Identifying Experimental Risks Using Surveys

Stuart Schechter Cristian Bravo-Lillo

2014

We surveyed 3,539 workers on Amazon's Mechanical Turk to gauge their response to five scenarios describing scientific experiments—including one scenario based on Facebook's emotional contagion experiment. Respondents[1] who reported being already aware of Facebook's experiment responded very differently to the scenario based on it than those who reported being unaware, so we focused on 2,102 respondents who reported being unaware. We asked these respondents whether they would want someone they cared about to be included as a participant, interpreting an answer of 'no' as indicating concern for participants. A greater fraction of respondents were concerned about the two of the four scenarios inspired by university-approved experiments than expressed concern for Facebook's experiment. We also asked whether the experiment should be allowed to proceed, interpreting a 'no' answer as disapproval of the experiment. A similar or greater fraction of respondents disapproved of the two more controversial scenarios based on university-approved studies as disapproved of the Facebook-experiment scenario. We found a statistically significant reduction (for $\alpha = 0.05$) in disapproval and concern for participants in a group of respondents shown a hypothetical variant of Facebook's experiment in which the manipulation performed by researchers was to insert extra positive posts into users' news feeds—instead of removing positive or negative posts based on treatment group.

1 Introduction

In evaluating the ethicality of an experiment, researchers and ethics boards must weigh the benefits of the study against potential risks—many of which are borne by participants. Alas, there is a great deal of guesswork in anticipating how participants and others will react to an experiment. The information that researchers and ethics boards require in order to make sound judgements is hard to come by; researchers rarely share, or even measure, participants' feelings, concerns, and opinions about the ethicality of the experiments they take part in. The rare instances in which we learn about the ethical consequences of experiments typically occur when concerns or harms are so serious as to come to the attention of the public.

In 2012, we began using surveys to identify disapproval and concern with experiments *before* exposing participants to them. We presented respondents[1] with a series of short descriptions of experimental scenarios and asked questions to gauge their ethical response. We wrote these short summaries with the goal of presenting the information salient to evaluating the ethicality of a study into a form that could be read and understood by a general audience within a minute. Our first such survey caused us to re-evaluate how participants might react to a study we had planned (and received approval to conduct). In light of our survey data, we concluded that the benefits to society of running the experiment no longer appeared to outweigh the risks. We began publicly advocating for the prophylactic use of ethical-response surveys in 2013 [3].

In this new work, we ask what researchers at Facebook would have learned had they had the opportunity to prophylactically perform an ethical-response survey prior to commencing their 2012 emotional contagion experiment [12]. In Facebook's experiment, researchers used an algorithm to remove posts from users' news feeds in order to determine whether a reduction in positive or negative posts from participants' friends would impact the emotional mood of posts made by participants themselves. This experiment, which was published in June 2014, quickly became controversial—attracting criticism that the researchers and those overseeing their work presumably had not anticipated.

We performed an ethical-response survey on a convenience sample of 3,539 workers on Amazon's Mechanical Turk who were based in the United States from July 2–4, 2014. Of these, 2,102 reported not yet being aware of Facebook's emotional contagion experiment; these participants presented us with an opportunity to gauge opinions not yet influenced by media coverage and the evolving public reaction that has followed the publication of Facebook's experiment.

We presented participants with five experimental scenarios: one about the Facebook experiment and four about other experiments. The details of the scenario related to the Facebook experiment varied between respondents whereas the other four did not. We wrote our control scenario to describe the experiment based on our understanding of Facebook's experiment from reading their paper [12]. We created nine other treatments in which we change facts about the scenario

[1] We refer to the group of Amazon Mechanical Turk workers who completed our survey as *respondents*, as opposed to *participants*, to prevent confusion between them and the participants in the experimental scenario our respondents were asked about.

In this survey, you will be ask you about five hypothetical scientific experiments. For each, we will ask you to:

- *Carefully* read an abstract description of the experiment (350 words or fewer).
- Answer 4 multiple-choice questions about each experiment.
- Optionally provide short explanations of your answers.

Finally, we will ask you some brief demographic questions at the end of the study. All personal information (e.g., age) is optional.
Your responses will be kept anonymous, though we reserve the right to copy or quote the responses you provide.
The entire survey should take **under 10 minutes of your time** and pays **$1.00**.
This survey is part of a research project being conducted by *anonymized*.
If you have any questions, please feel free to contact us at: team@ethicalresearch.org

Figure 1: We provided the following description to prospective respondents. The error "you will be ask you" in place of "we will ask you" is in the original and not a transcription error. Fortunately, we described the task again correctly in the first page of the survey.

based on Facebook's experiment; we modified facts such as what manipulations Facebook's researchers had performed or even which company had performed the research.

Of the other four experimental scenarios, two described deception experiments that members of our team had led in the past and for which we had worked to measure participants' response to ethics questions asked after debriefing (though we elided the use of consent for these studies). The final two scenarios summarized research from the past decade that had been the subject of ethical debate within the research community (without these elisions as the studies had been performed without consent). All four experiments that these scenarios were based on had been conducted with approval from university ethics boards.

For each of the five abstracts we presented to each respondent, we asked two questions designed to gauge concern for participants and disapproval with allowing the study. The participant *concern* question asked whether the respondent would want someone they cared about to be included as a participant. The *disapproval* question asked whether the respondent believed the experiment should be allowed to proceed or not.

Respondents who reported being previously aware of Facebook's experiment reported more concern for participants, and greater disapproval of proceeding with the experiment, than those who reported not being previously aware of it.. However, those who were not previously aware of Facebook's experiment expressed greater concern and disapproval when confronted with scenarios describing two controversial university experiments from the past decade than they did for Facebook's experiment. We also identified variants of the Facebook-experiment scenario that reduced disapproval and concern, the most successful of which had the researchers manipulating news feeds by adding positive posts instead of removing (both positive and negative) posts.

2 Experimental procedure

We used a single Human Intelligence Task on Amazon's Mechanical Turk to prevent the same worker account from taking the survey twice (though we cannot guarantee some workers with multiple accounts did not do so). We restricted workers to those coming from the United States.

After brief instructions, we presented five experimental scenarios in random order (randomized for each participant). Four of the scenarios were the same for each respondent, but we randomly assigned each respondent one of ten variants of the Facebook experiment. We then asked follow-up questions.

2.1 Recruiting and instructions

We offered a Human Intelligence Task (HIT) on Mechanical Turk in which we presented prospective respondents with the offer in Figure 1.

We presented participants who accepted the HIT with the following instructions:

> Each of the following five pages will contain a description of a hypothetical scientific experiment, followed by questions about that experiment. In order to answer the questions, please read the description of each experiment carefully.

2.2 Questions for each scenario

We randomized the order of the experimental scenarios, but kept the ordering of questions and response options consistent.

The first question that followed the description of each scenario was one we designed to measure respondents' *concern* for those participating in the experiment. We asked: "If someone you cared about were a candidate participant for this experiment, would you want that person to be included as a participant?"

We asked respondents about someone they care about, as opposed to themselves, because they might be more comfortable imagining others to be vulnerable and needing protection, whereas they might not want to admit being vulnerable

themselves. We provided the option to respond "Yes", "I have no preference", or "No". We designed these options to be ordinal: from least concerned to most concerned. We asked this concern question first in hopes that it would give respondents a chance to humanize potential participants and think about the consequences of the experiment on them.

We designed the second question to gauge whether respondents would disapprove of the experiment. We asked, "Do you believe the researchers should be allowed to proceed with this experiment?" We offered four options, again ordered from most approving to least approving with the first option being "Yes" (on the left) and the last "No" (the fourth option, on the right). We included the second option, "Yes, but with caution", for respondents who did not want to disapprove of a experiment but feared that an unambiguous "yes" would relieve researchers to their duty to take their ethical duties seriously. The option in the third position from the left, between the two "Yes" options and "No", was "I'm not sure." We treat this an ordinal value between the yes and the no options as the respondent is unable to commit to either and is therefore likely to be somewhere in between.

For each of the first two questions, we gave respondents a free-response field in which to explain their answers.

We also asked respondents "Are you aware of having ever participated in such a study?" and "Are you aware of a study like this one having been performed by researchers in the past? (For example, have you have heard about it in the news or learned about it in a class?)". The answers options, from left to right, were "Yes" and "No".

2.3 Closing questions

After collecting respondents' responses to the five experimental scenarios, we asked the following questions about respondents' demographics and about factors that might influence their opinions:

- What year were you born?
  (please use a four-digit year, or 'd' if you decline to answer)
- What is your gender?
  {Male;Female;I'm uncomfortable answering}
- What is your occupation?
- Have you ever purchased goods advertised via an unsolicited marketing email?
  {Yes;No;I'm uncomfortable answering}
- Have you ever participated in a study that involved deception?
  {Yes; No; I'm uncomfortable answering}
- Prior to participating in this study, had you heard about Facebook's 'mood' study (the experiment that has the subject in many recent news stories).
  {Yes; No}

We placed the question about prior knowledge of Facebook's experiment at the very end of our survey so as to avoid having this question taint responses to earlier questions.

2.4 Payment

We paid all respondents $1 for the HIT regardless of their level of effort, answer quality, or time spent. We also calculated a wage for each participant based on their time spent responding to the survey at an hourly wage of $9.32 (the highest minimum wage of any state in the US), up to a maximum of $3.11 for 20 minutes of time. If the wage exceeded the $1 paid for the HIT, we paid a bonus equal to the difference. We paid bonuses after all surveys were complete—had we paid immediately and word spread, some respondents might have delayed completion.

3 Experimental scenarios

We created two scenarios for experiments from the past decade that were the subject of ethical debate in the research community, two scenarios for experiments that we had run and gauged participants' ethical response to at the time of the experiment, and one scenario (with ten variants) for the recent Facebook experiment. In no description of these experimental scenarios did we mention that the experiment described was a real experiment or, in the case of the university studies, that it had been approved by an ethics board.

A Social phishing

We wrote this experimental scenario around the "Social Phishing" experiment performed by researchers at Indiana University [9]. In their experiment, researchers sent students phishing emails to see if they could be deceived into revealing their passwords on a website that impersonated a university system. Some of the emails researchers sent were customized based on participants' public Facebook profiles. The researchers collected passwords from those who entered them and tested them against a university password database to determine if they were valid. The exact wording of this scenario is in Appendix A.

We did not mention that participants were exposed to the experiment without their consent.

B Spam infrastructure infiltration & analysis

The second experimental scenario describes an experiment to measure the economics of spam performed by researchers at the University of California [10]. In this experiment, the researchers allowed a computer to be infected with software used to send spam. The researchers then modified the spam to direct recipients to servers controlled by the researchers, instead of the spammers. Thus, recipients of attackers' spam became unwitting participants in this study. The exact wording of this scenario is in Appendix B.

As with the previous study, we did not explicitly state that spam recipients did not opt into the study via a consent form, though we did indicate that spam recipients who visited the impersonated store would not be informed that it was not the genuine store run by spammers.

C Password-dialog spoofing

This scenario describes an experiment by researchers at Carnegie Mellon University and Microsoft Research to determine whether malicious websites can trick users into revealing their device (computer) password by mimicking (spoofing) security dialogs that are normally generated by the device's operating system [4]. The researchers presented the experiment to participants as an evaluation of online gaming websites. When participants visited a website run by the researchers, the researchers mimicked the operating system window used to download a software component. The window indicated that it required the user's (participant's) device username and password to install the software component. The researchers observed whether participants could be deceived to enter that information. (Unlike the Indiana University phishing study, the researchers did not actually collect passwords without participants consent.) The exact wording of this scenario is in Appendix C.

The experiment on which this scenario was run by a team that includes two authors of our ethical-response survey (and the paper you are reading now). The experiment, which was led by Carnegie Mellon University and performed in collaboration with Microsoft Research, was was approved by the Institutional Review Board of Carnegie Mellon University.

Participants in the actual experiment had received a consent form explaining that they were part of a University experiment, though the consent form did not disclose that security was the focus of the experiment. The researchers informed study participants of the deception during a debriefing at the end of the experiment. We elided the presence of the consent form in order to make the scenario more similar to the other, more controversial, experiments described in this survey.

D Spoofed-warning deception

This scenario describes an experiment by researchers at Carnegie Mellon University and Microsoft Research to improve security warning dialogs [2]. Like the previous study, it is a deception experiment in which researchers led participants to believe that online games were the focus of the study. Unlike the previous study, users were not tricked into typing passwords. Rather, they were shown a warning about the risk of installing software and the researchers tested to see whether participants could identify signs of danger in the warning. Regardless of how participants responded to the install warning, no harm would come to them. The exact wording presented of the scenario is in Appendix D.

As with the previous scenario, the experiment on which this scenario was run by a team that includes two authors of our ethical-response survey (and the paper you are reading now). The experiment, which was led by Carnegie Mellon University and performed in collaboration with Microsoft Research, was was approved by the Institutional Review Board of Carnegie Mellon University. Participants in the actual experiment had received a consent form explaining that they were part of a university experiment, though the consent form did not disclose that security was the focus of the experiment. Further, the researchers collected data to monitor participants' ethical responses during the study to ensure harm was minimal. We elided these facts in order to make the scenario more similar to the more controversial experiments described in this survey.

F Facebook's emotional contagion experiment

This scenario, presented in Figure 2, describes Facebook's emotional contagion experiment, based on our understanding of the experiment from reading their paper. The scenario focuses on facts about the experimental goals and methodology and so avoids touching on many issues that have been a subject of public debate. Specifically, it does not discuss oversight, terms of service, or the participation of university researchers in the experiment. As is consistent with the other scenarios, we do not explicitly state that the researchers did not obtain consent from participants.

However, many respondents did not receive this exact scenario (our control), but instead received one of the variants (treatments) that are described in the next section.

Researchers at Facebook want to study whether users are more likely to share positive (happy) thoughts if their friends have been posting positive thoughts, and whether they are more likely to share negative (unhappy) thoughts if their friends have been sharing negative thoughts.

- To increase the proportion of positive posts in some users' news feeds, the researchers will randomly exclude some fraction of friends' negative posts each time the news feed is loaded.
- To increase the proportion of negative posts in some users' news feeds, the researchers will randomly exclude some fraction of friends' positive posts each time the news feed is loaded.
- The researchers will use an automated algorithm to measure whether users' posts are of a positive or negative mood.
- The researchers will publish the anonymized aggregate results of the experiment in a scientific paper.
- Participants will not be identified and will remain anonymous.

If the researchers are not allowed to perform this experiment, they will not be able to make a valid scientific determination of whether users' moods are affected by the moods of their friends' posts. Therefore, the researchers will not be able to produce features that might protect the moods of psychologically-vulnerable users.

Figure 2: The experimental scenario description we used for Facebook's emotional contagion experiment.

4 Treatments

We created ten variants of the experimental scenario for the Facebook experiment. We assigned respondents to scenario variants (treatments) at random with uniform probabilities assigned to each.

F0 Control

The control does not diverge from the facts of Facebook's experiment as we understood them, described in Section 3.F and detailed in Figure 2.

F1 Only remove positive posts

We designed this scenario to test the hypothesis that respondents would be more disapproving and concerned by the removal of negative posts than by the removal of positive posts. We thought respondents might see more harm in missing out on an opportunity to provide support to a friend in need, who had posted a negative post.

To construct this scenario we deleted the references to removing *negative* posts for the purpose of increasing the proportion of *positive* posts in the feed. From the first paragraph, we removed the string: "are more likely to share positive (happy) thoughts if their friends have been posting positive thoughts, and whether they". We also removed the first bullet point, which had stated: "To increase the proportion of positive posts in some users' news feeds, the researchers will randomly exclude some fraction of friends' negative posts each time the news feed is loaded."

F2 Only remove negative posts

We designed this scenario to test the hypothesis that respondents might be particularly concerned with participants missing out on good news. In this treatment, participants would only miss out on negative posts.

We deleted references to removing *positive* posts for the purpose of increasing the proportion of *negative* posts in the feed. From the first paragraph, we removed the string: "and whether they are more likely to share negative (unhappy) thoughts if their friends have been sharing negative thoughts". We also removed the second bullet point, which had stated: "To increase the proportion of negative posts in some users' news feeds, the researchers will randomly exclude some fraction of friends' positive posts each time the news feed is loaded."

F3 Remove mention of publication

We created this scenario to test whether respondents would feel more or less favorably if the mention of a scientific publication were removed. Specifically, we deleted the second-to-last bullet point of the scenario, which had stated "The researchers will publish the anonymized aggregate results of the experiment in a scientific paper."

F4 Remove mention of product improvement

We created this scenario to test whether respondents would feel less favorably about the experiment if there were no mention of potential for product improvement that might benefit users. We removed the last sentence of the scenario, which had stated that a consequence of not allowing the research would be that "the researchers will not be able to produce features that might protect the moods of psychologically-vulnerable users."

F5 Promise not to use for advertising

To test the hypothesis that respondents might respond more favorably to the experiment if the results would not be used for advertising, we created a scenario in which researchers promised this. We appended one item to the list of bullet points. It stated: "The researchers promise in writing that the research findings will be used only to further science and improve the product for users. The results will not be used to improve Facebook's advertising algorithms."

F6 Insert posts instead of hiding

We hypothesized that respondents might be less concerned about researchers manipulating news feeds if the researchers had only added extra (bonus) posts, as opposed to removing that had been deemed relevant by Facebook's existing algorithms.

We changed the description of the study design so that, instead of hiding posts, the researchers would add negative or positive posts that otherwise would not have been deemed worthy of display on the news feed. We rewrote the first two bullet points as follows:

- To increase the proportion of positive posts in some users' news feeds, the researchers will randomly include additional positive posts that would otherwise have been deemed insufficiently relevant or unimportant.
- To increase the proportion of negative posts in some users' news feeds, the researchers will randomly include additional negative posts that would otherwise have been deemed insufficiently relevant or unimportant.

F7 Insert posts, and only positive ones

We hypothesized that respondents might be even less concerned if the added posts were only positive posts. We started with the prior treatment (F6), and deleted from the first paragraph the string: "and whether they are more likely to share negative (unhappy) thoughts if their friends have been sharing negative thoughts". We kept the first bullet point from the prior treatment (F6), which described increasing the proportion of positive posts, but deleted the second one, which had described increasing negative posts.

F8 Replace 'Facebook' with 'a social network'

To test whether respondents' opinions would change if the experiment were not identified as being conducted by Facebook, we replaced the third word of the scenario, "Facebook", with the phrase "a social network".

F9 Replace 'Facebook' with 'Twitter'

To test whether respondents might be have responded differently to the experimental scenario had it been conducted by Twitter, we replaced the third word of this scenario, "Facebook", with "Twitter".

5 Results

After piloting on July 1, we began surveying respondents at 12:00AM EDT the morning of Wednesday July 2, 2014. We removed from our data 31 responses in which respondents spent under 150 seconds (30 seconds per scenario) completing the survey.

5.1 Awareness

We had intended to filter out those respondents who reported being aware of the Facebook experiment before our survey when making comparisons between different treatments; we would be unable to separate their response to their opinion about the hypothetical experiment described in the survey from their response to the actual experiment, which may have been influenced by the opinions of friends or the media.

1,437 of 3,539 respondents (41%) answered 'yes' when asked if they had been aware of Facebook's 'mood' study. As expected, these respondents expressed significantly greater disapproval and concern about this experiment, as can be seen in the top rows of Tables 1a and 1b. The result of differences in both the question about whether the experiment should proceed ($\chi^2(1) = 17.71$), and whether respondents would want those they cared about to participate in the experiment ($\chi^2(1) = 11.84$), were highly-statistically significant. We present the results of these comparisons in Table 4.

A tempting explanation for these highly-significant differences is that respondents' opinions were strongly swayed by opinions of the media and other sources of information about Facebook's experiment. It's also possible that disapproval for aspects of the experiment not described in our hypothetical experiment description carried over to their evaluation of the ethics of the hypothetical scenario, and that they would have disapproved without the influence of others' opinions had we only presented those facts. For example, the difference may be due to the fact that those who were aware of the study

		Responded 'no'			
		All respondents		Only respondents assigned control Facebook scenario (F0)	
	Experiment described in abstract	Aware	Unaware	Aware	Unaware
F	Facebook's emotional contagion experiment	682/1,437 (47%)	481/2,102 (23%)	67/145 (46%)	50/207 (24%)
A	Social phishing	504/1,437 (35%)	603/2,102 (29%)	52/145 (36%)	54/207 (26%)
B	Spam infrastructure infiltration & analysis	438/1,437 (30%)	518/2,102 (25%)	44/145 (30%)	58/207 (28%)
C	Password-dialog spoofing	213/1,437 (15%)	326/2,102 (16%)	21/145 (14%)	32/207 (15%)
D	Spoofed-warning deception	106/1,437 (7%)	138/2,102 (7%)	14/145 (10%)	15/207 (7%)

(a) "Do you believe the researchers should be allowed to proceed with this experiment?"

		Responded 'no'			
		All respondents		Only respondents assigned control Facebook scenario (F0)	
	Experiment described in abstract	Aware	Unaware	Aware	Unaware
F	Facebook's emotional contagion experiment	812/1,437 (57%)	667/2,102 (32%)	78/145 (54%)	72/207 (35%)
A	Social phishing	723/1,437 (50%)	950/2,102 (45%)	75/145 (52%)	83/207 (40%)
B	Spam infrastructure infiltration & analysis	742/1,437 (52%)	961/2,102 (46%)	71/145 (49%)	91/207 (44%)
C	Password-dialog spoofing	404/1,437 (28%)	592/2,102 (28%)	42/145 (29%)	55/207 (27%)
D	Spoofed-warning deception	203/1,437 (14%)	306/2,102 (15%)	28/145 (19%)	28/207 (14%)

(b) "If someone you cared about were a candidate participant for this experiment, would you want that person to be included as a participant?"

Table 1: We examine the proportion of respondents who responded 'no' in order to express disapproval for the study proceeding (1a) and answered 'no' to having someone they cared about participate in the study (1b). We break these numbers down by whether respondents reported being previously aware of Facebook's experiment. A far greater proportion of respondents who reported already being aware of Facebook's were critical of the scenario based on it than those participants who reported having been unaware.

had learned explicitly that the researchers did not receive consent from participants, whereas those shown the hypothetical scenario were not told explicitly whether the researchers had or had not obtained consent.

Yet another alternate hypothesis is that those who are most likely to disapprove of the ethics of the Facebook experiment, or of research studies in general, were more likely to seek out hear about it from friends or see coverage of it in the media. To examine the hypothesis that those aware of the Facebook experiment were more disapproving of experimental scenarios in general, we examine in Table 1 respondents' 'no' answers to the same to questions for the four unrelated experiments. While there are differences in the more controversial experiments, they are much smaller than for our abstracted versions of Facebook's experiment.

Given the differences between those aware and unaware of the Facebook experiment at the time of the survey, we exclude from the remainder of our analysis those 1,437 of 3,539 respondents (41%) who reported being already aware it.

5.2 Comparison to IRB-approved studies

Some critics of Facebook's emotional contagion experiment have argued that it should have received the same level of scrutiny that would be required of an experiment run at a university [1, 5, 6, 8]. Without stepping into the debate of whether or when an institutional review board is *necessary*, we believe our results may provide insight to the question of whether such review is guaranteed to be *sufficient*.

Among respondents who reported being aware of Facebook's experiment and whose opinions may have been influenced by media coverage, the scenario based on it received the highest proportion of disapproval and concern for participants. This was not the case for those who reported being previously unaware of the experiment.

Rather, for those respondents who did report being unaware of Facebook's experiment, the Facebook control scenario came in third in our measure of concern for participants, behind the social phishing scenario (A) and the spam infrastructure scenario (B). See the rightmost column of Table 1b. The two experimental scenarios that caused our respondents the greatest concern for participants were performed by universities under their ethical oversight.

For the disapproval question (whether the experiment should be allowed to proceed), shown in the rightmost column of Table 1a, the Facebook scenario received no more disapproval than the two controversial university scenarios.

The two less controversial university experiments (C and D), which we had potentially made more controversial by eliding the presence of consent, received less concern and disapproval from our respondents than the other three.

	Experiment described in abstract *(order of presentation randomized for each respondent)*	Response: **No**	I'm not sure	Yes, but with caution	Yes	*Total*
Facebook's emotional contagion experiment *(each respondent saw one variant, selected at random)*						
F0	Control	50 (**24%**)	22 (11%)	62 (30%)	73 (35%)	207
F1	Only remove positive posts	68 (**31%**)	27 (12%)	47 (21%)	78 (35%)	220
F2	Only remove negative posts	48 (**26%**)	19 (10%)	47 (26%)	70 (38%)	184
F3	Remove mention of publication	52 (**24%**)	24 (11%)	66 (30%)	78 (35%)	220
F4	Remove mention of product improvement	54 (**25%**)	29 (14%)	61 (29%)	68 (32%)	212
F5	Promise not to use for advertising	70 (**32%**)	24 (11%)	42 (19%)	83 (38%)	219
F6	Insert posts instead of hiding	36 (**17%**)	25 (12%)	57 (27%)	94 (44%)	212
F7	Insert posts, and only positive ones	33 (**14%**)	28 (12%)	45 (20%)	124 (54%)	230
F8	Replace 'Facebook' with 'a social network'	37 (**19%**)	20 (10%)	49 (25%)	87 (45%)	193
F9	Replace 'Facebook' with 'Twitter'	33 (**16%**)	24 (12%)	51 (25%)	97 (47%)	205
F	*Total*	481 (**23%**)	242 (12%)	527 (25%)	852 (41%)	2,102
Other experiments *(all respondents saw all experiments)*						
A	Social phishing	603 (**29%**)	240 (11%)	721 (34%)	538 (26%)	2,102
B	Spam infrastructure infiltration & analysis	518 (**25%**)	316 (15%)	739 (35%)	529 (25%)	2,102
C	Password-dialog spoofing	326 (**16%**)	169 (8%)	848 (40%)	759 (36%)	2,102
D	Spoofed-warning deception	138 (**7%**)	132 (6%)	644 (31%)	1,188 (57%)	2,102

(a) "Do you believe the researchers should be allowed to proceed with this experiment?"

	Experiment described in abstract *(order of presentation randomized for each respondent)*	Response: **No**	Indifferent	Yes	*Total*
Facebook's emotional contagion experiment *(each respondent saw one variant, selected at random)*					
F0	Control	72 (**35%**)	74 (36%)	61 (29%)	207
F1	Only remove positive posts	85 (**39%**)	79 (36%)	56 (25%)	220
F2	Only remove negative posts	57 (**31%**)	76 (41%)	51 (28%)	184
F3	Remove mention of publication	77 (**35%**)	70 (32%)	73 (33%)	220
F4	Remove mention of product improvement	79 (**37%**)	73 (34%)	60 (28%)	212
F5	Promise not to use for advertising	91 (**42%**)	69 (32%)	59 (27%)	219
F6	Insert posts instead of hiding	55 (**26%**)	81 (38%)	76 (36%)	212
F7	Insert posts, and only positive ones	47 (**20%**)	92 (40%)	91 (40%)	230
F8	Replace 'Facebook' with 'a social network'	49 (**25%**)	75 (39%)	69 (36%)	193
F9	Replace 'Facebook' with 'Twitter'	55 (**27%**)	72 (35%)	78 (38%)	205
F	*Total*	667 (**32%**)	761 (36%)	674 (32%)	2,102
Other experiments *(all respondents saw all experiments)*					
A	Social phishing	950 (**45%**)	493 (23%)	659 (31%)	2,102
B	Spam infrastructure infiltration & analysis	961 (**46%**)	634 (30%)	507 (24%)	2,102
C	Password-dialog spoofing	592 (**28%**)	624 (30%)	886 (42%)	2,102
D	Spoofed-warning deception	306 (**15%**)	735 (35%)	1,061 (50%)	2,102

(b) "If someone you cared about were a candidate participant for this experiment, would you want that person to be included as a participant?"

Table 2: For each of the five experiments we presented abstracts for, we asked respondents two questions to gauge their level of disapproval and concern. We boldface the percent who responded 'no' to these questions as this answer indicates disapproval or concern. Note that we exclude from these tables those respondents who reported being aware of Facebook's study.

5.3 Treatment effects

For each treatment scenario, we tally all respondents' possible answers to our question about whether each experiment should be allowed to proceed and present them in Table 2a. We present in in Table 2b the tallies for whether respondents

Hypothesis (observed difference between treatments is product of chance)	"proceed with this experiment?"					"included as a participant?"				
		$\chi^2(1)$		Mann Whitney U			$\chi^2(1)$		Mann Whitney U	
	Stat.	p	HB(P)	p	HB(P)	Stat.	p	HB(P)	p	HB(P)
(1) F0 vs. F1	2.107	0.14662	1.00000	0.29320	1.00000	0.526	0.46846	1.00000	0.31072	1.00000
(2) F0 vs. F2	0.104	0.74656	1.00000	0.93231	1.00000	0.477	0.48967	1.00000	0.77997	1.00000
(3) F0 vs. F3	0.000	0.99045	1.00000	0.94079	1.00000	0.000	1.00000	1.00000	0.66750	1.00000
(4) F0 vs. F4	0.040	0.84231	1.00000	0.42626	1.00000	0.183	0.66923	1.00000	0.63698	1.00000
(5) F0 vs. F5	2.833	0.09236	0.89750	0.42191	1.00000	1.788	0.18117	1.00000	0.22285	1.00000
(6) F0 vs. F6	2.879	0.08975	0.89750	0.04333	0.43330	3.467	0.06261	0.62610	0.05067	0.48990
(7) F0 vs. F7	6.188	0.01287	0.15444	0.00028	0.00336	10.606	0.00113	0.01356	0.00133	0.01596
(8) F0 vs. F8	1.179	0.27748	1.00000	0.06198	0.55782	3.744	0.05300	0.58300	0.04899	0.48990
(9) F0 vs. F9	3.670	0.05538	0.60918	0.01251	0.13761	2.694	0.10074	0.90666	0.03684	0.40524
(10) F1 vs. F2	0.915	0.33882	1.00000	0.29095	1.00000	2.253	0.13336	1.00000	0.19070	1.00000
(11) F6 vs. F7	0.398	0.52813	1.00000	0.09838	0.78704	1.588	0.20758	1.00000	0.22188	1.00000
(12) F8 vs. F9	0.453	0.50084	1.00000	0.56823	1.00000	0.045	0.83148	1.00000	0.88518	1.00000

Table 3: We designed our experiment to test for twelve potential differences between treatments. For each comparison between two treatment groups, we present tests of both the proportion of participants who expressed disapproval or concern (an answer of 'no' to each of our two questions, using the χ^2 test) and of the ordinal score calculated from all possible answers (using the use Mann Whitney U test). HB(p) represents p values adjusted for the 12 planned comparisons using the conservative Holm-Bonferroni method of multiple-testing correction. (We do not correct for using two to questions per pair of treatments, nor two tests per comparison, as these tests are used to cross-validate each other and show the same trends.)

would want someone they cared about to participate.

The treatment that led the lowest disapproval and concern compared to the control treatment was F7, in which we had modified the experimental scenario so that Facebook's researchers would add (rather than hide) posts—and only add positive posts. The nearly-as-low disapproval and concern for F6, in which both negative posts were added, provides additional evidence that adding posts from below the relevance threshold might have been less objectionable than removing posts that were deemed relevant. The difference between F1 and F2 provides some weak additional evidence supporting the notion that respondents became more uncomfortable when manipulations made participants' news feeds more negative.

In contrast, modifying the goals of Facebook's experiment had little effect on our respondents' disapproval and concern for participants. Neither removing the goal of scientific publication (F3) nor removing the goal of product improvements benefiting users (F4) had much effect. Attempting to disabuse the notion that the research would be used for advertising may backfire, as respondents shown a statement that the research would not be used for advertising actually had higher levels of concern and disapproval than the control (though the difference did not cross the threshold of significance).

A small proportion of those respondents shown the scenario in which the experiment was run by 'a social network' (F8) or 'Twitter' (F9) disapproved or were concerned for participants than of those respondents shown the control (in which the experiment was run by Facebook). The effects would not constitute significance after our conservative correction for multiple testing, but they may well have been had we tested these two groups combined against the control. If this result is indeed the product of a true difference, proponents' of Facebook's experiment might see it as evidence that Facebook's research receives unequal criticism. On the other hand, critics might claim the effect to be the result from concerns about the company's history, market position, or other factors they consider to be legitimate sources of respondent concern.

5.4 Prediction value

One way to evaluate the predictive value of ethical-response surveys is to compare results derived from survey respondents with the answers of participants who have experienced experiments firsthand.

As part of the design of the warning experiment that inspired Scenario D [2], the researchers (which included authors of this paper) directed some participants to a post-deception survey [7] immediately after debriefing. A total of 780 participants responded. All but 11 (769 total) opted to share their feedback with ethics researchers. All were offered the opportunity to withhold their data if they found the experiment sufficiently unethical. 750 consented to the use of their data by the experiment's researchers, 15 found the experiment objectionable but allowed researchers to still use their data, and four chose to withhold their data from final results (but allowed researchers to use it to verify that their published results would not have been different had the data been included). None chose to withhold their data entirely. In total, 19/769 (2%) registered objection to the experiment in response to the question about withholding their data.

A total of 764 participants in the warning experiment had responded to a question asking whether the experiment should proceed, which was similar to the question we presented in this paper but had different response options. In all, 11 (1%) of participants answered that the experiment should 'definitely not' proceed, 15 (2%) 'probably not', and 25 (3%)

	All treatments		Facebook control (F0)	
Question	$\chi^2(1)$	p-value	$\chi^2(1)$	p-value
Disapproval	232.56	< 0.00001	17.71	0.00005
Concern	214.34	< 0.00001	11.84	0.00058

Table 4: χ^2 comparisons of the proportion of 'No' responses to the disapproval ("proceed") and concern for participants ("included as a participant") questions between those who reported being aware of the Facebook mood study and those who were not. We present results across all treatments, and considering only the Facebook control treatment (F0). P-values were corrected using the Holm-Bonferroni method.

'prefer not to answer', 177 'probably proceed' (23%) and 536 (70%) 'definitely proceed'.

Given the relatively large number of participants who preferred not to answer or who didn't want to allow ethics researchers to use their responses, the range of participants who participated in the experiment and who believed the experiment should not have been allowed to proceed ranged from between 3% and 8%. Given that, the 7% of respondents in our survey who disapproved of Scenario D seems to be a reasonable estimate.

On the other hand, of the 3,539 respondents in our current survey, including those aware of Facebook's experiment, 135 (4%) reported that they had participated in the warning deception experiment (D). Only one of the 135 ($< 1\%$) reported that it should not proceed, as compared to 7% (243/3,404) of all other participants (summed from Table 1). Of the other 134 respondents, 7 (5%) were 'unsure', 30 (22%) answered 'Yes, with caution', and 97 (72%) answered 'Yes'. Similarly, only 10 of the 135 (7%) would prefer that someone they cared about not participate in the study, while 52 (39%) would be indifferent, and 73 (54%) would want the person they care about to participate. Thus, concern is roughly half that of those who had not participated. These figures suggest that, at least for some scenarios, our methodology may have been overly conservative. One possible explanation for the difference is that those who had participated in the study were aware of the use of consent; another is that being part of the experiment made participants more comfortable that it was performed in a beneficent manner that was respectful of participants.

6 Limitations

Our survey had a number of limitations that are important to consider when examining our results.

Our experiment was designed to gauge differences in how respondents felt about experimental manipulations, and excluded the question of consent by eliding discussion of it. We did this because we believed respondents would be unlikely to disapprove of studies that participants knowingly consented to. Our decision to elide the lack of consent from all scenarios facilitates *relative* comparisons between experimental scenarios (such as sending phishing messages, removing items from news feeds, spoofing operating systems dialogs) and among variants of the Facebook scenario. However, the consequence of this elision is that absolute levels of disapproval and concern that we report may underestimate the actual levels of concern that would be present had we explicitly stated that researchers did not obtain participant consent. (Had our scenarios mentioned that researchers did not obtain consent, we would have also raised the expectation of consent among those who might otherwise not expected it.)

The process of compressing an experimental scenario into a short description introduces a number of other possible sources of error. In crafting these descriptions, we may have failed to anticipate which facts would be influential in respondents ethical decision making. We may have incorrectly interpreted information about an experimental design. As two authors were researchers on two of the studies described, we may have been subject to subconscious biases (or, a skeptical reader may reasonably suspect, conscious ones).

In order to reach a large number of respondents in a very short time, our survey relied on a convenience sample: workers on Amazon's Mechanical Turk crowdsourcing service. These individuals tend to be more tech savvy than the rest of the population. They also likely find themselves participating in far more research experiments, and interacting with researchers, than members of the general population. Some may be reliant on research studies for income and more forgiving of transgressions so long as they are paid. While these workers are an excellent group to reach out to in order to gauge the response research studies in which participants will be workers on Amazon's Mechanical Turk (e.g., studies C and D), the demographic differences are more problematic for examining research in which participants will be drawn from other populations.

Even if survey respondents closely resemble those who would be participants in research scenarios, there's no way to be certain that their responses to hypothetical questions about an experimental design will match how they would feel if they were to actually participate. While the respondents of the survey had reasonable resemblance to the self-reported responses of experimental participants for Scenario D, this is no way guarantees surveys will be predictive for others studies.

Finally, respondents were not required to have any prior background in ethics, ethics training, or knowledge of laws and regulations that govern research ethics (e.g., the common rule); nor did we provide them with any such background

or training. This was by design. Ethical controversies can occur when there is a disconnect between what cutting-edge research can be approved within the existing regulatory regime and what the public considers acceptable. Further, the rules give ethics boards considerable discretion to determine whether waiving rules such as participant consent is in the public interest.

7 Discussion

Much of the debate over Facebook's emotional contagion experiment has focused on rules and process. Following the controversy, the lead author of Facebook's experiment promised that the company's ethical-compliance process for research had evolved and improved [11]. Some industrial research labs, including Microsoft Research (which employs one of the authors of this paper) already have established experimental review boards (on which the author serves).

Regardless of what processes evolve to govern the set of individuals who must decide whether research is approved or rejected, those tasked with making the decisions will have tough choices. Most of the rules that govern research, such as the requirement for participant consent, give review boards considerable discretion. Ethics boards often have very little data with which to understand the implications of exercising their discretion.

Thus, despite the limitations of ethical-response surveys, we would have to imagine that those making decisions – whether for industrial research or research at universities – would prefer to have data with known limitations to no data at all. The costs of running such studies can be amortized over a number of experimental designs. It is our hope that ethical-response surveys become a standard tool for use when researchers propose new types of experiments about which the the reaction of participants and the public is unknown.

References

[1] Albergotti, R. Facebook Experiments Had Few Limits. `http://online.wsj.com/articles/facebook-experiments-had-few-limits-1404344378`, July 2014.

[2] *anonymized*. *anonymized*. In anonymized, *anonymized*, *anonymized* (*anonymized*, July 2013).

[3] *anonymized*. *anonymized*. anonymized (*anonymized* 2013).

[4] *anonymized*. *anonymized*. In anonymized (*anonymized* 2012), *anonymized*.

[5] Facebook emotion study examined by privacy commissioner. `http://www.cbc.ca/news/business/facebook-emotion-study-examined-by-privacy-commissioner-1.2695145`, July 2014. Retrieved on July/08/2014.

[6] Corbett, J. New questions, few answers in Cornell's Facebook experiment. `http://ithacavoice.com/2014/07/new-questions-answers-cornells-facebook-experiment/`, July 3, 2014. Retrieved on July/08/2014.

[7] The ethical research project. `https://www.ethicalresearch.org/`.

[8] Fung, B. The journal that published Facebook's psychological study is raising a red flag about it. `http://www.washingtonpost.com/blogs/the-switch/wp/2014/07/03/the-journal-that-published-facebooks-psychological-study-is-raising-a-red-flag-about-it/`, July 3, 2014. Retrieved on July 08, 2014.

[9] Jagatic, T. N., Johnson, N. a., Jakobsson, M., and Menczer, F. Social phishing. *Communications of the ACM 50*, 10 (2007), 94–100.

[10] Kanich, C., Kreibich, C., Levchenko, K., Enright, B., Voelker, G. M., Paxson, V., and Savage, S. Spamalytics: an empirical analysis of spam marketing conversion. *ACM Conference on Computer and Communications Security* (2008), 3–14.

[11] Kramer, A. Facebook wall post of June 29 at 1:05pm Pacific. `https://www.facebook.com/akramer/posts/10152987150867796`, June 2014. Retrieved on July 08, 2014.

[12] Kramer, A., Guillory, J., and Hancock, J. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Science 111*, 24 (2014), 8788–8790.

Full text of scenarios

A Social phishing

Phishing is an attack in which users are sent emails with a link to a fraudulent website in order to trick them into divulging their passwords. For example, some phishing emails appear to come from a user's bank and contain a link to a website that also appears to be the user's bank, but is actually controlled by the attacker. When the user types the password into the fake site, the attacker takes the password and can now login to the user's account.
University researchers want to quantify how much the success of a phishing attack would increase if the email its targets received appeared to come from someone the target user trusted—-a friend:

- The researchers will send phishing emails to students with a link to a website that impersonates one of the university's websites.
- The researchers will send half of the students an email that appears to be from one of the student's friends, who the researchers will identify by examining the student's Facebook profile. The researchers will send the other half of students an email that appears to be sent by someone the student does not know.
- If students enter passwords into the researchers' site, the researchers will, with the permission of the university, use the university's systems to verify that the passwords entered were valid passwords.
- Afterwards, the researchers will notify students that this was a research study. They will inform offer students the opportunity to ask to have their data excluded from the study and to comment about the study on a blog.
- The researchers will publish the anonymized aggregate results of the experiment in a scientific paper.
- Participants will not be identified and will remain anonymous.

If the researchers are not allowed to perform this experiment, they will not be able to measure how often users fall victim to phishing attacks. Therefore, the researchers will not be able to publish recommendations to help users better learn to recognize such attacks.

B Spam infrastructure infiltration & analysis

Computer security researchers, seeking to understand the economic infrastructure that enables email spam, want to measure the rate at which spam emails result in purchases.
Conducting such research is challenging. Researchers would not want to send spam. Spammers are unlikely to divulge how successful their emails are in attracting purchases.

- The researchers will allow one of their computers to become infected with software that is controlled by spammers, while the researchers maintain sufficient control of the computer to monitor how attackers are using it.
- The researchers will alter the commands that the spammers send to the researchers' infected computer, replacing the link to the spammer's store with a link to a website run by the researchers that mimics the appearance of the spammer's store.
- Without collecting payments or other personal information about those users who respond to the spam email seeking to make a purchase from the spammers, the researchers record the number of attempts made to purchase products from the store advertised by the spam.
- The researchers will not inform users who receive the spam sent by attackers using the infected computer as this might cause users to behave differently or otherwise compromise the validity of the results.
- The researchers will not inform users who visit the store to make a purchase that the store has been disabled or that their choice to make a purchase is being recorded.
- The researchers will publish the anonymized aggregate results of the experiment in a scientific paper.
- Participants will not be identified and will remain anonymous.

If the researchers are not allowed to perform this experiment, they will not be able to empirically measure the effectiveness of spam emails and may not be able to produce or publish well-informed recommendations for technical or policy approaches to stopping spam.

C Password-dialog spoofing

Computer security researchers want to learn the fraction of Internet users who fall for the tricks used by hackers to steal users' passwords. Conducting such research is challenging because if research participants know the attack is coming, or even that the study is about computer security, they may be less likely to fall for the tricks. The researchers thus plan to deceive participants as to the purpose of the human intelligence task (HIT) they will be asked to complete:

- During the task the researchers will replicate the techniques that hackers use to trick users into typing their passwords.
- Unlike criminal hackers, the researchers will not actually steal, collect, or store the passwords that users type.
- Afterwards, the researchers will present a detailed explanation of the deception to participants, reveal the true purpose of the study, and reassure participants that no passwords were actually stolen during the study.
- The researchers will publish the anonymized aggregate results of the experiment in a scientific paper.
- Participants will not be identified and will remain anonymous.

If the researchers are not allowed to perform this experiment, they will not be able to measure how often users fall victim to attacks that target users' passwords. Therefore, the researchers will not be able to produce or publish recommendations that help users better learn to recognize such attacks.

D Spoofed-warning deception

Computer security researchers want to measure different techniques for presenting security warnings.
One challenge in studying security decision making is that if participants are made aware that researchers are studying their security behavior, or become aware of it, they are likely to behave differently than they normally would. The researchers thus plan to deceive participants as to the purpose of the human intelligence task (HIT) they will be asked to complete:

- The researchers will give participants a task unrelated to security, but that will cause participants to encounter a security warning.
- While the warning will create the illusion that the participant is facing a security risk, the researchers will not actually expose participants to any real security risks.
- The researchers will measure how different ways of presenting a warning may make that warning more or less effective in convincing users to avoid a risk.
- At the conclusion of the experiment, the researchers will present a detailed explanation of the deception to participants, reveal the true purpose of the study, and reassure participants that they were never at any real risk.
- The researchers will publish the anonymized aggregate results of the experiment in a scientific paper.
- Participants will not be identified and will remain anonymous.

If the researchers are not allowed to perform this experiment, they will not be able to measure the effectiveness of different designs for computer security warnings. Therefore, the researchers will not be able to produce or publish recommendations to improve the effectiveness of future security warnings.