Minds and Machines
https://doi.org/10.1007/s11023-023-09644-y

The Ethics of Online Controlled Experiments (A/B Testing)

Andrea Polonioli[1,2] · Riccardo Ghioni[2,3] · Ciro Greco[4] · Prathm Juneja[5] · Jacopo Tagliabue[4,6] · David Watson[7] · Luciano Floridi[2,8]

Received: 12 November 2022 / Accepted: 20 August 2023

Abstract
Online controlled experiments, also known as A/B tests, have become ubiquitous. While many practical challenges in running experiments at scale have been thoroughly discussed, the ethical dimension of A/B testing has been neglected. This article fills this gap in the literature by introducing a new, soft ethics and governance framework that explicitly recognizes how the rise of an experimentation culture in industry settings brings not only unprecedented opportunities to businesses but also significant responsibilities. More precisely, the article (a) introduces a set of principles to encourage ethical and responsible experimentation to protect users, customers, and society; (b) argues that ensuring compliance with the proposed principles is a complex challenge unlikely to be addressed by resorting to a one-solution response; (c) discusses the relevance and effectiveness of several mechanisms and policies in educating, governing, and incentivizing companies conducting online controlled experiments; and (d) offers a list of prompting questions specifically designed to help and empower practitioners by stimulating specific ethical deliberations and facilitating coordination among different groups of stakeholders.

Keywords A/B tests · Research ethics · Digital ethics · Governance · Applied ethics

✉ Andrea Polonioli
apolonioli@coveo.com

1 Coveo Solutions Inc, 3175 des Quatre-Bourgeois Suite 200, Quebec City, G1W2k7, Canada

2 Department of Legal Studies, University of Bologna, Via Zamboni, 27, 40126 Bologna, IT, Italy

3 Department of Statistics, Computer Science and Applications "G. Parenti", University of Florence, Viale Morgagni 59, 50134 Florence, IT, Italy

4 Bauplan Labs, New York, USA

5 Oxford Internet Institute, University of Oxford, 1 St Giles', Oxford OX1 3JS, UK

6 New York University, New York, USA

7 King's College London, London, UK

8 Digital Ethics Center, Yale University, 85 Trumbull Street, New Haven, CT 06511, USA

Published online: 22 September 2023

1 Introduction

Running online controlled experiments, also known as A/B tests, has become an increasingly important aspect of data-driven decision-making for companies. It is also a powerful tool to evaluate the impact of changes made to software products and services (Jiang et al., 2019; Kohavi et al., 2020; Rajkumar et al., 2022; Siroker & Koomen, 2013; Thomke, 2020). A/B testing is an umbrella term and covers a wide spectrum of use cases and best practices for experimentation (see Kohavi et al., 2020 for a review). The key in any A/B test, or online controlled experiment, is that it is "controlled." Users are randomly exposed to one of two variants: Control (A), or Treatment (B). Therefore, online controlled experiments, exactly like offline controlled experiments, adopt randomization as the best scientific design for establishing a causal relationship between changes and their influence on behavior.[1][2] To better understand the phenomenon and share the same technical vocabulary, let us consider a typical scenario for an A/B test run by a social network *SN*, in which a team of data scientists is testing a new feed ranking algorithm. Simplifying a bit, to run an experiment at *SN*, the data science team must answer the following questions:

(1) What is the team's business goal? That is, what is *SN* trying to achieve? In this case, let us assume that it is an increase in advertising revenue.
(2) What is the evaluation criterion? That is, what metrics are considered good proxies for *SN*'s business goal? In this case, let us assume that it is engagement with the news feed, as measured by likes. This is typically known as a key performance indicator (KPI).

After a standard power analysis to check whether there are enough users available to detect the effect of interest in a given amount of time t, data scientists at *SN* launch the experiment. Users are randomly exposed to different conditions—that is, "treatment (T)" (the new algorithm) and "control (C)" (the old algorithm)—at some fixed ratio. Ideally, the two groups C and T should both be representative of the wider *SN* userbase. At the end of the trial period, researchers at *SN* review the results and estimate the average treatment effect (ATE), the average difference in KPI between the treatment and control groups. The ATE will determine whether the intervention could be beneficial for *SN*, significantly improving the KPI. To estimate it, researchers will resort to a combination of computations and statistical analyses (Veytsman, 2020). Although this example is based on an A/B test conducted by a fictional social network, it represents the fundamental methodology adopted by organizations across multiple domains.[3]

[1] It should be noted, however, that in psychology and neighboring experimental sciences randomization is not always used, and that quasi-experimental approaches are often used instead. For a discussion see Field & Hold (2003, p. 67).

[2] Other methodologies, such as naive pre-post analysis, cannot test for causal relationships.

[3] For instance, one of the most paradigmatic examples of A/B testing in history comes from Google's fifty shades of blue experiment, showing that switching the shade of blue used on advertising links in Gmail and Google search highlighted major differences (Kohavi et al., 2013).

While participant protection protocols are considered the norm in behavioral, medical, and social research, the situation is different when it comes to company-sponsored A/B testing. Research institutions must review ethical evaluation and informed consent procedures through Institutional Review Boards (IRBs) whenever research involves experimentation with human subjects (Burris & Moss, 2006). Yet, companies are not required to meet the same standards for A/B tests. Occasionally, experiments undergo internal reviews. They rarely undergo an ethics review.

Many practical challenges in running experiments at scale have been discussed (Kohavi et al., 2020). In the context of social media research, Grimmelmann (2015) highlights potential ethical risks in widespread A/B testing by private companies, including IRB laundering (the sidestepping of ethical reviews by academic researchers through collaborations with corporate partners) and the waiving of informed consent to obtain unbiased results. The author argues that already existing oversight mechanisms such as the Common Rule, IRBs and academic journals can be leveraged to mitigate these risks. Benbunan-Fich (2017) distinguishes between front-end A/B testing, i.e. changes in User Interface, and back-end C/D experimentation, where code is altered to intentionally deceive users. In the latter case, several recommendations are proposed to address the resulting ethical concerns, namely the development of an ethical code of conduct for online experiments, the design of a tool to obtain explicit consent from participants (similar to the one already in place for website cookies) and the creation of an independent user advocacy board to promote education and receive complaints about unethical conduct. However, so far the ethical implications of A/B testing have been rarely acknowledged, let alone thoroughly investigated. Providing clear, actionable, and principled ethical guidelines for responsible A/B testing is therefore especially timely and relevant. Overlooking risks and ignoring fundamental safeguards for the protection of human subjects who unknowingly become participants in A/B tests can be dangerous. After a decade in which tech companies were celebrated for empowering ordinary users, problems have been mounting over the past few years. Many digital companies have been shown to exploit behavioral biases, deception, and addictive tendencies (Costa & Halpern, 2019). While such manipulation has long been central to the business model of gambling and gaming industries (Dow Schüll, 2012), these practices are becoming more widespread (Wendel, 2020). In interface design on web pages or in games, this manipulation uses what are called "dark patterns" (Mathur et al., 2021; Waldmann, 2020). It is reasonable to conjecture that many of the established dark patterns now commonly found online have been accepted after A/B tests involving human subjects (Kramer et al., 2014).

In this article, we argue that the rise of an experimentation culture in industry brings unprecedented opportunities to businesses—but also significant responsibilities that have been overlooked for too long. We do not maintain that all online controlled experiments raise ethical concerns, but rather that it is always important to ask whether any ethical risks are involved.[4] To facilitate the assessment of

[4] Although some of the most ethically salient A/B tests concern the domain of social networks (e.g. Kramer et al., 2014; Rajkumar et al., 2022), ethically salient examples can be found in other domains too (e.g. media, Ecommerce). In the remainder of our paper, some of the examples we discuss showcase the breadth of the domain of A/B testing ethics.

the ethical risks involved, we propose a set of principles that should be adopted to encourage ethical and responsible experimentation, protecting users, customers, and society (for ease of reference we call it the DEC methodology, from the Digital Ethics Center).

Notably, practitioners regularly consider, assess, measure and report on the impact that their experiments have on their business. In this paper it is argued that practitioners should, in a similar fashion, also identify, manage and mitigate potential ethical risks. The prompting questions added in the Appendix are meant to provide practitioners with support and guidance.

The recommendations and analyses provided in this article are relevant to different audiences, from practitioners conducting online experiments to other corporate stakeholders (e.g. executives involved in ESG reporting), from scholars in the field of ethics of technology (both as authors and gatekeepers) to policy-makers and legislators.

This article aims to establish a new field of research on the ethics of A/B testing, stimulate further questions, and encourage more work on how to operationalize principles and recommendations. Although not all details are fleshed out in this article, it also already offers a wide range of recommendations that are practical in nature.

The paper is structured as follows. In Sect. 2, we introduce a new soft ethics and governance framework that organizations should strive to implement throughout the A/B test lifecycle, from planning to the communication of results. In Sect. 3, we discuss the effectiveness of several relevant mechanisms in educating, governing, and incentivizing companies conducting online controlled experiments. In the Appendix, we offer a complementary list of questions specifically designed to help empower practitioners by stimulating specific ethical deliberations. This constitutes a starting point in the development of an ethical code of conduct for A/B testing as recommended in Benbunan-Fich (2017). Section 4 concludes the article. Different audiences are expected to have diverse levels of familiarity with the content covered in different sections. For example, ethicists may be already familiar with many of the concepts and principles presented in Sect. 2, whereas industry practitioners may find them less familiar. At the same time, the article aims to engage all the relevant audiences by making the discussion of principles, policies, and mechanisms accessible to all relevant stakeholders.

2 A/B Testing Meets Ethics

A/B testing should strive to respect a critical set of ethical principles, which are general considerations that must be applied when doing research. An obvious approach to develop such a framework is offered by Beauchamp and Childress (Beauchamp & Childress, 2001, henceforth B&C). The moral framework laid out in their *Principles* has had an enormous impact on academics and practitioners across a wide variety of disciplines. B&C define common morality as the set of universal and constant norms shared by all persons committed to morality, while acknowledging that issues of moral status may vary significantly over time and between cultures. Following B&C, the key ethical principles that should guide any A/B testing are:

(1) *Autonomy*—Respect the right for an individual to make their own choice.
(2) *Fairness*—Treat individuals with fairness and equality.
(3) *Non-maleficence*—Do not harm individuals.
(4) *Beneficence*—Be beneficial to people and the environment.

Before commenting on each principle in the following sections, it is worth explaining here that a framework inspired by B&C's work is valuable for a number of reasons.

First, it corrects the error often made in the literature around ethics and technology, of equating ethics with "fairness". This is a common reductionist approach, arguably attributable to Rawls (1971) and the contract tradition in political philosophy, that is also encountered in the context of AI ethics. For instance, it is easy to associate AI Ethics with initiatives from major tech companies to issue so-called "fairness toolkits" such as IBM's "AI Fairness 360", Meta's "Fairness Flow" and Google's "What-if Tool" (Bellamy et al., 2018). However, as we shall see below, ethical concerns involved in A/B testing go beyond fairness, and so do the actions that practitioners should undertake. A benefit of adopting B&C's principles is that they provide us with a framework for acknowledging, understanding and systematizing the breadth of the ethical concerns involved in A/B testing. However, since theirs is not the only framework available which does not equate ethics with fairness, this benefit of their approach is best considered in combination with other reasons.[5]

Second, a pragmatic reason for adopting B&C's principles is that they have been successfully used in the adjacent, burgeoning literature on AI ethics. In particular, Floridi and Cowls (2019) report the results of a fine-grained analysis of several of the highest-profile sets of ethical principles for AI. Based on their comparative analysis, the authors argue that the four above-mentioned bioethical principles adapt well to the fresh ethical challenges posed by AI, although a new principle is needed in addition: explicability. Further, a paper by Jobin et al. (2019) reviewed 84 documents produced by several actors in the field and offered a classification of AI ethics guidelines. The classification proposed by Jobin et al. (2019) focused on a seemingly broader set of values than the one considered by B&C: transparency, justice and fairness, non-maleficence, responsibility and privacy. However, both the papers by Floridi and Cowls (2019) and Jobin et al. (2019) highlight that the core of B&C's framework is effectively applicable beyond the field of bioethics. Floridi and Cowls (2019) introduce a fifth principle, but this is justified by the specific demands arising in the field of AI. Further, the range of ethical considerations discussed by Jobin et al. (2019) is only apparently broader. In fact, as will become evident in the remainder of this paper, concerns about trust and transparency are related to the four general principles introduced above. Overall, this constitutes another reason, albeit preliminary, for considering B&C's principles when mapping the ethical domains of A/B testing.

Third, since the principles identified by B&C and at the heart of their proposal seek to be representative of different ethical perspectives, traditions, paradigms, and

[5] We wish to thank a reviewer for highlighting this point.

moral beliefs, they provide an ideal starting point for discussion among relevant stakeholders (Beauchamp, 2003). The greatest appeal of their approach lies precisely in its ecumenical and pluralistic nature: it is based on a set of norms affirmed by people from a wide variety of traditions, and it combines many of the most plausible elements of different theories into a clear and commonsensical framework. Importantly, however, the list of relevant ethical principles is not set in stone. B&C's framework offers a very fruitful starting point for the ethics of A/B testing and to categorize concerns and expectations, but the list of principles can be revised if needed, precisely as Floridi and Cowls (2019) did when they introduced a fifth principle in the field of AI ethics.

Having reviewed some reasons for adopting B&C's principles, it is also worth clarifying that the framework offered by the four principles is coherent with the *soft ethics* approach adopted in this article. "Soft ethics" (like soft law) is post-compliance ethics (Floridi, 2018): it applies after the application of the relevant legislation, such as the General Data Protection Regulation (GDPR) in the European Union. It differs from *hard ethics*, which may not be aligned with—and indeed may even oppose—some legislation. In this article, we are not concerned with hard ethics and how it may shape, criticize, or contrast with particular pieces of legislation. We assume that current law, at least in the EU, may be ethically acceptable and ask what more can and should be done over and above it. For instance, we acknowledge that non-discrimination law offers potential pathways to deal with instances of company-sponsored experimentation. An interesting example is posed by European non-discrimination law. Article 21 of the EU Charter of Fundamental Rights (European Union, 2012) establishes that "any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited." Similarly, we also welcome initiatives such as California's ban on "dark patterns" (Akhtar, 2021). What we argue is that, even assuming that current legislation is ethically sound, much more can and needs to be done, ethically speaking, whenever we deal with actions and practices that are legally uncharted, about which the law needs to be interpreted ethically, or where compliance is insufficient.

Before starting our discussion of each of B&C's principles, it is worth discussing two possible objections to the suggested approach. They concern the merits and applicability of B&C's principles. Let us consider them in turn.

2.1 Other Frameworks are Better

The first objection argues that despite the appeal of B&C's principles, there are alternative frameworks available which should be preferred. As it turns out, however, the presence of alternative frameworks does not really challenge the soundness of the approach suggested in this article. First, it is not obvious that adopting an alternative option would lead to different conclusions. For instance, consider the influential approach proposed by Friedman et al.'s (2013). The authors propose a list of thirteen values that are important for the design of information systems. In

their list of values, we encounter human welfare, privacy, freedom from bias, trust, autonomy, informed consent, accountability, all elements that, as we will see, can be fruitfully explored using B&C's principles. The few elements in their list that seem to be less obviously captured within B&C's framework, such as environmental sustainability, seem to be orthogonal to the present discussion of A/B testing. Admittedly, there are further relevant accounts that one could consider. For instance, one could object that human rights provide instead a better framework than the one put forth by B&C (Fukuda-Parr & Gibbons, 2021), where human rights represent an approach that highlights human dignity as the ground for our moral status. But alternative frameworks tend to converge upon closer scrutiny, meaning that other frameworks will likely accord with the conclusions of this paper (Baker, 2001). Second, there is no framework that has not attracted objections in the literature. Consider for example another influential approach, such as the capabilities approach originally proposed by Sen (e.g. Sen, 1985) and more recently further developed by Nussbaum and others. This approach stresses the importance of evaluating human welfare using the metric of what people are able to do and be. Sen emphasizes capabilities broadly, whereas, Nussbaum proposed a more specific list of capabilities that are required for a human life to be "not so impoverished that it is not worthy of the dignity of a human being" (Nussbaum, 2000, p. 72). This framework has influenced a number of policies (Bondi et al., 2021). But even this approach has been criticized (Jaggar, 2006; Nelson, 2008) and charged, for instance, with being paternalistic (Claassen, 2014).

2.2 B&C's Principles are not Useful in Practice

The second objection is that B&C's principles may seem appealing but are not useful in practice for two main reasons. First, because they are too broad to be action-guiding for practitioners, as only principles that are narrower and more specific are likely to be useful in practice (Whittlestone et al., 2019). Second, because B&C's framework does not really help us resolve conflicts between principles. Let us examine them in turn. Indeed, high-level principles are difficult to translate into practice. However, as it will become more apparent in the remainder of this article, the goal is not to merely provide high-level principles, but to unpack them, teasing apart different notions and conceptualizations, and to further enable and assist practitioners in their decision making by providing them with a detailed list of prompting questions in the Appendix. More work will be needed and is encouraged to further operationalize different ethical principles and recommendations, for instance by introducing a set of *ethical guardrails* to complement the more traditional metrics used as *organizational guardrails* to protect the business and as *technical guardrails* to ensure the internal validity and trustworthiness of the results. These *ethical guardrails* will also need to be contextualized as they will arguably be specific to different industries, contexts and use cases. The objection that B&C's approach does not in itself provide a universalizable method for prioritizing the four principles has been raised a number of times in the literature (e.g. Clouser & Gert, 1990), and might appear more scathing. More precisely, B&C's latest approach was committed

to reflective equilibrium as a methodology (Rauprich, 2008). This is a process by which our considered responses to actual cases influence our moral principles, and those principles then provide guidance for our response to further cases. The reality, however, is that although the framework does not in itself provide a universalizable method for prioritizing the four principles, this is not a shortcoming but an advantage. More precisely, given that there is no widely acceptable universalizable method for prioritizing these principles (*pace* the competing claims to the contrary), it is a positive feature of this approach that it allows to give different weight to these different principles when they conflict, also in relation to different circumstances, ethically, legally and culturally. Further, noting a tension between two values does not necessarily mean we are forced to choose between them: often, we may be able to find a tradeoff to get more of both things we value. In the Appendix, the paper offers prompting questions that also inquire as to whether any conflicts between different ethical principles are observed, and in case what weight was given to different considerations. This helps better capture any tensions that arise when high-level principles are applied to concrete cases. Though most of these tensions cannot be resolved straightforwardly, articulating them more clearly and explicitly will help further operationalize principles.

Let us now turn to the four principles.

2.3 Autonomy

Individual autonomy refers to the capacity to be one's own person, to live one's life according to reasons and motives that are one's own and not the product of manipulative or distorting external forces (Calvo et al., 2020). The growth of unregulated A/B testing may undermine human autonomy, as it can result in widespread, systematic omission of appropriate information about the risks, benefits, and alternatives of experiments, and in deceptive, opaque, and unintelligible practices that do not have the individual's best interest in mind.

Informed consent is closely related to the concepts of "autonomy" and "autonomous choice". The requirement to secure informed consent is the cornerstone of human subject protection (Resnik, 2018). The main objective of informed consent is to make prospective participants aware of the research and give them the option to opt out of the study. However, online companies automatically acquire implicit consent for research when a user accepts the terms of service (TOS), whereby these agreements are complex and difficult to read and thus raise doubts on the validity of "informed consent" (Luger et al., 2013). Consider the example of Facebook's experiments in 2014 (Kramer et al., 2014). Their study set out to test whether emotions were contagious via online social networks. For a week, Facebook showed people fewer positive or negative posts to people in the News Feed, and then measured how many positive or negative words they included in their own posts. People who saw fewer positive posts (a more depressing feed) posted 0.1% fewer positive words in their posts—their status updates were slightly less happy. People who saw fewer negative posts (a happier feed) posted 0.07% fewer negative words—their updates were slightly more positive. Technically, Facebook

had consent from all users. Yet, that was arguably a weak form of consent, as participants did not know that they were in the experiment, were not provided with any way to opt out, and were not informed about its scope or intent, its potential risks, or whether data would be kept confidential. This is in stark contrast with the consent required by offline experiments. Because of their length, most people fail to read TOS agreements and are unaware of their content (Obar & Oeldorf-Hirsch, 2020). The result was that participants were de facto uninformed and prevented from opting out, and the manipulation of emotions risked causing psychological harm to some users exposed to these practices, without any efforts to ensure their well-being.

Further, in line with Turilli & Floridi (2009), we do not consider *transparency* to be an ethical principle per se but rather a *moral enabler*, part of an ethical infrastructure *or infraethics* (Floridi, 2017). In particular, informational transparency *and* opacity can both be at odds with the concept of autonomy introduced above. Of course, in many cases, "complete understanding" of the systems with which we interact is neither desired nor required: we are perfectly happy to use technology by adopting "intentional" or "design stances" (rather than the more complete but cumbersome "physical stance") so long as the system functions correctly (Dennett, 1987). But given that we increasingly and preferentially trust and interact with software and intelligent systems, and that we rely on them to make decisions in a variety of socially significant and morally weighty contexts, the call for transparency in the design of services and goods has acquired an ethical dimension too. Importantly, the notion of transparency is highly relevant to many discussions around A/B tests, as these can be opaque when run with little or no human control or oversight. On this point, it is worth noting that not all A/B tests are the same.

First, some kinds of manipulations raise serious problems of transparency, especially when related to what we call "back-end A/B testing". Back-end A/B testing can be defined as manipulations that do not affect the outline or the design of web application like traditional surface level testing aimed at improving the design of websites or user interface (UI). Instead, back-end A/B testing modifies the inner workings of the algorithms that power certain portions of the UI, e.g. recommender systems. Consider the hypothetical data science team at *SN* which we introduced in Sect. 1. There is a difference between a scenario in which the team decides to test the impact of some new functionality on the outline or design of their website/app, such as adding a new "double thumb" option to rate content, and one in which they seek to introduce new machine learning algorithms to power the news feed. These latter types of changes are more opaque, as they operate largely beneath the surface of what the user can immediately detect. While users are becoming better at detecting and handling "dark patterns," (Shaw, 2019), they are nevertheless ill-equipped to scrutinize back-end changes and to hold companies accountable. We argue that cases of back-end testing raise unique ethical challenges, which means that informed consent should be a priority.

Second, it is important to make a distinction between two different scenarios:

(i) Treatment and control involve two different models with different parameters, but maximizing the same function.

(ii) Treatment and control involve two different models, e.g. one for engagement and one for long term value (LTV).

The majority of experiments tend to fall under type (i), as A/B tests are the paradigmatic example of exploitation (in the exploration/exploitation jargon of reinforcement learning): once the data science team at *SN* realizes that a personalized news feed is a valuable feature for engagement, it is "just" a matter of finding the right "knobs" to turn in the relevant algorithm. For technical reasons ranging from data drifts to causal inference (users are more likely to click what is recommended, ceteris paribus), A/B tests are essential to determine the optimal parameter configuration. Whatever harm or user benefit the relevant model is causing, type (i) testing would typically make it a bit higher or lower, but not structurally different. A more interesting ethical case is type (ii) experiments, which involve deep changes in either the problem framing, or the entire experience. For example, data scientists at *SN* may shift their attention from predicting likes to predicting lifetime value, that is, the total years on the platform. The new algorithm will be less readily comparable to the existing one, making A/B testing harder to interpret and arguably have a much higher risk of causing significant disadvantage in a portion of users, even if temporarily. A special case for type (ii) is when a new model introduces a new UI experience altogether. For example, Yu and Tagliabue (2020) introduced a model-based query refinement tool as an alternative to a standard search bar with no pre-existing functionality. Since bigger changes may tamper with users' intention in a new way, it is imperative that information and communication be handled more responsibly.

2.4 Fairness

Justice and fairness are closely related terms, often used interchangeably. In the literature on A/B testing, justice is mainly expressed in terms of fairness (Saint-Jaques et al. 2020). Fairness is a critical concern in the context of A/B testing due to the high stakes and risks involved. Considering how A/B tests can drive software changes with serious impact on society and daily practices, mediating personal and professional interactions, ensuring fairness in A/B tests becomes critical. For example, if left unregulated, experiments can end up reinforcing existing social (dis) advantages or stigmatization in targeted groups. Several different definitions of the concept of fairness have been provided in the literature (Saxena et al., 2019; Verma & Rubin, 2018).

It is unlikely that there will be a universal definition of fairness that is appropriate across all applications and experiments. However, many recent studies have investigated primarily two notions of fairness. *Group fairness* focuses on some sort of statistical parity (e.g. between outcomes) for members of different groups (e.g. gender), whereas *individual fairness* focuses on whether people who are similar with regards to the task receive similar outcomes (Dwork et al., 2012; Pedreschi et al., 2008). We agree with recent literature showing that these characterizations of fairness may be mutually compatible (Binns, 2020). Specifically, within the context of A/B testing, individual and group fairness constraints become important and relevant at different

stages of the A/B testing cycle. Individual fairness concerns the distribution of benefits and risks and should be preserved at the time of traffic allocation by relying on hashing that randomly assigns visitors to groups A and B. The protection of members of different protected groups becomes more relevant at the stage of final acceptance. Let us examine these points in turn.

In healthcare and clinical contexts, the principle of equipoise (a.k.a. the "uncertainty principle") holds that a user should partake in an A/B test only if there is uncertainty (see the value of opacity stressed above) about which condition is most likely to benefit the participant (Friedman & Nissenbaum, 1996; MacKay, 2018). Slightly more formally, to be in equipoise between two conditions A and B is to be cognitively indifferent between the statement "A is strictly more effective than B" and its negation. Equipoise regarding A and B is often considered sufficient for an assignment to be fair. However, the original definitions of equipoise introduced in the context of medical randomized control trials (RCTs), if literally interpreted, would substantially impede A/B testing. The definition should therefore be softened in the context of A/B testing, in the following way. A/B testing departs from principles of fairness whenever there is clear available evidence that a condition would lead to better outcomes and most users would be indifferent regarding the condition to which they are assigned.

Issues of bias do not obviously arise at the stage of traffic allocation. They become more relevant at the stage of final acceptance. Fairness is closely related to the concept of bias. A biased system "systematically and unfairly discriminates against certain individuals or groups of individuals in favor of others" (Friedman & Nissenbaum, 1996). When we ask whether a new algorithm improves a KPI (e.g., number of likes), it is imperative that we consider all the relevant groups within a population. For example, consider the context of the ranking and information retrieval, whereby ranking systems have a responsibility to their users and to the items that are being ranked. Importantly, people in current information retrieval systems are not only the ones issuing search queries, but increasingly they are also the ones being searched. This is especially important, as a number of studies have shown that ranked lists produced by a biased machine learning model can result in unfairly limited visibility for an already disadvantaged group (Geyik et al., 2019; Imana et al., 2021). Given the important role that ranking systems have come to play on websites such as Airbnb and Uber, or on human resource matchmaking platforms, such as LinkedIn, changes of rank can have a tangible impact on people's lives. Thus, controlling for key socioeconomic variables when evaluating the treatment effect of A/B tests may prove a sensible approach to uncover these asymmetric effects.

Considering the sensitive nature of the topic, it is no surprise that there are no published papers outlining evidently biased and discriminatory conditions included in A/B tests. Yet, these concerns are not at all far-fetched. Take recent arguments to the effect that Google's search engine is algorithmically biased (Noble, 2018). It has been shown that searching keywords like 'Black girls' directs to adult sites where women of color are hypersexualized. Other minorities and even religious groups are often associated with harmful stereotypes when searched with keywords on commercialized search sites. As it turns out, biases become a threat to fair treatment of

users also in the context of A/B testing. The soft ethics framework introduced here contends that instances of A/B testing that amplify biased behavior are unethical and must be avoided.

Discrimination refers to a difference in how individuals are treated based on their membership in a group. Instances of discrimination are broader than instances of bias. Further, while bias is a dimension of the process, discrimination describes the effects of the process. Notably, in many cases, discrimination is neither illegal nor obviously problematic. Yet, it can frequently raise concerns. A case in point is represented by personalized pricing, a form of discrimination in which costs are scaled to an individual's (predicted) willingness to pay. To economists, personalized pricing can be a desirable feature, with the potential to improve allocative efficiency (Inderst & Shaffer, 2009), although it should be noted that allocation efficiency does not necessarily entail social welfare (Bergemann et al., 1996). However, people's perceptions of fairness are often at odds with dynamic and personalized pricing, as shown in seminal work by Kahneman et al. (1986) and confirmed more recently by Inderst and Shaffer (2009).

If these practices are conducted using opaque means, there is also a risk that they reduce trust and create a perception of unfairness. While our framework does not specifically recommend against experimenting on personalized pricing, there is a general recommendation here that experimenters should be wary of A/B tests that would be deemed unfair and unethical by users should they become public.

2.5 Non-maleficence

This principle emphasizes the importance of not harming individuals. A/B testing should minimize the risk of causing psychological or emotional harm. For example, consider experiments on engagement on social networking websites. Addiction to social networks is not formally recognized as a diagnosis (Moqbel & Kock, 2018), but psychological dependency on such sites may interfere with important duties and activities. Notably, social media use has been associated with negative consequences such as reduced productivity, unhealthy relationships, and reduced life-satisfaction (Ponnusamy et al., 2020).

In the context of experimental research with human subjects, it is customary to accept that additional safeguards must be included in experiments involving vulnerable subjects such as children, prisoners, pregnant women, mentally disabled persons, or economically or educationally disadvantaged persons (Resnik, 2018). For instance, adults with mental disabilities or diseases that impair decision-making need additional protections in research because they may have compromised ability to consent to research participation. In the context of A/B testing, experimentation should be governed thoughtfully to protect the most vulnerable populations and additional safeguards must be included to protect the rights and welfare of these subjects. The issue and principle have gained special importance considering recent controversies based on the claim that Facebook (Meta) knew that Instagram was proving toxic for teenagers (Wells et al., 2021). It is critical that companies involved in A/B testing put in place screening practices to help exclude from experiments

members of a vulnerable group, in the same way in which vulnerable subjects are screened and excluded from clinical RCTs (see for instance the United States Code of Federal Regulations Title 45, Part 46, subparts B, C and D). Protections for vulnerable populations should be put in place in addition to, not in lieu of, overall protections for all users, as vulnerability may be context-specific.

While physical harm might not seem to be a primary concern in the context of traditional cases of A/B testing, it is worth noting that the relevance of this ethical concern should not be dismissed too quickly. For example, companies such as Lyft have introduced A/B testing in the context of hardware (Drayna et al., 2021), raising further kinds of ethical concerns about the safety and physical protection of participants.

2.6 Beneficence

The promotion of beneficial A/B testing can arguably be perceived as placing an unreasonable expectation in the context of company-sponsored experimentation. On this view, non-maleficence is sufficient for ethical A/B testing. This is because tech industry settings are different from academic and medical research settings in several ways. For instance, medical practice is bound by the Hippocratic Oath but there is no equivalent industry-wide oath for technology. However, considering the increasingly important role of Corporate Social Responsibility (CSR), it can be argued that a light, basic duty of beneficence should be understood as relevant to today's experimentation practices. After all, discussions on the importance of the so-called "triple bottom line"—the need to care for not just profit but also people and the planet—are not new (Elkington, 1997). In the broadest sense, the term "beneficence" refers here to the principle of considering and advancing the well-being of users. Beneficence generally means doing good or engaging in acts of kindness. Over and above refraining from harming others, the principle of beneficence thus requires companies to promote their welfare.

3 Incentive Mechanisms for a Soft Ethics Framework

Companies should not be left alone in trying to elevate their standards of ethical experimentation. Engineers and developers often involved in experiments are not systematically trained in ethics, may perceive ethical considerations as unnecessary red tape, and need to grapple with unavoidable conflicts of interests due to the close link between business and science. To foster compliance with ethical principles, companies need to be properly educated, governed, and incentivized. Arguably, however, ensuring ethical treatment of human subjects in the context of A/B testing is too complex to be addressed with a single, simplistic solution. On the contrary, several strategies need to be in place and several players need to be involved. Our suggestions are not meant to conclusively resolve these debates, but rather serve as a starting point in what will surely be a long ethical journey for the community. As complex problems typically require complex solutions, we submit several candidate

mechanisms that should prove helpful in fostering the adoption of *soft ethics*. What follows is a non-exhaustive outline of plausible recommendations.

3.1 Institutional Review Boards

A first mechanism that needs to be considered to promote a framework of *soft ethics* is Institutional Review Boards (IRBs). An IRB is essentially a panel of experts who review proposed research and determine whether any potential ethical concerns it might pose are sufficiently mitigated by the methodology or nature of the specific project. In the United States, IRBs in medicine were introduced to manage the ethical risks commonly faced in human subjects research. Some of that unethical conduct was particularly horrific. One notorious example is the Tuskegee experiments, in which doctors refrained from treating Black men with syphilis, despite the availability of penicillin, so that they could study the disease's unmitigated progression (Alsan & Wanamaker 2018). More generally, the goals of an IRB include upholding the core ethical principles of respect for persons, beneficence, and justice. IRBs carry out their function by approving, denying, and suggesting changes to proposed research projects.

Some countries like the United States require research institutions to have an IRB as a condition for federal funding. Before conducting a given study, a researcher submits it to the IRB board at her university, and only after IRB approval may the research begin. If the IRB declines to authorize the study, the researcher must work with the IRB to alter its nature or methods to address the IRB's concerns. If the researcher is unable to meet the IRB's demands, then the research, in theory, must not be conducted. This does not apply to company-sponsored research.

IRBs can help identify and mitigate ethical risks in A/B testing. Just as in medical research, IRBs can not only play the role of approving and rejecting various proposals but should also make ethical risk-mitigation recommendations to researchers and product developers.

While an A/B testing IRB would obviously be welcome as an important mechanism to minimize ethical risk, this solution may raise new dilemmas. Companies such as Microsoft and Meta have been launching internal IRBs over the past years, but it is unclear to what extent these boards can be truly independent (Wong & Floridi, 2023). The issue is especially relevant considering that the social contagion study mentioned above was approved by Facebook's IRB (Kramer, 2014). Some have plausibly argued that letting Facebook conduct and approve an ethical review of the study was like leaving the fox to guard the henhouse (Boesel, 2014). At the same time, opting for external IRBs would raise another set of worries. Lengthy review times for IRBs are a well-known barrier to research, and A/B tests are often time-sensitive (Liberale & Kovach, 2017). Unsurprisingly, there have been calls to improve the efficiency of the review process (Spellecy et al., 2018). But arguably the reform should be very ambitious, as a recent study of IRBs revealed that only 6% had tools sufficient for considering the area of internet research (Zimmer, 2020).

Considering this, we welcome the rise of IRBs internal to corporations and believe that these mechanisms can become critical to approve research that

involves more than a minimum risk. We are aware that corporate IRBs may not benefit from the same degree of independence that other academic IRBs do. Yet this is the issue that should be addressed. The current IRB turnaround times, policies and procedures are generally perceived to be hardly compatible with corporate expectations, business needs, expected agility and the ubiquity of A/B tests. In light of the above-mentioned concerns, we outline some tentative solutions aimed at increasing accountability while not overlooking feasibility. To begin with, companies could be required to make IRB reviews and deliberations public upon request, or perhaps publish reviews of what experiments were conducted alongside their context and rationale. Arguably, this can add an extra layer of accountability should the IRB approve some dubious experiment, or potentially expose internal pressures that drove poor decisions. Further, it is advisable that IRBs include company employees and external members (e.g. from academia) in order to at least partially mitigate the "foxes shouldn't guard henhouses" problem and "IRB laundering" mentioned in Grimmelmann (2015). This would help ensure that unethical experiments are not approved in the first place. In all, IRBs are critical when research involves more than a minimum risk. Neither a completely internal or a completely external review board would seem to be ideal. The former would likely face conflicts of interest, the latter could lack the agility and speed required in industry settings. A mixed board that includes both internal and external members appears to be a promising avenue. However, solutions must also be feasible and realistic to gain wide adoption. Hence, to further reduce turnaround times with external board members, it is suggested that internal members should be granted some kind of preemptive right to approve an A/B test if some conditions are met (e.g. review time exceeds *n* months, external members are not cooperating) and if there is agreement among all internal members.

3.2 *Soft Ethics* from the Perspective of ESG

Soft ethics can also be explored from the angle of Environmental, Social, and Corporate Governance (ESG). ESG has increasingly become a CEO-/CFO-level topic, underlining its importance for the entire organization and its implications for risk management as well as differentiation. According to IDC (2021), almost 75% of companies have already integrated or are currently in the process of integrating ESG considerations into their business approach. While this suggests how there might be a compelling business case for adopting a *soft ethics* framework that goes beyond the mere purpose of "doing good", ethical experimentation is not mentioned in CSR or ESG reports when accounting for companies' social footprint. Interestingly, however, the literature on AI ethics has been recently approached from the ESG angle (Owe & Baum 2021). We maintain that there is an opportunity here for companies that want a first-mover advantage in differentiating themselves in the marketplace by adopting a *soft ethics* framework in the context of A/B testing.

3.3 The Role of Conferences, Journals and Editorial Guidelines

Journals and associations can also play a critical role in facilitating compliance with ethical principles of experimentation. In particular, scientific publication has increasingly become popular among researchers in big tech, resulting in a growing number of corporate-affiliated papers published every year in journals or in conference proceedings (e.g. SIGIR, KDD, WSDM, WWW, WSDM, RecSys, CIKM). The fact that researchers frequently communicate via peer-reviewed publications matters substantially because research can be significantly shaped by journals' editorial decisions and policies. A few well-intentioned guidelines have been published. For example, bodies such as the Association of Computing Machinery (ACM) have long maintained ethical guidelines. However, the impact of these guidelines has been modest, and they have remained virtually invisible to a large part of the A/B testing community. It is good news that things have started to change. In our experience, an increasing number of conferences (e.g. NeurIPS, EMNLP) are encouraging reviewers to raise ethical concerns. This is important, as journals' and conferences' editorial decisions end up influencing the kind of projects that researchers will be carrying out (Horvat et al., 2015; Polonioli, 2017). Journal and conference guidelines are thus relevant as the prospect of manuscript rejection or article retraction may be an important drive to comply with different ethical standards. More generally, we argue for an increasingly important role of conferences and journals in promoting ethical A/B testing. Besides educating practitioners and acting as gatekeepers by adding relevant information about the experiments involved, conference and journal editors could also become members of mixed IRBs (as suggested in Sect. 3.1) and host the release of proceedings through which companies disclose information regarding the online experiments conducted.

3.4 The Use of Participant Compensation

Another incentive mechanism is to require companies to compensate study participants, a common practice in academic research with human subjects. For instance, in fields such as Experimental Economics, monetary rewards for participants have been the norm for decades (Hertwig & Ortmann, 2001). However, compensation need not be monetary. Access to a platform's premium features or new designs could be a viable alternative (similar to what is done with beta testing in software engineering).[6] This could be helpful in at least two ways: first, participants would need to be made aware that they are part of the experiment, thus truly enforcing the informed consent principle (and giving them the option to opt-out). Second, companies would need to compensate participants, thus potentially reducing the number of unnecessary and potentially harmful A/B tests. To mitigate problems with the latter consequence, it could be argued that compensation should be required only when

[6] Notably, while companies such as Twitter, Meta and Netflix typically allow users to enroll for "experimental versions" to access the latest features, users occasionally complain that it is not easy or possible to opt out (Spotify 2022).

assessment by the IRB reveals that the test poses significant risks (e.g., Kramer et al.,'s 2014 experiment would require compensation and informed consent while a change in website layout could be carried out without the need for compensation). Compensation has raised several issues in RCTs before, such as the exploitation of vulnerable populations (Pandya & Desai, 2013). Although we do not deny that the use of compensation may raise issues as well, it seems to be an interesting avenue to explore to mitigate and regulate the use of online experiments, which as of now is entirely unregulated. Further, we have already argued in Sect. 2.5 that measures should be in place to try to protect vulnerable populations.

3.5 Prompting Questions for an Ethical Use of A/B Testing in Industry Settings: A DEC Methodology

While all the mechanisms discussed so far may play a role in elevating the ethical standards of online experimentation, we also wish to offer a complementary tool specifically designed for practitioners to stimulate ethical deliberations. In the Appendix, we provide a list of questions to employ across the experimentation lifecycle, as it motivates deeper reflection to question whether A/B testing is used responsibly and ethically. We suggest that companies follow this list of questions and start compiling documentation that provides a concise, holistic picture of an A/B test. Documentation should be aimed at both internal and external audiences. More precisely, we encourage the creation of "A/B Test Cards" inspired by Google's "Model Cards". A recent paper by Mitchell et al. (2019) introduced the idea of Model Cards, a "one-pager" summing up what we know about a given model: input and output of course, but also the accuracy on a test set, biases and limitations, best practices for its use, as well as further relevant information. In line with the data-centric AI movement, the importance of extending documentation to the context surrounding a model—data, training, operations—has led to the creation of Directed Acyclic Graph (DAG) Cards, which emphasize "documentation as code" as a best practice for developers, and causal lineage for reporting and debugging (Tagliabue et al., 2021). We suggest that companies draw inspiration from this approach to reporting and start documenting their online experiments by offering brief summaries of their findings in clear and simplified formats, focusing on scoping, design, implementation, and dissemination. In the Appendix, we provide the full "A/B Test Card" in line with the ethical analysis developed in this article. We propose a set of sections that an A/B test card should ideally include and we provide a list of 15 questions that comes with a description of each one's relevance and supporting guidance on how to answer it. We acknowledge that not all sections and questions have the same importance. In particular, those regarding the preservation of autonomy and avoidance of harm are more important than others, for instance the ones concerning pre-registration. Further, we appreciate that the weight attributed to each question and section might to some extent depend on the specific use case. For example, for experiments involving particularly new and ethically risky hypotheses, it becomes even more important to conduct thorough research and formulate detailed hypotheses.

Although this inevitably introduces some elements of subjectivity, we believe that the A/B test card greatly facilitates the evaluation process and fosters compliance with best practices to mitigate ethical risks and increase accountability. We encourage practitioners to not only rate their compliance with best practices but also assess the relevance and weight of all checklist items.

3.6 Other Mechanisms

A possible objection to our recommendations and analyses accepts the overall sensibility of the approach but argues that this is unlikely to have a tangible impact, since not all companies will be inclined to follow our recommendations, and the external mechanisms discussed so far are not strong enough. To be sure, we agree that other external mechanisms could play an important role in driving compliance. For instance, one of the biggest tools users and consumers have is the ability to scrape, monitor and inspect platforms. ProPublica's Facebook Political Ad Collector is a case in point. ProPublica (a non-profit newsroom focusing on investigative journalism) developed a browser extension to collect political ads on Facebook in a crowdsourced manner. The investigative journalists at ProPublica were able to purchase housing ads that specifically excluded African Americans, mothers of high school kids, people interested in wheelchair ramps, Jews, expats from Argentina and Spanish speakers (Larson, 2017). Notably, many of these forms of discrimination are not only ethically challenging but even illegal, insofar as they involve protected categories of ads (housing and job) and persons (ethnicity and disability). We believe that similar initiatives, and more generally the possibility of crawling, scraping and auditing platforms (Jiang et al., 2019) will also be powerful mechanisms to empower users and society.

4 Conclusion

Protection protocols have become the norm in both medical research and the social and behavioral sciences. However, the use of human subjects in research that is not federally or publicly funded—such as in the case of privately funded A/B testing, often affecting millions of potentially unaware people—has remained unregulated. Unfortunately, the growing literature on A/B testing has not paid sufficient attention to the practice's ethical dimensions. This article fills the gap by introducing a new governance framework that explicitly recognizes how the rise of an experimentation culture in industry settings brings not only unprecedented opportunities to businesses but also responsibilities. The ethical framework recommended in this article is meant to be actionable, reasonable, scalable, and legally compliant. We know that the ethics of A/B testing and responsible causal inference is a nascent area of research, and we encourage readers to take a critical view to our assertions and use them as a point of departure for further thought and exploration.

Appendix: A/B Test Card—A DEC Methodology

Scoping: Checking the Relevance of A/B Testing

1. What is the experimental hypothesis that you are aiming to test?

Experimental hypotheses should be thoroughly and clearly articulated. This means not only that the relevant business goals and metrics should be stated clearly, but also that relevant mechanisms should be discussed. We acknowledge that this is hard to enforce and that experiments can be useful even in contexts in which no mechanistic explanation is present. However, wherever possible, it is important to try and link hypotheses to causal models. This is because when we lack causal models and well-formulated hypotheses, it is hard to rule out the hypothesis that what led to an effect represents a violation of the above-mentioned ethical principles. Specifically, while the importance of causal information has been stressed by many authors in a wide range of disciplines, including computer science (Pearl, 2009), and philosophy (Glymour, 1998; Mackie, 1974; Woodward, 2005), in the context of A/B testing, experimental hypotheses are often conceptually unsophisticated and lack information about the causal mechanisms responsible for the effect under consideration. For example, in an e-commerce context, practitioners may test which of two models for product recommendations performs better on a given website, whereby such A/B tests are typically silent regarding the causal texture and relevant mechanisms involved. The example of different models for product recommendations may sound innocuous and hardly worrisome from an ethical standpoint, but it is easy to think of contexts and examples that raise a range of ethical questions. Imagine A/B tests on the impact that different types of social media content may have on engagement metrics. Glossing over the actual causal mechanisms involved may prevent experimenters from appreciating that what is being impacted is participants' addiction. In fact, paradigmatic tests in the industry—such as Google testing 50 shades of blue for the hyperlinks in the search page, as discussed by Kohavi et al. (2020)—perform small perturbations to the context of choice for the users. The particular shade of blue has arguably no causal link to the semantic relevance of the hyperlink, given a user search. The A/B test is exploiting a spurious correlation somewhat hidden in our cognition. A mature experimentation framework should strive to improve the sophistication of experimental hypotheses and improve the understanding of causal mechanisms involved.

2. Is A/B testing the only way to infer causal effects?

One should not explore causal relationships using RCTs when the latter are considered unethical. To take a famous example, the decades-long debate on whether smoking causes cancer could have ended with a single experiment—provided we were willing to randomly expose half of the subjects to a suspected carcinogen. Because such a study was rightly deemed unethical, we had to make do with observational data (Loeb et al., 1984). Similarly, causal relationships in industry settings should also be explored with other approaches when A/B testing exposes users to unnecessary harm. This helps underline the importance in

such contexts of exploring alternative research designs. A large and growing body of literature in statistics and computer science has emerged to investigate causal relationships with limited or no experimental data (Pearl, 2009; Imbens & Rubin, 2015; Angrist et al., 2015; Peters et al., 2017). These tools can be used to avoid unethical A/B tests without compromising business objectives. Arguably, practitioners with limited familiarity with causal inference methods might be less inclined to explore the feasibility of these alternative approaches. However, a recent and welcome development is that causal inference methods have seen increased adoption in industry settings (e.g. Yao et al., 2021).

3. Was the A/B test approved by an institutional review board (IRB)?

A/B tests that impact a sizable number of users or pose non-trivial ethical risks must be ethically approved. Ideally, practitioners should resort to an institutional review board (IRB) that has A/B testing competencies that allow them to understand and consider the ethical considerations unique to online experiments.

Design: Ensuring an Ethical A/B Test

4. What protocols have been set up to identify ethical risks?

Practitioners should consider potential externalities of the intervention and develop metrics to evaluate them. To assist in this process, practitioners should always strive to refer to existing literature, as this may help them anticipate unintended results or consequences. Exploring what ethical considerations other A/B testing practitioners or researchers encountered during their projects is a valuable contribution to the quality and outcome(s) of your A/B test design.

5. What measures have been taken to obtain consent?

Visitor consent should be requested. Obtaining explicit consent is generally the preferred option (e.g. collecting consent may be similar to what is now the accepted practice regarding cookies). However, sometimes it may be acceptable to conduct an A/B test without prior consent (Gelinas et al., 2016), but only if it is impracticable to obtain consent and research involves minimal risk. In those cases, two alternatives are available. "Implied consent'" acquired through TOS, with an added option to opt-out from experiments at any time, may be sufficient for experiments with low risk. The "notify after" approach can instead be adopted to inform users retroactively about the performed study and is similarly applicable to situations with minimal risk.

6. What measures have been taken to limit the use of deception and disinformation?

A/B tests that include the use of deception should ideally demonstrate that the practitioners are aware of, seeking to minimize, and have a plan to address the possible negative impacts on participants. Importantly, several disciplines have argued in favor of minimizing the use of deception in experimental settings. For instance, deception has traditionally been used in psychological experiments (Bortolotti & Mameli, 2006), yet this has started to change (Kimmel, 2012) as it is now common in IRBs for experimental psychology to limit the use of deception (Oates et al., 2021).

7. What precautions have been taken to prevent unfairness in the treatment of the sample population and in randomization?

Setting up relevant protocols to prevent unfairness is paramount. For instance, offline evaluations can represent important measures to prevent unfairness and discrimination. There are two ways to conduct experiments: online, using A/B tests to compare different approaches by their effect on logged user interactions, and offline, using some log of historical user interactions. Offline evaluation prior to online evaluation can offer some indication of which experiments are likely to perform well online and which are not. Notably, it has become customary to run a preliminary offline evaluation to avoid losing money or breaking a system (Gilotte et al., 2018). This is important considering that to gather enough data to reach statistical significance and be able to study periodic behaviors (the signal can be different from one day to the next), an A/B test is usually implemented over several weeks. Organizations should run a preliminary offline evaluation also for ethical purposes and reserve A/B tests only for those cases where there is genuine uncertainty regarding the causal effects of the experiment.

8. What precautions have been taken to protect vulnerable populations?

A/B testing should be governed thoughtfully to protect the most vulnerable populations and additional safeguards must be included to protect the rights and welfare of these subjects. It is critical that companies involved in A/B testing put in place screening practices to help exclude from experiments members of a vulnerable group, in the same way in which vulnerable subjects are screened and excluded from clinical RCTs.

9. Are there any tensions between ethical principles and considerations?

In the context of A/B testing, the different principles discussed in Section 2 (i.e. autonomy, fairness, non-maleficence and beneficence) may come into conflict. For example, tensions can arise between privacy and the personalization of services for users (Awad & Krishnan, 2006). It is unclear how frequently tensions between principles can be observed in practice. Arguably, however, documenting any such conflicts, the weight assigned to different principles, as well as the ways in which conflicts are resolved will constitute an important step towards a more mature field of A/B testing.

Implementation: Preserving Ethics when Implementing A/B Tests

10. What are the planned timetables for the A/B tests, and will new ethical concerns arise if the tests are extended?

Ethical concerns with A/B testing may be time variant, and practitioners should consider the repercussions of extending tests. This requires practitioners to consider both the possible impact of lengthier exposure to the A/B test, as well as external factors that may shift over time. Take, for example, an A/B test that changes the frequency of political news posts on user's social media platforms. If an election is approaching, an extension of the A/B test may have more salient implications for the political behavior of the subjects.

11. What plans are there to monitor A/B tests and pause or terminate them if negative side effects unexpectedly occur?

Plans for experiments to be paused or terminated if negative side effects occur are critical and industry practitioners should aim to establish these as clearly and early as possible to avoid instances where such decisions may be subjected to a higher degree of subjectivity and dispute. Planning fixed and recurring check-ins can help ensure ethics is being considered, monitored, and assessed throughout the project, and ensure appropriate suspension or termination of the experiment if ethical concerns should arise. Setting some ethical guardrail metrics is also critical. Kohavi (2020) argues in favor of the importance of guardrail metrics. In particular, organizational guardrails are used to protect the business, while technical guardrails ensure the internal validity and trustworthiness of the results. We agree on the importance of guardrail metrics and further support the introduction of a different class of guardrails to complement the former business-centered ones. We call this class ethics-based guardrail metrics. Its purpose is to ensure that users are treated ethically. For instance, recall the example of a fictitious A/B test provided in Sect. 1. In that case, data scientists were testing a new feed ranking algorithm. Crucially, besides monitoring organizational and business guardrail metrics to avoid disastrous treatment effects and make sure the experiment is valid, it is important to consider ethics-based guardrails too. In that example, this would entail monitoring the sentiment of the users, whether it is too negative due to the new feed, or perhaps monitoring the presence of fake news or toxic comments.

Dissemination: Communicating A/B Test Results

12. Are research questions, hypotheses, and methods pre-registered, before observing the outcomes?

Pre-registration promotes transparency and accountability to the public. This is not limited to academic settings. For instance, in 2021, the OECD launched a pre-registration portal allowing practitioners to publish the essential information about their different research projects and trials, even before starting projects. This is something that, for example, Booking.com is doing already: they pre-register the desired outcome and null hypothesis (Katsimerou, 2020). Companies can follow OECD's and Booking's practices to promote transparency and accountability.

13. Is there any plan to share all results, including negative and null results?

Publishing A/B test results—e.g., through pre-prints, conference proceedings, or journal publications—can advance research and industry processes and contribute to the exchange of knowledge and best practices among A/B testing practitioners and experts.

14. Is there any plan to share anonymized data?

If possible and under the condition of anonymization and relevant legal compliance, practitioners may want to make their data publicly available. Open data allows replication of results and encourages accessibility and dissemination. Openness, namely the sharing of data, materials, methods, and results, is essential for scientific collaboration and progress (Shamoo and Resnik, 2009). Anonymization helps

researchers protect confidentiality and privacy when sharing or publishing data or samples. However, special caution is required. In the past decade, statistical methods for re-identifying individuals have been developed. In particular, linkage attacks have helped achieve information by correlating multiple data sources. To cite a famous example, Netflix, the world's largest online movie streaming service, publicly released a dataset containing movie ratings of 500,000 Netflix subscribers. The dataset was intended to be anonymous, and all personally identifying information had been removed. However, researchers demonstrated that an attacker who knew only a little bit about an individual subscriber could easily identify this subscriber's record if it was present in the dataset, or, at the very least, identify a small set of records which include the subscriber's record (Narayanan & Shmatikov, 2006). We acknowledge that sharing datasets for the purpose of reproducing the causal analysis is rare and hard to enforce. In fact, to the best of the authors' knowledge, there is only one dataset of experiments that has been released (Matias et al., 2021). Other players, such as ASOS, claim to share datasets (Liu et al., 2021), but they only report outputs, not inputs of experiments, meaning that it is impossible really to know what has worked and why. Our argument, however, is that since data availability is now increasingly considered best practice (Tedersoo et al., 2021), corporate sponsored A/B testing should also try and follow. Differential privacy tools can prevent against attacks of the sort used against the Netflix dataset (Dwork et al., 2012).

15. Have new ethical concerns emerged from scaling and adapting the results of an A/B test to new contexts?

When adapting the results of an A/B test from one environment into another, it is important to understand the factors that contributed to the success of the intervention and identify whether the same conditions exist or need to exist in a new or altered environment for the intervention to be successful.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement.

References

Akhtar, A. (2021). California is banning companies from using 'dark patterns,' a sneaky website design that makes things like canceling a subscription frustratingly difficult. Retrieved January 2, 2021 from https://www.businessinsider.com/what-are-dark-patterns-2021-3?r=US&IR=T

Alsan, M., & Wanamaker, M. (2018). Tuskegee and the health of black men. *The Quarterly Journal of Economics, 133*(1), 407–455. https://doi.org/10.1093/qje/qjx029

Angrist, J., & Pischke, J. (2015). *Mastering 'metrics: The path from cause to effect*. Princeton University Press.

Awad, N. F., & Krishnan, M. S. (2006). The personalization privacy paradox: An empirical evaluation of information transparency and the willingness to be profiled online for personalization. *MIS Quarterly*. https://doi.org/10.2307/25148715

Baker, R. (2001). Bioethics and human rights: A historical perspective. *Cambridge Quarterly of Healthcare Ethics, 10*(3), 241–252.

Beauchamp, T. (2003). A defense of the common morality. *Kennedy Institute of Ethics Journal, 13*, 259–274. https://doi.org/10.1353/ken.2003.0019

Beauchamp, T., & Childress, J. (2001). *Principles of biomedical ethics*. Oxford University Press.

Bellamy, R., Kuntal, D., Hind, M., Hoffman, S., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., & Mojsilovic, A. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *Artificial Intelligence*. https://doi.org/10.48550/arXiv.1810.01943

Benbunan-Fich, R. (2017). The ethics of online research with unsuspecting users: From A/B testing to C/D experimentation. *Research Ethics, 13*(3–4), 200–218.

Bergemann, D., Brooks, B., & Morris, S. (1996). The limits of price discrimination. *American Economic Review, 105*, 921–957.

Binns, R., (2020). On the apparent conflict between individual and group fairness. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 514–524).

Boesel, W. E. (2014). *Facebook's Controversial Experiment: Big Tech Is the New Big Pharma*. Retrieved January 2, 2021 from https://time.com/2951726/facebook-emotion-contagion-experiment/.

Bondi, E., Xu, L., Acosta-Navas, D. and Killian, J.A., (2021). Envisioning communities: a participatory approach towards AI for social good. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 425–436).

Bortolotti, L., & Mameli, M. (2006). Deception in psychology: Moral costs and benefits of unsought self-knowledge. *Accountability in Research, 13*(3), 259–275.

Burris, S., & Moss, K. (2006). US health researchers review their ethics review boards: A qualitative study. *Journal of Empirical Research on Human Research Ethics, 1*(2), 39–58.

Calvo, R., Dorian, P., Vold, K., & Ryan, R. (2020). Supporting human autonomy in AI systems: A framework for ethical enquiry. *Ethics of digital well-being: A multidisciplinary approach* (pp. 31–54). Springer.

Claassen, R. (2014). Capability paternalism. *Economics & Philosophy, 30*(1), 57–73.

Clouser, K. D., & Gert, B. (1990). A critique of principlism. *The Journal of Medicine and Philosophy, 15*(2), 219–236.

Costa, E., & Halpern, D. (2019). *The behavioural science of online harm and manipulation, and what to do about it* (pp. 1–82) [Technical Report]. https://www.cxmlab.com/wp-content/uploads/2019/07/BIT_The-behavioural-science-of-online-harm-and-manipulation-and-what-to-do-about-it_Single-2.pdf

Dennett, D. (1987). *The intentional stance*. MIT Press.

Dow Schüll, N. (2012). Addiction by design. *Princeton University Press*. https://doi.org/10.1515/9781400834655

Drayna, G, Chen, CJ & Schulte, M. (2021). A/B tests for Lyft Hardware. Lyft (March 2021). Retrieved January 2, 2021 from https://eng.lyft.com/a-b-tests-for-lyft-hardware-570330b488d4

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214--226. https://doi.org/10.1145/2090236.2090255

Elkington, J. (1997). *Cannibals with forks. The triple bottom line of 21st century*. New Society Publishers.

European Union. (2012). Charter of Fundamental Rights of the European Union. https://www.refworld.org/docid/3ae6b3b70.html

Floridi, L. (2017). Infraethics–on the Conditions of Possibility of Morality. *Philosophy & Technology, 30*(4), 391–394.

Floridi, L. (2018). Soft ethics, the governance of the digital and the general data protection regulation. *Philosophical Transactions of the Royal Society a: Mathematical, Physical and Engineering Sciences, 376*(2133), 20180081.

Floridi, L., & Cowls, J. A. (2019). Unified framework of five principles for AI in society. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.8cd550d1

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. In N. Doorn, D. Schuurbiers, I. van de Poel, & M. Gorman (Eds.), *Early engagement and*

new technologies: Opening up the laboratory philosophy of engineering and technology (Vol. 16, pp. 55–95). Springer. https://doi.org/10.1007/978-94-007-7844-3_4

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS), 14*, 330–347. https://doi.org/10.1145/230538.230561

Fukuda-Parr, S., & Gibbons, E. (2021). emerging consensus on 'Ethical AI': Human rights critique of stakeholder guidelines. *Global Policy, 12*, 32–44.

Gelinas, L., Wertheimer, A., & Miller, F. G. (2016). When and why is research without consent permissible? *Hastings Center Report, 46*(2), 35–43.

Geyik, S., Ambler, S., & Kenthapadi, K. (2019). Fairness-aware ranking in search\& recommendation systems with application to linkedin talent search. *Proceedings of the 25th Acm Sigkdd International Conference on Knowledge Discovery & Data Mining*, pp. 2221–2231. https://doi.org/10.1145/3292500.3330691

Gilotte, A., Calauzènes, C., Nedelec, T., Abraham, A. & Dollé, S. (2018). Offline a/b testing for recommender systems. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (pp. 198–206).

Glymour, C. (1998). Learning causes: Psychological explanations of causal explanation. *Minds & Machines, 8*(1998), 39–60.

Grimmelmann, J. (2015). The law and ethics of experiments on social media users. *Colo. Tech. LJ, 13*, 219.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, *24*(3), 383–403.

Horvat, M., Mlinaric, A., Omazic, J., & Supak-Smolcic, V. (2015). An analysis of medical laboratory technology journals' instructions for authors. *Science and Engineering Ethics, 22*, 1095–1106.

IDC. (2021). *Why organizations should care about responsible AI & digital ethics*. IDC.

Imana, B., Korolova, A., & Heidemann, J. (2021). Auditing for discrimination in algorithms delivering job ads. *Proceedings of the Web Conference, 2021*, 3767–3778. https://doi.org/10.1145/3442381.3450077

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Inderst, R., & Shaffer, G. (2009). Market power, price discrimination, and allocative efficiency in intermediate-goods markets. *The RAND Journal of Economics, 4*(2009), 658–672.

Jaggar, A. (2006). Reasoning about well-being: Nussbaum's methods of justifying the capabilities. *Journal of Political Philosophy, 14*, 301–322.

Jiang, S., Martin, J., & Wilson, C. (2019). Who's the Guinea Pig? Investigating online A/B/n tests in-the-wild. *Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019*, 201–210. https://doi.org/10.1145/3287560.3287565

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.

Kahneman, D., Knetsch, J., & Thaler, R. (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *The American Economic Review, 76*, 728–741.

Katsimerou, C. (2020). There's more to experimentation than A/B. Booking. https://booking.ai/theres-more-to-experimentation-than-a-b-223fba846876.

Kimmel, A. J. (2012). Deception in research. In S. J. Knapp (Ed.), *APA handbook of ethics in psychology* (pp. 401–421). American Psychological Association.

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y. and Pohlmann, N., (2013), August. Online controlled experiments at large scale. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1168–1176).

Kohavi, R., Tang, D., & Xu, Y. (2020). *Trustworthy online controlled experiments: A practical guide to a/b testing*. Cambridge University Press.

Kramer, A., Guillory, J., & Hancock, J. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, 111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Larson J, Angwin J and Valentino-DeVreis J (2017) How We are Monitoring Political Ads on Facebook. ProPublica, 5 December. Available at: www.propublica.org/article/howwe-are-monitoring-political-ads-on-facebook (Accessed 5 May 2021).

Liberale, A. P., & Kovach, J. V. (2017). Reducing the time for IRB reviews: A case study. *Journal of Research Administration, 48*(2), 37–50.

Liu, C. H., Cardoso, Â., Couturier, P., & McCoy, E. J. (2021). Datasets for online controlled experiments. *Databases*. https://doi.org/10.48550/arXiv.2111.10198

Loeb, L. A., Emster, V. L., Warner, K. E., Abbotts, J., & Laszlo, J. (1984). Smoking and lung cancer: An overview. *Cancer Research, 44*(12_Part_1), 5940–5958.

Luger, E., Moran, S., & Rodden, T. (2013). Consent for all: Revealing the hidden complexity of terms and conditions. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/2470654.2481371

MacKay, D. (2018). The ethics of public policy RCTs: The principle of policy equipoise. *Bioethics, 32*(11), 59–67. https://doi.org/10.1111/bioe.12403

Mackie, J. (1974). *The cement of the universe: A study of causation*. Clarendon Press.

Mathur, A., Kshirsagar, M., & Mayer, J. (2021). What makes a dark pattern... Dark? Design attributes, normative considerations, and measurement methods. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–18)

Matias, J. N., Munger, K., Le Quere, M. A., & Ebersole, C. (2021). The upworthy research archive, a time series of 32,487 experiments in US media. *Scientific Data, 8*(1), 1–6.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. & Gebru, T., (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229).

Moqbel, M., & Kock, N. (2018). Unveiling the dark side of social networking sites: Personal and work-related consequences of social networking site addiction. *Information & Management, 55*(1), 109–119. https://doi.org/10.1016/j.im.2017.05.001

Narayanan, A., & Shmatikov, V. (2006). How to break anonymity of the Netflix prize dataset. *Cryptography and Security*. https://doi.org/10.48550/arXiv.cs/0610105

Nelson, E. (2008). From primary goods to capabilities: Distributive justice and the problem of neutrality. *Political Theory, 36*, 93–122.

Noble, S. A. (2018). *Algorithms of oppression*. New York University Press.

Nussbaum, M. D. (2000). *Women and human development: The capabilities approach*. Harvard University Press.

Oates, J., Kwiatkowski, R., & Coulthard, L. M. (2021). *Code of human research ethics* (pp. 5–30). UK British Psychological Society Psychol Soc.

Obar, J. A., & Oeldorf-Hirsch, A. (2020). The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society, 23*(1), 128–147.

Owe, A., & Baum, S. (2021). The ethics of sustainability for artificial intelligence. *Sustainability, 13*(15), 8503.

Pandya, M., & Desai, C. (2013). Compensation in clinical research: The debate continues. *Perspectives in Clinical Research, 4*(1), 70.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pedreschi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 560--568)

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms* (p. 288). MIT Press.

Polonioli, A. (2017). New issues for new methods: Ethical and editorial challenges for an experimental philosophy. *Science and Engineering Ethics, 23*(4), 1009–1034.

Ponnusamy, S., Iranmanesh, M., Foroughi, B., & Hyun, S. S. (2020). Drivers and outcomes of Instagram addiction: Psychological well-being as moderator. *Computers in Human Behavior, 107*, 106294. https://doi.org/10.1016/j.chb.2020.106294

Rajkumar, K., Saint-Jacques, G., Bojinov, I., Brynjolfsson, E., & Aral, S. (2022). A causal test of the strength of weak ties. *Science, 377*(6612), 1304–1310.

Rauprich, O. (2008). Common morality: Comment on Beauchamp and Childress. *Theoretical Medicine and Bioethics, 29*, 43–71.

Rawls, J. (1971). A theory of justice. *Harvard University Press, Belknap Press.* https://doi.org/10.2307/j.ctvjf9z6v

Resnik, D. (2018). *The ethics of research with human subjects Protecting people, advancing science, promoting trust*. Springer.

Saint-Jacques, G., Sepehri, A., Li, N., & Perisic, I. (2020). Fairness through experimentation: Inequality in A/B testing as an approach to responsible design. *Social and Information Networks*. https://doi.org/10.48550/arXiv.2002.05819

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D., & Liu, Y. (2019). How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness.In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 99–106. Conference on AI, Ethics, and Society

Sen, A. K. (1985). *Commodities and Capabilities*. Oxford: Elsevier Science Publishers.

Shamoo, A. E., & Resnik, D. B. (2009). *Responsible conduct of research*. Oxford University Press.

Shaw, S. (2019). Consumers Are Becoming Wise to Your Nudge. Retrieved January 2, 2021 from https://behavioralscientist.org/consumers-are-becoming-wise-to-your-nudge/

Siroker, D., & Koomen, P. (2013). *A/B testing: The most powerful way to turn clicks into customers*. John Wiley.

Spellecy, R., Eve, A., Connors, E., Shaker, R., & Clark, D. (2018). The real-time IRB: A collaborative innovation to decrease IRB review time. *Journal of Empirical Research on Human Research Ethics, 13*(4), 432–437.

Spotify, (2022). Allow Pro Users to Opt-Out of A/B Testing. https://community.spotify.com/t5/Closed-Ideas/All-Platforms-Allow-Pro-Users-to-Opt-Out-of-A-B-Testing/idi-p/5092429.

Tagliabue, J., Tuulos, V., Greco, C. & Valay D. (2021). DAG Card is the new Model Card. 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney. https://datacentricai.org/neurips21/papers/43_CameraReady_neurips_data_centric_2021_DAG_CARDS_camera_ready.pdf

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., & Kogermann, K. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data, 8*(1), 1–11.

Thomke, S. (2020). *Experimentation works: The surprising power of business experiments*. Harvard Business Press.

Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology, 11*(2), 105–112.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Ieee/ACM International Workshop on Software Fairness (Fairware)*. Workshop on software fairness (fairware). Doi https://doi.org/10.23919/FAIRWARE.2018.8452913

Veytsman, B. (2020). *Computational Causal Inference*. https://arxiv.org/pdf/2007.10979.pdf

Waldman, A. E. (2020). Cognitive biases, dark patterns, and the 'privacy paradox.' *Current Opinion in Psychology, 31*, 105–109.

Wells, G., Horwitz, J., & Seetharaman, D. (2021). Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show. Wall Street Journal. *Wall Street Journal*. Retrieved 2 February 2022, from https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documentsshow-11631620739

Wendel, S. (2020). *Designing for behavior change: Applying psychology and behavioral economics*. O'Reilly Media.

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 195–200).

Wong, D., & Floridi, L. (2023) Meta's oversight board: A review and critical assessment. *Minds & Machines*, *33*, 261–284. https://doi.org/10.1007/s11023-022-09613-x.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford University Press.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., & Zhang, A. (2021). A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD), 15*(5), 1–46.

Yu, B & Tagliabue, J. (2020). Blending search and discovery: Tag-based query refinement with contextual reinforcement learning. In Workshop on Natural Language Processing in E-Commerce (EcomNLP). https://arxiv.org/abs/2010.09495

Zimmer, M. & Chapman, E. (2020). Ethical Review Boards and Pervasive Data Research: Gaps and Opportunities. In *AoIR Selected Papers of Internet Research*.