. Technical Report. University of St. Gallen.
18. F. Maxwell Harper, Yan Chen, Joseph Konstan, and Sherry Xin Li. 2010. Social comparisons and contributions to online communities: A field experiment on movielens. *The American economic review* (2010), 1358–1398.
19. Bart P. Knijnenburg and Alfred Kobsa. 2013. Helping Users with Information Disclosure Decisions: Potential for Adaptation. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI '13)*. ACM, New York, NY, USA, 407–416. DOI: http://dx.doi.org/10.1145/2449396.2449448
20. Bart Piet Knijnenburg, Alfred Kobsa, and Hongxia Jin. 2013. Counteracting the negative effect of form auto-completion on the privacy calculus. (2013).
21. Robert E. Kraut, Paul Resnick, and Sara Kiesler. 2012. *Evidence-based social design: Mining the social sciences to build online communities*. MIT Press.
22. Erin Krupka, Steve Leider, and Ming Jiang. 2015. *A meeting of the minds: informal agreements and social norms*. Working paper. University of Michigan.
23. Erin Krupka and Roberto A. Weber. 2009. The focusing and informational effects of norms on pro-social behavior. *Journal of Economic Psychology* 30, 3 (2009), 307–320.
24. Erin L. Krupka and Roberto A. Weber. 2013. Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association* 11, 3 (2013), 495–524.
25. James E. Reinmuth and Michael D. Geurts. 1975. The collection of sensitive information using a two-stage, randomized response model. *Journal of Marketing Research* (1975), 402–407.
26. Jen Shang and Rachel Croson. 2009. A Field Experiment in Charitable Contribution: The Impact of Social Information on the Voluntary Provision of Public Goods. *The Economic Journal* 119, 540 (2009), 1422–1439. DOI: http://dx.doi.org/10.1111/j.1468-0297.2009.02267.x
27. Sarah Spiekermann, Jens Grossklags, and Bettina Berendt. 2001. E-privacy in 2nd generation E-commerce: privacy preferences versus actual behavior. In *Proceedings of the 3rd ACM Conference on Electronic Commerce (EC '01)*. ACM, New York, NY, USA, 38–47. DOI:http://dx.doi.org/10.1145/501158.501163
28. Fred Stutzman and Jacob Kramer-Duffield. 2010. Friends Only: Examining a Privacy-enhancing Behavior in Facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 1553–1562. DOI: http://dx.doi.org/10.1145/1753326.1753559
29. Fred Stutzman, Jessica Vitak, Nicole Ellison, Rebecca Gray, and Cliff Lampe. 2012. Privacy in Interaction: Exploring Disclosure and Social Capital in Facebook. In *Proceedings of the International AAAI Conference on Web and Social Media*. AAAI. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4666
30. Abhay Sukumaran, Stephanie Vezich, Melanie McHugh, and Clifford Nass. 2011. Normative influences on thoughtful online participation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3401–3410.
31. Bob Tedeschi. 2012. E-commerce report; everybody talks about online privacy, but few do anything about it. (June 3 2012).
32. Richard H. Thaler and Cass R. Sunstein. 2008. *Nudge*. Yale University Press.
33. Roger Tourangeau and Ting Yan. 2007. Sensitive questions in surveys. *Psychological bulletin* 133, 5 (2007), 859.

34. Janice Y. Tsai, Serge Egelman, Lorrie Cranor, and Alessandro Acquisti. 2011. The effect of online privacy information on purchasing behavior: An experimental study. *Information Systems Research* 22, 2 (2011), 254–268.

35. Yang Wang, Pedro Giovanni Leon, Alessandro Acquisti, Lorrie Faith Cranor, Alain Forget, and Norman Sadeh. 2014. A Field Trial of Privacy Nudges for Facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 2367–2376. DOI: http://dx.doi.org/10.1145/2556288.2557413

36. Yang Wang, Pedro Giovanni Leon, Kevin Scott, Xiaoxuan Chen, Alessandro Acquisti, and Lorrie Faith Cranor. 2013. Privacy Nudges for Social Media: An Exploratory Facebook Study. In *Proceedings of the 22Nd International Conference on World Wide Web Companion (WWW '13 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 763–770. http://dl.acm.org/citation.cfm?id=2487788.2488038

37. Rick Wash. 2010. Folk Models of Home Computer Security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security (SOUPS '10)*. ACM, New York, NY, USA, Article 11, 16 pages. DOI: http://dx.doi.org/10.1145/1837110.1837125

38. David B. Wiseman and Irwin P. Levin. 1996. Comparing risky decision making under conditions of real and hypothetical consequences. *Organizational Behavior and Human Decision Processes* 66, 3 (1996), 241–250.

39. Sean D. Young and Alexander H. Jordan. 2013. The influence of social networking photos on social norms and sexual health behaviors. *Cyberpsychology, Behavior, and Social Networking* 16, 4 (2013), 243–247.