Anticipating Experimental Risks Using Surveys

Stuart Schechter
Microsoft Research
stuart.schechter@microsoft.com

Cristian Bravo-Lillo
Universidad de Chile
cristian.bravo@gmail.com

We introduce a survey instrument for anticipating otherwise-unforeseen risks resulting from research experiments. We present experiments hypothetically, then ask: "If someone you cared about were a candidate participant for this experiment, would you want that person to be included as a participant?" (Q1) and "Do you believe the researchers should be allowed to proceed with this experiment?" (Q2). Having honed this approach over multiple studies, and multiple years, we have aborted proposed studies due to survey respondents' concerns. In this paper, we test this instrument by presenting five past (real) experiments, posed as hypotheticals, to 3,539 workers on Amazon's Mechanical Turk. These experiments include Indiana University's social phishing study, University of California's 'spamalytics' study, and Facebook's emotional contagion experiment. We reveal what researchers behind controversial experiments might have foreseen had our instrument been available to them prior to conducting their experiments.

INTRODUCTION

In evaluating the ethicality of a proposed experiment, researchers and ethics boards must weigh the benefits of the study against potential risks. Alas, there is a great deal of guesswork in anticipating risks. Researchers and their ethics boards may fail to foresee harm to participants or may misjudge public reaction. Such failures may occur even when prior research provides precedent the experimental design. Authors of prior research are unlikely to have used the scarce space allotted for their paper to inform future researchers of ethical results [17] such as participants' reactions, feelings, concerns, and opinions regarding the ethics of the experiment in which they took part; few ethics boards or publication venues would expect researchers to collect and report such information.

In 2012, we began a concerted effort to apply the same level of scientific inquiry to the ethics of our experiments as to our primary research agenda. When running online experiments that involved deception, we followed our debriefings with questions to elicit participants' concerns.

We also created and began using an *ethical-response survey instrument* to identify unforeseen risks, concerns, and other reactions to experiments *before* exposing participants to these experiments. In these surveys, we would present respondents[1] with a series of short descriptions of experimental scenarios and would ask the same questions for each one. We wrote these summaries to be accessible to a general audience, sufficiently short so as not to lose the reader's attention (each no more than 350 words), yet to pack in the information salient to the ethical challenges unique to each experiment.

Our first such survey caused us to abandon an experiment we had received approval to conduct and completed all preparations for. We had planned to email victims of password breaches to ask questions about the appropriateness of security researchers' use of their breached data. We had assumed that email recipients would be forgiving of unsolicited communications that were intended to protect the interests of recipients, that had no commercial purpose, that benefited science, and that were used sparingly—surveying hundreds to protect the interests of millions. However, when we used an early version of our instrument to survey workers on Mechanical Turk, a greater fraction were concerned with the email study of breach victims (sending unsolicited email to a small fraction of victims) than with researchers' use of all victims' publicly-released breach data (which required no contact with victims). In addition to the statistical differences, participants' strongly-worded responses forced us to reconsider the assumptions under which we had justified our use of unsolicited email. The risks of emailing breach victims now appeared to outweigh the benefits. We began publicly advocating for the prophylactic use of ethical-response surveys in 2013 [3].

In this paper, we further demonstrate the value of ethical-response surveys by using this instrument to examine controversial experiments from the past decade, so as to discover what the researchers who conducted these experiments might have learned if given the opportunity to use the survey instrument beforehand.

One of our experimental scenarios was written to describe an experiment in which researchers sent phishing messages to students who had not signed up to be part of a study or given their consent. In another scenario, researchers knowingly allowed spammers to compromise their computers, then used the spammers' in-

[1]To avoid confusion, we refer to those who complete our survey as *respondents* to prevent the reader from confusing them with participants in the experimental scenarios that we described to respondents.

This paper is a submission draft currently under anonymous review.
This draft was generated on October 30, 2014.

frastructure to measure spam response rates. A third scenario, which we rushed to add to our survey in June 2014, describes Facebook's controversial emotional contagion experiment [14], which was published earlier that month. The final two scenarios described deception studies, led by members of our team, which were conducted with ethical-response questions integrated into a post-debriefing survey—providing participants' perspective into the experiments they had just completed.

We performed our ethical-response survey from July 2–4 2014 on 3,539 US-based workers on Amazon's Mechanical Turk—a convenience sample not necessarily representative of the participant pool for the experiments described in the survey, or of the general public, but mirroring the pool of research participants available to researchers who require short turn-around in advance of an IRB application or other approval process.

We present respondents' insights into the ethics of these studies and differences in their overall evaluating of which studies should have been allowed to proceed.

EXPERIMENTAL PROCEDURE

We offered our survey as a $1.00 Human Intelligence Task (HIT) on Amazon's Mechanical Turk crowdsourcing service, requiring that workers come from the US.

After brief instructions, we presented participants with five experimental scenarios, randomizing the order in which they were presented to participants. We asked participants to "please read the description of each experiment carefully." We presented each scenario on its own page, followed by four multiple-choice questions about the scenario and optional free-response text fields for explanations. While we presented four of the five scenarios verbatim to all respondents, we created ten variants of the Facebook experiment and randomly assigned each participant one variant. After participants responded to all five scenarios, we concluded with follow-up questions—mostly demographic.

Questions for each scenario

We designed the first question that followed the description of each scenario (Q1) 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 say that participants expressed *concern* for participants iff they answered "No".

We always asked about concern for participants first to give respondents a chance to humanize potential participants, and to think about the consequences of the experiment on them, before we asked Q2: "Do you believe the researchers should be allowed to proceed with this experiment?"

We offered four response options for Q2, again ordered from most approving to least approving with the first option being "Yes" (on the left) and the last "No" (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 absolve researchers of all other ethical responsibilities. The option in the third position from the left, between "Yes, but with caution" 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. We say that respondents exhibited *disapproval* of an experiment iff they responded "No".

For each of these first two questions, we gave respondents a free-response field in which they could optionally 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 response options, from left to right, were "Yes" and "No".

Closing questions

After respondents completed the five experimental scenarios, we asked for their year of birth, gender, occupation (free response), whether they had ever purchased goods advertised via unsolicited email (for insight into the spam experiment), whether they had participated in a study involving deception, and, lastly, whether they had heard about "Facebook's 'mood' study".

Payment

We paid all respondents $1.00 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 $1.00 paid for the HIT, we paid a bonus equal to the difference.[2] We paid bonuses after all surveys were complete—had we paid immediately and word spread, some respondents might have delayed completion to increase their bonus.

EXPERIMENTAL SCENARIOS

We presented the following scenarios as hypotheticals, not as historical case studies.

[2]For this we received effusive thank-you emails, reviews on Turkopticon [19] (a workers' forum), and the practice was cited in a recommendation for treating workers fairly [15].

Social-phishing emails

This experimental scenario was based on the "social phishing" experiment performed by researchers at Indiana University [11].

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.

Spammer infiltration

The second experimental scenario describes an experiment to measure the economics of spam performed by researchers at the University of California [13].

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.

Password-dialog spoofing

This scenario describes an experiment led by one of the members of our team that appeared at *anonymized* [4].

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.

During the real experiment, the researchers presented participants with a consent form explaining that this was a university experiment, though deceiving participants by eliding that its goal was to study participants' security behavior. The researchers debriefed participants at the conclusion of the experiment, explaining the nature and necessity of the deception. This was the first study in which members of our team, who led the experiment, began inserting ethical-response questions at the end of the debriefing process. The initial questions were quite

rudimentary, simply inviting participants the opportunity to volunteer concerns via free-response fields.

Spoofed-warning deception

This experiment, also led by a member of our team, was similar to the one above but did not trick users' into typing their passwords. It was performed after the other study and the post-deception ethics questions were refined. We included a question about whether the study should have been allowed to proceed similar to the one we now use in our ethical-response survey.

> 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.

Emotional contagion

This final scenario describes Facebook's emotional contagion experiment.

> 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.

As the scenario focuses on facts about the experimental goals and methodology, it does not make salient numerous issues that were subject of public debate, such as the reliance on (and accuracy of) Facebook's terms of service, ethical oversight, 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.

Variants

Many respondents did not receive the emotional contagion scenario exactly as it appears above (our control), but instead received one of nine of the variants described below. We assigned respondents to either the control or one of the scenario variants (treatments) uniformly at random (with a 10% chance of being assigned to each variant). We created some of these variants to explore different ways to run the experiment in order to optimize the trade-off between risks and benefits—a step we expect some researchers might want to take when performing a prophylactic ethical-response survey. We also included other variants to test hypotheses outside the scope of this paper (e.g., would the study have been less controversial if Twitter had run it?)

Only remove + posts: We elided references to the researchers increasing the proportion of *positive* posts by hiding *negative* posts.

Only remove − posts: We elided references to the researchers increasing the proportion of *negative* posts by hiding *positive* posts.

No publication: We elided references to the researchers publishing the results in a scientific paper.

No prod. improvement: We elided that, if unable to run the experiment, "the researchers will not be able to produce features that might protect the moods of psychologically-vulnerable users."

Promise no advertising: We added a bullet stating that "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."

Insert posts (+ & −): We changed the experiment so that the researchers would adjust the ratio of positive to negative posts by adding posts that otherwise would not have been deemed worthy of display on the news feed (not removing posts).

Insert posts (+ only): We replicated the above treatment, but with the researchers adding only positive posts.

Not Facebook: To make the company performing the research ambiguous, we replaced "Facebook" with "a social network".

'Facebook'→'Twitter': We replaced "Facebook" with "Twitter".

LIMITATIONS

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

While we designed our survey instrument to anticipate the risks and concerns of future experimental participants, all the experiments we examined in this use of the instrument took place in the past.

As with any compression process, some fidelity will inevitably be lost when complex experiments are simplified for presentation. In describing experiments, we may have failed to anticipate which facts would be most salient to respondents. We may also 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.[3]

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 more frequently, 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 are drawn from workers on Amazon's Mechanical Turk (e.g., the spoofed-warning deception study), 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 guarantee that respondents will correctly anticipate how they would feel about the experiment were they to be a participant.

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 ethics boards will approve and what members of the public consider acceptable.

RESULTS

After piloting on July 1, we survey 3,539 respondents during a three-day period starting 12:00AM EDT Wednesday July 2, 2014. An additional 31 respondents, who we did not count toward that total or include in our results, responded to our survey but spent spent less than 150 seconds (30 seconds per scenario) completing it.

Of our 3,539 respondents, 1,745 (49%) reported being female, with 27 (1%) declining to answer; 1,127 (32%) reported having participated in a deception study, with 91 (3%) declining to answer; and 120 (3%) reported having purchased goods advertised via an unsolicited marketing email, with another 20 (1%) declining to answer.

As Facebook's emotional contagion experiment was receiving extensive press and social-media coverage debate during our survey, we anticipated that this might impact respondents' perceptions of our hypothetically-posed description of the experiment, as well as its variants. We asked participants if they were already aware of Facebook's 'mood' study, requiring a "yes" or "no" answer; 1,437 respondents (41%) answered "yes", with a greater fraction providing this affirmative answer as the three-day survey period progressed.

The survey's predictive 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 we are principally concerned with avoiding concerns about participants and disapproval of the experiment, we focus our analysis on respondents who expressed concern for participants (Q1:"No") or disapproval of the experiment (Q2:"No")—see Table 1.

Participants vs. non-participants

For each experiment, we asked respondents whether they believed they had been part of an experiment like the one described in the scenario. Respondents would be unlikely to know if they had received spam as part of spammer infiltration experiment, and so only 11 of our 3,539 participants, less than a third of a percent, reported believing they had been a participant. Similarly, only 12 reported having been one of the unwitting participants in the social phishing experiment, two of whom also claimed to have been part of the spammer infiltration experiment (though neither claimed to have participated in all five experiments). While Facebook did not disclose the list of unwitting participants in its emotional contagion experiment, we find it more reasonable for respondents to consider themselves participants, as 55 (1.5%) did.

As the password-dialog spoofing and spoofed-warning deception experiments were both run on Amazon's Mechanical Turk, the same platform on which we ran our survey, it was quite likely that some of our respondents had been part of that experiment. Indeed, when asked to describe why they would believe they had been part of these experiments, some even recalled details such as the institution that had performed the study.

A total of 52 of our respondents (1.4%) reported that they had participated in the password-dialog spoofing study and 135 (4%) reported having participated in the

[3]A skeptical reader may even reasonably suspect conscious bias. This is one of the reasons why we presented full text of each scenario in this paper.

experimental scenario	*response*	*self-reported participants*		*self-reported* ***non****-participants*		*two-tailed test* $\chi^2(1)$	p
Password-dialog spoofing	concern for participants	9/52	(17%)	987/3,487	(28%)	2.544	0.111
	disapproval of experiment	2/52	(4%)	537/3,487	(15%)	4.441	0.035
Spoofed-warning deception	concern for participants	10/135	(7%)	499/3,404	(28%)	4.972	0.026
	disapproval of experiment	1/135	(1%)	243/3,404	(7%)	7.313	0.007

Table 1: The proportion of respondents who expressed concern for participants (Q1) or disapproval for the experiment (Q2), separated by whether respondents reported having been a participant in the experiment. We include only the two experiments performed on Mechanical Turk – the same platform we used for our survey – as participants would be very unlikely to have participated (and even less likely to know they participated) in the other experiments.

larger, and more recent, spoofed-warning deception experiment. In Table 1 we compare the responses of those who reported being part of these experiments to those who did not, focusing on the proportion of participants who exhibited concern for participants or disapproval of the experiment.

Among the 52 self-reported participants of the password-dialog spoofing study, 9 (17%) expressed concern for participants. This proportion was smaller than the 28% of self-reported non-participants who expressed concern. (19 of the self-reported participants expressed indifference and 24 would want the person they cared about to be included as a participant.)

Only two of the 52 respondents who reported having participated in the password-dialog spoofing experiment (4%) exhibited disapproval of the experiment, answering that it should not proceed. This proportion is less than a third of the 537/3,487 (15%) self-reported participants who expressed disapproval. (3 of the self-reported participants were unsure whether it should proceed, 20 answered "Yes, but with caution", and 27 answered an unconditional "Yes".)

Of the 135 respondents who reported having participated in the spoofed-warning deception, 10 (7%) expressed concern for participants. This proportion is less than half of the 499/3,404 (15%) of self-reported non-participants who did so. Comparing these two proportions with a two-tailed χ^2 test yields $\chi^2(1) = 4.972$, $p = .0258$ without correction for multiple testing. (52 of the self-reported participants expressed indifference and 73 would want the person they care about to participate.)

Only one of the 135 self-reported participants in the spoofed-warning experiment (0.7%) expressed disapproval of the experiment, as compared to 243/3,404 (7%) of self-reported non-participants who said the experiment should not proceed. Comparing these two proportions with a two-tailed χ^2 test yields $\chi^2(1) = 7.313$, $p = .0068$ without correction for multiple testing. (7 of the self-reported participants were unsure, 30 answered "Yes, but with caution", and 97 answered and unconditional "Yes".)

One possible explanation for the consistently lower concern and disapproval among respondents who reported being past participants is that they were aware that the experiments were performed with consent, contained extensive debriefings, and, in some cases, included ethical follow-up questions—none of which were detailed in the scenario presented to respondents. If this hypothesis is correct, our ethical-response survey may have been overly conservative, at least for these two experiments. A less optimistic hypothesis is that the participants who had disapproved of deception (and may have found other requester behaviors more objectionable than other workers did) would have been more likely to abandon Mechanical Turk before our ethical-response survey was conducted; the two deception experiments took place in 2012 and 2013, whereas our ethical-response survey was conducted in July 2014.

Comparison to post-experiment responses

As part of the design of the later iterations of the spoofed-warning deception experiment [2], we and our collaborators directed some participants to a post-deception debriefing and survey hosted by The Ethical Research Project [7]. 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 spoofed-warning deception experiment responded to a question asking whether the experiment should proceed, which was similar to the question we now use in our surveys, but which had different response options. In all, 11 (1%) of participants answered that the experiment should "definitely not" proceed, 15 (2%) "probably not", 25 (3%) "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% to 8%. Given that, the 7% of respondents in our survey who disapproved of this experiment in the survey may indeed be a reasonable estimate.

COMPARING EXPERIMENTS

We present a side-by-side comparison of all five experimental scenarios in Table 2, examining both concern for participants (Table 2a) and disapproval of the experiment (Table 2b).

As Facebook's emotional contagion experiment received extensive press and social media attention, we present alongside our aggregate results those for the 2,102 respondents who reported being unaware of that study. Indeed, combining all variants of the emotional contagion experiment, there were large and statistically significant differences between those respondents who had and had not already heard about the experiment, both in terms of concern for participants ($\chi^2(1) = 11.84, p = 0.00058$) and disapproval of the experiment ($\chi^2(1) = 17.71, p = 0.00005$). Those who reported being aware of the experiment were more concerned for participants and more disapproving of the experimental scenario based on it.

A tempting explanation for these highly-significant differences is that respondents' opinions were strongly swayed by opinions of the media. It's also possible that respondents assumed that certain facts about the real experiment, which we had not included in our hypothetical descriptions, applied to the hypothetical experiment. For example, those who were aware of the study had learned explicitly that the researchers did not receive consent from participants, whereas the hypothetical scenario did not indicate whether consent had been obtained or not.

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 hear about it from friends or see coverage of it in the media. Indeed, a smaller proportion of respondents who reported being unaware of the emotional contagion experiment expressed disapproval and concern with our two more-controversial university experiments (see Table 2). However, the difference was much smaller for the experiments than the emotional contagion experiment, suggesting that this hypothesis would not explain the entire difference.

While Facebook's emotional contagion experiment has likely received the most media coverage of any recent controversial experiment, the spammer infiltration and social-phishing scenarios received a greater level of disapproval from our respondents. Examining all respondents, the pairwise differences were not significant. However, the differences were larger for those who reported being unaware of the social contagion experiment. Whereas 72 (35%) of the 207 respondents given the emotional contagion (control) scenario answer that it should not be allowed to proceed, 950 of 2,102 respondents (45%) felt that the the social-phishing email experiment should not proceed ($\chi^2(1) = 8.226$, $p = 0.004$).

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, 9]. Without stepping into the debate of whether or when an institutional review board is *necessary*, we believe our results indicate that such review is in no way guaranteed to be *sufficient*—both the similarly-controversial social-phishing email and spammer-infiltration experiments received university approvals.

The two experimental scenarios that registered the least objection and concern in Table 2 were those based on experiments that we conducted. To make the conflict-of-interest clear, we wrote the scenario descriptions that describe and attempt to justify the ethical basis our own experiments, whereas the researchers behind the other experiments did not have the opportunity to do so.

ANTICIPATING RISKS AND CONCERNS

We have presented ethical-response surveys as a means to anticipate risks and concerns that researchers may not considered the existence of, or may have failed to appreciate the importance of. For the purpose of anticipating experimental risks, we examine the optional explanations provided by respondents who expressed concern for participants (Q1) or disapproval of experiments (Q2).

Social-phishing emails

In a retrospective, Indiana University's IRB chair and the social phishing experiment's principal investigator presented the difficulty position that ethics boards find themselves in: "the ethical issues relating to waiving aspects of informed consent are controversial and there is little consensus among IRB members and ethicists." [8] Indeed, opinions on such broad questions as whether deception experiments should ever be allowed vary across disciplines [10]. In approving the social-phishing email experiment, the IRB concluded "the experiment would cause minimal to no real harm to the participants" [12]. Yet after the experiment was performed, with 921 recipients and 810 spoofed senders, the debriefing blog received 440 comments, mostly negative, before the researchers disabled the commenting due to an "overwhelming portion of non-constructive posts" [8].

The responses to our hypothetically-posed scenario overlap with many of the reactions that the authors of the social phishing study received from the comments following their study. For example, *R3168* wrote "I understand the necessity of studies like this, but I would feel taken advantage of and victimized." *R2574* feared "this would cause unnecessary stress" and *R2823* felt that being phished "would be traumatizing, experiment or not."

experimental scenario	*all respondents to our survey*			*those who reported being unaware of the emotional contagion experiment*		
	No	Indifferent	Yes	**No**	Indifferent	Yes
Spoofed-warning deception	509 (**14%**)	1,271 (36%)	1,759 (50%)	306 (**15%**)	735 (35%)	1,061 (50%)
Password-dialog spoofing	996 (**28%**)	1,091 (31%)	1,452 (41%)	592 (**28%**)	624 (30%)	886 (42%)
Social-phishing emails	1,673 (**47%**)	804 (23%)	1,062 (30%)	950 (**45%**)	493 (23%)	659 (31%)
Spammer infiltration	1,703 (**48%**)	1,039 (29%)	797 (23%)	961 (**46%**)	634 (30%)	507 (24%)
Emotional contagion	150 (**43%**)	108 (31%)	94 (27%)	72 (**35%**)	74 (36%)	61 (29%)
Emotional contagion variants: Only remove + posts	171 (**48%**)	116 (32%)	70 (20%)	85 (**39%**)	79 (36%)	56 (25%)
Emotional contagion variants: Only remove − posts	147 (**43%**)	115 (34%)	78 (23%)	57 (**31%**)	76 (41%)	51 (28%)
Emotional contagion variants: No publication	172 (**47%**)	99 (27%)	94 (26%)	77 (**35%**)	70 (32%)	73 (33%)
Emotional contagion variants: No prod. improvement	154 (**44%**)	119 (34%)	79 (22%)	79 (**37%**)	73 (34%)	60 (28%)
Emotional contagion variants: Promise no advertising	181 (**51%**)	98 (27%)	78 (22%)	91 (**42%**)	69 (32%)	59 (27%)
Emotional contagion variants: Insert posts (+ & −)	129 (**37%**)	118 (34%)	99 (29%)	55 (**26%**)	81 (38%)	76 (36%)
Emotional contagion variants: Insert posts (+ only)	116 (**31%**)	144 (39%)	112 (30%)	47 (**20%**)	92 (40%)	91 (40%)
Emotional contagion variants: Not Facebook	133 (**38%**)	111 (32%)	104 (30%)	49 (**25%**)	75 (39%)	69 (36%)
Emotional contagion variants: 'Facebook'→'Twitter'	126 (**36%**)	119 (34%)	105 (30%)	55 (**27%**)	72 (35%)	78 (38%)

(a) **Concern for participants.** Q1: "If someone you cared about were a candidate participant for this experiment, would you want that person to be included as a participant?"

experimental scenario	*all respondents to our survey*				*those who reported being unaware of the emotional contagion experiment*			
	No	I'm not sure	Yes, but with caution	Yes	**No**	I'm not sure	Yes, but with caution	Yes
Spoofed-warning deception	244 (**7%**)	212 (6%)	1,044 (29%)	2,039 (58%)	138 (**7%**)	132 (6%)	644 (31%)	1,188 (57%)
Password-dialog spoofing	539 (**15%**)	307 (9%)	1,395 (39%)	1,298 (37%)	326 (**16%**)	169 (8%)	848 (40%)	759 (36%)
Social-phishing emails	603 (**31%**)	407 (12%)	1,186 (34%)	839 (24%)	603 (**29%**)	240 (11%)	721 (34%)	538 (26%)
Spammer infiltration	956 (**27%**)	550 (16%)	1,189 (34%)	844 (24%)	518 (**25%**)	316 (15%)	739 (35%)	529 (25%)
Emotional contagion	117 (**33%**)	43 (12%)	85 (24%)	107 (30%)	50 (**24%**)	22 (11%)	62 (30%)	73 (35%)
Emotional contagion variants: Only remove + posts	143 (**40%**)	41 (11%)	79 (22%)	94 (26%)	68 (**31%**)	27 (12%)	47 (21%)	78 (35%)
Emotional contagion variants: Only remove − posts	119 (**35%**)	42 (12%)	79 (23%)	100 (29%)	48 (**26%**)	19 (10%)	47 (26%)	70 (38%)
Emotional contagion variants: No publication	140 (**38%**)	32 (9%)	90 (25%)	103 (28%)	52 (**24%**)	24 (11%)	66 (30%)	78 (35%)
Emotional contagion variants: No prod. improvement	117 (**33%**)	47 (13%)	100 (28%)	88 (25%)	54 (**25%**)	29 (14%)	61 (29%)	68 (32%)
Emotional contagion variants: Promise no advertising	143 (**40%**)	41 (11%)	65 (18%)	108 (30%)	70 (**32%**)	24 (11%)	42 (19%)	83 (38%)
Emotional contagion variants: Insert posts (+ & −)	96 (**28%**)	43 (12%)	93 (27%)	114 (33%)	36 (**17%**)	25 (12%)	57 (27%)	94 (44%)
Emotional contagion variants: Insert posts (+ only)	83 (**22%**)	51 (14%)	87 (23%)	151 (41%)	33 (**14%**)	28 (12%)	45 (20%)	124 (54%)
Emotional contagion variants: Not Facebook	106 (**30%**)	35 (10%)	76 (22%)	131 (38%)	37 (**19%**)	20 (10%)	49 (25%)	87 (45%)
Emotional contagion variants: 'Facebook'→'Twitter'	99 (**28%**)	39 (11%)	84 (24%)	128 (37%)	33 (**16%**)	24 (12%)	51 (25%)	97 (47%)

(b) **Disapproval of the experiment.** Q2: "Do you believe the researchers should be allowed to proceed with this experiment?"

Table 2: For each of the five experiments we described to respondents, we asked two questions to gauge their concern for participants and disapproval of the experiment. As experiments are written from the perspective of researchers, we asked these questions in the order shown so as to give respondents the opportunity to think about the perspective of participants before considering whether the experiment should be allowed to proceed. We use boldface to highlight the percent who responded "no" to these questions as this answer indicates concern for participants in Q1 (a) and disapproval of the experiment in Q2 (b).

Respondents were concerned that this stress and sense of exploitation might cause distrust in the university. *R1775* was concerned the experiment would "undermine their trust in the university" and *R837* wrote "Using the University as a foil for the attack could reduce the participants trust of the organization itself ." Has they been available to the researchers and IRB, such responses might have suggested the potential for a backlash.

The blog comments posted by participants in the actual experiment revealed that "a large number of subjects... believed that either they or their friends had been affected by malware" [8]. Concerns about such side effects were reported in our survey, suggest that these concerns might have been foreseeable. For example, *R2183* wrote: "If I was sent a phishing email from a friend, I would immediately contact them after I discovered it. My friend would probably worry/delete their email account."

Even in their retrospective, the IRB chair and principal investigator took the position that participants were "exhibiting a lack of appreciation for the fact that personal data that is put on publicly accessible forums no longer is private" when they complained about the researchers mining their relationships on Facebook. Regardless of whether the data was public, many of our survey respondents explained that researchers should not be crawling this public information for the purposes not intended by those who made it public, even if the information was ac-

cessible. For example, *R1857* referred to the harvesting of information as "creeping" Facebook pages.

Finally, while the researchers promised that they had handled passwords carefully, those in our survey reported being uneasy that they could never be sure the passwords hadn't been misused or that the passwords had been sufficiently protected from hackers. *R3002*: "I would be concerned that the participant's personal information gathered by the researchers could end up in the wrong hands. For example, this could happen if a third party found out about the study and hacked the researchers' data."

Emotional contagion

Out of the 352 respondents who received the unmodified (control) description of Facebook's emotional contagion experiment, 207 reported being previously unaware of the study (respondents denoted '-U') and 145 reported being aware of it ('-A'). Of those, 48 unaware and 53 aware participants provided explanations for either their concern for participants or disapproval of the experiment.

Regardless of the actual risk of harm to the emotionally-vulnerable or mentally ill, 8 of the 207 participants not previously aware of the emotional contagion experiment (3.9%) reported concern for participants from these populations. For example, *R1892-U* was concerned for participants who were "depressed and unstable" and *R1893-U* explained that "people's emotional states can be very fragile".

Facebook's researchers may have assumed that, since their algorithms already filter posts for relevance on behalf of users, a small amount of additional filtering would be effectively harmless. However, respondents were concerned about filtering unambiguously-relevant posts would cause participants to miss out on important information and lead to misunderstandings. *R3393-U* observed that "this kind of research could cause family and friend problems due to lack or misrepresentation of information." *R1373-U* wrote "I wouldn't want [users] to miss out on potentially important things. Like my grandmother is sick right now. If negative things were excluded, I wouldn't see my Uncle's updates about her."

Respondents were also concerned that removing negative posts might prevent users from getting help from friends in time of need. *R2006-U* wrote that "posting negative posts serves a purpose. You get support from friends and then in turn show friends that support will be offered when it is needed." *R664-A* asked "what if one of those 'depressing posts' was a cry for help?"

While the research may not have broken the Facebook newsfeed's *promised* level of reliability and filtering accuracy, had the researchers had accesses to responses such as these they may have come to realize just how much users have come to rely on Facebook's best effort. Respondents *expected* to receive Facebook's best effort regardless of what's promised in the fine print. Thus, some participants felt that even if such research is allowed by Facebook's terms of service, "it exploits participants and doesn't treat them with respect" (*R627-A*).

Spammer infiltration

In creating our description of the spammer-infiltration study, we asked the lawyer who vetted the experiment to verify the accuracy of our scenario description. Despite this effort, the survey responses we received indicates that this experiment is one we could have done a much better job summarizing.

We failed to describe a salient feature of the experiment: that those who responded to spam and attempted to make a purchase would see a warning message that prevented them from completing a purchase. As a result of our elision, a great number of respondents wondered, in the words of *R578*, "what happens when the person never gets the merchandise they ordered?" We addressed this in future iterations of our survey, but still see similar concern for participants (well above 40%) and disapproval rates (well above 20%) while the responses are less insightful than for the other studies.

The nature of this experiment was still difficult for respondents to grasp, with some confused about whose computers would be infected, with respondents focusing more on the infection than the resulting spam. For example, *RP985* wrote: "I'm not sure if I fully understand the study but it seems like we could be helping people whose computers are infected but won't be doing so which seems unfair."

Many respondents questioned why spam recipients who tried to purchase a product were not debriefed that they were part of an experiment. *R1696* wrote "people should be informed if they were included in the study, maybe on the "checkout" page." *R272* was concerned that that "the researchers won't inform the users, and no awareness is being raised until the research is published." Indeed, given the methodology and the low probability that the unwitting participants would know each other, we are not aware of a compelling reason not to debrief spam recipients who attempted to make a purchase. (For a discussion of arguments against forgoing debriefing, see Sommers and Miller [18].)

Many respondents worried that researchers might find themselves unable "to control the experiment" (*R2739*) and that the experiment could backfire. In the words of *R1548*, "you're trying to hijack a spammer. What is your plan when they find you?" *R3127* feared that the spammers might "launch an attack against the researchers." While we would likely agree with the researchers if they concluded such concerns are exaggerated, had the experiment included a debriefing these responses indicate that this is a topic worth addressing.

DISCUSSION

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 make these tough choices, leading to a movement to increase the use of empirical research methods to address questions of research ethics [16]. While our ethical-response survey instrument has limitations, especially when compared to post-facto measures of those participating in approved experiments, researchers and ethics boards would benefit from identify risks and sources of controversy before approving experiments.

To reduce the burden and cost of performing ethical-response surveys, we have made our instrument available to the community in two different forms. The first is an open-source tool that other researchers can copy and use on it on their own. The second is a service; we integrate other researchers' scenarios into a survey that we field periodically. We work with researchers to ensure their scenario descriptions are written in a manner that facilitates comparison with our existing scenarios, facilitating cross-scenario comparison.

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* (July 2014).
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. Finn, P., and Jakobsson, M. Designing ethical phishing experiments. vol. 26 (Spring 2007), 46–58.
9. 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.
10. Hertwig, R., and Ortmann, A. Deception in experiments: Revisiting the arguments in its defense. *Ethics & Behavior 18*, 1 (2008), 59–92.
11. Jagatic, T. N., Johnson, N. a., Jakobsson, M., and Menczer, F. Social phishing. *Communications of the ACM 50*, 10 (2007), 94–100.
12. Jakobsson, M., Johnson, N., and Finn, P. Why and how to perform fraud experiments. *IEEE Security & Privacy 6*, 2 (2008), 66–68.
13. 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.
14. 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.
15. Salehi, N., Irani, L., Bernstein, M., Ogbe, E., and Alkhatib, A. Dynamo: Guidelines for academic requesters. http://wiki.wearedynamo.org/index.php/Guidelines_for_Academic_Requesters.
16. Sieber, J. E. Empirical research on research ethics. *Ethics & Behavior 14*, 4 (2004), 397–412.
17. Sigmon, S. T., Boulard, N. E., and Whitcomb-Smith, S. Reporting ethical practices in journal articles. *Ethics & Behavior 12*, 3 (2002), 261–275.
18. Sommers, R., and Miller, F. G. Forgoing Debriefing in Deceptive Research: Is It Ever Ethical? *Ethics & Behavior 23*, 2 (2013), 98–116.
19. Turkopticon. http://turkopticon.ucsd.edu/.