Who's the Guinea Pig? Investigating Online A/B/n Tests in-the-Wild

Shan Jiang
Northeastern University
sjiang@ccs.neu.edu

John Martin
Northeastern University
martin.john@northeastern.edu

Christo Wilson
Northeastern University
cbw@ccs.neu.edu

ABSTRACT

A/B/n testing has been adopted by many technology companies as a data-driven approach to product design and optimization. These tests are often run on their websites without explicit consent from users. In this paper, we investigate such online A/B/n tests by using Optimizely as a lens. First, we provide measurement results of 575 websites that use Optimizely drawn from the Alexa Top-1M, and analyze the distributions of their audiences and experiments. Then, we use three case studies to discuss potential ethical pitfalls of such experiments, including involvement of political content, price discrimination, and advertising campaigns. We conclude with a suggestion for greater awareness of ethical concerns inherent in human experimentation and a call for increased transparency among A/B/n test operators.

CCS CONCEPTS

• **Security and privacy** → Privacy protections; • **Human-centered computing** → User studies; Empirical studies in HCI; • **Social and professional topics** → Codes of ethics;

KEYWORDS

online controlled experiments; A/B/n testing; personalization

ACM Reference Format:
Shan Jiang, John Martin, and Christo Wilson. 2019. Who's the Guinea Pig?, Investigating Online A/B/n Tests in-the-Wild. In *FAT* '19: Conference on Fairness, Accountability, and Transparency (FAT* '19), January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287565

1 INTRODUCTION

It has been almost a century since Sir Ronald Fisher developed the theory and practice of controlled experiments. Since then, these techniques have been widely adopted by businesses and marketers to optimize everything from the layout of assembly lines, to the messaging and targeting of advertising campaigns.

A/B/*n* testing (a.k.a. split testing or bucket testing) in particular has been eagerly adopted by technology companies as a data-driven approach to product design and optimization. An A/B/*n* test is a straightforward between-subjects experimental design where the users of a website are split into n groups, with one serving as the control and the other $n-1$ as treatments. The control group is shown the product as-is, while the treatments groups are shown variations. After some time has passed, the users' behaviors are analyzed to determine which treatment, if any, had a desirable effect (e.g., more clicks, higher conversion rates, etc.).

Many large technology companies are open about their advocacy of *Online Controlled Experiments* (OCEs) like A/B/*n* testing. In 2000, Google famously used A/B tests to experiment with the presentation of results in Google Search, and by 2011 Google engineers claimed to be running over 7,000 A/B tests per year [21]. Google and other large companies like Facebook, Microsoft, and LinkedIn have published papers describing their infrastructure for supporting and scaling OCEs on their platforms [6, 34, 40, 43, 66, 67, 71, 72].

Although it is known that OCEs are widely used by the biggest technology companies, critical questions remain unanswered about their use in practice. First, are OCEs used by websites beyond the largest platforms, and if so, at what scale (e.g., how many treatments and experiments)? Second, what is the substance of the treatments that users are subject to? This second question is especially pertinent, because OCEs are a form of human experimentation. Thus, we must consider the ethics of these experiments, especially because they (1) may be conducted without explicit consent,[1] and (2) may cause a variety of harms, depending on the design being tested. Indeed, companies are not bound by frameworks such as the Belmont Report [13], and evidence suggests that the executives, engineers, and marketers who conduct OCEs may not be versed in, or even aware of, experimental ethics [50, 58].

In this study, we take a first step towards filling this knowledge gap by using Optimizely [53] as a lens. Optimizely is a service that enables website operators to build, manage, and analyze OCEs (such as A/B and multivariate tests) on their websites. It is the most popular of several services that offer similar functionality, as of mid-2018 [22]. Crucially for us, Optimizely is implemented using a client-side JavaScript library: when a user visits a website, the library dynamically determines what *audiences* (i.e., treatment groups) the user is part of, and executes any *experiments* (i.e., treatments) associated with those audiences. This design enables us to record **all audiences and experiments** that are available on a website that uses Optimizely.

We crawled 10,584 sites drawn from the Alexa Top-1M that were previously observed including resources from Optimizely [9]. Of these sites, 575 were running experiments at the time of our crawls from January 29 to March 19, 2018. Using this dataset, we examine: what kinds of websites conducts OCEs, how many audiences do

FAT '19, January 29–31, 2019, Atlanta, GA, USA*

ACM ISBN 978-1-4503-6125-5/19/01…$15.00
https://doi.org/10.1145/3287560.3287565

[1] We are not aware of any major website that prominently discloses the existence of Optimizely experiments to users or asks for affirmative consent, although it is possible that websites may disclose this practice in their terms of use or privacy policy.

they construct, what features define these audiences, and how many experiments do they run? In total, our analysis considers 1,143 audiences and 2,001 experiments.

Measurement Results. We found that the usage of Optimizely is heavily distributed over top-ranked websites. Most of these websites were conducting ≤ 5 experiments on ≤ 5 audiences, while a small number of websites were running dozens of experiments with complicated audience structure (e.g., Optimizely itself, The New York Times, and AirAsia). We analyze how websites segment audiences overall, and present detailed results for popular attributes such as *location, device,* and *browser*. Furthermore, we analyze the experiments captured in our dataset, and find that most were "dummies" (i.e., no variations) or were A/B tests (i.e., a single variation) that targeted "everyone" or audiences from a single group.

Case Studies. In addition, we also qualitatively investigate the substance of a subset of experiments (since Optimizely experiments can make arbitrary changes to web pages, there is no way to analyze them at scale). We focus on three case studies: political content, e-commerce, and advertising. Each case study is motivated by specific ethical concerns. For example, prior work has shown that changing the partisan valence of online content can influence voting behavior [28, 29]; if an OCE were to manipulate the valence of political content (e.g., news headlines), this could increase political polarization. Similarly, prior work has uncovered numerous instances of online price discrimination, including A/B testing on Expedia [35]; OCEs could also be used to manipulate consumer behavior by altering prices, changing the presentation of products to emphasize more or less expensive options, or tailor discounts to specific users.

Our hope is that this study raise awareness of the scope of OCEs in-the-wild, and fosters a conversation about the ethics of these experiments. Today, users do not affirmatively consent to the vast majority of OCEs. Instead, users are unaware, experiments are not transparent, and experimenters are not accountable to their subjects. Although our study focuses on Optimizely, we note that they are just a tool vendor; ultimately it is up to companies using this tool (and others like it) to grapple with the ethics of experiments and to obey best practices.

2 RELATED WORK

We begin by briefly surveying existing work on OCEs.

Systems. Research groups from several major tech companies have published papers describing the systems they use internally to deploy and manage OCEs. Facebook, Booking.com, Microsoft in general, and Bing specifically, have all developed systems to help their software developers design, deploy, and analyze the results of OCEs [6, 24, 34, 40, 43, 66]. These systems focus on the challenge of scaling experiments to very large platforms. Facebook's system, PlanOut, is open-source [56]. Google also has a system for managing OCEs, and their work focuses on managing and avoiding conflicts between concurrent experiments [67]. Xu et al. from LinkedIn discuss the challenges of changing corporate culture to embrace OCEs [72], and describe a system for deploying OCEs within mobile apps [71].

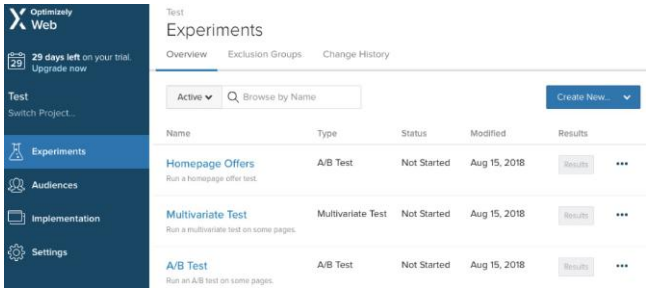

Figure 1: Optimizely Experiments UI. A user can create new experiments to run, which include A/B tests, multivariate tests, etc. A user can further specify experiment details, including design variations, targeted audience, etc.

Methods. There is a body of work focused on practical methods for conducting OCEs. The founders of Optimizely published a book that introduces the topic of A/B testing to marketers and web designers [63]. Kohavi et al. presented a denser and more academically-minded guide to implementing and analyzing A/B tests on the web [45], as well as a follow-up focused on univariate versus multivariate experimental designs [47]. Kohavi et al. published two subsequent papers that scrutinize real-world experimental case studies that generated unexpected results to illustrate subtle issues with experiment design that can impact the interpretation of results [42, 46]. For example, one case study from Bing highlighted the difficulty of choosing success criteria for experiments (a.k.a. the Overall Evaluation Criterion) that balances short- and long-term business goals. Kohavi et al. later summarized their insights into seven rules of thumb for successful OCEs [44].

A separate thread of work focused on statistical methods for increasing the power of OCEs. For example, Deng et al. presented a new variance estimation metric that relaxes assumptions about the independence participant randomization [25], while Hill et al. focus on measuring causality in experiments involving online display ads [36].

Law and Ethics. Previous work has highlighted ethical issues with specific OCEs. For example, Facebook's "emotional contagion" news feed experiments [48] and OKCupid's manipulation of their matching algorithms [41] have been criticized for their absence of user consent, as well as their lack of protection for human subjects from the perspectives of law [33] and research ethics [14].

Although the prior work we highlight here often presents case studies of specific experiments that were conducted by large companies, none present a comprehensive overview of OCEs in practice, such as how audiences are segmented, or the broad range of treatments that are tested.

3 BACKGROUND

In this study, we use the service provided by Optimizely as a lens to analyze OCEs at-scale across the web. In this section, we introduce Optimizely and describe the functionality offered by their service.

Optimizely is an American software company that provides a platform for conducting OCEs on websites. Founded in 2010, Optimizely is one of several companies that offer tools and platforms for conducting OCEs, including Google Optimize, Adobe Target, AB

Tasty, and Visual Website Optimizer. As of early 2018, Optimizely was the most widely used OCE platform (as measured by the number of websites embedding their JavaScript) [22]. Therefore, we choose Optimizely as the subject of our study.

Tool Description. At a high-level, Optimizely offers tools that enable the *operator* of a website (e.g., the developer, designer, or maintainer) to deploy A/B/*n* and multivariate tests to their websites. Optimizely offers a one-month free trial, after which the operator must pay a monthly fee to continue using the service. Additionally, Optimizely offers various tiers of functionality, with the most expensive tier unlocking more advanced forms of experimentation and larger population sizes.

After an operator signs up, the Optimizely web interface allows it to define *audiences* (i.e., treatment groups) and *experiments* (i.e., treatments).[2] Figure 1 shows a screenshot of the Optimizely interface, including several experiments that we created in our account. The left-hand rail provides links for creating audiences and experiments.

Optimizely offers operators a broad set of attributes for segmenting users into audiences. This includes: how and when a user is accessing the operator's website (e.g., their choice of browser); localization (e.g., a user's language and IP address); whether the user was funneled to the operator's website via an ad campaign or an affiliate; specific tracking and/or session cookies in the user's browser; or even custom JavaScript functions defined by the operator that execute in the user's browser. Tracking identifiers like cookies can be defined by the operator, or drawn from third-party Data Management Platforms like BlueKai. Operators may use Boolean logic to combine attributes, in order to define narrowly segmented audiences.

As shown in Figure 1, operators define experiments using Optimizely's tools. The term "experiment" is a bit of a misnomer, because it encompasses true experiments like A/B/*n* and multivariate tests, as well as statically defined personalization (e.g., a custom homepage banner shown to specific audiences in perpetuity). That said, for the sake of consistency, we adopt the term "experiment" throughout this paper. Operators can configure any number of *variations* per experiment, as well as choose to divide users randomly across variations, or to manually assign specific audiences to variations. Optimizely allows many-to-many assignment between experiments and audiences, i.e., a given experiment can include many audiences, and a given audience may be assigned to many experiments.

Operators have complete control over the treatment effect of each experimental variation. Operators define treatment effects using arbitrary snippets of HTML, CSS, and JavaScript, giving them complete freedom to manipulate the DOM of their website. Further, operators specify which pages on their website a given experiment applies to using fully qualified URLs, partial paths, or regular expressions. Optimizely provides a WYSIWYG interface for designing experiments that is powerful enough for non-programmers (e.g., marketers) to construct relatively sophisticated experiments.

Optimizely allows operators to configure the success metrics for their experiments, and their tools automatically take care of collecting the necessary telemetry (e.g., user clicks) to calculate the metrics. Optimizely provides reporting functionality that allows operators to evaluate the efficacy of their experiments.

Low-Level Implementation. For the purposes of our study, it is important to understand how Optimizely integrates with operators' websites. Once an operator has defined their audiences and experiments, Optimizely saves this information as JSON-encoded configuration files and stores them on Optimizely's CDN. The operator must then configure its website to include Optimizely's JavaScript library into their website (e.g., using a `<script src="...">` tag).

When a user visits a page on the operator's website, the Optimizely JavaScript library is downloaded and executed by the user's browser. The library then (1) fetches the JSON files that define the operator's audiences and experiments, (2) evaluates which audiences (if any) this user is a member of, (3) looks up the experimental treatments associated with these audiences, and (4) injects the associated treatment code into the web page. In the case of A/B/*n* experiments that are randomized (i.e., not associated with a particular audience), the library randomly determines which bucket to place the current user into, and injects the associated treatment code into the web page.

The way that Optimizely integrates with operators' websites is critical to our study. Because all of the audiences and experiments defined for a given website are in the JSON configuration files, and these files are available client-side, we are able to write a crawler that identifies and records these files for later analysis.[3]

4 DATA COLLECTION

In this section, we describe the process we used to gather data for this study. Recall that our goal is to analyze the audiences and experiments that website operators define in practice using Optimizely. To collect this data, we periodically crawled a large number of popular websites, detected the presence of Optimizely's JavaScript code, and if it was present, recorded the JSON configuration file containing the definitions for the audiences and experiments on that website.

Crawl Strategy. To select websites for our crawls, we relied on data from a prior crawl of the Alexa Top-1M domains that was conducted in 2016 [9]. One of the products of this crawl was a list of all third-party resources included by the Alexa Top-1M domains,[4] including references to Optimizely. 10,584 domains included at least one resource from Optimizely: these are the websites we targeted in our crawls.[5]

We implemented our crawler in PhantomJS [55], which is a headless-browser designed for automation and crawling that is based on Chromium. We configured PhantomJS to present a valid `User-Agent` for Chrome, as well as a standard-sized desktop viewport. We scripted PhantomJS to visit a web page, wait 30 seconds for the web page to load completely, then attempt to access the

[2] In the rest of the paper, we use the terms "audience" and "experiment" to refer to the objects defined by Optimizely.

[3] There are also browser extensions that allow users to interact with Optimizely experiments within their own browser. Examples include [26, 37].

[4] This crawl visited 20 pages per domain, thus providing reasonable coverage of included third-party resources.

[5] BuiltWith [16], a website profiler, estimated that there were 15,000 websites using the Optimizely library in 2016. Thus, the list from Bashir et al. [9] achieves reasonable coverage of websites that include Optimizely.

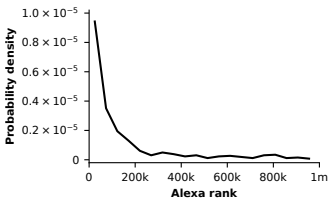

Figure 2: Optimizely usage over Alexa rank. The usage of Optimizely is heavily distributed over top-ranked websites. Lower-ranked websites were less likely to use ($r_{pb} = -0.028^{***}$).

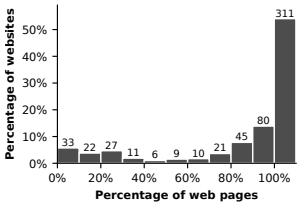

Figure 3: Web pages per site including the Optimizely library. Most websites (311 of 575, 54.1%) include Optimizely in all of their web pages in our sample.

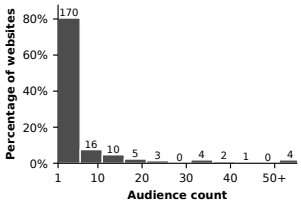

Figure 4: Distribution of audiences per website. Most websites (170 of 221, 76.9%) have defined ≤ 5 audiences, although Optimizely itself has defined 114.

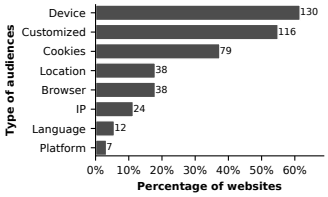

Figure 5: Popularity of audience attributes. A majority of websites (116 of 221, 52.5%) are using customize JavaScript to define audiences.

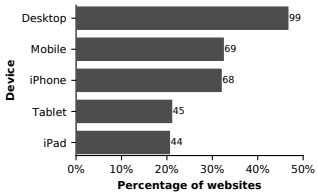

Figure 6: Device types targeted for audiences.

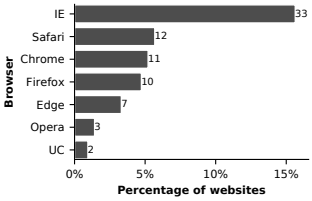

Figure 7: Browser types targeted for audiences.

Optimizely JSON configuration data by executing the following function that is made available by the Optimizely JavaScript library:

```
exp_json = optimizely.get("data");
```

If the Optimizely JavaScript library is included in the current web page, and if the website's operator has configured experiments, then `exp_json` will contain a reference to the JSON configuration data, and the crawler saves this to a file. Otherwise, an exception is generated and we record that the web page did not contain any Optimizely experiments.

Our crawler visited all 10,584 target domains once per week from January 29 to March 19, 2018. We visited the websites periodically to collect longitudinal data about their usage of Optimizely. The crawler visited the homepage for each website, as well as ≤ 19 randomly selected links on the homepage that pointed to the first-party domain.

Note that for any given domain, it is possible for zero or more web pages to include Optimizely experiments. One reason we might observe zero usage is that the website used Optimizely in 2016, but no longer does. Another potential reason is that the website uses Optimizely for analytics (e.g., as an alternative to Google Analytics), but not for experimentation; in this case the JSON configuration data will not exist. Finally, as we will show, some websites only include Optimizely within a subset of their web pages.

Ethics. We were careful to obey standard ethical practices during our data collection. The audience and experiment data we collected are high-level rule-based groups and contains no personal information. We only visit at most 20 pages per website per week, therefore our impact on the websites is minimal. Our data collection also had no impact on the experiments running on the website.

5 MEASUREMENT

In this section, we present a broad overview of Optimizely usage across popular websites, with a focus on the audiences and experiments defined by websites.

5.1 Overview

Of the 10,584 websites that including resources from Optimizely in our historical dataset (collected in 2016) [9], we found that 575 (5.4%) included the Optimizely JavaScript library in at least one of their web pages during our crawls. There are several possible reasons for this discrepancy. First, websites may have stopped using Optimizely between 2016 and 2018. Second, websites may include tracking beacons from Optimizely (e.g., for analytics purposes) but not the experimental tools that we are interested in. Third, websites may use Optimizely on portions of their site that our crawler failed to reach, due to the random nature of our crawls. For the remainder of this study, we focus on these 575 websites.

In Figure 2, we examine how the usage of Optimizely is distributed across websites sorted by popularity according to Alexa. We observe that the likelihood of using Optimizely decreases drastically with Alexa rank (Point Biserial $r_{pb} = -0.028^{***}$); 93.7% of the usage we observe occurs on the top 20% most popular websites. This observation suggests that even though Optimizely's tools are designed for non-experts, only websites with significant budgets and staff are able or willing to engage in OCEs.

Next, we examine the percentage of web pages per website that include the Optimizely JavaScript library in Figure 3. We observe that most websites (311 of 575, 54.1%) in our sample include Optimizely on all of their pages. We hypothesize that this could be the result of templated web design. The remaining websites in our sample either include Optimizely on a few of their pages (82 of 575 websites include Optimizely in < 30% of their pages) or most of their pages (146 of 575 websites include Optimizely in > 70% of their pages). The former observation suggests that there may be false negatives in our dataset, i.e., websites that use Optimizely on a small number of pages that our crawler happened to miss. The latter observation is unsurprising, since even templated website designs sometimes have exceptional pages that do not follow the design language (e.g., terms of use and privacy policies).

Note that just because a website includes the Optimizely library does not necessary mean it is actively running experiments with Optimizely. We examine the distribution of experiments in § 5.3.

5.2 Audiences

We now turn our attention to the audiences that website operators have defined for their experiments. We observe that 221 of 575 (38.4%) websites in our sample defined audiences for their experiments. Note that this does not necessarily mean the remaining 354

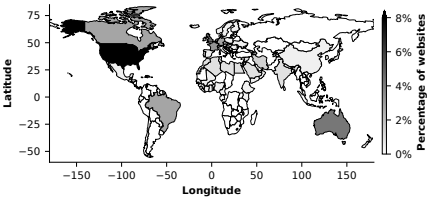

Figure 8: Countries targeted for audiences. The targeted countries are, in descending order of website usage: *US, UK, Australia, France, Germany, Canada*, etc., with 28 of 211 (13.3%) websites targeting US users.

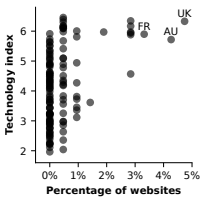

Figure 9: World location usage versus technology index. Positive correlation is found ($r = 0.370^{***}$).

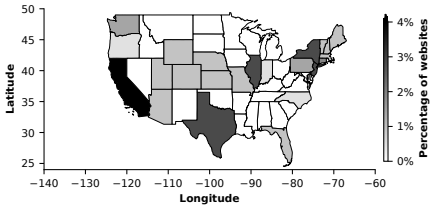

Figure 10: US states targeted for audiences. The targeted states are, in descending order of website usage: *California, Texas, New York, Illinois, New Jersey, Connecticut, Washington*, etc., with 9 of 211 (4.3%) websites targeting Californians.

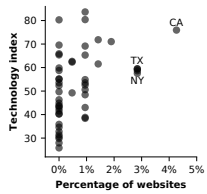

Figure 11: US location usage versus technology index. Positive correlation is found ($r = 0.426^{**}$).

websites were not running experiments; recall that Optimizely allows simple A/B/n experiments to be run with random assignment of users, i.e., no manually defined audiences are required.

Figure 4 shows the distribution of audiences per website. We observe that most websites in our sample (170 of 221, 76.9%) have defined ≤ 5 audiences. This suggests that the dominant use case for predefined audiences on Optimizely is experimentation, rather than fine-grained audience segmentation. Websites wishing to micro-target users (e.g., for advertising and recommendation) are better off using machine learning tools that automatically infer user segments, rather than manually defining complex conditions using Optimizely predefined audiences.

That said, we do observe four websites that each defined a large number of audiences: Optimizely (114 audiences), The New York Times (90), AirAsia (79) and CREDO Mobile (64). We examine two of these in-depth in § 6.

Recall that audiences are defined using one or more attributes. Figure 5 presents the percentage and count of websites in our sample that have defined at least one audience with specific attributes. The eight most popular attributes are, in descending order of website usage: *device, custom JavaScript, cookies, location, browser, IP address, language,* and *platform*. That a majority of websites in our sample (116 of 221, 52.5%) are leveraging custom functions to define audiences suggests a high level of sophistication by these operators.

We now investigate audience segmentation by *device, browser,* and *location*. Figure 6 shows the percentage and count of websites targeting different types of *devices*. We observe that operators are primarily interested in separating desktop and mobile users, and relatively less interested in subdividing smartphone and tablet users. We hypothesize that iPhone and iPad users are frequently segmented because it is easier to identify them as classes (the hardware is more homogeneous), versus the Android device ecosystem which is heavily fragmented across manufacturers. Alternatively, website operators may view Apple users as "high value" (i.e., as a proxy for affluence), and thus segment them for specialized targeting [35, 50]. The website that uses device targeting most heavily is CDW, a company that sells these technology products, which has seven audiences of devices.

Figure 7 presents the popularity of targeting different *browsers*, were we make two interesting observations. First, it is surprising that Internet Explorer is the most frequently targeted browser. We suspect that operators are using Optimizely to selectively apply compatibility patches to their websites for people using the defunct Internet Explorer browser. We manually inspected the two websites targeting UC browser users and found they are also doing so for compatibility reasons. Second, it is surprising that Safari is more frequently targeted than Chrome, since Chrome is the most popular web browser. Unlike Internet Explorer, Safari is a modern and standards-compliant browser, so compatibility patching is unlikely to be the reason why it is targeted. Instead, Safari may be highly targeted for the same reason as iPhone and iPad users, i.e., as a means to segment the valuable Apple-user population.

We observe that segmenting audiences by *location* is also popular. Optimizely allows a website operator to segment users at different levels of geographic granularity, including: continents, countries, states, and cities. We first examine how the audiences in our sample are distributed at the country-level. As shown in Figure 8, audiences are primarily localized within the following countries, in descending order of popularity: *US, UK, Australia, France, Germany, Canada*, etc., with 28 of 211 (13.3%) of websites in our sample targeting a US audience. As one possible explanation for why specific countries are more targeted than others, we cross reference our audience distribution with the 2017–2018 *technology index* of each country from The World Bank [8], which is a measure of geographic technological adoption. The resulting scatterplot is shown in Figure 9. We find a significant and positive correlation between the technology index and audience distribution (Pearson $r = 0.370^{***}$), which supports our hypothesis that operators may focus on localized audiences from technologically developed countries.

Next, we focus on how audiences are segmented within the US, since it is the most popular country in our dataset. As shown in Figure 10, the most popular states are, in descending order: *California, Texas, New York, Illinois, New Jersey, Connecticut, Washington*, etc., with 9 of 211 (4.3%) websites in our sample segmenting Californians. Similar to above, we cross reference this distribution with the technology index of each state from the 2016–2017 State Technology and Science Index [68]. The scatterplot is shown in Figure 11, and again we find a significant and positive correlation between technology index and audience distribution (Pearson $r = 0.426^{**}$). Websites that heavily target locations include PG&E with 20 audiences and Teacher Certification Map with 19 audiences.

5.3 Experiments

Next, we shift focus to the experiments *per se* that are being run by websites. Of the 575 websites that used Optimizely in our sample, we observed 297 (51.7%) of them running experiments. The remaining websites may just be using Optimizely for analytics purposes, or they may have had a lull in experiments at the time of our crawls.

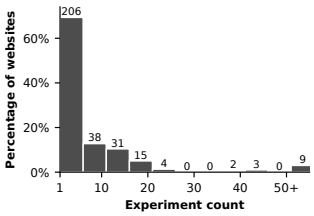

Figure 12: Distribution of experiments per website. 206 of 297 (69.4%) websites are running ≤ 5 experiments. Nine websites are running ≥ 50 experiments, with the max being 127 by Optimizely itself.

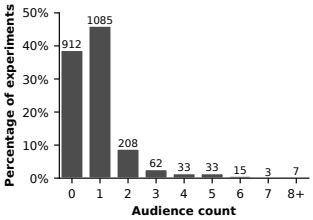

Figure 13: Distribution of audiences per experiment. 912 of 2358 (38.7%) of experiments have no audiences, which corresponds to all site visitors. 1085 of 2358 (46.0%) have a single audience.

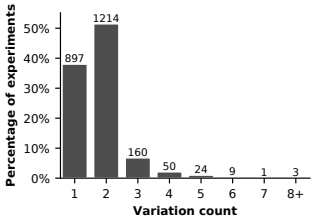

Figure 14: Distribution of variations per experiment. 897 of 2358 (38.0%) of experiments have a single variation, which corresponds to "dummy" experiments. 1214 of 2358 (51.5%) are two variation A/B tests.

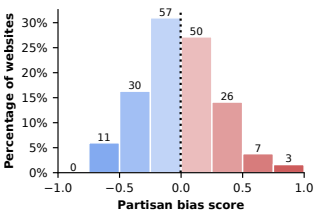

Figure 15: Partisan bias scores of websites using Optimizely. 184 of 575 (32.0%) websites are scored. The distribution is normal (D'Agostino test, $p = 0.702$) and centered around 0 ($\mu = -0.004$, T-test, $p = 0.876$).

Figure 12 shows the number of experiments per website. We observe that the majority of websites in our sample (206 of 297, 69.4%) are running ≤ 5 experiments. This is unsurprising: experiments take time to plan, develop, and evaluate. Although the largest tech companies like Google, Facebook, and Microsoft run thousands of experiments per year [6, 21, 43], it is unreasonable to expect smaller companies to match their pace.

We do observe nine websites running ≥ 50 experiments each. Four are the same as the websites with the most audiences groups: Optimizely (127 experiments), The New York Times (91), AirAsia (87), and CREDO mobile (76). Each of them runs a large number of experiments corresponding to their many audiences. One notable exception is Teespring, which we observe running 76 experiments with only 2 audience groups. Most (74 of 76, 97.4%) of Teespring's experiments are running without any audiences, meaning all site visitors are eligible to be experimental subjects.

Next, we investigate the number of audiences per experiment. As shown in Figure 13, 912 of 2358 (38.7%) experiments in our sample target no audiences, which means all site visitors are eligible to be experimental subjects. Note that these experiments may still have multiple variations, in which case users would be randomly divided across them. 1085 of 2358 (46.0%) experiments have a single audience, which corresponds to a two variation A/B test where the B group is the audience members, and the A group is all other site visitors. The remaining experiments (361 of 2358, 15.3%) include multiple audiences. The experiments with the most complex audience specifications (≥ 6, totally 25) in our sample are mostly personalizations from AirAsia (11 of 25, 44.0%) and The New York Times (9 of 25, 36.0%).

Finally, we investigate variations per experiment, i.e., the number of different treatments shown to subjects in each experiment. As shown in Figure 14, 897 of 2358 (38.0%) of experiments in our sample have a single variation, which corresponds to "dummy" experiments were all site visitors experience the same treatment. These "dummies" may represent the final stage of successful experiments, e.g., the operator determined that a specific treatment in an A/B/n experiments was most effective, and they now apply it to all site visitors. 1214 of 2358 (51.5%) have two variations, which may correspond to A/B tests. The remaining experiments (247 of 2358, 10.4%) have more variations, which may correspond to multivariate tests. The experiments with the most variations (≥ 7, totally 4) in our sample are run by the websites iCracked,VapeWorld, Shutterstock,and the Department of Motor Vehicles (DMV).

6 CASE STUDIES

As we observe in § 5, Optimizely is used across many popular websites that receive millions of visitors. The operators of these websites frequently segment users into audiences, and conduct thousands of experiments on users without their knowledge or explicit consent. Although it is likely that many of these experiments are utterly banal and completely benign (e.g., simple changes to the layout and styling of the website), Optimizely is a powerful tool that can be used to implement arbitrary changes to a website. Thus, we must consider whether the experiments that operators are conducting in practice have the potential to harm participants.

Because the capabilities of Optimizely's tool are so general, we cannot feasibly analyze the treatment effects of experiments at scale. Instead, in this section we present case studies that highlight specific types of experiments we have observed in our data, along with potential risks associated with these designs. We divide our case studies into three areas: *political content, price discrimination,* and *advertising campaigns.*

6.1 Political Content

There are numerous examples demonstrating that the confluence of technology and politics can lead to undesirable consequences. Eli Pariser famously pointed out that personalization algorithms could potentially lead to the formation of partisan "filter bubbles" that reinforce each person's preexisting political beliefs [30, 54]. Two studies have demonstrated that partisan information shown in search results can impact the voting preferences of users [28, 29]. Finally, exposure to charged political content has been found to correlate with negative expressions of emotion from users, suggesting that it may also impact their internal emotional state [39].

In our dataset, we observe many overtly political websites running experiments using Optimizely. To quantify this, we adopt a mapping of websites to partisan bias scores developed by Robertson et al. [61]. These scores are calculated based on the relative frequency that a given website is shared on Twitter by registered Democrats and Republicans. A score of -1 (1) indicates that a given website is shared exclusively by Democrats (Republicans). Although there are other partisanship scoring metrics available, such as Media Bias/Fact Check [20], AllSides [2], and several from prior papers [7, 15, 60], we adopt the Robertson et al. mapping because it

Table 1: Partisan bias scores and corresponding examples of websites using Optimizely. The most partisan websites tend to have overtly political connotations, such as LGBTQ rights (e.g., Human Rights Campaign), abortion (e.g., Planned Parenthood), and environmentalism (e.g., Sierra Club) on the left, and conservativism (e.g., The Heritage Foundation) and religion (e.g., Compassion in Jesus' Name) on the right. Several left-leaning news sources (e.g., The New York Times and CNN) also use Optimizely. All of these websites may be able to shape visitor's political beliefs through experimentation.

Party	Bias score	Examples		
Left	[-1, -0.75)	-	-	-
	[-0.75, -0.5)	Human Rights Campaign (www.hrc.org)	Sierra Club (www.sierraclub.org)	Planned Parenthood (www.plannedparenthood.org)
	[-0.5, -0.25)	The New York Times (www.nytimes.com)	Springer (www.springer.com)	edX (www.edx.org)
	[-0.25, 0)	CNN (www.cnn.com)	Flipboard (flipboard.com)	ABC News (abcnews.go.com)
Right	(0, 0.25]	History (www.history.com)	Legacy (www.legacy.com)	ABC Shows (abc.go.com)
	(0.25, 0.5]	ESPN (www.espn.com)	A&E (www.aetv.com)	United Service Organizations (www.uso.org)
	(0.5, 0.75]	PGA Championship (www.pga.com)	6 Pack Bags (www.sixpackbags.com)	Safelite AutoGlass (www.safelite.com)
	(0.75, 1]	The Heritage Foundation (www.heritage.org)	Heritage Action (heritageaction.com)	Compassion in Jesus' Name (www.compassion.com)

covers the largest number of unique websites.[6] Using this dataset, we are able to score 184 of 575 (32.0%) of the websites in our sample.

Table 1 shows examples of scored websites. Not all scored websites using Optimizely are overtly political: for example, links to ESPN just happens to be shared more frequently on Twitter by registered Republicans, possibly reflecting a different set of cultural interests than registered Democrats. These non-political websites tend to have partisan bias scores in the range $[-0.5, 0.75]$. In contrast, websites with extreme bias scores (< -0.5 or > 0.75) are overtly political, such as Human Rights Campaign, Planned Parenthood, and Sierra Club on the left, and The Heritage Foundation and Compassion in Jesus' Name on the right. These websites, together with major news sources (e.g., The New York Times and CNN), have the potential to influence visitor's political opinions using OCEs.

We find that partisan bias scores for websites using Optimizely are normally distributed (D'Agostino normality test, $p = 0.702$) and centered around 0 (mean $\mu = -0.004$; no evidence for non-zero mean, T-test, $p = 0.876$), as shown in Figure 15. This suggests that the use of Optimizely is relatively balanced across the political spectrum.

Next, we discuss a specific class of experiments being run by websites in our sample to highlight potential ethical issues.

The New York Times. The New York Times (NYT) is a newspaper with a slight left lean (partisan bias score, -0.260) that uses Optimizely heavily, running 91 experiments with 90 defined audiences in our sample.

We observe that the NYT frequently conducts randomized A/B/n tests for headlines of news stories on their homepage [17]. For example, at 1pm on August 20, 2018, the NYT experimented with two headlines for a story about the Pope: "Pope Condemns 'Atrocities' of Abuse in Letter to Catholics" and "In Letter to 1.2 Billion Catholics, Pope Cites Sex Abuse 'Atrocities'". These experiments come and go rapidly; this example was gone by 2pm, and all visitors were shown the latter headline.

The practice of A/B/n testing headlines encapsulates competing priorities between the business and editorial sides of the newsroom. On one hand, headlines that receive more clicks generate more advertising revenue. On the other hand, the "clickiest" headline may not be the most informative or nuanced from a reporting perspective. The rise of "click-bait" headlines vividly illustrates this tension [19, 57]. Depending on what metric is being optimized for, A/B/n testing headlines may actually prove detrimental to the long term credibility of the newsroom [42] by privileging inflammatory, decontextualized, or highly partisan headlines.

Another issue with A/B/n testing headlines is that it complicates the idea of mass media as a shared frame for the public to contextualize events. Prior work demonstrates that people frequently do not read news articles, preferring instead to scan the headlines [31, 32]. Further, studies have shown that even when people do read an article, the framing of the headline shapes their opinion of the whole story [27, 31]. Combined, these two effects suggest that even when readers engage with a single news outlet, they may walk away with widely diverging understandings of events.

To be clear: we do not observe the NYT experimenting with click-bait headlines in our data, nor are we accusing the NYT of intentionally crafting partisan headlines. Our goal with this case study is simply to highlight that (1) A/B/n testing news headlines may have unintended consequences, and that (2) the tools for conducting these experiments are widely available and accessible to non-technical website operators. Unscrupulous news outlets could easily test unethical headlines, and users would be unaware. Even scrupulous news outlets may inadvertently fan the flames of partisanship if they uncritically allow the results of A/B/n tests to determine the framing of their news headlines.

6.2 Price Discrimination

Another area of concern associated with personalization is online price discrimination. This refers to the practice of offering different prices to different customers for the same product. In most cases, price discrimination is not illegal; to economists it is actually desirable, as it allows businesses to capture additional value from consumer surplus. However, consumers have shown discomfort with online price discrimination when the practice is non-transparent, i.e., people feel cheated when they discover that others had access to better prices than they did [58]. Several studies have identified price discrimination across a broad swath of e-commerce websites [35, 51, 52].

To identify websites in our sample that are potentially conducting price discrimination experiments, we searched for the keywords "$," "price," and "sale" in the experimental treatments.[7] In total, we observed 40 websites with 117 experiments containing at least one of these keywords. Next, we delve into several specific websites that implement price discrimination through Optimizely experiments.

[6] Robertson et al. report that their scores have high correlation with the other scoring systems mentioned here [61].

[7] OCEs involving price discrimination usually contains at least one of these keywords, however, containing these keywords does not necessarily indicate price discrimination. Therefore the statistics we report should be treated as upper bounds.

PolicyGenius. PolicyGenius is an online insurance policy company that we observe running 13 experiments with 13 defined audiences in our dataset. A group of its experiments, called "Marketing Life Landing Page Test," targets audiences based on Urchin Tracking Module (UTM) parameters, which are parameters passed by Google's analytics and advertising services. Audiences with different UTMs are shown different insurance advertisements, including several of the form "Term Life Insurance As Low As X Per Month," where X can be \$9.99, \$10, or \$29.[8] Other treatments use text of the form "Compare term life insurance quotes and save up to 40%" and "Compare and apply for term life insurance online," i.e., with and without discount offers.

It is possible that some of the treatments being tested by PolicyGenius are merely marketing differences, rather than price discrimination. For example, maybe all customers can receive 40% discounts, but only some customers are shown advertisements highlighting the deal. However, even in this case there may be adverse impact on customers: those who are unaware of the possibility for discounts may be less likely to claim them, and thus end up paying more.

The pricing treatments are a clearer case of price discrimination (e.g., offering the same insurance policy for \$10 and \$29 per month), but without deeper investigation it is unclear whether this particular instance is unethical. For example, the higher-priced policies may be targeted to senior citizens. However, prior investigations have found systematic racial discrimination in insurance [4] and ridesharing [18, 38, 49, 69] markets linked to geolocation. If the audience segments used by PolicyGenius are localized, or include other sensitive demographic characteristics, this price discrimination would be similarly troubling.

VapeWorld. VapeWorld is an e-commerce website that sells vaporizers to adults. We observe it running 19 experiments with 6 defined audiences in our dataset. A group of its experiments offer lower prices on specific merchandise to audiences in California, or from within the US. This case is similar to geolocation-based online price discrimination that was discovered on Staples that privileged customers from affluent urban areas [70].

6.3 Advertising Campaigns

Numerous studies have documented the potential for discrimination in online advertising [64]. This may occur due to exclusion (e.g., not showing ads for jobs [23] or housing to specific populations [3, 5]), or through association (e.g., showing criminal background check ads tied to the names of African American individuals [65]).

Online advertising-related experiments are the single most common treatment we observe in our dataset. We identified experiments related to advertising by searching for a list of domains (gathered from prior work [9]) known to be involved in tracking users and serving ads (e.g., doubleclick.net) in all experimental treatments in our dataset. We find that 123 of 575 (21.4%) websites in our sample are conducting experiments that involve injecting ad and tracking-related resources into their web pages.

Figure 16 shows the distribution of advertising companies per website in our sample. We observe that most websites (85 of 123, 69.1%) are only experimenting with a single company, while a quarter (33 of 123, 26.8%) are experimenting with two or three.

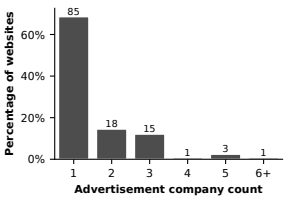

Figure 16: Distribution of advertising companies per website.

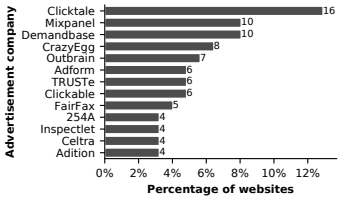

Figure 17: Top advertising companies observed in Optimizely experiments.

Figure 17 shows the top advertising companies that appear in Optimizely experiments. The advertising companies used by less than four websites are not shown. Across all 575 websites in our sample, we observe 78 unique advertisers out of 984 (7.9%) in our full list of advertisers. None of the companies shown in Figure 17 are massive players in the industry (as compared to Google, Facebook, etc. [12]), which suggests that websites in our sample are experimenting with alternative or highly specialized monetization strategies, possibly to supplement their primary sources of ad revenue.

Optimizely. We observe that Optimizely itself is running 127 experiments with 114 defined audiences in our sample. 75 of 127 (59.1%) of these experiments are using advertising services from Demandbase. These experiments target audiences from specific companies, including Petco, Adidas, Target, Centers for Medicare & Medicaid Services, etc. Each treatment places a different banner picture on the homepage that is specifically designed to attract users from each company, e.g., the banner picture for Petco is a dog, while for Centers for Medicare & Medicaid Services it is a picture of doctor/patient communication, etc.

To the best of our knowledge, none of these organizations were customers of Optimizely at the time of our data collection. Rather, it appears that Demandbase works with Optimizely to identify visitors to their website that demonstrate an interest in buying the service. Optimizely then crafts custom, branded homepages for these potential customers.

While the experiments we highlight on Optimizely are not obviously unethical, they do exemplify just how highly targeted modern online advertising can be. It is easy to envision contexts where such micro-targeting could be creepy (e.g., on a health care website) or ethically problematic (e.g., on mortgage company website).

CNN. CNN is a televised news service that we observe running 20 experiments with 11 audiences in our dataset. In these experiments, users with different cookies are shown ads from two different *Content Recommendation Networks* (CRNs): Outbrain and Taboola. One possible explanation for this behavior is that CNN is evaluating the revenue it earns from each CRN before deciding on a permanent partner. Another possibility is that CNN directs each user towards the CRN that has the most extensive tracking profile on that individual, since this would maximize revenue per impression (more specific targeting typically yields higher profits). Prior work has pointed out that Outbrain and Taboola both have a history of acting unethically by failing to prominently disclose paid advertising [10].

[8] Unfortunately, UTM parameters are opaque IDs, therefore we cannot deduce what audience segments are being targeted.

7 DISCUSSION

In this study, we leverage Optimizely as a lens to present the first large-scale observational study of OCEs in-the-wild. Of the Alexa Top-1M domains, we found 575 websites that included Optimizely's JavaScript library. The majority of these websites were conducting ≤ 5 experiments on ≤ 5 audiences at the time of our crawls, but we observe a small number of websites with extensive suites of experiments. We also analyze the most common attributes used to target audiences, such as by geolocation, device, browser, etc.

We delve into the specific treatments we observe in our dataset through a series of three case studies. These case studies highlight problematic classes of experiments (e.g., news headline optimization and price discrimination) that may raise ethical concerns, since they have the potential to harm experiment subjects. That said, it bears repeating that we do not observe **any** websites engaging in overtly unethical behavior in our dataset. The intent of our case studies is not to shame bad actors. Rather, the point is to highlight real-world use cases for A/B testing tools like Optimizely, and facilitate a discussion about the social consequences of these experiments.

Towards Transparency and Consent. We observe that Optimizely is used by many extremely popular websites. However, to our knowledge, visitors to these sites are never asked to explicitly consent to these experiments. Even the existence of experiments is rarely, if ever disclosed. For example, the NYT does not mention Optimizely or OCE's in their privacy policy or terms of use.[9]

An explicit goal of our work is to raise awareness among the public about the existence of OCEs, and push website operators towards greater transparency. We argue that transparency is one of the best defenses against harm to users, as exemplified by laws like the General Data Protection Regulation [62]. Additionally, companies should follow the same norms as academia and obtain informed consent before running substantive experiments. Alternatively, in cases where pre-disclosure may confound an experiment (e.g., by priming users), users should be debriefed after the fact. Tool providers like Optimizely could encourage these best practices by asking their users to self-certify that they are obeying ethical norms. Any operator found to not be following these practices could be cut-off from Optimizely's service.

Experimental Ethics for Practitioners. Optimizely's tools (and others like it) are designed to be accessible to a broad range of people (e.g., web designers and marketers). On one hand, it is nice to see tools that democratize powerful capabilities beyond software engineers. On the other hand, with power comes responsibility. There is never a guarantee that the people who use Optimizely or similar tools will have received training in the ethics of experimentation. As tools like Optimizely make experimentation accessible, the likelihood of situations where untrained or careless operators conduct experiments that harm subjects increases.

It may be incumbent on tool providers like Optimizely to provide ethics training to their userbase. For example, Optimizely could require new users to go through online training modules that introduce basic concepts like beneficence, justice, respect for persons, and informed consent. This would integrate seamlessly with Optimizely's existing training modules [1] and synergize nicely with self-certification of compliance with best practices.

Limitations. Our work has two major limitations. First, although Optimizely has the largest share of the OCE market, it is not used by large platforms like Google or Facebook. These services are visited by billions of users every day and they are running experiments, which unfortunately we cannot analyze. Second, we only examined the use of Optimizely's built-in audiences. However, 53% of websites in our sample are targeting customized audiences, which are opaque to us. These customized audiences could potentially target individuals using sensitive attributes or inferred "interests" from third-party data brokers [11], which may also raise ethical concerns.

Future Work. We hope that future work will extend our study by (1) auditing other A/B testing platforms (e.g., Google Optimize), (2) conducting more focused analysis on the experimental treatments being implemented by specific website verticals (e.g., e-commerce), and (3) surveying website operators to understand if and how they grapple with ethical issues in OCEs.

[9] As of August 2018 they do mention headline tests in a blog post [17].

ACKNOWLEDGMENTS

We thank Dillon Reisman and Arvind Narayanan for the inspiration for this work [59], as well as the anonymous reviewers for their helpful comments. This research was supported in part by NSF grant IIS-1553088. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

[1] Optimizely Academy. 2018. (2018). https://www.optimizely.com/academy
[2] AllSides. 2018. (2018). https://www.allsides.com/media-bias/media-bias-ratings
[3] Julia Angwin and Terry Parris Jr. 2016. Facebook Lets Advertisers Exclude Users by Race. ProPublica. (Oct. 2016). https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race
[4] Julia Angwin, Jeff Larson, Lauren Kirchner, and Surya Mattu. 2017. Minority Neighborhoods Pay Higher Car Insurance Premiums Than White Areas With the Same Risk. ProPublica. (April 2017). https://www.propublica.org/article/minority-neighborhoods-higher-car-insurance-premiums-white-areas-same-risk
[5] Julia Angwin, Ariana Tobin, and Madeleine Varner. 2017. Facebook (Still) Letting Housing Advertisers Exclude Users by Race. ProPublica. (Nov. 2017). https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin
[6] Eytan Bakshy, Dean Eckles, and Michael S. Bernstein. 2014. Designing and Deploying Online Field Experiments. In *Proc. of WWW*.
[7] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.
[8] The World Bank. 2018. (2018). https://data.worldbank.org
[9] Muhammad Ahmad Bashir, Sajjad Arshad, William Robertson, and Christo Wilson. 2016. Tracing Information Flows Between Ad Exchanges Using Retargeted Ads. In *Proc. of USENIX Security Symposium*.
[10] Muhammad Ahmad Bashir, Sajjad Arshad, and Christo Wilson. 2016. Recommended For You: A First Look at Content Recommendation Networks. In *Proc. of IMC*.
[11] Muhammad Ahmad Bashir, Umar Farooq, Maryam Shahid, Muhammad Fareed Zaffar, and Christo Wilson. 2019. Quantity vs. Quality: Evaluating User Interest Profiles Using Ad Preference Managers. In *Proc of NDSS*.
[12] Muhammad Ahmad Bashir and Christo Wilson. 2018. Diffusion of User Tracking Data in the Online Advertising Ecosystem. In *Proc. of PETS*.
[13] Belmont Report 1979. The Belmont Report. U.S. Office for Human Research Protections. (1979). https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html
[14] Raquel Benbunan-Fich. 2017. The ethics of online research with unsuspecting users: From A/B testing to C/D experimentation. *Research Ethics* 13, 3-4 (2017),

200–218.
[15] Ceren Budak, Sharad Goel, and Justin M Rao. 2016. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly* 80, S1 (2016), 250–271.
[16] BuiltWith. 2018. Optimizely Usage Statistics. (2018). https://trends.builtwith.com/analytics/Optimizely
[17] Mark Bulik. 2016. Which Headlines Attract Most Readers? The New York Times. (June 2016). https://www.nytimes.com/2016/06/13/insider/which-headlines-attract-most-readers.html
[18] Joe Castiglione, Tilly Chang, Drew Cooper, Jeff Hobson, Warren Logan, Eric Young, Billy Charlton, Christo Wilson, Alan Mislove, Le Chen, and Shan Jiang. 2016. TNCs Today: A Profile of San Francisco Transportation Network Company Activity. *San Francisco County Transportation Authority Report* (June 2016).
[19] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media. In *Proc. of ASONAM*.
[20] Media Bias/Fact Check. 2018. (2018). https://mediabiasfactcheck.com
[21] Brian Christian. 2012. The A/B Test: Inside the Technology That's Changing the Rules of Business. Wired. (April 2012). https://www.wired.com/2012/04/ff-abtesting/
[22] Datanyze 2018. Market Share Category: A/B Testing. Datanyze. (2018). https://www.datanyze.com/market-share/testing-and-optimization/
[23] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. In *Proc. of PETS*.
[24] Alex Deng, Pavel Dmitriev, Somit Gupta, Ron Kohavi, Paul Raff, and Lukas Vermeer. 2017. A/B Testing at Scale: Accelerating Software Innovation. In *Proc. of SIGIR*.
[25] Alex Deng, Jiannan Lu, and Jonthan Litz. 2017. Trustworthy Analysis of Online A/B Tests: Pitfalls, Challenges and Solutions. In *Proc. of WSDM*.
[26] dill.reisman. 2016. Pessimizely. Chrome Web Store. (2016). https://chrome.google.com/webstore/detail/pessimizely/kkkmbamdihcpdhgckbkgmiffhaejmdji
[27] Ullrich K. H. Ecker, Stephan Lewandowsky, Ee Pin Chang, and Rekha Pillai. 2014. The effects of subtle misinformation in news headlines. *Journal of Experimental Psychology: Applied* 20, 4 (2014), 323–335.
[28] Robert Epstein and Ronald E Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences* 112, 33 (2015), E4512–E4521.
[29] Robert Epstein, Ronald E Robertson, David Lazer, and Christo Wilson. 2017. Suppressing the search engine manipulation effect (SEME). *Proceedings of the ACM: Human-Computer Interaction* 1 (2017), 42.
[30] Seth Flaxman, Sharad Goel, and Justin M Rao. 2016. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly* 80, S1 (2016), 298–320.
[31] Renae Franiuk, Jennifer L. Seefelt, and Joseph A. Vandello. 2008. Prevalence of Rape Myths in Headlines and Their Effects on Attitudes Toward Rape. *Sex Roles* 58, 11–12 (June 2008), 790–801.
[32] M. Glenski, C. Pennycuff, and T. Weninger. 2017. Consumers and Curators: Browsing and Voting Patterns on Reddit. *IEEE Transactions on Computational Social Systems* 4, 4 (Dec 2017), 196–206.
[33] James Grimmelmann. 2015. The law and ethics of experiments on social media users. *J. on Telecomm. & High Tech. L.* 13 (2015), 219.
[34] Somit Gupta, Sumit Bhardwaj, Pavel Dmitriev, Lucy Ulanova, Aleksander Fabijan, and Paul Raff. 2018. The Anatomy of a Large-Scale Online Experimentation Platform. In *Proc. of IEEE ICSA*.
[35] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. 2014. Measuring price discrimination and steering on e-commerce web sites. In *Proc. of IMC*.
[36] Daniel N. Hill, Robert Moakler, Alan E. Hubbard, Vadim Tsemekhman, Foster Provost, and Kiril Tsemekhman. 2015. Measuring Causal Impact of Online Actions via Natural Experiments: Application to Display Advertising. In *Proc. of KDD*.
[37] jaggli. 2016. Controllizely. Chrome Web Store. (2016). https://chrome.google.com/webstore/detail/controllizely/iihlmdifgondmobahflhldmbpfelggoe
[38] Shan Jiang, Le Chen, Alan Mislove, and Christo Wilson. 2018. On Ridesharing Competition and Accessibility: Evidence from Uber, Lyft, and Taxi. In *Proc. of WWW*.
[39] Shan Jiang and Christo Wilson. 2018. Linguistic Signals under Misinformation and Fact-Checking: Evidence from User Comments on Social Media. *Proceedings of the ACM: Human-Computer Interaction (PACMHCI)* 2, CSCW (November 2018).
[40] Katja Kevic, Brendan Murphy, Laurie Williams, and Jennifer Beckmann. [n. d.]. Characterizing Experimentation in Continuous Deployment: A Case Study on Bing. In *Proc. of ICSE: Software Engineering in Practice Track (ICSE-SEIP '17)*.
[41] Emil OW Kirkegaard and Julius D Bjerrekær. 2016. The OKCupid dataset: A very large public dataset of dating site users. *Open Differential Psychology* 46 (2016).
[42] Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, and Ya Xu. 2012. Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained. In *Proc. of KDD*.
[43] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online Controlled Experiments at Large Scale. In *Proc. of KDD*.
[44] Ron Kohavi, Alex Deng, Roger Longbotham, and Ya Xu. 2014. Seven Rules of Thumb for Web Site Experimenters. In *Proc. of KDD*.
[45] Ron Kohavi, Randal M. Henne, and Dan Sommerfield. 2007. Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO. In *Proc. of KDD*.
[46] Ron Kohavi and Roger Longbotham. 2010. Unexpected Results in Online Controlled Experiments. In *Proc. of KDD*.
[47] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery* 18, 1 (2009), 140–181.
[48] Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* (2014), 201320040.
[49] Kyungmin Brad Lee, Marcus Bellamy, Nitin Joglekar, Shan Jiang, and Christo Wilson. 2018. Surge Pricing on a Service Platform under Spatial Spillovers: Evidence from Uber. *SSRN* 3261811, 10 (2018).
[50] Dana Mattioli. 2012. On Orbitz, Mac Users Steered to Pricier Hotels. The Wall Street Journal. (2012). https://www.wsj.com/articles/SB10001424052702304458604577488822667325882
[51] Jakub Mikians, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. 2012. Detecting price and search discrimination on the internet. In *Proc. of HotNets*.
[52] Jakub Mikians, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. 2013. Crowd-assisted search for price discrimination in e-commerce: First results. In *Proceedings of the ninth ACM conference on Emerging networking experiments and technologies*. acm, 1–6.
[53] Optimizely. 2018. (2018). https://www.optimizely.com
[54] Eli Pariser. 2012. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin Books.
[55] PhantomJS. 2018. (2018). http://phantomjs.org
[56] PlanOut 2017. PlanOut: A framework for online field experiments. Facebook, Inc.. (2017). https://facebook.github.io/planout/.
[57] Martin Potthast, Tim Gollub, Kristof Komlossy, Sebastian Schuster, Matti Wiegmann, Erika Patricia Garces Fernandez, Matthias Hagen, and Benno Stein. 2018. Crowdsourcing a Large Corpus of Clickbait on Twitter. In *Proc. of COLING*.
[58] Anita Ramasastry. 2005. Web sites change prices based on customers' habits. CNN. (June 2005). http://edition.cnn.com/2005/LAW/06/24/ramasastry.website.prices/
[59] Dillon Reisman. 2016. A Peek at A/B Testing in the Wild. Freedom to Tinker Blog. (May 2016). https://freedom-to-tinker.com/2016/05/26/a-peek-at-ab-testing-in-the-wild/
[60] Filipe Nunes Ribeiro, Lucas Henrique, Fabrício Benevenuto, Abhijnan Chakraborty, Juhi Kulshrestha, Mahmoudreza Babaei, and Krishna P Gummadi. 2018. Media Bias Monitor: Quantifying Biases of Social Media News Outlets at Large-Scale.. In *ICWSM*. 290–299.
[61] Ronald E. Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM: Human-Computer Interaction (PACMHCI)* 2 (November 2018). Issue CSCW.
[62] Adam Satariano. 2018. What the G.D.P.R., EuropeâĂŹs Tough New Data Law, Means for You. The New York Times. (May 2018). https://www.nytimes.com/2018/05/06/technology/gdpr-european-privacy-law.html
[63] Dan Siroker and Pete Koomen. 2013. *A/B Testing: The Most Powerful Way to Turn Clicks Into Customers* (1 ed.). Wiley.
[64] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P Gummadi, Patrick Loiseau, and Alan Mislove. 2018. Potential for Discrimination in Online Targeted Advertising. In *Proc. of FAT**. 5–19.
[65] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery. *ACM Queue* 11, 3 (April 2013).
[66] Chunqiang Tang, Thawan Kooburat, Pradeep Venkatachalam, Akshay Chander, Zhe Wen, Aravind Narayanan, Patrick Dowell, and Robert Karl. 2015. Holistic Configuration Management at Facebook. In *Proc. of SOSP*.
[67] Diane Tang, Ashish Agarwal, Deirdre OâĂŹBrien, and Mike Meye. 2010. Overlapping Experiment Infrastructure: More, Better, Faster Experimentation. In *Proc. of KDD*.
[68] State Technology and Science Index. 2017. (2017). http://statetechandscience.org
[69] Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. 2017. Toward a geographic understanding of the sharing economy: Systemic biases in UberX and TaskRabbit. *ACM Transactions on Computer-Human Interaction (ToCHI)* 24, 3 (2017), 21.
[70] Jennifer Valentino-DeVries, Jeremy Singer-Vine, and Ashkan Soltani. 2012. Websites Vary Prices, Deals Based on Users' Information. The Wall Street Journal. (Dec. 2012). https://www.wsj.com/articles/SB10001424127887323777204578189391813881534
[71] Ya Xu and Nanyu Chen. 2016. Evaluating Mobile Apps with A/B and Quasi A/B Tests. In *Proc. of KDD*.
[72] Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin. 2015. From Infrastructure to Culture: A/B Testing Challenges in Large Scale Social Networks. In *Proc. of KDD*.