PRIVACY AND A/B EXPERIMENTS

EDWARD W. FELTEN*

The debate about Facebook's mood manipulation experiment[1] has focused mainly on Facebook's manipulation of what users saw, rather than the "pure privacy" issue of which information was collected and how it was used. It might be tempting to conclude that because Facebook did not change its data collection procedures, the experiment couldn't possibly have affected users' privacy interests. But that reasoning would be incorrect, because the fact that Facebook manipulated users' news feeds increased the impact on users' privacy beyond the impact of simply gathering information.

More generally, the privacy implications of A/B experiments on users have not received much attention in the policy literature. This essay explains the distinctive privacy implications of A/B experiments, explores some hypotheticals, and applies the resulting framework to Facebook's experiments.

The term "A/B experiment" refers to an experiment in which users of a site or service are divided at random into groups, each group is shown different content or functionality, and users' responses are recorded to see whether there are statistically significant differences between the groups.[2] These experiments are often called "A/B tests" in the technology industry, and "randomized user tests" in much of the published literature.

REVIEW: CORRELATION AND CAUSATION

The usual reason to go beyond simply observing users' behavior and begin conducting A/B experiments is that observation typically detects *correlation* only, whereas A/B experiments can show *causation.*[3]

* Director, Center for Information Technology & Policy; Robert E. Kahn Professor of Computer Science and Public Affairs, Princeton University. Thanks to Yannis Avramopoulos, Josh Kroll, Johan Ugander for comments on drafts of this work.

1. Adam D.I. Kramer, Jamie E. Guillory & Jeffrey T. Hancock, *Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks*, 111 PROC. NAT'L ACAD. SCI. 8788, 8788 (2014).

2. *See, e.g.*, Brian Christian, *The A/B Test: Inside the Technology That's Changing the Rules of Business*, WIRED (Apr. 25, 2012), http://www.wired.com/2012/04/ff_abtesting/.

3. A more general account of when and how one can distinguish causality from correlation is beyond the scope of this paper. Although it is sometimes possible to infer

Recall that when we say that two variables X and Y are *correlated*, this means that X and Y tend to occur together more often than they would by chance if the variables were independent.[4] Correlations can appear in a data set by coincidence, but we will restrict our discussion to cases where the possibility of coincidental correlation can be ruled out with very high confidence.

If X and Y are correlated with very high confidence, then one of three things must be true: X causes Y, or Y causes X, or there is some other factor (or factors) that causes both X and Y.[5] Causation can be direct or indirect.[6] When we say that X causes Y, this means that if we intervene to make X true, this will make Y more likely on average.[7] For example, when we say that smoking causes lung cancer, this means that starting smoking will make lung cancer more likely on average, and stopping smoking will make lung cancer less likely.

Knowing that X causes Y is more valuable than knowing that X and Y are correlated, for two reasons. First, causation can act as a guide to action because it predicts what will happen if we take a particular action, whereas simple correlation does not enable such predictions. Second, causation is useful in constructing generalizable theories about the world.

With this background in place, let us proceed to discuss the privacy consequences of A/B experiments.

A HYPOTHETICAL A/B EXPERIMENT

To simplify the discussion, consider a hypothetical social network called Wo. Wo lets users set up accounts and establish mutual friend relationships, as on Facebook.[8] Rather than letting users post detailed status updates, Wo only lets a user set their status to one of two values: Happy or Sad. Users viewing the service are shown icons depicting their friends, in the form of a happy or sad face, depending on the friend's status. Wo keeps records of users' status changes and when each user views their friends' statuses.

Wo can learn certain things about its users by collecting and analyzing these records. Wo can tell how often a given user is Happy,

causality from observation alone, A/B experiments are the primary method for determining causal relationships. *See, e.g.*, JUDEA PEARL, CAUSALITY: MODELS, REASONING, AND INFERENCE (2d ed. 2009).

4. *See, e.g.*, *id.* at 126-32.
5. *See, e.g.*, *id.* at 2-5.
6. *See, e.g.*, *id.* at 126-32.
7. *See, e.g.*, *id.*; *cf.* Orin S. Kerr, *A Theory of Law*, 16 GREEN BAG 2d 111 (2012).
8. *Finding Friends & People You May Know*, FACEBOOK, https://www.facebook.com/help/433894009984645/ (last visited Apr. 21, 2015).

how long a user's Happy and Sad states persist, and to what extent a user's status correlates with the status of each friend and with the statuses of friends in aggregate.

A/B experiments allow Wo to learn things about its users that it could not learn by observation alone. Suppose Wo wants to study "emotional contagion" among its users. In particular, it wants to know whether seeing more Sad faces causes a user to be more likely to set their own status to Sad. Wo could try to learn about contagion by observation, for example by measuring whether a user's status tends to be correlated with her friends' statuses. But correlation is not causation. Alice and Bob might be Sad at the same time because Alice's sadness causes Bob's sadness, or because Bob's sadness causes Alice's sadness, or because of a common causal factor. Perhaps Alice and Bob are both Sad because something bad happened to their mutual friend Charlie, or because they live in the same town and the weather there is bad.

Wo can study emotional contagion by doing an A/B experiment in which it artificially manipulates what users see. In some fraction of cases, chosen at random, Wo shows the user a random status for one friend (rather than that friend's actual status at the time), and Wo measures whether the user is more likely to set their status to Sad when the false friend status is Sad. (Wo must attend to some methodological issues that we will not discuss here, but the idea should be clear.)

INFERENCES ABOUT GROUPS AND INDIVIDUALS

This kind of experiment allows Wo to learn about users in aggregate and about individual users. If Wo's user population is very large, it will probably be able to determine with some precision the average amount of emotional contagion over the whole user population. That will be of scholarly interest and might help Wo improve its products or help its users.

The same experiment will also give Wo information about individual users. Every time Wo, as part of an A/B experiment, modifies what user Alice sees, Wo learns a little bit about what *causes* Alice's behavior. For example, if Wo changes Alice's experience four times, each time by telling Alice that her friend Bob is Sad when Bob is Happy in reality, then Wo will learn a little bit about the causal effect of Bob's mood on Alice's mood. As a side-effect of doing an A/B experiment on the user population, Wo is also doing a small A/B experiment on Alice, observing how Alice behaves at different times when given manipulated information about Bob's mood. This is true even if Wo's primary motivation for doing the

experiment is only to learn about users in aggregate.

The information gleaned about an individual user will typically have low statistical confidence, because there will be relatively few observations of that one user. But probabilistic information is still information, and the small A/B experiment that will have been performed on Alice will have at least a small impact on how much Wo knows about her.

A notable feature of this hypothetical experiment is that Alice probably could not tell that the experiment occurred. She would know that she was revealing her emotional state to Wo by setting her status, and it would be evident to her that Wo could easily determine correlations, but she would not know that Wo was learning *causal* information about how manipulable her emotions were. Importantly, the privacy impact of an interaction like this depends not only on which types of information are gathered, but also on which prompts were given to the user and how those prompts were chosen. Experimenting on users affects their privacy.

RELATION TO FACEBOOK'S EMOTIONAL CONTAGION EXPERIMENT

What does this hypothetical teach us about the privacy impact of Facebook's experiment?

There are some differences between the Wo hypothetical and Facebook's real experiment. The most obvious difference is that Facebook status updates are not simply Happy or Sad but are more complex unstructured text written by the users.[9] Facebook used a particular methodology to classify the mood conveyed by updates,[10] and this methodology was not (and could not hope to be) perfect. However, to the extent that Facebook's mood classifying algorithm was accurate on average, this difference from the Wo hypothetical is unlikely to be significant from a policy standpoint.

Another difference is that the Facebook experiment did not modify the content of any status update, but instead suppressed updates that were classified as having a certain mood. Because Facebook, unlike Wo, selects only a subset of friends' updates to display, Facebook can modify a user's experience by changing the selection of accurate updates, rather than by modifying the content of any particular update. The effect is to modify the aggregate mood of the updates shown to a particular user.[11] Although this difference

9. *How Do I Post a Status Update?*, FACEBOOK, https://www.facebook.com/help/436937729656469?sr=4&query=status&sid=08Y1ZHkXGop74roH4 (last visited Apr. 21, 2015).

10. Kramer et al., *supra* note 1.

11. *Id.*

arguably affects the ethical calculus surrounding the studies, it does not significantly affect the privacy analysis.

It follows that Facebook, by doing its emotional contagion study using an A/B experiment, did learn some non-zero amount of information about the manipulability of individual users' emotions. Given the published results of the study, the information learned about individual users was probably very weak, in the statistical sense of being correlated with the truth but only very weakly correlated, for the vast majority of users or perhaps for all users. Because the study was not disclosed to the affected users, these users would not have been able to discover that Facebook had learned this information.[12]

BASIC ETHICS OF A/B EXPERIMENTS

In light of the privacy impact of A/B experiments, it is clear that the ethics of A/B experiments are an important and interesting topic. This section is a first-cut attempt at thinking through some basic ethical questions relating to A/B experiments. We are considering A/B experiments in general, and not just experiments done for academic research purposes; what is ethical rather than what is legal or what is required by administrative bodies such as Institutional Review Boards; and considering how people should act rather than observing how they do act.

Let us start with an obvious point: Some uses of A/B experiments are clearly ethical. For example, if a company wants to know which shade of blue to use in their user interface, they might use an A/B experiment to try a few shades of blue and measure users' responses. This is ethical because no user is harmed, especially if the only result is that the service better serves users. There is little if any privacy impact to learning whether a particular user prefers one shade of blue over another slightly different shade.

Another point that should be obvious is that some uses of A/B experiments are clearly unethical. Consider a hypothetical study in which a service falsely tells teens that their parents are dead, or that tries to see if a service can incite ethnic violence in a war-torn region. Both studies are unethical because they cause significant harm or risk of harm to users.

So the question is not whether A/B experiments are ethical, but rather where we should draw the line between ethical and unethical uses. A consequence of this is that any argument that implies that A/B experiments are always ethical or always unethical must be

12. *Id.*

wrong.

Here is an example argument: Company X does A/B experiments all the time; this is just another type of A/B experiment; therefore this is ethical. Here is another: Company X already uses an algorithm to decide what to show to users, and that algorithm changes from time to time; this is just another change to the algorithm; therefore this is ethical. Both arguments are invalid, in the same way that it is invalid to argue that Chef Bob often cuts things with a knife, therefore it is ethical for him to cut up anything he wants. The ethical status of the act depends on what exactly Chef Bob is cutting, or exactly which A/B experiment is being done, or which exact algorithm is being used. (At the risk of stating the obvious: the fact that these sorts of invalid arguments are made on behalf of a practice does not in itself imply that the practice is bad.)

Another argument goes like this: Everybody knows that companies do A/B experiments of type X; therefore it is ethical for them to do A/B experiments of type X. This is also an invalid argument, because knowledge that an act is occurring does not imply that the act is ethical.

But the "everyone knows" argument is not entirely irrelevant, because we can refine it into a more explicit argument that deserves closer consideration. This is the implied consent argument: User Bob knows that if he uses Service X he will be subject to A/B experiments of Type Y; Bob chooses to use Service X; therefore Bob can be deemed to have consented to Service X performing A/B experiments of Type Y on him.

Making the argument explicit in this way exposes two potential failures in the argument. First, there must be general knowledge among users that a particular type of experiment will happen. "Everyone knows" is not enough, if "everyone" means everyone in the tech policy commentariat, or everyone who works in the industry. Whether users understand something to be happening is an empirical question that can be answered with data; or a company can take pains to inform its users, that is, it can actually inform users rather than providing minimal the-information-was-available-if-you-looked notification theater.

Second, the consent here is implied rather than explicit. In practice, User Bob might not have much real choice about whether to use a service. If his employer requires him to use the service, then he would have to quit his job to avoid being subject to the A/B experiment, and the most we can infer from his use of the service is that he dislikes being a subject of the experiment less than he would dislike losing his job. Similarly, Bob might feel he needs to use a service to keep tabs on his children, to participate in a social or

religious organization, or for some other reason. The law might allow a legal fiction of implied consent, but what we care about ethically is whether Bob's act of using the service really does imply that he does not object to being an experimental subject.

Both of these caveats will apply differently to different users. Some users will know about a company's practices but others will not. Some users will have a free, unconstrained choice whether to use a service but others will not. Consent can validly be inferred for some users and not others; and in general the service will not be able to tell for which users the conditions for ethically sufficient consent exist. So if an experiment is run on a randomly selected set of users, it is likely that consent can be inferred for only a subset of those users.

Where does this leave us? Where the risks are minimal, A/B experiments without consent are unobjectionable, as in the shades-of-blue example. Where risks are extremely high or there are significant risks to non-participants, as in the ethnic-violence example, the experiment is unethical even with consent from participants. In between, there is a continuum of risk levels, and the need for consent would vary based on the risk. Higher-risk cases would merit explicit, no-strings-attached consent for a particular test. For lower-risk cases, implied consent would be sufficient, with a higher rate of user knowledge and a higher rate of unconstrained user choice required as the risk level increases.

Where exactly to draw these lines, and which processes a company should use to avoid stepping over the lines, are topics beyond the scope of this paper. What is clear is that ethics discussions about A/B experiments must account for the fact that A/B experiments impact the privacy of users.

MEASURING CONSENT

In most cases, ethical questions can be addressed by securing informed consent from participants. In the case of an A/B experiment, this would mean informed consent by any user whose experience would be modified in the experiment. This type of consent is clearly feasible in the case of Wo, and would have been feasible for Facebook.

Two main arguments have been raised against the viability of consent for certain experiments. First, some have argued that informing users about a study in advance would influence their behavior during the study, thereby confounding the study's validity. Second, some have argued that the time and annoyance involved in the informed consent process outweighs the benefit to users of

requiring consent, especially because an A/B experiment (absent a consent dialog) does not inconvenience users in any way. Both arguments are supported, at least implicitly, by a claim that the vast majority of users would have consented to a specific experiment, had they been asked. Ethically, the case for skipping the consent step is stronger when that claim is true, that is, when the vast majority would indeed have consented.

The assertion that at least (say) 95% of users would have consented, or equivalently that each user in a randomly selected subgroup would have consented with at least 95% probability, can itself be tested by experiment: simply select a random subgroup of users and ask for their consent to participate in the primary study. This preliminary consent-measurement experiment will either support or contradict claims about the consent rate for the overall user population.

Performing a separate consent experiment has one obvious disadvantage, compared to asking for individual consent from participants in the primary study: some users will end up participating in the primary study even though they would not have consented to do so. We cannot know who those users are, but we know that they exist and we can estimate their number statistically.

Consent experiments do have countervailing advantages, in addressing the two previously stated arguments against consent. First, by doing the consent experiment on one set of users and the primary experiment on a separate set, we can minimize the likelihood that participants in the primary study will behave differently due to knowing about the study. Because participants in the primary study did not themselves participate in the consent process, they are unaware that the study exists and that they are part of it. Second, if the user population is very large, it might be possible to perform a consent experiment on a smaller number of users and the primary experiment on a much larger number, thereby reducing the inconvenience imposed on the user community by the consent process while still collecting enough evidence to support performing the primary study. Each of these benefits accrues because consent is not sought from the primary study participants; so these benefits are not separable from the disadvantage of the consent experiment approach.

Consent experiments seem to have some merit for online A/B experiments, although they would have made less sense for traditional in-person social science studies. In the in-person case, users already know they are subjects in a study and are already being inconvenienced; and the user population is not so large that a separate consent experiment is feasible. This is one example of how

different experimental and procedural options are available for online experiments.

More needs to be done to understand how to protect the privacy and other interests of users who are subjects of online A/B experiments. Although the options and the necessary procedures may differ from those seen in traditional experiments, experiments' responsibility to protect users from harm remains as strong as ever.