You Needn't Build That:
Reusable Ethics-Compliance Infrastructure for Human Subjects Research

Cristian Bravo-Lillo
Carnegie Mellon University

Serge Egelman
UC Berkeley

Cormac Herley, Stuart Schechter, Janice Tsai
Microsoft Research

Introduction

Just as security is often a secondary task when users sit down to accomplish something on their computers, ethics tends to be a secondary task for the security researchers who study these users. Both security and ethics rules are often viewed as an inconvenience to those whose productivity is reduced by demands to comply. For researchers, ethics requirements such as informed consent and debriefing are just one of many sources of friction that stand in the way of their research goals.

In this paper, we describe how shared tooling could assist in three different research functions related to ethical compliance: obtaining informed consent, debriefing, and the surveying of surrogate participants when consent cannot be obtained from actual participants. Having invested the time to exceed ethical compliance standards in our recent security experiments, we believe this increased attention to ethical design has benefited participants. We are building services to perform these compliance tasks with the goal of reducing the cost of compliance to researchers and obtaining a level of attention to participant protection that would be unreasonable to expect from researchers for whom this is not a primary goal.

While we are in part motivated to build reusable ethics-compliance tools because they serve a social good, we too stand to benefit; we plan to build these tools as services that facilitate the sharing of ethics-related behavioral data with the ethics research community. As members of that community, we hope to aggregate the behavioral observations flowing from myriad experiments' ethics infrastructure and use these data to iteratively improve the design of our tools. We also hope to run experiments and analyses using these data that benefit the research community as a whole. We hope that, as the flow of data on ethics-related interactions grows, other researchers will also use these data to advance the state of research ethics.

In the remainder of this paper, we describe proposed improvements to three ethics-compliance tasks that could be achieved improve reusable tooling. These tasks are the obtaining of informed consent, the debriefing participants along with monitoring participants reactions during debriefings, and the surveying of surrogate participants.

1 Obtaining Informed Consent

Human-subjects guidelines require researchers to obtain informed consent from research participants (or obtain an exemption). However, the consent experiences created to meet this requirement often provide little more than legal protection for the researchers, without actually ensuring that the consent obtained is truly informed. Compounding this problem is the growing popularity of online studies, where researchers are not physically present to visually observe the reading of consent forms or answer questions about them.

Ensuring consent is informed is difficult even for those dedicated to this goal. Pedersen *et al.* [3] asked 260 students to read a consent form that contained the following two sentences:

> In the questionnaire, you will be asked to recall information from this form. For example, you will be asked to recall the phrase lucky charms when completing the questionnaire.

Only about a quarter of participants were able to recall the phrase 'lucky charms.' Those who read the form online were half as likely to recall it as those who had read the form on paper.

We have also observed challenges of obtaining informed consent in our own research. Figure 1 illustrates how little time participants spent reading a consent form in a study we conducted on Amazon's Mechanical Turk [currently under submission]. Almost all the participants finished the consent in a small fraction of the time that would be required to read every word.

Obtaining informed consent may require us to overcome rational reasons to ignore consent forms. Participants may reasonably expect that any researcher diligent enough to comply with consent requirements, and who has maintained a positive rating on Mechanical Turk, probably isn't going to cause them significant harm. For research participants recruited via Mechanical Turk, many of whom encounter human-subjects experiments as just another form of work in a marketplace in which they are paid per task completed, spending time reading consent forms reduces their hourly wage.

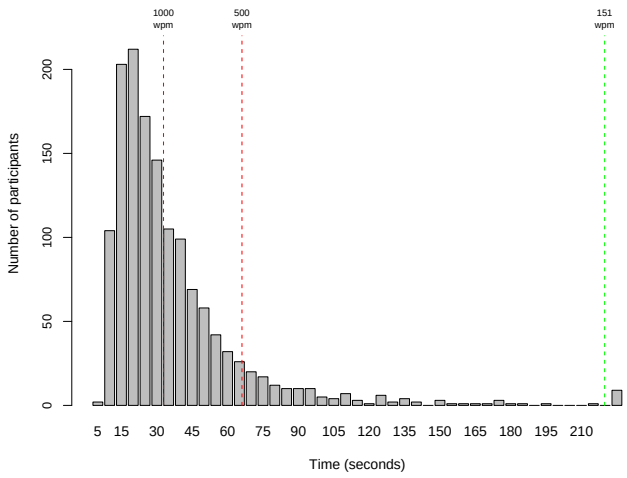

Figure 1: We presented a 552-word consent page to 1406 participants and measured their completion times. For each value on the x axis, a corresponding bar represents the number of participants who completed the page in between $x - 5$ and x seconds. Vertical lines represent the speed required to read 552 words in the period of time on the x axis. The vertical line at 151 words-per-minute represents the average English on-screen reading speed as measured by Dyson and Haselgrove [2].

A reusable consent experience design and presentation tool, informed by research, could start to address some of these challenges.

In designing such a tool, we would incorporate the findings of Varnhagen *et al.* [6], who found that participants spend more time reading, and recall more information from, consent forms that are divided into multiple succinct pages, as opposed to those contained in a single long page.

We would also propose using a design that makes clear to participants that the consent process is an important part of the study that requires their focus and attention. A consent tool could assist with this by making it easy to design and incorporate questions that test whether participants read and understand the risks. In addition, the tool could monitor speed with which participants can progress through each page or even limit the rate of progression based on models of the time required to process the page's contents.

We, as ethics researchers, would also benefit by providing a service that facilitates the design and delivery of consent experiences. Such an infrastructure would provide a regular flow of data on the terms and conditions common to human-subjects studies, how participants read consent forms, which consent terms are most likely to result in participant abandonment, and more.

Such uses of the human-subjects data arising from the consent process would need to be approved by the ethics board(s) of all researchers using the infrastructure. However, these ethics boards would have strong reason to do so. Consent forms built using this infrastructure would likely provide a higher level of care to ensuring true informed consent than a consent experience built from scratch by researchers not focused on ethics. Furthermore, both ethics boards and participants stand to benefit from the findings of quantitative ethics research that can lead to improved consent experiences.

2 Debriefing & Monitoring Harm

Deception is common in security studies because informing prospective participants that a study is about security may compromise its ecological validity; participants may make attending to security a primary goal and thus exhibit security behavior far different from that which they would exhibit had security been their secondary goal.

Studies that employ deception are almost always required to end with a *debriefing*, which serves a number of essential functions in addressing deception's potential hazards. The act of revealing the deception, known as 'dehoaxing' in the ethics literature, brings the deception to an end and informs that participant about aspects of the study that could not be revealed during the initial consent process. Explaining why the deception was necessary to the experiment, and the potential social benefits of the experiment, can help desensitize participants to any anxiety that they may experience upon learning they have been deceived. The debriefing process may also give the researcher an opportunity to empathize with participants or recognize if unexpected harms have arisen, especially if it is performed in person. Finally, researchers need to debrief participants so that, if participants would no longer consent to their behavioral data being used, they can withdraw their previously-uninformed consent. While regulations allow for forgoing debriefing in some circumstances, Sommers and Miller argue that the reasons for doing so often fail to hold up to ethical scrutiny [5].

Alas, as more behavioral research studies take place online, researchers are less likely to witness debriefings, to be available to answer participant questions, and to gauge how well these debriefings succeed in ameliorating anxiety. Furthermore, few researchers will invest the time to measure the efficacy of debriefings or iteratively improve them. Even when researchers do invest the time to ensure they have developed an effective debriefing, few will invest the space in their publications to share what they've learned about debriefing with other researchers. For example, Sharpe and Faye found that fewer than a quarter of studies in the Journal of Personality and Social Psychology report on whether a debriefing was performed and, when they do, they usually don't provide any additional details [4].

To achieve all the goals of debriefing in an online study, a debriefing mechanism needs to do more than just displaying the text to dehoax and desensitize participants. It should also:

- encourage the reporting of unexpected causes of harm, or reasons for feeling harmed,

- attempt to measure participants' perceptions of whether the experiment's harm outweigh the its benefits,
- offer participants the opportunity to withdraw consent for the use of their data, as they will be more informed about the experiment than they were when they initially consented.

A debriefing mechanism should also allow both researchers, and those charged with ensuring the ethical standards of their institution(s) (*e.g.*, members of their ethics boards), to monitor the ethical responses gathered by this system–and do so in real time. Having the ability to monitor the ethical impacts of experiments in progress, and halt or modify experiments in response to feedback from early participants, could enable ethics boards to approve experiments they might otherwise deem too risky.

With any debriefing survey, there exists the possibility that participants will fail to report harms because they are unwilling to entrust their feelings with the very researchers who just admitted to deceiving them. Thus, a final requirement for a good debriefing mechanism is that it is offered by an entity that has not been tainted by the deception, and in which further investigation will support the entity's trustworthiness (*e.g.*, web searches should reveal that researchers at reputable institutions are affiliated with this entity).

Progress To Date

We have built a reusable third-party service for performing debriefing and post-debriefing surveys, hosted at `https://www.ethicalresearch.org/Ethics.aspx`.

Our survey is informed by our experience iteratively building post-debriefing surveys and monitoring participant responses over multiple experiments. We first monitored participant responses to a deception study in which participants were asked to play and rate games, but during which we presented a spoofed operating system dialog box that requested their device password [1]. We were surprised by the extent to which participants overwhelmingly supported the use of deception, and wondered if our methodology might have consciously or subconsciously biased participants underreport harm. We iterated the methodology in a follow-up study, in which we used a similar OS-window spoofing ruse as in our previous experiment to test the efficacy of certain warning-design elements. The design iterations during these experiments, and additional design iterations following additional feedback from other researchers, resulted in a number of methodological refinements.

We not only randomize the order of questions in our post-deception survey, but we also select from one of two different variants of the same question so that we can detect if choices in wording bias responses. Each participant is only shown one of the variants of a question, which allows for between-subjects comparisons. For example, some participants are asked directly whether they feel the experiment they just participated in should or should not have been allowed to proceed, whereas others are asked the same question in the context of allowing future experiments "like this one." We employ two variants because we hypothesized that some participants may believe there's little to be gained by expressing objections to a study that has already been approved and is underway. Another concern we attempt to address is the potential for participants to underreport harm to themselves, which might occur for such reasons as not wanting to admit to themselves that they had been harmed. Some participants are asked about how they felt after performing the study, whereas others receive variants that ask how the participant would advise someone they cared about who was considering becoming a participant in the study.

There is a high fixed cost to designing a debriefing survey that attempts to account for so many of the potential behavioral confounds that arise when respondents are still processing the revelation that they have been deceived. The time investment over the initial iterations and the construction of the final infrastructure is similar to the time one would invest to design, build, and publish an independent experiment—far beyond the effort researchers could otherwise invest in ethical compliance.

With each experiment that uses the tool, we will grow a corpus of data on how research participants respond to different types of experiments. We offer participants the opportunity to opt out from having their responses made available to ethics researchers (ourselves included). We also allow participants to choose not to share their responses with researchers' ethics boards (perhaps out of concern that their honest opinions might get researchers in trouble with their ethics boards) or with the researchers themselves, who participants may not trust following the revelation of a deception.

As with consent tools, researchers employing our debriefing tool will require approval from their ethics boards to use this infrastructure and allow us to collect participants' responses. Again, we expect ethics boards to welcome researchers' use of infrastructure designed with ethics as a primary goal. We also expect ethics boards to appreciate the potential value of the aggregate research data for improving the quality of debriefing experiences for future participants, and determining what types of deceptions may simply be unacceptable regardless of debriefing.

3 Surrogate-Participant Surveys

Researchers cannot always obtain the consent of those they observe. In some studies it may be impractical to collect consent from everyone whose behavior might have impacted the data being collected. Consider, for example, a study of network traffic flowing through a large ISP. The aggregate traffic may be the result of millions of users' individual behavioral interactions. Even if the researchers wanted to contact every individual involved for consent, the network traffic itself does not contain sufficient information to allow researchers to identify and contact them.

In other studies, especially those involving observations of crime or victimization, those observed may consider being contacted for consent more harmful than simply allow-

ing researchers to use the data. For example, victims of password data breaches may approve of researchers' use of breach data to perform aggregate studies of password behavior, but might not approve of having their inbox cluttered with a request for consent from everyone who wanted to analyze the breach data–frequent and perhaps unnecessary reminders of their past victimization.

In situations where consent from participants cannot be obtained, ethics boards may require researchers to use *surrogate* participants to determine whether participants would likely consent if asked, or to measure how participants might react to information. Researchers may instruct surrogate participants to imagine themselves to be in the position of an actual participant and answer the question on that participant's behalf. Surrogate participants can also provide advance feedback for studies in which consent may be obtainable, but where deception will render it uninformed and unacceptable levels of harm could occur before feedback can be obtained.

There are high fixed costs to developing survey tools for reaching out to surrogate participants and collecting and analyzing data. Surveying surrogate participants about multiple experimental designs at once significantly reduces the marginal cost for each additional experimental design. Furthermore, when a single survey is used to investigate multiple experimental designs, researchers can compare the *relative* ethical acceptability or repulsion of these designs.

Progress To Date

We have developed a survey with which to measure surrogate participants' responses to multiple experiment abstracts. We undertook our first survey with the goal of evaluating how victims of a password breach, whose passwords had been made public by attackers, would feel if researchers performed studies on those breached passwords. In our survey, we show participants abstracts describing a study, starting with the scientific question/goal of the research and providing background (if needed). The abstract then uses bullet points to summarize steps in the experimental flow that may impact respondents' perceptions of the ethicality of the experiment. Finally, a short statement summarizes the consequences if the researchers are not able to perform the experiment. A sample is illustrated in Appendix A. We have used this survey tool to gauge surrogate participant responses to our own experiments.

Now that we have built our survey, there is but a small marginal cost to re-run it when researchers present new experimental designs for which feedback is needed. Being part of a larger survey enables researchers to understand how participants respond to their experimental abstract in comparison to others. We envision performing this survey online frequently, incorporating abstracts for proposed experiments for which researchers or ethics boards would like feedback.

Again, we as ethics researchers can benefit from providing this service. Whereas researchers who submit experiment abstracts are primarily interested in discovering surrogate participants' concerns with their experiment, we are interested in comparing across experiments and evaluating how subtle changes in the design and description of an experiment might have significant impacts on its perceived ethical acceptability. Since researchers may not want abstract descriptions of their experimental designs made public in advance of publishing the experiment itself, we would offer a non-disclosure period before the inclusion of an experimental abstract or data from it would be made available to the ethics research community.

Summary

For most researchers, complying with rules and norms of ethics will always be a secondary task. Reusable ethics infrastructure can save researchers time while increasing the level of ethical care in their experiments. Reusable infrastructure can also help those of us in the ethics research community measure the efficacy of consent forms, understand participants' reactions to debriefings, and analyze the comparative acceptability of experimental designs and practices as perceived by surrogate participants. Recognizing this, we have created an effort to undertake the provision of improved infrastructure for ethical compliance in human subjects experiments.

References

[1] Cristian Bravo-Lillo, Lorrie F. Cranor, Julie Downs, Saranga Komanduri, Stuart Schechter, and Manya Sleeper. Operating system framed in case of mistaken identity. In *The 19th ACM Conference on Computer and Communications Security (CCS)*, October 16–18 2012.

[2] Mary C. Dyson and Mark Haselgrove. The influence of reading speed and line length on the effectiveness of reading from screen. *International Journal of Human-Computer Studies*, 54:585–612, 2001.

[3] Eric R. Pedersen, Clayton Neighbors, Judy Tidwell, and Ty W. Lostutter. Do undergraduate student research participants read psychological research consent forms? Examining memory effects, condition effects, and individual differences. *Ethics and Behavior*, 21(4):332–350, 2011.

[4] Donald Sharpe and Cathy Faye. A Second Look at Debriefing Practices: Madness in Our Method? *Ethics & Behavior*, 19(5):432–447, 2009.

[5] Roseanna Sommers and Franklin G. Miller. Forgoing Debriefing in Deceptive Research: Is It Ever Ethical? *Ethics & Behavior*, In process of publication, 2013.

[6] Connie K. Varnhagen, Matthew Gushta, Jason Daniels, Tara C. Peters, Neil Parmar, Danielle Law, Rachel Hirsch, Bonnie Sadler Takach, and Tom Johnson. How informed is online informed consent? *Ethics and Behavior*, 15(1):37–2013, 2005.

A Sample experimental abstract from our surrogate survey.

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 task (HIT) they will be asked to complete:

- The participants will be given a task unrelated to security, but will encounter a security warning during the task.
- 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 aggregate results of the experiment will be used to publish a scientific paper. Participants' identities will remain anonymous.

If they are not allowed to collect this data, they cannot measure the effectiveness of different designs for computer security warnings. Therefore, they cannot publish recommendations to help improve the effectiveness of future security warnings.