9

Ethics in Controlled Experiments

> The progress of science is far ahead of man's ethical behavior
> – *Charlie Chaplin (1964)*

> . . .testing where changes in program CODE induce user DECEPTION. . . [we] call this new approach C/D experimentation to distinguish it from. . .A/B testing
> – *Raquel Benbunan-Fich (2017)*

Why you care: *Understanding the ethics of experiments is critical for everyone, from leadership to engineers to product managers to data scientists; all should be informed and mindful of the ethical considerations. Controlled experiments, whether in technology, anthropology, psychology, sociology, or medicine, are conducted on actual people. Here are questions and concerns to consider when determining when to seek expert counsel regarding the ethics of your experiments.*

Background

A broad definition of ethics is the set of rules or morals that govern what we should or should not do. Ethics, as applied to research, govern the rules of conduct that ensure the integrity of the results, the values essential for collaborative work, public accountability, as well as moral and social values, including both public safety and the protection of human subjects (Resnick 2015). The application of ethics to research can change over time, reflecting the changing world, culture, and human responses to the unexpected ramifications of research studies over time. As Charlie Chaplin wrote in the quotation above, rules and regulations for ethical behavior are developing and lagging the science.

This subject is too deep to delve into fully here, so we only give an overview of the research ethics of controlled experiments. For a deeper study, we recommend several references (Loukides, Mason and Patil 2018, FAT/ML 2019, ACM 2018, King, Churchill and Tan 2017, Benbunan-Fich 2017, Meyer 2015, 2018), which present key principles, checklists, and practical guides. While experimenters are often not experts, we should ask ourselves questions, critically examine our practices, and consider the long-term best interests of our users and the business. Note that we are writing this in our capacity as individuals and not as representatives of Google, LinkedIn, or Microsoft.

Two recent examples from technology illustrate the need for these questions.

1. Facebook and Cornell researchers studied emotional contagion via social media (Kramer, Guillory and Hancock 2014) to determine whether randomly selected participants exposed to slightly more negative posts posted more negative content a week later and, conversely, whether other randomly selected participants exposed to slightly more positive posts had more positive posts themselves a week later.
2. OKCupid ran an experiment where they enrolled pairs of customers whom the algorithm said were 30%, 60%, and 90% matches, and, for each of these three groups, told a third of them that they were 30% matches, a third of them that they were 60% matches, and a third of them that they were 90% matches (The Guardian 2014, Meyer 2018).

Given these examples, and many others, how do we assess and evaluate which A/B experiments to run?

We can first turn to the Belmont Report, released in 1979 (The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1979) that establishes principles for biomedical and behavioral studies, and to the Common Rule (Office for Human Research Protections 1991) that establishes actionable review criteria based on these principles (Meyer 2012). These were established after several examples, including the Tuskegee Syphilis study from the 1930s (CDC 2015) and the Milgram experiment in the 1960s (Milgram 2009) in the medical domain, where the risk of substantial harm is commonly much higher than that in online experiments. Based on these guidelines, we now ask questions about whether this clinical trial is justified (Hemkens, Contopoulos-Ioannidis and Ioannidis 2016), and there are situations where conducting randomized controlled trials (RCTs) is unrealistic or perceived as unethical (Djulbegovic and Hozo 2002).

The Belmont report and the Common Rule provide three key principles in the context of biomedical and behavioral human subjects research:

- **Respect for persons:** Treat them with respect, that is, treat people as autonomous agents when they are and protect them when they are not. This translates to a focus on transparency, truthfulness, and voluntariness (choice and consent).
- **Beneficence:** Protect people from harm. While the Belmont Report states that beneficence means minimizing risks and maximizing benefits to participants, the Common Rule recognizes the challenge in doing so and focuses instead on properly assessing the risks and benefits, and balancing those appropriately when reviewing proposed studies.
- **Justice:** Ensure that participants are not exploited and that there is a fair distribution of risks and benefits.

Because of the complexity, the Common Rule lays out provisions that balance not just the benefits and risks of the study itself but also informs the necessity of transparency, truthfulness, and voluntariness for participants in the study, including waivers.

While these questions are a useful framework from a discipline – medicine – in which substantial harms could occur, there are rarely unambiguous right or wrong answers, so assessing these principles with regards to specific online A/B experiments requires judgment, thought, care, and experience. Here are key areas to consider.

Risk

In your study, what **risk** does a participant face? Does the risk exceed that of minimal risk, defined by the Common Rule as "the probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests." The harm could be physical, psychological, emotional, social, or economic.

One useful concept is *equipoise* (Freedman 1987): whether the relevant expert community is in equipoise – genuine uncertainty – with respect to two treatments.

In evaluating online controlled experiments, one useful litmus test is whether you could ship a feature to all users without a controlled experiment, given the organizational standards. If you could make the change to an algorithm, or to the look-and-feel of a product without an experiment, surely you should be able to run an experiment and scientifically evaluate the change first; perhaps you will uncover unexpected effects. Shipping code is, in fact, an experiment. It may not be a controlled experiment, but rather an inefficient

sequential test where one looks at the time series; if key metrics (e.g., revenue, user feedback) are negative, the feature is rolled back.

The resistance to an online controlled experiment when giving everyone either Control or Treatment would each be acceptable is sometimes referred to as the "A/B illusion" (Meyer 2015, Meyer et al. 2019). When you decide to ship something, you are assuming what effect will result, and that assumption may or may not hold. If you are willing to ship something to 100% of users, shipping to 50% with the intent of going to 100% as an experiment should also be fine. In an example Meyer wrote (Meyer 2015):

> . . .the head of a company is concerned some of her employees are failing to save enough for retirement. . . She decides that from now on, when she sends out 401(k) mailings, she will include a statement about how many co-workers within five years of the employee's age have signed up for automatic enrollment. She hypothesizes that the minority of employees who haven't enrolled may be influenced to do so by knowledge of the majority's contrary behavior.

While the head of company is well-intentioned, and studies have shown the benefits of peer effects, when the controlled experiment was run, it resulted in oppositional reaction and decrease in savings (Beshears et al. 2011).

Benefits

The other side of risk is to understand the benefits of the study. Oftentimes for online controlled experiments, benefits are considered in terms of improving the product, which can be directly for users in the Treatment, for all users who benefit from the results, or even indirectly in terms of building a sustainable business so that users can continue benefitting from the service. Improvements to user productivity might fall in the first two buckets, while improvements to ads revenue might fall in the last bucket of indirect benefit.

One situation where assessing the benefits may be trickier is when running experiments that knowingly provide participants a worse experience with a goal of ultimately improving the experience for all users, often by being able to quantify tradeoffs. Examples include running experiments that slow user experience (see Chapter 5), showing more ads to understand long-term effects (see Chapter 23) or disabling features such as recommendations to assess their value. These cases violate equipoise in that there is general agreement that the Treatment is not beneficial but has minimal risk to users. The benefit of running these experiments involves establishing tradeoffs that can be used for more informed decision making and ultimately help improve the user experience for all. Importantly, there is no deception of users in these cases. While there is a higher

risk profile with a greater potential for harm than most online controlled experiments, there is a medical analogy for these types of experiments in drug toxicity studies: at some point, too much of a drug can be bad, but without running the studies, we could not know how much or how bad the effects are.

One point to emphasize is the major difference between running experiments to try out new features, new text, new algorithms and infrastructure, even to establish tradeoffs, versus running *deception* or *power-of-suggestion* experiments that focus on behavioral experimentation and relationships between people (Benbunan-Fich 2017). Deception experiments carry higher ethical risk and raise questions about whether participants are respected.

When thinking about respect for participants, the first questions we should ask are around transparency and expectation. Products set user expectations about what they provide by both what is in the UI and what is broadly communicated. Experiments should follow those expectations.

Alongside several other ways of ensuring transparency, informed consent is a key ethical concept where participants agree to participate in the study after they are fully informed about risks and benefits, the process, any alternative options, and what data is being gathered and how it is handled. Note that here, we are discussing consent in terms of its general meaning rather than specific to any legal definition, such as under Europe's General Data Protection Regulation (European Commission 2018). Most medical experiments have informed consent for each participant, and those that do not are typically minimal risk and meet other conditions, thus qualifying for a waiver of consent under the Common Rule. In contrast, experiments by online service providers usually involve a far lower level of risk to the participants, although as online services start impacting offline experiences, such as with shipping physical packages, ride sharing, and so on, the risk and consequentiality can increase. In addition, given the scale of experiments, obtaining informed consent is both prohibitively expensive and annoying to users. Instead, consider the range of possibility from experiments where consent is needed to those where the risk and potential harm to users is very low and consent is not needed. One alternative towards the middle of that spectrum is *presumptive consent*, where a smaller but representative group of people are asked how they would feel about participating in a study (or class of studies) and, if they agree, assuming that this sentiment would generalize to all participants (King et al. 2017).

Provide Choices

Another consideration is what **choices** do participants have? For example, if you are testing changes to a search engine, participants always have the

choice to use another search engine. Switching costs for other online services may be higher in terms of time, money, information sharing, and so on. These factors should be considered when assessing the choice offered to participants and the risks and benefits to be balanced. For example, in medical clinical trials testing new drugs for cancer, the main choice most participants face is death, making it allowable for the risk to be quite high, given informed consent.

Data Collection

One prerequisite for running A/B experiments is that data instrumentation is present for experiment analysis and for making decisions. Often, this data must be collected to measure and provide a high quality of service to users. As a result, data collection consent is often included in the Terms of Service for online services. While other references discuss data collection in more detail (Loukides et al. 2018), and while it is of course a pre-requisite that any experiments comply with all applicable privacy and data protection laws, experimenters or engineers should be able to answer these key questions about data collection:

- What data is being collected and what do users understand about that collection, with privacy by design being one useful framework in this area (Wikipedia contributors, Privacy by Design 2019).
  - Do users understand what data is being collected about them?
  - How sensitive is the data? Does it include financial or health data? Could the data be used to discriminate against users in ways that infringe human rights?
  - Can the data be tied to the individual, that is, is it considered personally identifiable (see Sidebar later in this chapter)?
  - For what purpose is the data collected and how can the data be used, and by whom?
  - Is it necessary to collect the data for the purpose? How soon can the data be aggregated or deleted to protect individual users?
- What could go wrong with the data collection?
  - What harm would befall users if that data or some subset be made public?
  - Consider harm to their health, psychological or emotional state, social status, or financials.

- What are user's expectations of privacy and confidentiality, and how are those expectations being guaranteed?
  For example, if participants are being observed in a public setting (such as, a football stadium), there is a lower expectation of privacy. If the study is on existing public data, then there is also no expectation of further confidentiality. If the data is not personally identifiable (see Side bar on page 103), then privacy and confidentiality are not necessarily a concern (NSF 2018). Otherwise:
  - What level of confidentiality can participants expect?
  - What are the internal safeguards for handling that data? Can anyone at the company access the data, especially if it's personally identifiable, or is the data secure with access logged and audited? How are breaches to that security caught, communicated, and managed?
  - What redress will happen (will participants be informed) if these guarantees are not met?

Culture and Processes

Many issues we address are complex and nuanced. It can be tempting to just rely on experts to make all judgments and set principles. However, to ensure that the ethical considerations are met, it is important that your corporate culture, everyone from your leadership down, understands and considers these questions and implications. Introspection is critical.

Companies – leaders – should implement processes to ensure that this level of understanding reaches across the board to:

- Establish cultural norms and education processes to keep employees familiar with the issues and ensure that these questions are asked at product and engineering reviews.
- Create a process that fulfills the purpose of Institutional Review Boards (IRBs). IRBs review possible human subjects research, assess the risks and benefits, ensure transparency, provide processes, and more to ensure the integrity and respect for participants. The IRB approves, requires alternatives, or denies studies. They provide questions for experimenters to consider that ensure thorough review and adequate introspection and establish just-in-time processes for educational purposes.
- Build tools, infrastructure, and processes so that all data, identified or not, is stored securely, with access time limited to those who need it to complete their job. There should be a clear set of principles and policies for what data

usage is acceptable and what is not acceptable. You should ensure that all data use is logged and regularly audited for violations.
- Create a clear escalation path for how to handle cases that have more than minimal risk or data sensitivity issues.

These questions and processes around the ethics of experiments are not an item to check off, but rather discussions that improve the design of the product and experiments for end users.

SIDEBAR: User Identifiers

One frequently asked question is what is the difference between identified, pseudonymous, and anonymous data? While the precise definitions may shift based on context or applicable laws and are still being discussed, an overview of the high-level concepts associated with these concepts are:

- **Identified** data is stored and collected with personally identifiable information (PII). This can be names, IDs (such as a social security number or driver's license), phone numbers, and so on. A common standard is HIPAA (Health and Human Services 2018b, Health and Human Services 2018c), which has 18 identifiers (HIPAA Journal 2018, Health and Human Services 2018a) that are considered personally identifiable. Device ID (such as, a smartphone's device ID) is also considered personally identifiable in many instances. In Europe, GDPR (General Data Protection Regulation) holds an even higher standard, and considers any data to be personal data if it can be linked to an individual (European Commission 2018).
- **Anonymous** data is stored and collected without any personally identifiable information. This data is considered **pseudonymous** if it is stored with a randomly generated ID, such as a cookie, that is assigned to some event, such as the first time a user opens an app or visits website and does not have an ID stored. However, simply stating that data is pseudonymous or anonymous does not mean that re-identification cannot happen (McCullagh 2006). Why? We must distinguish between anonymous data and anonymized data. Anonymized data is identified or anonymous data that has been looked at and guaranteed in some way that the re-identification risk is low-to-nonexistent, that is, given the data it almost impossible for someone to determine which individual this data refers to. Often, this guarantee is done via the Safe Harbor method or other methods such as k-anonymity (Samarati and Sweeney 1998) or differential privacy (Dwork and Roth 2014). Note that many of these methods do not guarantee that anonymous data will not have re-identification

risk, but rather try to quantify the risk and the constraints, such as limiting queries or adding noise with additional queries (Abadi et al. 2016).

In EU-based privacy literature, the current high bar globally with respect to privacy, they no longer discuss anonymous data as a separate category, but instead simply talk about personal data and anonymized data.

So, for the data being gathered, collected, stored, and used in the experiment, the questions are:

- How sensitive is the data?
- What is the re-identification risk of individuals from the data?

As the sensitivity and risk increases, you must increase the level of data protection, confidentiality, access control, security, monitoring and auditing, and so on.