Off-Policy Learning of Content Promotions: Optimizing Digital Distribution Channels

Joel Persson
ETH Zurich, jpersson@ethz.ch,

Stefan Feuerriegel
LMU Munich, feuerriegel@lmu.de,

Cristina Kadar
Neue Zürcher Zeitung (NZZ), cristina.kadar@nzz.ch,

A common decision-making problem for digital content publishers is deciding which content to promote on their distribution channels (e.g., website front page, email newsletters, and social media pages), where personalization is often undesirable or impossible but where, instead, a small set of items is shown to the entire user base. In this paper, we develop an off-policy learning framework for this decision problem. Our objective is to learn an optimal policy per channel, which, given information about content, decides which items to promote. We show that counterfactual policies are non-parametrically identified from historical data and that the optimal policy selects the content with the largest conditional average treatment effect. Building on methods from causal machine learning, we present a fully data-driven estimation procedure that is scalable and doubly robust. To evaluate our framework, we partnered with an international newspaper, where the goal is to optimize the selection of content to promote on its online front pages. Using real-world data, we show that our optimal policy outperforms common baselines and that it significantly improves over the company's current practice. Altogether, our work provides a framework for data-driven promotions of digital content and thereby supports content publishers in optimizing their distribution channels.

Key words: content promotion; digital distribution channels; data-driven decision-making; off-policy learning; causal machine learning

1. Introduction

Crucial for the success of content is how decision-makers distribute it to audiences.[1] This is especially important in online settings where some content receives vastly more views, clicks, and

[1] A study from 2020 found that US chief marketing officers considered distribution to be the single most important content marketing activity (EMarketer 2020a). This is supported by the fact that content marketing received 66.3%

engagement than others (Berger and Milkman 2012, Upworthy 2012, Robertson et al. 2023). As such, decision-makers are faced with the task of selecting which content to promote on the front page of their website, their social media pages, and other owned digital distribution channels. Here, the objective of the decision-maker is to select a few items out of a larger content pool, so that key performance measures for the business (e.g., traffic, engagement, conversions) are maximized.

Decisions around which content to promote on digital distribution channels are oftentimes made without personalization. There are several reasons for this. First, for many channels, personalization is not sought. Publishers may want to control which content is promoted to users to align it with their brand image or marketing strategy. For example, news websites known for their investigative reports may want to promote these reports even if many users are more interested in sports. As a case in point, the New York Times even stated: "*Curation of [the website front page] page remains a core value of the product. [...] Trying to personalize this surface, we kept finding resistance to monkeying with the format because it actually had curatorial value that our users valued*" (Rockwell 2019). Similarly, several of the largest tech companies (e.g., Facebook, Apple, and LinkedIn) nowadays commonly rely on human-in-the-loop decision-making for curating their news products (Rockwell 2019). Second, for some channels, personalization is not even possible due to technical reasons. For example, owners of Facebook, Instagram, and LinkedIn pages cannot even personalize their content to visitors as the same articles are shown to the entire user base. Third, personalization may not be possible due to a lack of user-level data. On the one hand, many websites do not track users for privacy reasons unless they are logged in, and, typically, websites refrain from requiring logging in as doing so may eventually reduce the number of users. On the other hand, for many websites, the vast majority of users visit infrequently. Even for high-traffic websites like that of the Guardian, more than 70% of readers visit less than once per day (Guardian 2006). Hence, user-level data for personalizing content promotion are oftentimes lacking, so that decision-making is made at the content level.

Learning which content to promote on digital distribution channels is subject to empirical challenges. First, content publishers may not want to experimentally test which content is best to promote. This is because randomization has a high risk of making sub-optimal decisions, which, for many content publishers (e.g., a reputable newspaper), is not desired. Instead, learning which content to promote must be done from historical, non-randomized data. Second, the extent to which content benefits from promotion is highly heterogeneous (e.g., weather information on a

of all investments and, thus, more than other activities such as advertising, search engine optimization, and paid search (EMarketer 2020b).

news website may be visited by most users, regardless of whether it is promoted or not). Hence, many marketing decisions such as content promotions should not be based on outcome prediction but on predicting the incremental effects of decisions on the outcomes (Bodapati 2008, Ascarza 2018, Hitsch and Misra 2018, Lemmens and Gupta 2020). However, incremental effects cannot be observed in historical, non-randomized data because one can only observe the outcome for whether the content is promoted or not, but not the counterfactual. This is often referred to as the fundamental problem of causal inference (Rubin 1974, Holland 1986). Third, content publishers generate vast amounts of data on content that can inform the decisions of which content to promote. Hence, methods are needed that can deal with high-dimensional data.

In this paper, we present an off-policy learning framework to optimize the selection of which content to promote across digital distribution channels such that an outcome of interest (e.g., traffic, engagement, conversions) is maximized. We first show that the performance of counterfactual policies – and thus the optimal policy – is non-parametrically identified from the historical data. That is, we show that the average outcome we would have under a counterfactual policy can be expressed as a function of the joint distribution of the historical data, without making any parametric assumptions. This formally motivates our off-policy learning framework based on historical data and non-parametric estimation via machine learning. Next, we derive the closed-form solution for the optimal policy and show that it is a threshold decision rule that promotes the items with the largest conditional average treatment effect (CATE). Our estimation procedure builds on recent work from causal machine learning and is fully data-driven, robust, and computationally scalable. Furthermore, we present a tailored approach to post-hoc explainability based on post-Lasso inference, which allows us to interpret how our optimal policy makes promotion decisions.

Our framework directly addresses the above-mentioned empirical challenges. First, the challenge of using historical data is addressed by non-parametric identification for off-policy learning. This is a crucial benefit of our framework over alternative methods such as online learning and uplift modeling, which instead rely upon randomization. Second, to learn the incremental effects of promotions, we take a causal inference approach and model the CATE. Specifically, we use doubly robust learning (DRL) (Dudík et al. 2014, Athey and Wager 2021, Kennedy 2020), which adapts the doubly robust estimation of average treatment effects via parametric models to CATE estimation via machine learning. Here, "doubly robust" refers to that only one of two models – that is, one for the conditional mean outcome from the promotion decision and one for the propensity score of promotions – needs to be correct for unbiased estimation of the CATE and thus the optimal

policy. Third, our framework accommodates high-dimensional data through machine learning, thus enabling a data-driven approach to learn predictors of the content which is best to promote.

To evaluate our method, we partnered with *Neue Zürcher Zeitung* (NZZ), a leading news media company that is active in Switzerland and Germany.[2] The goal of the company is to learn which content the editors shall promote so as to maximize the average of a proprietary performance score (which corresponds to traffic, engagement, and conversions). We use a unique, large-scale, and high-dimensional panel dataset with over 2000 news articles, more than 600 time periods, and a rich set of covariates. Importantly, our data contain all information shown to the editors at the time of their decisions, all of their decisions, and all outcomes from their decisions. We find that our optimal policy performs best overall. It improves upon the current practice of the company and, furthermore, outperforms common baselines from the literature. The gain of optimal policy is explained by that it makes decisions based on CATE prediction, not outcome prediction. This empirically confirms that, for practical applications, managers should make promotion decisions based on the increment in the outcome due to the decision, not the predictions of the outcome itself. Finally, using post-hoc explainability, we provide insights into which content characteristics contribute to the decisions of the optimal policy and thereby make managerial suggestions for content publishers of which content is best to promote on distribution channels.

Our contributions are as follows. First, we formally show that, in our marketing setting, optimal policies are non-parametrically identifiable from historical data where treatment decisions (i.e., promotions) were made according to a different policy. We then connect the non-parametric identification to machine learning estimation via doubly robust learning. To the best of our knowledge, previous related works using doubly robust off-policy learning have not explicitly shown how the historical data enables identification of an optimal policy (e.g., Ellickson et al. 2022, Huang and Ascarza 2023, Yang et al. 2023). Establishing such a connection is nonetheless useful, as non-parametric identification is a necessary criterion for the use of machine learning (i.e., non-parametric) estimation of counterfactual policies, especially when the historical treatment decisions were non-randomized. Second, our framework allows for post-hoc explainability. For this, the use of doubly robust learning is important to ensure valid inference during post-hoc explainability. Third, our work connects to a growing body of research that adopts the incremental-effect approach for marketing decision-making. Previous research has shown this for problems related to targeting and personalization (Bodapati 2008, Ascarza 2018, Hitsch and Misra 2018, Lemmens and Gupta 2020).

[2] German front page: https://www.nzz.ch/deutschland. Swiss front page: https://www.nzz.ch

We add by developing a tailored incremental-effect approach for a novel decision problem, namely, content promotions.

The rest of the paper is structured as follows. In Sec. 2, we review previous research related to digital content distribution and off-policy learning. In Sec. 3, we introduce the decision-making problem at our partner company. In Sec. 4, we present our framework for off-policy learning for optimizing content promotions. In Sec. 5, we provide the results of our off-policy evaluation. Finally, we discuss our contribution and managerial implications (Sec. 6).

2. Related Work

Our work contributes to previous research on optimizing digital content distribution (e.g., Agarwal et al. 2009, Hauser et al. 2009, Kale et al. 2010, Li et al. 2010, Garcin et al. 2013, Urban et al. 2014, Schwartz et al. 2017). Due to the breadth of the literature, we focus on a few selected studies relevant to ours.

One key difference in terms of methods is whether the optimal policy is learned online vs. offline, which is also known as on-policy vs. off-policy learning (see Sutton and Barto 2018). Online methods learn an optimal policy while making decisions, whereas offline methods learn an optimal policy from historical data (Sutton and Barto 2018, Ch. 5.4). Online methods have been used for web content optimization, news recommendation, and channel curation (e.g., Agarwal et al. 2009, Hauser et al. 2009, Kale et al. 2010, Li et al. 2010, Garcin et al. 2013, Urban et al. 2014, Schwartz et al. 2017, Dimakopoulou et al. 2019). Typically, online methods optimize a policy through exploration and exploitation (Sutton and Barto 2018). Online methods are, in general, suitable when the randomization of decisions has few to no negative implications. Hence, online methods are not appropriate for problems such as ours where randomization is neither possible nor wanted (e. g., a company would not pick random articles for their social media pages, as they may want to select articles in line with their brand strategy and as a random article is sub-optimal in terms of their business performance score). In this context, offline methods are preferred as they learn a policy from historical data prior to deployment. Prior research on content optimization from historical data has mostly relied on structural econometric models (Johar et al. 2014, Besbes et al. 2016, Song et al. 2019, Zhang et al. 2019). While structural econometric models allow for economic insights that can inform human decisions (Reiss 2011), such methods do generally not output a decision policy, which is the objective of our work.

Methods for policy learning also differ as to whether they make decisions by predicting outcomes or by predicting incremental effects of decisions on outcomes. Optimization based on predicted

outcomes is a common approach for recommender systems and structural econometric models (e.g., Agarwal et al. 2009, Hitsch and Misra 2018). However, it is now well known in marketing that this approach is generally sub-optimal for marketing decision-making (e.g., Ascarza 2018, Hitsch and Misra 2018). To this end, marketers seek to optimize against incremental effects of the decision on the outcome. This approach is also known as uplift modeling, and several works have demonstrated the effectiveness of the incremental-effect approach over the outcome-based approach (Bodapati 2008, Ascarza 2018, Lemmens and Gupta 2020). Motivated by this, we later adapt the incremental-effect approach to a new decision problem, namely, content promotions, and show that the optimal solution is to optimize for incremental effects.

Finally, our work relates to an emerging stream of papers in marketing (e.g., Hitsch and Misra 2018, Simester et al. 2019, Ellickson et al. 2022, Liu 2022, Smith et al. 2022, Yoganarasimhan et al. 2022, Huang and Ascarza 2023, Yang et al. 2023). These papers first estimate the CATE via machine learning and then use a suitable estimator for off-policy evaluation. To estimate the CATE, some of the papers use double machine learning (DML). DML is doubly robust but suffers from that subsequent inference on the CATE is valid only if the estimated CATE function contains the true CATE function. For off-policy evaluation, some works use inverse probability weighting (IPW), which, although it is unbiased, is not doubly robust and, moreover, is sample-inefficient. Motivated by these observations, we instead base our CATE estimation and off-policy evaluation on DRL (Dudík et al. 2014, Foster and Syrgkanis 2019, Kennedy 2020), which is doubly robust, sample-efficient, and allows for valid inference on heterogeneity in the CATE without making any assumptions on the heterogeneity and even if the estimated CATE is misspecified. In sum, we contribute to previous works by studying a new marketing decision problem, namely, content promotions. To this end, we developed a tailored framework based on DRL for marketing decision-making and then empirically demonstrate the performance of our framework.

3. Empirical Setting

3.1. Setup

To evaluate our off-policy learning framework, we partnered with *Neue Zürcher Zeitung* (NZZ), a leading newspaper in Switzerland and Germany. A key task at *Neue Zürcher Zeitung* is to optimize the promotion decisions for two of their distribution channels: the Swiss front page of the website and the German front page of the website. Our aim is thus to learn which content to promote per time period and front page so that a proprietary performance score (explained later) is maximized.

The front pages target different audiences from different countries, so separate promotion policies should be learned for each front page.

The decision-making process for promoting content at our partner company is as follows. At the start of every hour, the editors access an internal dashboard. The dashboard shows all news articles (hereafter called "items") that were (re)published on the website within the last 72 hours, including information about the articles' performance in the previous time periods (in terms of, e.g., the total number of clicks and average engagement time) and whether an article was previously promoted by the editors. Based on the information in the dashboard, the editors decide which of the news articles that have been published on the website shall be promoted on the Swiss and German front pages, respectively. Screenshots of the front pages are shown in Appendix A. The editors aim to fill around 30 slots per time period, but, in principle, up to around 60 articles can be featured on either front page at the same time.

3.2. Data

We received access to the internal data from our partner company for a time span of six weeks at the end of 2021. Our data include all articles that were published during this time period, all information that was shown to the editors at the time of their promotion decisions, their promotion decisions, and all outcomes from their decisions. We also have access to all metadata on the published content. Our dataset is an unbalanced panel, where each item $i = 1, \ldots, N$ is included for 72 consecutive time periods, starting at the time period of publication and ending at the time period where the newspaper no longer considers it relevant. We filter for observations from 6:00 a.m. until 11:00 p.m. as no promotions are made at night. Based on discussions with the partner company, we use the following comprehensive set of covariates (which is analogous to the information shown in the company dashboard):

- *Time information:* We construct dummy variables for the hour of the day and for the day of the week. For each item, we also construct variables with the total duration (in hours) since publication and total duration (in hours) on either front page.
- *Content characteristics:* We compute the content length (i. e., number of characters) of the item. Using the metadata, we construct dummy variables for the following: section (e. g., Sport, Business & Finance, International), type (e. g., Opinion, Comment, Interview), format (e. g., regular, visual, opinion), etc. Overall, there are 16 sections, 11 types, and 4 different formats. Inspired by earlier research in marketing and management science (Archak et al. 2011, Berger and Milkman 2012, Berger et al. 2020), we also estimate the sentiment of the

title shown on the website, the title shown in search engines (i. e., the ranked list obtained after a Google search), and the lead text summarizing the story on the front page. For this, we use a sentiment model based on the large language model BERT (Kenton and Toutanova 2019) which was pre-trained on 1.834 million German texts (Guhr et al. 2020, Guhr 2022). Specifically, we used the large language model to classify the four types of texts of each content as positive, neutral, or negative and then construct a dummy variable for each sentiment category with neutral sentiment being the reference category.

- *Past performance indicators:* These are time-varying variables of historical performance per item related to traffic, engagement, and conversions, both for the previous period alone and from the time period of publication up until the end of the previous time period. Traffic indicators include the number of clicks, whereas engagement indicators include average reading time, average scroll depth, and the recirculation rate (i.e., the proportion of item views that were followed by at least one more view, thus accounting for continued engagement within a user's session up until the start of the current time period). The conversion indicators attribute the view of an item to a conversion, and include cumulative last-touch registrations (i.e., the total count of times an item was the last view of a user before they registered an account) and subscription path (i.e., the proportion of times the article was among the 10 last articles a user viewed before they bought a subscription).
- *Past promotion decisions:* To control for a possible delay and decay in the effect of past promotion decisions, we construct the first lag of the past promotion decisions and their cumulative count since publication. To account for the two different front pages, we construct two binary variables indicating if an item featured on the website was promoted on the Swiss front page and on the German front page, respectively. We further construct variables indicating if an item was distributed via an email newsletter, a push notification, or on Twitter in the previous time period, as well as the cumulative count of these since publication.

All covariates are per time period measured *prior* to the promotion decisions. We thus preserve the chronological order in the decision process and ensure that there is no look-ahead bias (i.e., post-treatment bias) in our evaluation later. We one-hot-encode categorical variables. Altogether, our final dataset spans $T = 603$ time periods with over $\sim$116,000 item-hour observations across $N =$ 2189 items and 254 covariates. Summary statistics for the covariates are provided in Appendix C.

Our outcome of interest is the proprietary performance score (a larger performance score is better), which our partner company computes as a summary function of different performance

indicators related to traffic, engagement, and conversions. In particular, the partner company first scales the different dimensions of performance indicators for valid aggregation and then weighs them according to their internal importance (which we are not allowed to disclose to ensure that information on the different revenue streams remains confidential). The performance score is then non-negative unbounded with a mean of 1.40 (for the Swiss front page) and 1.29 (for the German front page) and a standard deviation of 1.30 and 0.90, respectively. The performance score may be interpreted as either the utility function of the partner company in choosing which content to promote or as a surrogate for the long-term revenue from display ads and subscriptions (which are only observed infrequently and not at an item level but at a company level).

3.3. Descriptives

We now report descriptives documenting the main sources of heterogeneity in the performance score, the promotion decisions, and the lift in the performance score from promotion. The descriptives demonstrate why the data alone are not sufficient for learning optimal promotion policies, but that a tailored method for causal inference from historical data is needed.

When studying heterogeneity across time, we find that the performance score is, on average, larger in the morning (Fig. 1). This may be explained by that most people read the news when they wake up. We also find that the items that were promoted had, on average, a substantially larger performance score, thus indicating a positive lift from the promotion decisions. The lift in outcomes from promotion varies over hours and days and across the two front pages. As such, the data shows that there is substantial heterogeneity in the outcome with or without promotion.

We also find that content characteristics such as the section, type, and format are important sources of heterogeneity in outcomes (Fig. 2). For instance, on the Swiss front page, opinion articles tend to have the highest performance score, whereas, on the German front page, articles about domestic news tend to have the highest performance scores. Thus, the data suggest that content characteristics are also important predictors of the outcome and effectiveness of promotions, and their effectiveness varies across the two front pages.

In addition, the fraction of content that was promoted per time period varies by the section, type, and format of the items (Fig. 3). For example, for both the Swiss and the German front page, items of type “Explained” (i.e., stories that explained news events such as COVID-19) have a high likelihood of being promoted. Moreover, Swiss domestic news were rarely promoted on the German front page, as they are not relevant for the German audience.

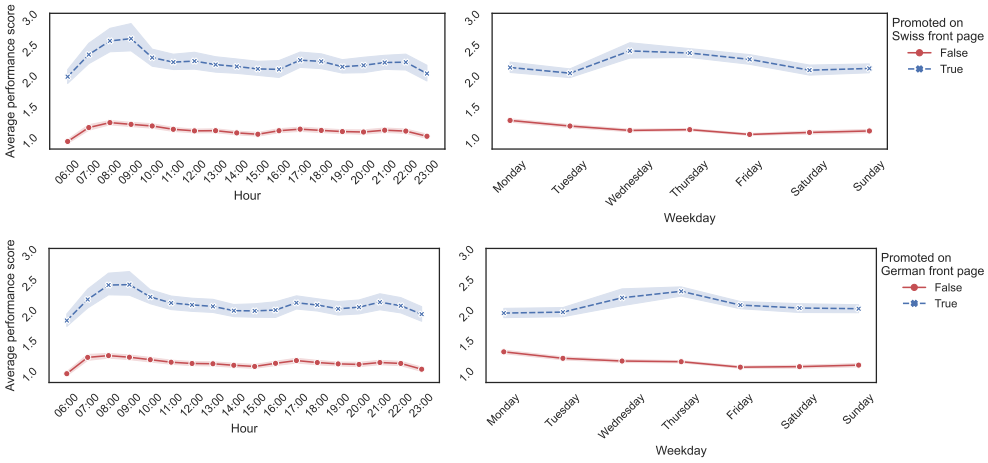


Figure 1 **Average performance score by hour and day of the week for items that were promoted or not. Shaded regions are 95% confidence intervals from 1000 bootstrap runs. Top: Swiss front page. Bottom: German front page.**

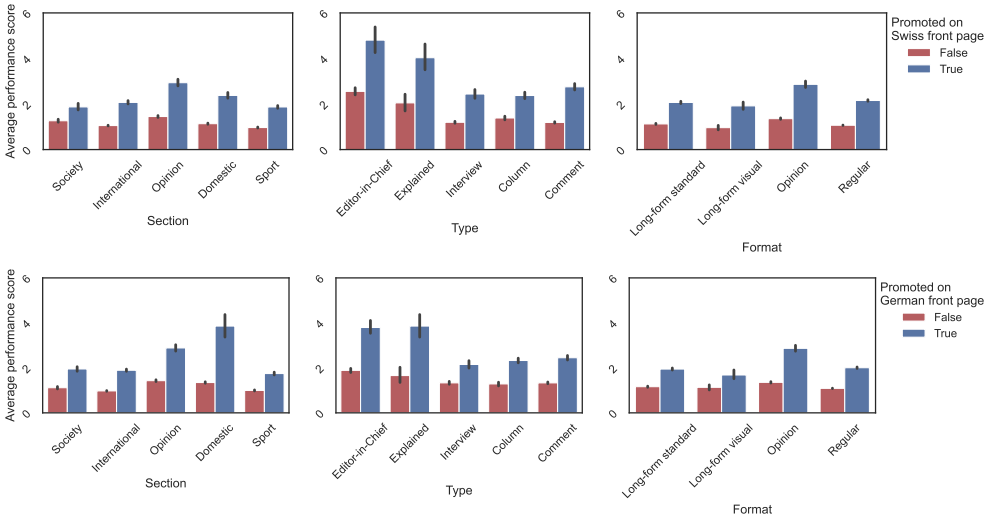


Figure 2 **Average performance score by the section, type, and format per promotion decision. Error bars are 95% confidence intervals from 1000 bootstrap runs. Top: Swiss front page. Bottom: German front page.**

To summarize, promoted items have a substantially larger performance score. However, the magnitude of the performance score and the likelihood of promotion both vary by time and content characteristics. Thus, the observed difference in performance scores between content that is promoted or not promoted is confounded by the selection of which content was promoted. Motivated by this, we develop an off-policy framework for content promotions where we leverage tailored methods for causal inference from historical data and thus account for selection bias.

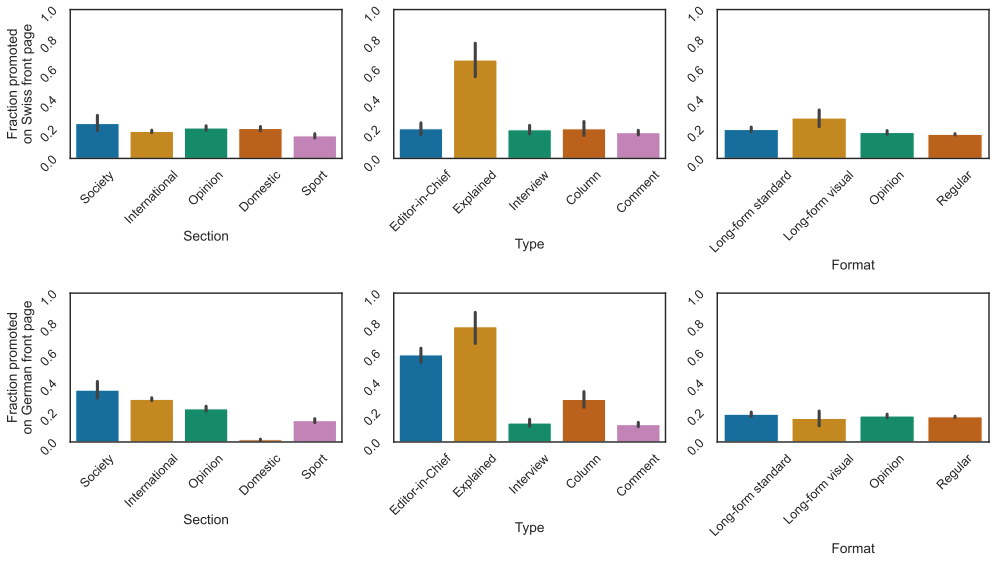


Figure 3 **Fraction of content promoted by section, type, and format of an item. Error bars are 95% confidence intervals across 1000 bootstrap runs. Top: Swiss front page. Bottom: German front page.**

4. The Off-Policy Learning Framework

Off-policy learning from historical data requires the following three components for reliable causal inference: (1) a model of the data-generating process (DGP); (2) establishing identification of counterfactuals given the model of the DGP; and (3) an estimation procedure given the identification with sound statistical properties (e.g, robustness to misspecification, sample efficiency, and scalability to large data). In the following, we present our off-policy framework encompassing all three components, each tailored to our setting.

4.1. Model of the DGP of Historical Data

We now present a simple mathematical model of DGP of our historical data. We consider a myopic decision-maker (e.g., editor) who, in each time period, shalldecide which items to promote across several distribution channels such that the mean outcome is maximized. The purpose of the model is twofold: (1) to illustrate the empirical challenges in this setting; and (2) to clarify how decisions and outcomes were generated, so that an optimal policy can be reliably identified from the historical data using off-policy learning.

Let $t = 1, \ldots, T$ denote past time periods, and let $\mathcal{I}_t$ denote the set of $N_t = |\mathcal{I}_t|$ published items that were relevant to users during time period t. Let $\mathcal{H}_t = \{(\mathbf{S}_{it}, \mathbf{A}_{it}, Y_{it}) \colon i \in \mathcal{I}_t\}$ be the historical data of items from time period $t = 1 \ldots, T$, which are independent and identically distributed (i.i.d.) realizations from a joint distribution P_π. We do not require this distribution to be known.

At the start of a time period t, the editor observes the state $\mathbf{S}_{it} \in \mathcal{S} \subset \mathbb{R}^p$ of each item $i \in \mathcal{I}_t$. The states are i.i.d. drawn from an unknown distribution $P_{\mathbf{S}}$. The state of an item is represented by p covariates (e.g., time information, time-invariant content characteristics, time-varying past performance indicators). The state then captures the observed heterogeneity across items.

Having observed the state $\mathbf{S}_{it}$, the editor decides whether to promote item $i \in \mathcal{I}_t$ or not on each channel $m = 1, \ldots, M$. Let $\mathbf{A}_{it} = (A_{it1}, \ldots, A_{itM}) \in \mathcal{A} = \{0,1\}^M$ be the M-dimensional vector of the promotions decisions for an item (i.e., the binary treatments). Here, $A_{itm} = 1$ denotes that item i was promoted on channel $m = 1 \ldots, M$ at time t, and $A_{itm} = 0$ denotes that it was not. For example, if there are $M = 2$ channels and an item i was promoted on both, then $\mathbf{A}_{it} = (1,1)$. In the literature, the promotion decisions are sometimes also called actions or treatment assignments.

To make decisions based on the state, the editor follows so-called behavior policies $\pi = (\pi_1, \ldots, \pi_M)$, where π_m is the behavior policy for channel m. Importantly, we cannot observe the behavior policies but only the editor's decisions. We thereby view each behavior policy π_m as a randomized function from the state space $\mathcal{S}$ to the space of probability distributions over the decisions $A_{itm} \in \{0,1\}$. The observed decisions $\mathbf{A}_{it}$ for an item i across the M channels are then i.i.d. drawn from a conditional distribution $P_{\mathbf{A}_t|\mathbf{S}_t}$, where, for each channel m, $\pi_m(a_{itm} \mid \mathbf{s}_{it}) = \mathbb{P}[A_{itm} = a_{itm} \mid \mathbf{S}_{it} = \mathbf{s}_{it}]$ denotes the probability of a promotion decision a_{itm} given an observed state $\mathbf{s}_{it}$.

Given capacity restrictions for the distribution channels, at most $C_{tm} < N_t$ items can be promoted on each channel $m = 1, \ldots, M$ per time period t. This reflects that the pool of relevant items is typically much larger than the capacity of a channel (e.g., in our empirical application, the newspaper has over 200 relevant articles at a time but can promote only a subset thereof). The constraint may vary by channel and over time since different channels may have more or fewer slots (e.g., a desktop web page versus an email newsletter). We denote the selected subset of items to promote by $\mathcal{C}_{tm} = \{i \in \mathcal{I}_t \colon A_{itm} = 1,\ \sum_i A_{itm} = C_{tm}\}$.

After the promotion decisions are made, the outcome $Y_{it} = Y(\mathbf{S}_{it}, \mathbf{A}_{it}) \in \mathbb{R}$ of each item $i \in \mathcal{I}_t$ is realized. In our setting, the outcome is the proprietary performance score of the partner company. The outcomes are i.i.d. draws from a conditional distribution $P_{Y_t|\mathbf{S}_t,\mathbf{A}_t}$. Thus, the outcome of an item depends on its realized state and whether or not it was promoted on one of the channels. Without loss of generality, the editor's decisions may be made in an attempt to optimize the conditional expectation of the potential outcome given the states, $\mu(\mathbf{S}_{it}, \mathbf{A}_{it}) \coloneqq \mathbb{E}[Y(\mathbf{S}_{it}, \mathbf{A}_{it}) \mid \mathbf{S}_{it}]$. The outcome is also called reward or payoff in the literature (Sutton and Barto 2018).

The so-called value of a policy at time t is the marginal mean outcome that is generated if the items that are relevant at time t would be subject to the decisions of the corresponding policy. The value of the editor's historical decisions in time period t is thus

$$\mathbb{E}_{P_\pi}[Y(\mathbf{S}_t, \mathbf{A}_t)] = \int y_t \, \mathrm{d}P_\pi = \int_{\mathcal{S}} \int_{\mathcal{A}} \int y_t \, \mathrm{d}P_{Y_t|\mathbf{S}_t,\mathbf{A}_t}(y_t) \, P_{\mathbf{A}_t|\mathbf{S}_t}(\mathbf{a}_t) \, \mathrm{d}P_{\mathbf{S}_t}(\mathbf{s}_t), \tag{1}$$

where the expectation is over the joint distribution P_π comprising the distribution of states, the conditional distribution of decisions, and the conditional distribution of outcomes.

Generally, the editor's decisions are sub-optimal, meaning they do not attain the maximum value. The reason is two-fold. First, outcomes are only observed for the decisions that are made. Hence, the incremental effects of promoting content cannot be directly observed by the editors. Second, the editor may not know which covariates in the state are truly predictive of the incremental effects of promotion and therefore relevant for decision-making. Thus, the decisions of which items to promote are instead based on some heuristic, and any observed difference in outcomes between items that were promoted or not promoted thus suffers from selection bias. Together, these challenges limit observational learning from historical data. To address these challenges, the next section introduces our objective of using off-policy learning to learn an optimal policy from the historical data generated under the editors' behavior policies.

4.2. Objective for Off-Policy Learning

To learn and evaluate a policy for optimal content promotions, we decompose the T time periods from the historical data into a training set $\mathcal{T}$ for learning the policy and an evaluation set $\mathcal{V}$ for evaluating it, where $T = |\mathcal{T}| + |\mathcal{V}|$ and $s > t$ for all time periods $s \in \mathcal{V}$, $t \in \mathcal{T}$.

Let $d_m : \mathcal{S} \to \{0, 1\}$ be a stationary deterministic policy, meaning a mapping from the state space to the space of promotion decisions for a channel $m = 1 \ldots, M$, where $d_m(\mathbf{S}_{it}) = 1$ means that the policy promotes an item with state $\mathbf{S}_{it}$ on channel m, and $d_m(\mathbf{S}_{it}) = 0$ that it does not. We consider separate policies for each channel, since the same items are not necessarily best to promote on all channels. The value of a policy d_m at time t is given by

$$V_t(d_m) = \mathbb{E}_{P_{d_m}}[Y(\mathbf{S}_{it}, A_{tm}, \mathbf{A}_t^{-m})] = \int Y(\mathbf{s}_t, A_{tm} = d_m(\mathbf{s}_t), \mathbf{A}_t^{-m}) \, \mathrm{d}P_{d_m}, \tag{2}$$

where we decompose the decisions into those for channel m, i.e., A_{tm}, and those for the other channels, i.e., $\mathbf{A}_t^{-m}$. This allows us to consider how outcomes depend on decisions for only a single channel. The distribution P_{d_m} refers to the joint distribution of the data if the decisions to channel m would be made according to a policy d_m, but, for the other channels, the decisions would be made according to their respective behavior policies $\{\pi_k \colon k \neq m\}$. The above estimand allows for

comparing the value of a policy over time, as we can expect there to be temporal heterogeneity in expected outcomes. To evaluate a policy overall, we use the long-run average value

$$V(d_m) = \frac{1}{|\mathcal{V}|} \sum_{t \in \mathcal{V}} V_t(d_m) \tag{3}$$

over the $|\mathcal{V}|$ time periods in the evaluation set $\mathcal{V}$.

To this end, the objective is, for each channel $m = 1, \ldots, M$, to use historical data from the training set $\mathcal{T}$ and learn an optimal policy for the evaluation set $\mathcal{V}$, meaning that one solves the following constrained optimization problem

$$d_m^* \coloneqq \arg\max_{d_m \in \mathcal{D}} V(d_m) \tag{4a}$$

$$\text{s.t.} \sum_{i \in \mathcal{I}_t} d_m(\mathbf{S}_{it}) \leq C_{tm}, \quad \forall t \in |\mathcal{V}|, \tag{4b}$$

where $\mathcal{D} \subseteq \{\mathcal{S} \to \mathcal{A}\}$ is a finite set of policies represented by a chosen class of machine learning models. Thus, the optimal policy attains the maximum value on the evaluation set among the considered policies, that is, $V(d_m^*) \geq V(d_m)$ for any policy d_m in $\mathcal{D}$.

4.3. Identification

We use the potential outcomes framework (Rubin 1974) for identification. We thus view $\pi_m(A_{itm} \mid \mathbf{S}_{it})$ as the propensity score of treatment assignment $A_{itm} = \{0, 1\}$ for channel m given state $\mathbf{S}_{it}$, and $Y(\mathbf{S}_{it}, A_{itm}, \mathbf{A}_{it}^{-m}) = Y(\mathbf{S}_{it}, \mathbf{A}_{it})$ as the corresponding potential outcome associated with channel m when the decisions to the other channels are held fixed as in the data, with expectation $\mu(\mathbf{S}_t, A_{itm}, \mathbf{A}_{it}^{-m}) \coloneqq \mathbb{E}[Y(\mathbf{S}_{it}, A_{itm}, \mathbf{A}_{it}^{-m}) \mid \mathbf{S}_t, \mathbf{A}_{it}^{-m}]$.[3] We make the following identification assumptions that are standard in causal inference (Imbens and Rubin 2015, Hernán and Robins 2020).

Assumption 1 **(Consistency)**. *The potential outcomes under any assignment a_{itm} given a state $\mathbf{S}_{it}$ equals the observed outcomes given that assignment and state, i.e., $Y_{it} = Y(\mathbf{S}_{it}, a_{itm}, \mathbf{A}_{it}^{-m})$ for all $a_{itm} \in \{0, 1\}$, $m = 1, \ldots, M$, and $t = 1, \ldots, T$.*

Assumption 2 **(No interference)**. *The potential outcomes of any item are conditionally independent of the current and past promotion decisions to other items given its state, i.e., $Y(\mathbf{S}_{it}, \mathbf{A}_{it}) \perp\!\!\!\perp \mathbf{A}_{js} \mid \mathbf{S}_{it}$ for all $i \neq j$ and $s \leq t \leq T$ such that $i \in \mathcal{I}_t$ and $j \in \mathcal{I}_s$.*

[3] In our empirical setting, the outcomes are the proprietary performance score, which is a known deterministic function of the states. However, as the states of the next time period are unknown at the time of the decisions, we treat the outcomes as a random function of the states and the promotion decisions in the current time period, both of which are known after the decisions are made. Technically, the outcome is a random variable whose distribution depends on the values of $\mathbf{A}_{it}$ conditional on $\mathbf{S}_{it}$ and possible exogenous factors. Following the standard potential outcomes setup, we take a state $\mathbf{S}_{it}$, $1 \leq t \leq T$, as fixed and given from the data, so that we only consider the set of potential outcomes defined by the possible values of $\mathbf{A}_{it}$ conditional on a state $\mathbf{S}_{it}$, that is, $\{Y(\mathbf{S}_{it}, \mathbf{A}_{it}) \colon \in \mathcal{A}\}$.

ASSUMPTION 3 **(Sequential unconfoundedness)**. *The potential outcomes of an item with respect to each single channel are independent of the current decision given the state, i.e.,* $Y(\mathbf{S}_{it}, a_{itm}, \mathbf{A}_{it}^{-m}) \perp\!\!\!\perp \mathbf{A}_{it} \mid \mathbf{S}_{it}$ *for all* $a_{itm} \in \{0,1\}$, $t = 1, \ldots, T$, *and* $m = 1, \ldots, M$.

ASSUMPTION 4 **(Positivity)**. *All relevant items have a non-zero probability of being promoted, i.e.,* $0 < \pi_m(a_{itm} \mid \mathbf{S}_{it}) < 1$ *for all* $m = 1, \ldots, M$, $\mathbf{S}_{it} \in \mathcal{S}$, $\mathbf{A}_{it} \in \mathcal{A}$, *and* $t \leq T$, *such that* $p(\mathbf{S}_{it}) > 0$.

Assumption 1 connects the potential outcomes to the observed outcomes, which is necessary for identifying counterfactual outcomes from data. By differentiating between promotion decisions from different channels, we satisfy the assumption by construction. Assumption 2 is reasonable in our setting since readers generally engage with items because precisely those items appeal to them. Relaxing the assumption would require learning the interference between current and past content, which would quickly become computationally infeasible in our setting. Assumption (3) states that, given the state, the outcome is independent of the current assignment.[4] This assumption is also reasonable in our setting, as we control for all information shown to editors at the time of their decisions as well as their previous decisions and the metadata. Hence, conditional on a state, selection bias is removed and historical decisions appear "as-if" random. Assumption 4 states that all items in the data have a non-zero probability of being promoted. In our empirical application, this seems reasonable given that decisions were historically based on heuristics and because, at the partner company, the decisions are made by multiple editors with different heuristics.

Our aim is to learn policies from historical data that can make decisions for new content in later time periods. To allow for this, we make four additional structural assumptions on the DGP. The first three assumptions are standard in the literature on off-policy learning when the policy shall be applicable over a possibly indefinite future horizon (see, e.g., the literature on dynamic treatment regimens; Ertefaie and Strawderman 2018, Liu et al. 2018, Liao et al. 2021), whereas the fourth is motivated by our setting of multiple decision variables for the different channels.

ASSUMPTION 5 **(Conditional stationarity)**. *Conditional on the state, the density of states and historical decisions are stationary, i..e,* $p(\mathbf{s}_{it}), \pi_1(a_{it1} \mid \mathbf{s}_{it}), \ldots, \pi_M(a_{itM} \mid \mathbf{s}_{it}) \perp\!\!\!\perp t \mid \mathbf{S}_{it}$ *for all* $\mathbf{s}_{it} \in \mathcal{S}$, $a_{itm} \in \{0,1\}$, *and* $t \leq T$.

ASSUMPTION 6 **(First-order Markov process)**. *The behavioral policy follows a Markov process, i.e.,* $a_{itm} \perp\!\!\!\perp \mathbf{S}_{i1}, \ldots, \mathbf{S}_{i,t-1} \mid \mathbf{S}_{it}$ *for all* $\mathbf{S}_{it} \in \mathcal{S}$, $a_{itm} \in \{0,1\}$, $t \leq T$, *and* $m = 1, \ldots, M$.

[4] The assumption is also known as sequential exogeneity, sequential ignorability, and selection on observables. In sequentially randomized trials such as dynamic A/B tests, Assumption 3 holds by design. In observational studies, treatments are generally not randomized, and, hence, sequential unconfoundedness is an assumption.

ASSUMPTION 7 **(Conditional contemporaneous cross-channel independence)**. *Given the state, we have that within the same time periods:*

1. *Decisions to different channels are independent, i.e.,* $\pi_m(A_{itm} \mid \mathbf{S}_{it}) \perp\!\!\!\perp \pi_k(A_{itk} \mid \mathbf{S}_{it}) \mid \mathbf{S}_{it}$ *for each pair* $(m,k) \subseteq \{1,\ldots,M\}$, $m \neq k$, *and all* $\mathbf{S}_{it} \in \mathcal{S}$, $t \leq T$.
2. *Treatment effects are independent across channels, i.e.,* $Y(\mathbf{S}_{it}, A_{itm}=1, \mathbf{A}_{it}^{-m}) - Y(\mathbf{S}_{it}, A_{itm}=0, \mathbf{A}_{it}^{-m}) \perp\!\!\!\perp \mathbf{A}_{it}^{-m} \mid \mathbf{S}_{it}$ *for all* $\mathbf{S}_{it} \in \mathcal{S}$, $\mathbf{A}_{it}^{-m} \in \{0,1\}^{M-1}$, $m=1\ldots,M$, $t \leq T$.

Assumption 5 implies that decisions and states are identically distributed over time. The assumption is required for deploying a policy learned from historical data to new content, since, then, the time periods for learning and deployment are different. Assumption 6 states that, given the current state, the current decisions are independent of the past. In our setting, the assumption holds for three reasons: (1) The current state of an item includes historical information via its age and performance since publication. (2) In the dashboard, only the current state is shown to editors. (3) The relevancy of content quickly decays after publication. Hence, past states do not provide additional information on outcomes and decisions beyond what is provided by the current state. Finally, Assumption 7 implies that, conditional on the states, there are no interactions between channels in treatment assignments and treatment effects. This still allows for marginal dependence between channels and even conditional dependence between channels given past treatment assignments.[5] Under the assumption, we have that (i) the joint conditional probability of the decisions across all channels decomposes into the product of the conditional probabilities of the decisions per channel, and that (ii) the incremental effect of promoting an item on one channel does not depend on the decisions for the item on the other channels. This is a standard assumption for double machine learning and doubly robust learning, which allows us to learn an optimal policy separately per channel (Chernozhukov et al. 2018a). The assumption is likely to hold in our application, as promotion decisions to the two front pages are managed by different teams and because users are directly assigned to one front page based on their geographic location.[6]

The above assumptions consider a stationary setting, which seems reasonable over a comparatively short horizon as in our empirical application. This also mimics how off-policy learning is used in practice, where the learned policy will be regularly re-estimated (e.g., once per month) to

[5] This means that, for a given time period and channel, the decisions on the other channels do not provide additional information beyond the information that is provided by the state.

[6] Relaxing the assumption would require a combinatorial approach, which would be computationally intractable. With 2 channels and around 200 items per time period of which up to 60 can be promoted, there would be $200!/((200-60)! \times 60!) = 7.04 \times 10^{51}$ combinations to search over per time period, of which possibly only one is optimal. Hence, Assumption 7 also facilitates the scalability of our framework.

accommodate possible non-stationarities in the DGP. Moreover, if some of the assumptions would not hold, then our optimal policy can be viewed as an approximate solution and will thus still be useful in the sense that it provides an upper bound on counterfactual policy performance.

To identify the value (i.e., mean outcome) of a counterfactual policy from historical data, we must be able to write its value as an expectation over the historical data, where the expectation is taken with respect to the joint distribution under the behavior policies. The following proposition shows that, under our model of the DGP of the historical data and our assumptions, any counterfactual policy – and thus the optimal policy – is non-parametrically identified from historical data.

PROPOSITION 1 **(Non-parametric identification)**. *Under Assumptions 1–7, the value of a counterfactual policy d_m, $m = 1 \ldots, M$, is non-parametrically identified from the DGP via the following oracle augmented inverse probability weighting (AIPW) formula*

$$V(d_m) = \frac{1}{T} \sum_{t=1}^{T} \int \left[\mu(\mathbf{S}_t, d_m(\mathbf{S}_t), \mathbf{A}_t^{-m}) + \frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\}}{\pi_m(A_{itm} \mid \mathbf{S}_t)} \times \Big(y_t - \mu(\mathbf{S}_t, A_{tm}, \mathbf{A}_{it}^{-m}) \Big) \right] \mathrm{d}P_\pi, \quad (5)$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, $\mu(\mathbf{S}_t, d_m(\mathbf{S}_t), \mathbf{A}_{it}^{-m}) \coloneqq \mathbb{E}[Y_t \mid \mathbf{S}_t, d_m(\mathbf{S}_t), \mathbf{A}_{it}^{-m}] = \int y_t \, \mathrm{d}P_{Y_t|\mathbf{S}_t, d_m(\mathbf{S}_t), \mathbf{A}_{it}^{-m}}(y_t)$, and P_π is the joint distribution of the data under the behavior policies for all channels.

Proof. The proof uses a change of measure, simplifying the likelihood functions of the data under the behavior and counterfactual policies given our model of the DGP, and then rewriting the expression; see Appendix B.1. □

Eq. (5) is the oracle AIPW formula (Robins et al. 1994, Robins 1999) adapted to our setting. Intuitively, Eq. (5) shows that, if the counterfactual and the behavior policies disagree, the counterfactual policy value is the conditional mean outcome of that policy. If the two policies agree, the counterfactual policy value is the conditional mean outcome of that policy augmented with the error in the conditional mean outcome inversely weighted with the conditional probability of the observed decision (Robins et al. 1994, Robins 1999). Thus, the augmentation term penalizes cases where the decisions of the counterfactual policy and the behavior policy coincide but (i) the conditional mean outcome the decision is far from the actual outcome; (ii) the conditional probability of the decision is small; or (iii) both (i) and (ii) hold.

Estimators based on AIPW have two benefits. Compared to (non-augmented) inverse probability weighting (IPW), AIPW estimators are statistically efficient (Robins and Rotnitzky 1995). This is because AIPW estimators use all observations, whereas IPW estimators only use those observations for which the behavior policy and counterfactual policy agree. The use of all observations

is important in our setting where only a small subset of items can be promoted per time period and, hence, where the likelihood that the policies agree is small. Moreover, AIPW estimators are doubly robust. This means that unbiased estimation of the policy value only requires one of two nuisance models – i.e., one for the conditional mean outcome and one for the propensity score – to be correctly specified (Robins et al. 1994). In contrast, IPW estimators are biased if only the model for the propensity score is incorrect. With parametric models, double robustness is not necessarily a benefit, as the likelihood of misspecifying both nuisance models is still high. This is where the benefit of non-parametric identification comes in. By showing that the policy value is non-parametrically identified, we can use machine learning to estimate the nuisance models. Hence, the requirement for double robustness is more likely to be fulfilled. Motivated by this, we later present an estimation procedure based on AIPW and machine learning.

4.4. Policies for Content Promotions

4.4.1. Optimal Policy. We now show that the optimal policy is a threshold rule that, for each time period, promotes the items with the largest CATE. The CATE is the average treatment effect as a function of covariates. It thus accounts for heterogeneity in treatment effects. The CATE is typically defined with respect to a single, binary decision variable (i.e., treatment). In our setting, we have multiple binary decision variables that jointly affect the outcome. To learn an optimal policy per channel, we note that, under Assumption 7.2, the CATE for a channel m is given by

$$\tau_m(\mathbf{s}_t) = \mathbb{E}[Y(\mathbf{S}_t, A_{tm} = 1, \mathbf{A}_t^{-m}) - Y(\mathbf{S}_t, A_{itm} = 0, \mathbf{A}_t^{-m}) \mid \mathbf{S}_t = \mathbf{s}_t]. \tag{6}$$

The CATE is thus the counterfactual difference in expected outcomes had an item been promoted on channel m or not given its state, where the conditional expectation of potential outcomes may depend on the decisions to all channels but their difference does not.

Let $\{\tau_m(\mathbf{s}_{it})\colon i \in \mathcal{I}_t\}$ be the set of CATEs for the items in time period t, and let $\tau_{tm}^{(C_{tm})}$ be the C_{tm}-th largest CATE among those. Given the constraint that at most C_{tm} items can be promoted on a channel at a time, the optimal policy is to promote the C_{tm} items with the largest CATE for the channel per time period. By statistical decision theory, such a policy is Bayes-optimal.[7]

THEOREM 1. *The oracle-optimal policy for the off-policy learning objective in Eq.* (4) *is*

$$d_m^*(\mathbf{s}_{it}) = \mathbb{1}\left\{\tau_m(\mathbf{s}_{it}) \geq \tau_{tm}^{(C_{tm})}\right\} = \begin{cases} 1, & \textit{if} \quad \tau_m(\mathbf{s}_{it}) \geq \tau_{tm}^{(C_{tm})}, \\ 0, & \textit{else.} \end{cases} \tag{7}$$

[7] Previous research has considered similar approaches for marketing decision-making based on incremental effects (e.g., Ascarza 2018, Hitsch and Misra 2018, Lemmens and Gupta 2020). Our approach is tailored to our setting and thus differs from these works in two ways: First, our outcome depends on multiple binary treatments, and so we define the CATE as the counterfactual expected change in the outcome with respect to one treatment holding the other fixed as in the data. Second, we embed the CATE estimation in an off-policy learning framework.

Proof. See Appendix B.2. □

COROLLARY 1. *The oracle-optimal policy promotes the optimal selection for its channel. That is, for each time period $t \in \mathcal{V}$ and channel $m = 1, \ldots, M$, the optimal selection is given by*

$$\mathcal{C}^*_{tm} = \left\{ i \in \mathcal{I}_t \colon d^*_m(\mathbf{s}_{it}) = 1, \sum_{i \in \mathcal{I}_t} d^*_m(\mathbf{s}_{it}) = C_{tm} \right\}. \tag{8}$$

Theorem 1 states that, if only, say, five items can be promoted on a channel at a time, then the optimal decision is to promote the five items with the largest change in the outcome if they would be promoted. In practice, we do not have the true CATE, and, so, the optimal policy must be estimated. Sec. 4.6 presents our procedure for this.

4.4.2. Baseline Policies. In our empirical evaluation, we later also compare other policies that act as baselines for our optimal policy. The following baseline policies are inspired by common methods in the literature on off-policy learning and which we carefully adapt to our setting, so that the capacity constraint is ensured. The baseline policies are:

- **Randomized policy:** This policy randomly selects items from the current pool. We randomize the selection via random sampling of C_{tm} items without replacement from the current pool $\mathcal{I}_t$, collecting them in a slate $\mathcal{C}_{tm}$, and then promoting an item if it is included in $\mathcal{C}_{tm}$. Thus

$$d_m(i; \mathcal{C}_{tm}) = \mathbb{1}\{i \in \mathcal{C}_{tm}\}. \tag{9}$$

  This policy represents a worst-case baseline, as it uses none of the information available.

- **Autoregressive policy:** The autoregressive policy promotes the items that had the largest outcomes in the previous time period. Formally, let $Y_{t-1}^{(C_{tm})}$ be the outcome of the item with the C_{tm}-th largest outcome in time period $t-1$. The policy is then given by

$$d_m(i; y_{i,t-1}) = \mathbb{1}\left\{ Y_{i,t-1} \geq Y_{t-1}^{(C_{tm})} \right\}. \tag{10}$$

  This policy is not based on a prediction of the outcome or the CATE, but simply the lagged observed outcome. It serves as a heuristic that requires no modeling or contextual information.

- **Naïve policy:** This policy promotes the items with the largest predicted promotion outcome, i.e.,

$$d_m(\mathbf{s}_{it}) = \mathbb{1}\left\{ \widehat{Y}(\mathbf{S}_{it}, A_{itm} = 1, \mathbf{A}_{it}^{-m}) \geq \widehat{Y}^{(C_{tm})} \right\}, \tag{11}$$

  where $\widehat{Y}^{(C_{tm})}$ denotes the C_{tm}-th largest predicted outcome once promoted, among the items in time period t. The naïve policy is generally sub-optimal, as it neglects that items that have a large predicted outcome in the case with promotion can also have a large predicted

outcome in the case without promotion. We nevertheless include the naïve policy because it is frequently used in marketing practice (e.g., Ascarza 2018, Hitsch and Misra 2018) and because it allows us to assess the performance gain of the incremental-effect approach over a policy that makes decisions based on predicted outcomes.

- **UCB policy:** The upper confidence bound (UCB) policy is similar to the optimal policy but, instead of ranking content based on the point prediction of the CATE, it focuses on the 95% upper confidence bound of the CATE. The policy is inspired by the original UCB algorithm (Lai et al. 1985) developed for online bandit algorithms that we adapt to our causal inference setting with capacity constraints and offline evaluation. Formally, our UCB policy is given by

$$d_m(\mathbf{s}_{it}) = \mathbb{1}\left\{\widehat{\tau}_m(\mathbf{s}_{it}) + 1.96 \cdot \mathrm{SE}_{itm} \geq (\widehat{\tau}_{tm} + 1.96 \cdot \mathrm{SE}_{tm})^{(C_{tm})}\right\}, \tag{12}$$

  where, for a given period t and channel m, SE_{itm} denotes the standard error of the CATE estimate for the item i and, by slight abuse of notation, $(\widehat{\tau}_{tm} + 1.96 \cdot \mathrm{SE}_{itm})^{(C_{tm})}$ denotes the C_{tm}-th largest 95% upper confidence bound on the estimated CATEs among the items in the time period.[8] The UCB policy is often referred to as optimistic in the sense that it makes decisions based on the best-case realizations of the CATEs of the items.

- **LCB policy:** The lower confidence bound (LCB) policy follows the same principle as the UCB policy but, instead, considers the 95% lower confidence bounds of their CATEs, i.e.,

$$d_m(\mathbf{s}_{it}) = \mathbb{1}\left\{\widehat{\tau}_m(\mathbf{s}_{it}) - 1.96 \cdot \mathrm{SE}_{itm} \geq (\widehat{\tau}_{tm} - 1.96 \cdot \mathrm{SE}_{tm})^{(C_{tm})}\right\} \tag{13}$$

  The LCB policy can be viewed as pessimistic in that it makes decisions based on the worst-case realizations of the CATEs.

The presented policies are intentionally myopic, so that they optimize the current decisions only with respect to the immediate outcomes. The use of myopic policies is motivated by three reasons. First, news articles are highly perishable. Hence, it is highly unlikely that there is a trade-off between promoting an item in the current time period and a future time period. Second, each time period will generally contain new items. Thus, the states of the relevant items in future time periods are unknown. This means that a non-myopic policy would require information that is unavailable at the time of decisions. Third, related research typically treats decisions around which content to show to users as myopic (Kale et al. 2010, Swaminathan et al. 2017, Dimakopoulou et al. 2019).[9]

[8] Depending on the model for predicting the CATE, the standard error may be obtained with a normal approximation (as in a Wald confidence interval) or with a suitable bootstrap method such as "bootstrap of little bags" (Athey et al. 2019).

[9] For additional details on optimality, Dirick and Jennergren (1975) provide sufficient and necessary conditions under which myopic policies are optimal for sequential decision problems.

4.5. Performance Measures

We empirically compare the performance of the different policies as follows. By Theorem 1, the optimal policy should attain the largest value. Hence, we compare the performance of the baseline policies in terms of their regret relative to the optimal policy. For an arbitrary policy d_m, the relative oracle-regret is defined as

$$R(d_m) = \frac{V(d_m^*) - V(d_m)}{V(d_m^*)}. \tag{14}$$

However, as the oracle-optimal policy is unknown, we instead evaluate the regret of a policy relative to an estimate $\widehat{d}_m^*$ of the optimal policy. The estimated regret $\widehat{R}(d_m)$ is thus obtained by replacing $V(d_m^*)$ and $V(d_m)$ in Eq. (14) with estimates $\widehat{V}(d_m^*)$ and $\widehat{V}(d_m)$, respectively.

We also compare the policies in terms of their relative gain over the current practice at our partner company. The relative gain of a policy d_m is defined as the relative improvement in the outcome over the behavior policy, that is,

$$G(d_m) = \frac{V(d_m) - V(\pi_m)}{V(\pi_m)}. \tag{15}$$

We also evaluate this quantity in terms policy estimates $\widehat{V}(d_m)$ and $\widehat{V}(\pi)$.

4.6. Estimation

4.6.1. Overview of Our Estimation Procedure. Our estimation procedure consists of four steps: (1) estimate potential outcomes; (2) estimate the CATE function; (3) estimate the optimal policy and its value; and (4) perform post-hoc explainability to generate insights for marketing decision-making. In the following, we detail the steps. Pseudocode for our estimation procedure is provided in Appendix E.

4.6.2. Step 1: Estimating the Potential Outcomes. Our first step is to estimate the potential outcomes, as they are later required to estimate the CATE. To do so, we use DRL (Dudík et al. 2014, Foster and Syrgkanis 2019, Kennedy 2020), which adapts the AIPW estimator of average treatment effects using parametric models to the estimation of so-called doubly robust scores of the potential outcomes using machine learning. DRL involves two sub-steps: (i) estimate nuisance models of the outcome and propensity scores; and (ii) use these models to predict the doubly robust scores of potential outcomes. In the following, we detail both sub-steps and their connection to our model of the DGP.

Sub-step (i). To get credible estimates of potential outcomes from historical data with non-randomized decisions, we must account for how potential outcomes depend on states and decisions

according to the DGP. This is provided by Propostion 1, which establishes that, given our identifying assumptions and model of the DGP, the potential outcomes depend on two models, namely, a propensity score model π_m and an outcome model μ. In the causal inference literature, these models are commonly referred to as nuisance models, reflecting that they are not of interest in themselves but are simply used to predict potential outcomes. As such, we seek to learn nuisance models that yield estimates

$$\widehat{\pi}_m(a_{itm} \mid \mathbf{s}_{it}) = \widehat{\mathbb{P}}[A_{itm} = a_{itm} \mid \mathbf{S}_{it} = s_{it}] \qquad \text{for } m = 1, \ldots, M \text{ and} \tag{16}$$

$$\widehat{\mu}(\mathbf{s}_{it}, a_{itm}, \mathbf{a}_{it}^{-m}) = \widehat{\mathbb{E}}[Y_{it} \mid \mathbf{S}_{it} = s_{im}, A_{itm} = a_{itm}, \mathbf{A}_{it}^{-m} = \mathbf{a}_{it}^{-m}]. \tag{17}$$

Thus, we need one propensity score model per channel but only a single outcome model. The latter follows from that $\mathbf{A}_{it}^{-m} = (A_{itm}, \mathbf{A}_{it}^{-m})$, and, so, we can use the same outcome model for each channel simply by using its decision variable as the treatment and the other channels' decisions variables as controls. The only assumptions for this are two-fold: the error term in the outcome model, $\varepsilon_{it} = Y_{it} - \mu(\mathbf{S}_{it}, A_{itm}, \mathbf{A}_{it}^{-m})$, must be conditionally exogenous within time periods (i.e., $\mathbb{E}[\varepsilon_{it} \mid \mathbf{S}_{it}, A_{it}] = 0$); and promotion decisions must enter both models additively without cross-channel interactions.[10] The former holds by Assumption 3, and the latter holds by Assumption 7.

Sub-step (ii). We then predict the so-called doubly robust scores (Athey and Wager 2021) of the potential outcomes where we use the fitted nuisance models from above. Recall that we define the CATE as a function of changing one of the multiple treatment variables while holding the others fixed at their observed values. We thus compute the doubly robust score for our setting via

$$\widehat{Y}(\mathbf{S}_{it}, a_m, \mathbf{A}_{it}^{-m}) \coloneqq \widehat{\mu}(\mathbf{S}_{it}, a_m, \mathbf{A}_{it}^{-m}) + \frac{\mathbb{1}\{A_{itm} = a_m\}}{\widehat{\pi}_m(A_{itm} \mid \mathbf{S}_{it})} \left(Y_{it} - \widehat{\mu}(\mathbf{S}_{it}, A_{itm}, \mathbf{A}_{it}^{-m})\right), \tag{18}$$

where A_{itm} and $a_m \in \{0, 1\}$ are the observed assignment and (possibly counterfactual) assignment for item i in time period t on channel m, respectively.

As the nuisance models are non-parametric, any machine learning model can, in principle, be used. We estimate the nuisance models as gradient-boosted trees. Gradient-boosted trees have many practical advantages in that they are scalable, accommodate non-linearities, and are robust to overfitting (Friedman 2001). The only requirement for attaining good statistical properties (e.g., semi-parametric efficiency) of their CATE estimates is that the nuisance models are estimated with

[10] Intuitively, the latter means that the CATE of promoting an item on one channel cannot depend on whether it is promoted on another channel. This also greatly simplifies the model, since, without interactions, one does not need to predict the possible outcomes from every possible combination of promotion decisions across the channels. This is also reasonable from a practical point of view since promotions for different geographical markets are managed by different teams (and visitors are directly assigned to one front page based on their geographic location).

sample-split cross-fitting (van der Laan and Rose 2011, Chernozhukov et al. 2018a, Kennedy 2020). At a high level, sample-split cross-fitting means that different splits of the training data are used to estimate the nuisance models vs. using their predictions to obtain the doubly robust scores. In particular, the doubly robust score is obtained using an outcome model and a propensity score model fitted on different splits, for observations on which neither was estimated. The procedure is repeated several times with different splits such that the prediction of the doubly robust score is not sensitive to the way in which the data was split. Implementation details (e.g., hyperparameter tuning and sample-split cross-fitting procedure) are provided in Sec. 4.7.

4.6.3. Step 2: Estimating the CATE Function. Our second step is to estimate the CATE function. To do so, we first obtain an AIPW plug-in estimate of the CATE in Eq. (6) by taking the difference in the doubly robust scores, i.e.,

$$\widehat{\tau}_{itm} = \widehat{Y}(\mathbf{S}_{it}, A_{itm} = 1, \mathbf{A}_{it}^{-m}) - \widehat{Y}(\mathbf{S}_{it}, A_{itm} = 0, \mathbf{A}_{it}^{-m}). \tag{19}$$

We then further model these CATE estimates for each channel m as a function h_m of the state covariates. This serves three purposes. First, it allows for more robust out-of-sample predictions than just using the nuisance models. The doubly robust scores given by Eq. (18) are non-linear functions of non-parametric nuisance models with unknown functional forms. The plug-in CATE estimates from their contrast inherit their properties and will thus be noisy. Projecting the estimated CATEs onto another model h_m allows us to obtain a single and simpler function that is more robust for out-of-sample predictions (Chernozhukov et al. 2018b, Kennedy 2020). Second, it improves scalability upon deployment in real-world settings. Having fitted a model h_m for the difference in potential outcomes (i.e., the CATE), we only need to evaluate a single model h_m to obtain the CATE predictions that are needed by the optimal policy. In contrast, using the nuisance models requires evaluating two models, namely, both a propensity score model and an outcome model, to first estimate the potential outcomes and then take their difference to estimate the CATE. Third, fitting a model h_m of the CATE in terms of the state covariates aids in post-hoc explainability (see Step 4 later). The CATE estimate obtained from the nuisance models of Step 1 generally depends on the state covariates in an unknown and highly non-linear way. This is because the CATE estimates are given by the difference in doubly robust scores, which are themselves non-linear functions of the non-parametric nuisance models.

Thus, we seek to learn a function of how the CATE depends on the state covariates, i.e.,

$$h_m(\mathbf{s}_{it}) = \widehat{\mathbb{E}}[\widehat{\tau}_m(\mathbf{S}_{it}) \mid \mathbf{S}_{it} = \mathbf{s}_{it}] \tag{20}$$

where h_m is a non-parametric function of the CATE for channel m. To do so in a data-driven manner, we fit a non-parametric regression model of the form $\widehat{\tau}_{itm} = h_m(\mathbf{S}_{it})$, where $\widehat{\tau}_{itm}$ is the predicted CATE from the nuisance models given by Eq. (19) from Step 1 and where h_m is a supervised machine learning model. By our non-parametric identification result, any non-parametric model of conditional means can, in principle, be used. We finally use the estimated h_m to make robust predictions of the CATE for new, unseen items.

We implement h_m as an orthogonal random forest (Oprescu et al. 2019), which is is a variant of causal forests (Wager and Athey 2018) and generalized random forests (Athey et al. 2019) that achieves a low estimation error by using a doubly robust moment equation. Following best practices in causal machine learning (van der Laan and Rose 2011, Chernozhukov et al. 2018a, Kennedy 2020), we estimate the nuisance models and the CATE function on different splits of the training data as part of the sample-split cross-fitting procedure. This reduces bias and increases efficiency in the estimated CATEs by making the predictions from the CATE function independent of those of the nuisance models. Implementation details are provided in Sec. 4.7.

4.6.4. Step 3: Estimating the Policy and its Value for Unseen Content. In Step 3, we obtain the decisions of the estimated optimal policy and then estimate its value. Importantly, whereas Step 1 and Step 2 of our estimation procedure were performed on the training set of our historical data, Step 3 is done on the evaluation set. We thus evaluate the performance of the policy when it makes decisions for content that was not used in the estimation of its underlying models. We repeat the following steps for all channels $m = 1, \ldots, M$. For each time period $t \in \mathcal{V}$:

(i) For each item $i \in \mathcal{I}_t$ with state $\mathbf{s}_{it}$, we predict the CATE $\widehat{\tau}_m(\mathbf{S}_{it})$ for channel m by evaluating the CATE function h_m from Eq. (20). Thereby, we predict the CATE for all new items that are available in the current time period.

(ii) We sort the items $i \in \mathcal{I}_t$ in decreasing order with respect to their predicted CATE and filter for the C_{tm} items that have the largest predicted CATE. For each item $i \in \mathcal{I}_t$, we obtain a plug-in estimate of the optimal policy as $d_m^*(\mathbf{S}_{it}) = \mathbb{1}\{\widehat{\tau}_m(\mathbf{S}_{it}) \geq \widehat{\tau}_{tm}^{(C_{tm})}\}$, where $\widehat{\tau}_{tm}^{(C_{tm})}$ is the C_{tm}-th largest predicted CATE for channel m among the relevant items at time t. We thus yield the decisions for our estimated optimal policy from Theorem 1. Here, $\widehat{\mathcal{C}}_{tm}^*$ for time period t and channel m is the estimated optimal selection of items as introduced in Corollary 1.

We then estimate the value $\widehat{V}(\widehat{d}_m^*)$ of the policy via the AIPW estimator given by

$$\widehat{V}(\widehat{d}_m^*) = \frac{1}{|\mathcal{V}|} \sum_{t \in \mathcal{V}} \frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} \left(\widehat{\mu}(\mathbf{S}_{it}, \widehat{d}_m^*(\widehat{\mathbf{S}}_{it}), \mathbf{A}_{it}^{-m}) + \frac{\mathbb{1}\{A_{itm} = \widehat{d}_m^*(\widehat{\mathbf{S}}_{it})\}}{\widehat{\pi}_m(A_{itm} \mid \mathbf{S}_{it})} \left(Y_{it} - \widehat{\mu}(\mathbf{S}_{it}, A_{itm}, \mathbf{A}_{it}^{-m})\right) \right). \tag{21}$$

The estimator in Eq. (21) is the plug-in sample analogue of Eq. (5) calculated based on the evaluation set using the fitted nuisance models from Step 1 and the fitted CATE function from Step 2. The following proposition and theorem provide the theoretical properties of our procedure.

PROPOSITION 2. *Under Assumption 3, an estimated optimal policy is unbiased, i.e.,*

$$\mathbb{E}[\widehat{d}_m^*(\mathbf{S}_{it})] = d_m^*(\mathbf{S}_{it}). \tag{22}$$

Proof. See Appendix B.3. □

THEOREM 2. *Under Assumptions 1, 3, and 4, the value estimator in Eq.* (21) *in unbiased, i.e.,* $\mathbb{E}[\widehat{V}(\widehat{d}_m^*)] = V(d_m^*)$*, in the doubly robust sense.*

Proof. See Appendix B.4. □

Proposition 2 states that an estimated optimal policy will, on average, make the same decisions as the oracle-optimal policy. Theorem 2 implies that, on average, our evaluation using the historical data will recover the true performance of any policy, including the optimal policy, if either the outcome model or the propensity score model (or both) is correct. Thus, our evaluation is doubly robust against model misspecification.

4.6.5. Step 4: Post-Hoc Explainaibility. Our fourth step is to explain how the optimal policy arrives at its decisions of which content to promote. To this end, our aim is to analyze the contributions of covariates to the heterogeneity in the CATE. Here, we consider two approaches: (i) Our first approach is based on the SHAP value method, which we use to check the variable importance with respect to non-linear CATE heterogeneity. (ii) Our second approach builds on valid post-Lasso inference with respect to marginal CATE heterogeneity. For both approaches, our use of DRL is crucial. In contrast to DML, inference on a CATE function estimated with DRL are valid over the space of CATE functions that are estimated, regardless of the form of the true CATE function and its heterogeneity (Battocchi et al. 2019). This is important for robust inference, as the true CATE function cannot be known in practice. This property of DRL thus allows us to apply data-driven approaches for post-hoc explainability to our estimated CATE function h_m.

Our first approach is to calculate the SHAP values of the CATE function. The SHAP value of a covariate measures its contribution to a prediction relative to the average prediction of the model

(Lundberg and Lee 2017). A key benefit of the SHAP value method in our work is that directly explains the logic of the CATE function h_m in Eq. (20). Another benefit of the SHAP value method is that it accounts for non-linearities in the treatment effect heterogeneity; however, it does not allow for valid statistical inference. For this purpose, we use our post-Lasso inference approach.

For our second approach, we adapt the method in (Chernozhukov et al. 2018b, Semenova et al. 2023), which estimates the best linear projection (BLP) of the CATE. This allows for discovering marginal heterogeneity in the CATE with respect to covariates, even when the true CATE is non-linear and unknown (Chernozhukov et al. 2018b). A straightforward approach to estimate the BLP of the CATE is to regress the CATE estimates on all state covariates subject to Lasso-regularization. However, Lasso induces regularization bias in point estimates on the selected covariates and does not readily allow for valid post-selection inference. To address this, we combine the Lasso-regularized estimation of the BLP of the CATE with sample-splitting (for details; see, e.g., Wasserman and Roeder 2009, Belloni and Chernozhukov 2013, Belloni et al. 2014, Zhao et al. 2021). Specifically, we proceed as follows: (i) we randomly split the training set into two non-overlapping sets. (ii) On one set, we fit a Lasso-regularized linear model where the dependent variable is the CATE estimate from the CATE function and where the independent variables are all state covariates. (iii) On the other set, we fit a linear model with the same dependent variable but where the independent variables are state covariates selected by the Lasso-regularization. We estimate the latter linear model without regularization using ordinary least squares (OLS) and heteroskedasticity-robust standard errors (HC2). Mathematical details are provided in Appendix D.4.

The sample-splitting in sub-step (i) ensures that the variable selection in sub-step (ii) is independent of the inference in sub-step (iii). Thus, standard confidence intervals on the parameter estimates from the model in sub-step (iii) are asymptotically valid without corrections, as long as a suitable multiple testing correction is applied (see, e.g., Pelger and Zou 2022). Recall that the decisions of the optimal policy are determined by the heterogeneity in the CATE among items. Hence, it follows that the state covariates selected in sub-step (ii) that have statistically significant parameter estimates in sub-step (iii) after a multiple testing correction are sufficient for the estimated optimal policy (see Appendix D.4 for details). We empirically validate this and find that the performance of the optimal policy is qualitatively the same when it only uses the covariates selected by our post-Lasso inference procedure to make decisions.

4.7. Implementation Details

We implement and tune the machine learning models as follows. In Step 1, we first perform model selection and hyperparameter tuning of the nuisance models on the training set. The nuisance models are merely used to predict the doubly roust scores and can thus be tuned with respect to predictive performance (Athey and Wager 2021). We implement the nuisance models via gradient-boosted trees (Friedman 2001) and follow best-practice in machine learning by performing hyper-parameter tuning using cross-validation. The propensity score model uses the log-loss, and the outcome model uses the mean squared error loss to account for the different distributions of the target variable. We also experimented with a random forest, which was slightly inferior in terms of performance but led to overall qualitatively similar conclusions.

In Step 2, we fit the hyperparameter-tuned nuisance models and the CATE function via cross-validated, sample-split cross-fitting on the training set. We use an orthogonal random forest (Oprescu et al. 2019) for the CATE function. We construct the orthogonal random forest from 500 non-parametric regression trees using the sub-sampled "honest" procedure in (Athey and Imbens 2016, Wager and Athey 2018) and tune its hyperparameters as part of the cross-validated, sample-split cross-fitting procedure. Details are provided in Appendix D.

In Step 3, our optimal policy and the baseline policies are evaluated with the doubly robust AIPW estimator from Eq. (21) using the same fitted and hyperparameter-tuned nuisance models. All evaluated policies thus only differ in how they make decisions, but neither in their estimation nor in their off-policy evaluation (given by the AIPW estimator based on the outcome model and the propensity score model). The only exception is the value of behavior policy, which, by definition, is given by the sample average of the outcomes in the data. To make decisions, both the UCB policy and LCB policy also require the lower and upper bounds on the confidence intervals of the predicted CATE of each item. For this, we construct 95% confidence intervals using "bootstrap of little bags" (Athey et al. 2019) on the predictions from the orthogonal random forest.

In Step 4, we fit the Lasso-regularized model using 5-fold cross-validation. For each fold, we run coordinate descent optimization for 1000 iterations with 100 different values of the tuning penalty λ along the regularization path. Among the 5 cross-validated models, the final Lasso-regularized model is selected as the one that best predicts the CATE function h_m.

The performance of the policies depends on the capacity constraint. We set C_{tm}, i.e., the number of items to promote by the evaluated policies per channel and time period, to the same number of items that were promoted by our partner company for each channel and time period (i.e., the

behavior policy). This way, any difference in performance between different cannot be due to that they promote more or fewer items. Later, we also provide a sensitivity analysis where we vary the capacity constraint, finding that our proposed framework performs consistently best.

Additional implementation details (including pseudocode for an off-policy learning and evaluation algorithm) are in Appendix D. The runtime for making decisions on unseen items is in the order of milliseconds using standard office hardware at our partner company.

5. Evaluation

5.1. Performance Comparison

We now empirically evaluate the performance gain of our optimal policy over the baseline policies. We follow best-practice in machine learning and emphasize that all policies are evaluated using the same machine learning models that were rigorously hyperparameter tuned to ensure a fair comparison. Using the fitted models, we evaluate the AIPW estimator for each policy on the same evaluation set, thus quantifying the ability of each policy to make decisions for unseen, out-of-sample items. Finally, all policies are subject to the same capacity constraint as the behavior policy. Therefore, our evaluation ensures that any difference in performance between policies can only be attributed to the differences in which items they promote.

Table 1 reports the performance. We have three main findings. First, we find that our optimal policy performs best. It achieves large gains over the behavior policy, which corresponds to the decision-making in current practice (value of 1.287). Specifically, our optimal policy improves the value over the behavior policy by 5.21% (Swiss front page) and 2.25% (Germany front page). To put that into context, we perform a simple back-on-the-envelope analysis. Recall that the value is the average of the proprietary performance score used by our partner company and that the performance score is a summary function of traffic, engagement, and conversions. Assuming that a 1% gain in value corresponds to a 0.1–0.5% gain in revenue from ads, sales, and subscriptions and that Assumption 7.2 holds, then a back-of-the-envelope calculation suggests that deploying the optimal policy for both front pages would lead to a 1.9–9.5 million USD increase in revenue relative to current practice.[11]

Second, our optimal policy outperforms the baseline policies by a clear margin. This confirms the effectiveness of making content promotion decisions based on the incremental effect of the decision on the outcome. As expected, the randomized policy is the least effective, as it makes decisions

[11] The back-of-the-envelope calculation is based on that the total revenue of our partner company was about 285 million USD in 2022 of which 89% is estimated to come from ads, sales, and subscriptions (DigiDay 2018, Center 2021, Statista 2022, DigiDay 2023).

independently of outcomes. It even performs worse than the behavior policy, thus implying that the current practice for selecting which content to promote is not completely sub-optimal. Nonetheless, we observe that all other baselines outperform the behavior policy. This can be explained by that the current practice is primarily based on heuristics, whereas all baselines except the randomized policy are based on data-driven models and optimize for outcomes.

Third, we observe that the performance improvement of our optimal policy over current practice is larger for the Swiss front page than for the German front page. Put differently, the current practice of decision-making is closer to the optimal decisions for the German market than for the Swiss market. In discussions with stakeholders from our partner company, we were informed that this may be due to the newspaper having only recently expanded to Germany. As a result of the expansion, the newspaper has a relatively small market share in Germany and, thus, a fairly homogeneous audience of early adopters, whereas, in Switzerland, the audience is well-established and more diverse. This makes it relatively more difficult for editors to anticipate the interests of the Swiss user base than the German user base, as the promotion decisions for the latter can be directed toward a relatively narrow range of user interests.

In sum, the above results confirm empirically that our optimal policy based on optimizing against the CATE performs best and leads to substantial improvements over the current practice. Hence, throughout the rest of our results, we focus on our optimal policy and the heterogeneity in both its performance and decisions.

Table 1 Policy value estimates on the evaluation set (out-of-sample).

	Swiss front page				*German front page*			
Policy	Value	SE	Regret	Gain	Value	SE	Regret	Gain
Randomized policy	1.188	0.0067	12.26%	−7.69 %	1.244	0.0070	5.47%	−3.34 %
Autoregressive policy	1.312	0.0074	3.10%	1.94%	1.298	0.0073	1.37%	0.85%
Naïve policy	1.329	0.0075	1.85%	3.26%	1.300	0.0073	1.22%	1.01%
UCB policy	1.345	0.0076	0.66%	4.51%	1.312	0.0074	0.30%	1.94%
LCB policy	1.300	0.0073	3.99%	1.01%	1.282	0.0072	2.58%	−0.39 %
Optimal policy	**1.354**	**0.0076**	**0.00%**	**5.21%**	**1.316**	**0.0074**	**0.00%**	**2.25%**

Notes. Value = average performance score across content and time periods. Larger is better. Largest value in bold. SE: standard error. Regret: decrease in value relative to our optimal policy. Gain: increase in value relative to behavior policy.

5.2. Heterogeneity across Time

We now seek to understand how the performance of our optimal policy varies over time. Our results across hour of day and day of the week are shown in Fig. 4. We find that the optimal policy performs consistently best. Moreover, we observe that the average value under our optimal policy is characterized by similar dynamics as the average value of the behavior policy, which adds to the

credibility of our results. For example, the overall news consumption is particularly pronounced in the early morning, and we thus also expect a large gain from optimizing content promotions for the early morning, which is confirmed by our results.

5.3. Heterogeneity across Section, Type, and Format of Items

We also find that, under the optimal policy, the average performance score varies by the section, type, and format of content (Fig. 5). The value is slightly higher for opinion and explanatory articles and substantially larger for articles written by the editor-in-chief. The results are similar across the two front pages but different in magnitude. Overall, promoting items has a positive effect on the average performance of the corresponding items, but the magnitude depends both on when the items are promoted and their characteristics.

5.4. Heterogeneity in Promotion Decisions

The optimal policy outperforms current practice by promoting different items. To further understand where these improvements stem from, we examine the heterogeneity in the categories of content that tend to be promoted by the optimal policy vs. the behavior policy.

Fig. 6 compares the fraction of items that are promoted by the optimal policy vs. the behavior policy per section, type, and format. Similar to current practice, the optimal policy promotes

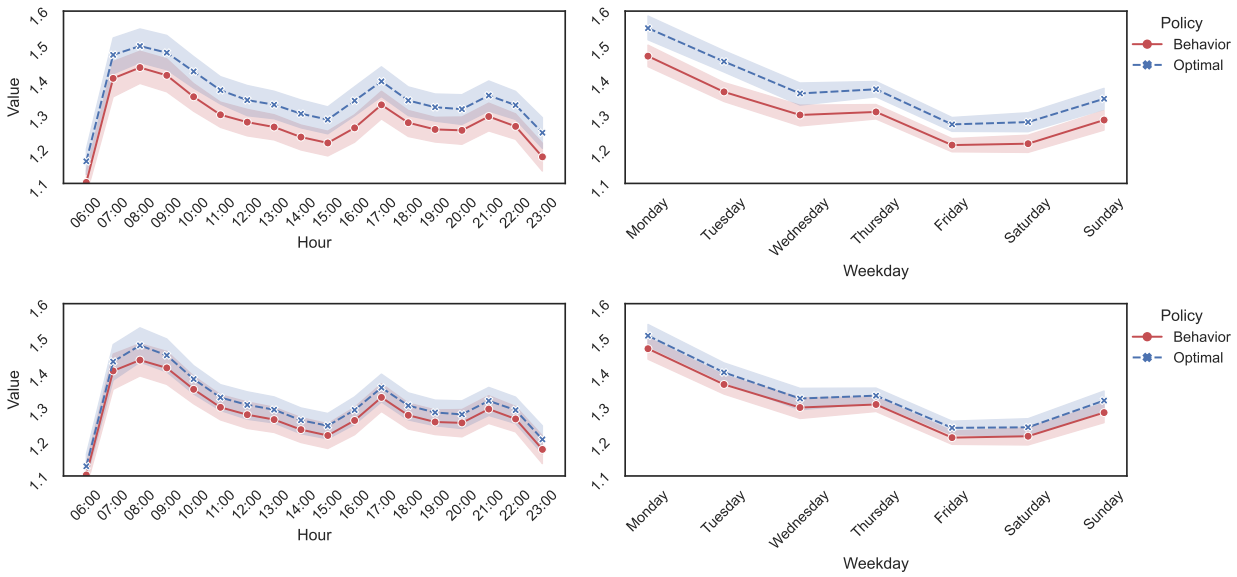


Figure 4 **Value of the behavior policy (current practice) and our optimal policy by hour and day of the week. Shaded regions are 95% confidence intervals across 1000 bootstrap runs. Top: Swiss front page. Bottom: German front page.**

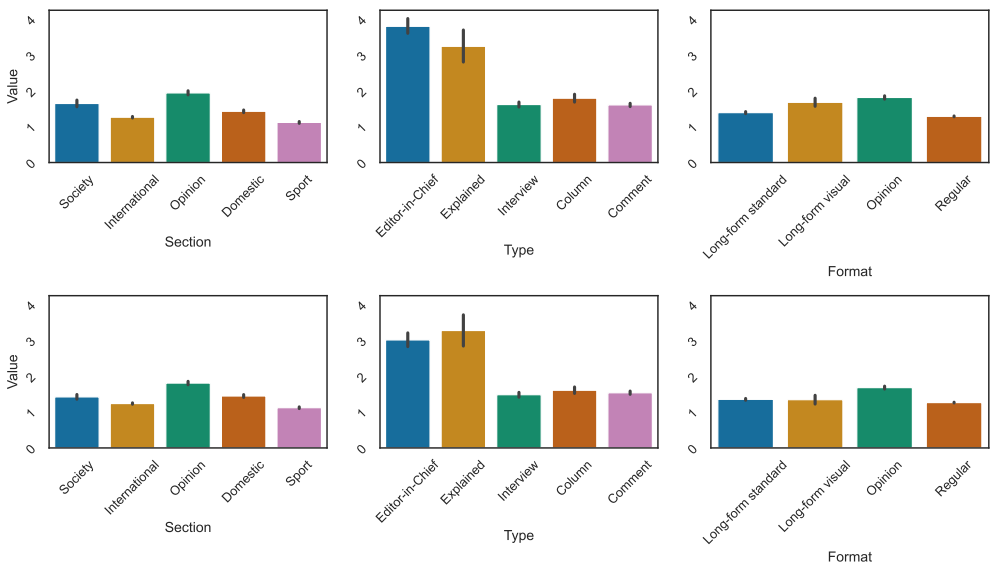


Figure 5 **Value of optimal policy by the sections, types, and formats of items. Error bars are 95% confidence intervals across 1000 bootstrap runs. Top: Swiss front page. Bottom: German front page.**

items from different categories at different rates. For example, our results suggest that our partner company should promote relatively more items belonging to the "Opinion" section but fewer belonging to the "International" section. We also find heterogeneity concerning what types of items should be promoted on the two front pages. Historically, about 40% of the articles written by the editor-in-chief were promoted. For the Swiss but not the German front page, the optimal policy recommends the rate should increase to around 95%. A possible explanation stems from that the newspaper is based in Switzerland. Thus, articles by the editor-in-chief may resonate more with the Swiss market. This is also supported by Fig. 6, which shows that those articles have a substantially higher CATE for the Swiss front page.

Previous research has found that the sentiment of content is a strong driver of views and engagement (Berger and Milkman 2012, Upworthy 2012, Robertson et al. 2023). Here, we find the optimal policy promotes positive, neutral, and negative content at a similar rate (top row of Fig. 7). The result for the optimal policy is explained by that the estimated CATE of promotion is roughly the same for different content (bottom row of Fig. 7). We later discuss this finding in Sec. 6.

5.5. Sensitivity Analysis of Varying Capacities

We examine the sensitivity of our optimal policy to different capacity constraints. We evaluate two scenarios: setting the capacity C_{tm} per channel and time period to $C_{tm} = 30$ or to $C_{tm} = 60$ items. The former corresponds to the historical average number of promotions per channel and

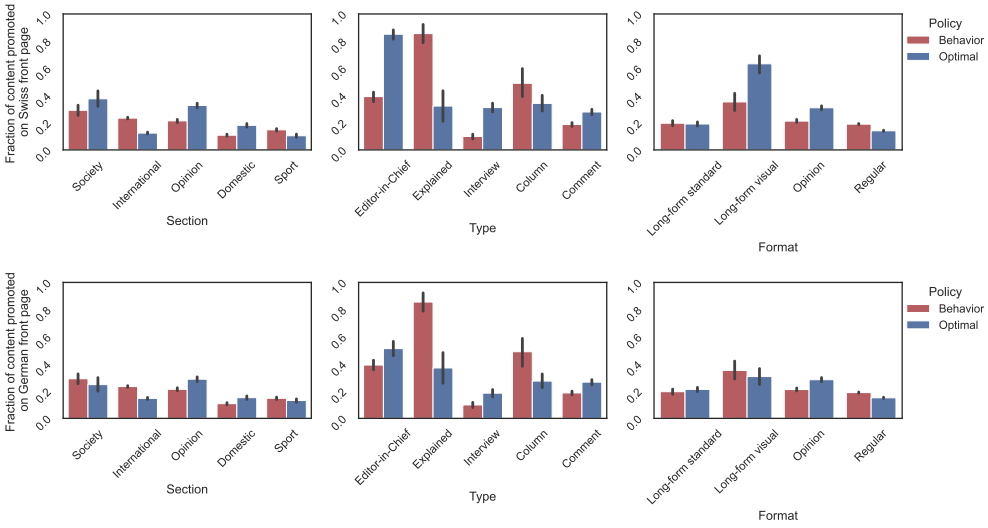


Figure 6 **Fraction of items promoted by optimal policy vs. behavior policy by content category. Error bars are 95% confidence intervals across 1000 bootstrap runs. Top: Swiss front page. Bottom: German front page.**

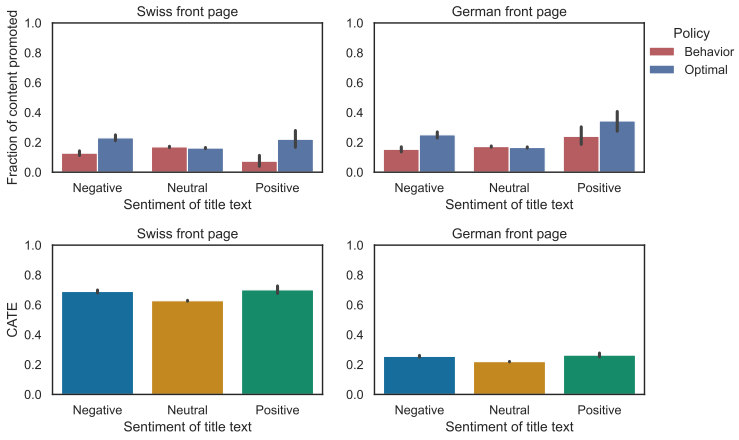


Figure 7 **Top: Fraction of content promoted by optimal policy vs. behavior policy by sentiment of content title. Bottom: Estimated CATE by sentiment of content title. Error bars are 95% confidence intervals across 1000 bootstrap runs.**

time period, whereas the latter is the largest number of items that can be shown on a front page at a time. The results are reported in Table 2. We find that the optimal policy performs best across both scenarios. For a capacity of 60 items, the gain of the optimal policy over current practice is 13.21% for the Swiss front page and by 5.05% for the German front page, respectively.

Table 2 Sensitivity of policy value when varying the capacity constraint.

	Capacity $C_{tm} = 30$ promotions				*Capacity $C_{tm} = 60$ promotions*			
Policy	Value	SE	Regret	Gain	Value	SE	Regret	Gain
Swiss front page:								
Randomized policy	1.188	0.0067	11.54%	−7.69 %	1.188	0.0067	18.46%	−7.69 %
Autoregressive policy	1.304	0.0074	2.90%	1.32%	1.400	0.0079	3.91%	8.78%
Naïve policy	1.320	0.0075	1.71%	2.56%	1.420	0.0080	2.54%	10.33%
UCB policy	1.335	0.0075	0.60%	3.73%	1.442	0.0081	1.03%	12.04%
LCB policy	1.292	0.0073	3.80%	0.39%	1.389	0.0078	4.67%	7.93%
Optimal policy	**1.343**	**0.0076**	**0.00%**	**4.35%**	**1.457**	**0.0082**	0.00%	**13.21%**
German front page:								
Randomized policy	1.244	0.0070	5.18%	−3.34 %	1.244	0.0070	7.99%	−3.34 %
Autoregressive policy	1.295	0.0073	1.30%	0.62%	1.324	0.0075	2.07%	2.87%
Naïve policy	1.296	0.0073	1.22%	0.70%	1.328	0.0075	1.78%	3.19%
UCB policy	1.308	0.0074	0.30%	1.64%	1.344	0.0076	0.59%	4.43%
LCB policy	1.278	0.0072	2.59%	−0.70 %	1.307	0.0074	3.33%	1.55%
Optimal policy	**1.312**	**0.0074**	**0.00%**	**1.94%**	**1.352**	**0.0076**	**0.00%**	**5.05%**

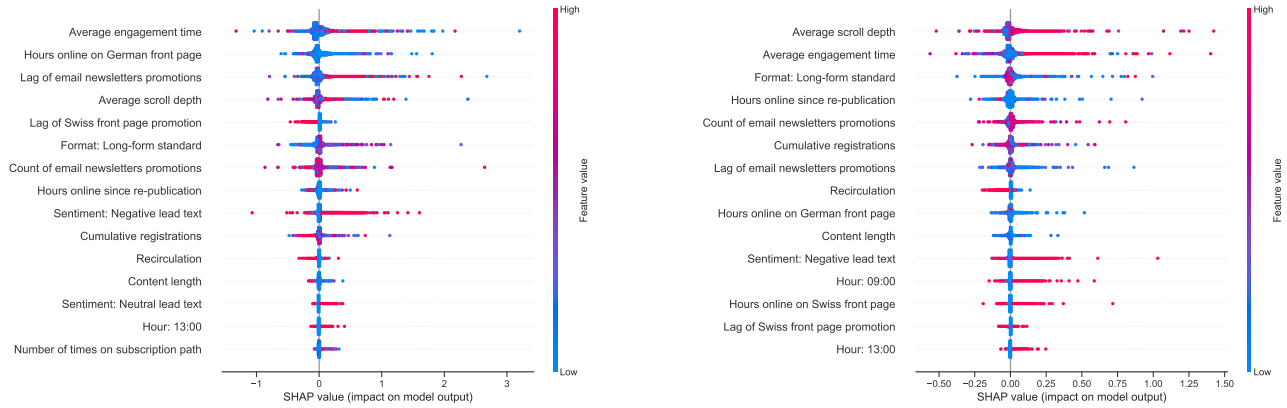
Notes. Value = average performance score across content and time periods. Larger is better. Best value in bold. SE: standard error. Regret: decrease in value relative to our optimal policy. Gain: Increase in value relative to behavior policy.

5.6. Post-Hoc Explainability of Our Optimal Policy

We now aim to understand which covariates are relevant for the decision-making of our optimal policy. We build upon our two approaches for post-hoc explainability, namely, the SHAP value method and the post-Lasso inference. We apply both methods to the CATE function to estimate which covariates predict the heterogeneity in the incremental effect of promotion, thus indicating which type of items benefit relatively more from promotion. The strength of the SHAP value method is that it accounts for complex non-linearties in the CATE heterogeneity, whereas the benefit of our post-Lasso inference approach is that it allows for valid statistical inference on marginal CATE heterogeneity.

We first turn to results from the SHAP value method. We use it to provide initial findings of the heterogeneity that can be further explained by the post-Lasso inference method. The results are shown in Fig. 8. The covariates are ordered from top to bottom in terms of their mean absolute SHAP values, and the dots represent the SHAP values of the covariates in terms of explaining heterogeneity in the CATE. Thus, the horizontal spread of dots for a given covariate corresponds to the variance in the CATE that is explained by that covariate, relative to the average CATE. Comparing the SHAP values of the CATE to the Swiss and the German front pages, we find that 14 out of the top-15 covariates are the same, but that their relative importance differs slightly. In particular, average engagement time, average scroll depth, and previous email newsletter promo-

tions are among the top-5 most important covariates in terms of explaining the heterogeneity in the CATE of promotion on both front pages.



(a) Swiss front page (b) German front page

Figure 8 SHAP values for the top-15 covariates with the largest feature importance on the predictions of the CATE function (in descending order).

We now move to the result from the post-Lasso inference approach (see Table 3). Recall that the sample splitting allows for asymptotically valid inference on the marginal treatment effect heterogeneity with respect to the selected covariates, even when the true CATE function is non-linear and unknown (Belloni and Chernozhukov 2013, Belloni et al. 2014, Battocchi et al. 2019). Moreover, we base our inference on heteroskedasticity-robust standard errors to account for that the variance in the marginal treatment effect heterogeneity with respect to a covariate may be non-constant. Our estimates are thus robust under minimal assumptions on the CATE.

First, within channels, the marginal effects of covariates on heterogeneity in the CATE vary substantially. On average, and all else equal, a one standard deviation increase in the average engagement time (relative to the number of characters of the content) is predicted to have a statistically significant increase on the CATE by 0.024 and 0.014 points for the Swiss and German front page, respectively. Moreover, the CATE of promoting an item on either front page is lower if the item has already been promoted in email newsletters. This can be expected since the users have then likely already seen the item. Among past performance indicators (up until the start of the current time period), we find that more views, longer engagement time, and more registrations predict a higher CATE. Among content types, articles written by the editor-in-chief benefit the most from promotion on the Swiss front page, but are not predicted to benefit from promotion

on the German front page. This is consistent with our previous results that our optimal policy promotes articles by the editor-in-chief relatively more often to the Swiss front page than the German front page. Second, we also find some differences in marginal treatment effect heterogeneity across the channels. For example, articles in a visual format are predicted to have a lower CATE for the Swiss front page and a higher CATE for the German front page, relative to content in standard format. Additional results evaluating the performance in the first-stage Lasso are in Appendix F.

Overall, we find that covariates with large feature importance in our analysis based on the SHAP value method (e.g., average engagement time and average scroll depth) tend to have a comparatively large absolute coefficient in the post-Lasso inference approach. Therefore, the results from both approaches for post-hoc explainability appear to be largely consistent.

6. Discussion

In this paper, we presented a framework for learning and evaluating decision-making policies that optimize the selection of which content to promote across digital distribution channels. Our framework addresses a common decision problem faced in practice by newspapers, blogs, and owners of social media pages, who aim to choose which of their content to show to users such that an outcome of interest like traffic, engagement, or conversions is maximized. Here, decisions are typically not personalized for technological and privacy reasons. For example, Facebook, Instagram, and LinkedIn do not allow owners of social media pages to personalize their content to visitors; instead, the same items are shown to the entire user base. Likewise, newspapers such as the New York Times and our partner company Neue Zürcher Zeitung see the curation of their distribution channels as a core value of their product (Rockwell 2019).

Our framework has several strengths for practice. First, our framework learns the optimal policy for content promotions from historical data. This is an advantage over uplift modeling (e.g., Devriendt et al. 2018), which leverages randomized experiments and may thus be costly or undesired in practice. In contrast, our framework relies on non-randomized, observational data, which is typically widely available in practice. Second, a challenge of learning from observational data is establishing non-parametric identification of counterfactuals, which we have shown above. This enables the use of machine learning, which makes it straightforward for practitioners to use off-the-shelf tools to select models, capture non-linearities, and accommodate high-dimensional data. Third, our framework is computationally efficient and thus scales to real-world settings as encountered in practice.

	CATE for Swiss front page			*CATE for German front page*		
Covariate	Coef.	SE	p-value	Coef.	SE	p-value
Past performance indicators:						
Article views	0.033	0.013	0.009	0.011	0.004	0.005
Average engagement time	0.024	0.007	0.001	0.014	0.004	0.000
Average scroll depth	−0.033	0.002	0.000	0.011	0.001	0.000
Recirculation	−0.009	0.001	0.000	−0.005	0.001	0.000
Cumulative registrations	0.059	0.007	0.000	0.023	0.002	0.000
Number of times on subscription path	0.004	0.003	0.154	0.001	0.001	0.292
Time-invariant content characteristics:						
Content length	−0.002	0.001	0.262	−0.004	0.001	0.000
Section: Business & finance	−0.062	0.007	0.000	−0.032	0.006	0.000
Section: Celebrities & events	0.066	0.012	0.000	−0.005	0.006	0.433
Section: Culture	−0.077	0.008	0.000	−0.035	0.006	0.000
Section: Domestic	−0.028	0.009	0.003	−0.029	0.006	0.000
Section: Education	0.097	0.016	0.000	0.036	0.009	0.000
Section: International	−0.066	0.007	0.000	−0.024	0.006	0.000
Section: International politics	−0.047	0.010	0.000	−0.009	0.009	0.292
Section: NZZ in English	0.053	0.012	0.000	0.059	0.008	0.000
Section: Opinion	−0.170	0.011	0.000	−0.022	0.008	0.005
Section: Science & technology	−0.051	0.010	0.000	−0.024	0.007	0.001
Section: Society	−0.002	0.035	0.950	−0.032	0.010	0.002
Section: Sport	−0.034	0.010	0.000	−0.026	0.006	0.000
Section: Visuals	−0.077	0.011	0.000	−0.077	0.010	0.000
Section: Zurich	−0.055	0.008	0.000	−0.030	0.006	0.000
Type: Breaking news	−0.175	0.006	0.000	−0.022	0.003	0.000
Type: Column	0.025	0.013	0.061	0.016	0.006	0.011
Type: Comment	0.088	0.010	0.000	0.034	0.005	0.000
Type: Editor-in-Chief	0.431	0.024	0.000	—	—	—
Type: Explained	0.106	0.022	0.000	0.008	0.007	0.295
Type: Guest comment	0.115	0.011	0.000	—	—	—
Type: Interview	0.052	0.011	0.000	0.014	0.004	0.000
Type: News in brief	−0.040	0.017	0.021	−0.008	0.004	0.054
Format: Long-form standard	−0.038	0.005	0.000	0.002	0.003	0.604
Format: Long-form visual	−0.040	0.011	0.000	0.013	0.006	0.022
Format: Opinion	0.023	0.009	0.017	0.018	0.005	0.000
Sentiment: Negative lead text	0.113	0.006	0.000	0.016	0.003	0.000
Sentiment: Positive lead text	—	—	—	—	—	—
Sentiment: Negative SEO title	0.029	0.008	0.001	0.007	0.003	0.010
Sentiment: Positive SEO title	0.035	0.018	0.053	0.006	0.004	0.075
Sentiment: Negative title	−0.023	0.006	0.000	0.004	0.002	0.073
Sentiment: Positive title	−0.031	0.009	0.001	—	—	—
Time information:						
Hours online on Swiss front page	0.002	0.001	0.000	0.001	0.000	0.000
Hours online on German front page	−0.001	0.001	0.189	0.001	0.000	0.000
Hours online since publication	−0.005	0.002	0.040	−0.008	0.001	0.000
Past promotions:						
Count of social media promotions	0.040	0.007	0.000	−0.026	0.001	0.000
Count of email newsletters promotions	−0.010	0.001	0.000	−0.007	0.001	0.000
Count of push notification promotions	−0.047	0.007	0.000	0.020	0.001	0.000
Lag of social media promotions	0.005	0.001	0.000	0.004	0.001	0.001
Lag of email newsletters promotions	0.006	0.001	0.000	0.002	0.001	0.031
Lag of push notification promotion	0.003	0.002	0.132	0.002	0.001	0.095
Lag of Swiss front page promotion	−0.017	0.002	0.000	−0.011	0.001	0.000
Lag of German front page promotion	−0.069	0.002	0.000	0.000	0.001	0.993
Hour fixed effects	Yes			Yes		
Weekday fixed effects	Yes			Yes		
R-squared	0.195			0.151		
Adjusted R-squared	0.194			0.150		
AIC	14 760			−51 260		
BIC	15 360			−50 690		

Coef.: point estimate of marginal effect. SE: robust (HC2) standard errors.

Table 3 Marginal effects of covariates (using pre-treatment values) for explaining heterogeneity in the CATE. The estimates are obtained from the BLP of the CATE function estimated with OLS in our post-Lasso inference procedure. Cells with a "—" denote that the corresponding covariate was dropped by the first-stage Lasso regularization. Continuous covariates are z-standardized to zero mean and unit variance for better interpretability.

Our work contributes to marketing science in three ways. First, one stream of research has studied content optimization, such as through the optimization of news recommendation, ad delivery, or website design at the user level (e.g., Agarwal et al. 2009, Hauser et al. 2009, Kale et al. 2010, Li et al. 2010, Garcin et al. 2013, Urban et al. 2014, Schwartz et al. 2017). We add to this stream and offer a tailored, data-driven tool for making optimal content promotions across digital distribution channels. Second, a different stream has demonstrated that decision-making based on the incremental-effects approach is optimal for personalized targeting (Bodapati 2008, Ascarza 2018, Hitsch and Misra 2018, Lemmens and Gupta 2020). Here, we add by studying a new decision problem, namely, optimal promotions at the content level for which we adapt the incremental-effects approach. Third, another stream in marketing has studied the use of off-policy learning for optimal targeting (e.g., Hitsch and Misra 2018, Simester et al. 2020, Ellickson et al. 2022, Huang and Ascarza 2023, Liu 2022, Yoganarasimhan et al. 2022, Yang et al. 2023). Many of these works are based on methods that are either not doubly robust or, if they are, use DML. In contrast, we use DRL, which, different from DML, allows for inference without making the assumption that the estimated CATE contains the true CATE. This allows us to directly apply data-driven approaches for post-hoc explainability to understand how to make optimal decisions. Furthermore, due to the use of DRL, our framework should be highly robust for use in practice.

Our work offers direct managerial implications. We show theoretically and empirically that decisions based on incremental-effects approach are highly effective and outperform common baselines for content promotions. Hence, managers seeking to optimize the distribution of content should not necessarily promote the content that is predicted to be successful in terms of outcomes but, instead, promote the content whose outcome is predicted to change the most if it is promoted. Moreover, our results suggest how editors should further improve their strategies for promoting different types of content. For example, a common phrase from the newsroom is *"If it bleeds, it leads"*, implying that news about negative events tends to generate more clicks. For example, a recent study based on large-scale field experiments found that negative words in news headlines increase click-through rates (Robertson et al. 2023). However, our results call for caution when following such common wisdom. In contrast, our results imply that also positive stories can be beneficial to promote as our partner company not only focuses on short-term clicks but it considers the long-term utility as measured by the proprietary performance score.

Finally, we foresee many valuable applications of off-policy learning across a variety of decision problems in marketing. In particular, our framework is relevant for applications with rich historical

data but where randomized experimentation (e.g., A/B-testing) and personalization are not wanted or not feasible. In principle, our off-policy framework can be applied to problems where the decision is not at the user level, but at the product and item level. For instance, e-commerce companies can use our off-policy learning to decide which products to show to new users on their landing pages.

7. Concluding Remarks

In this work, we leverage off-policy learning to optimize content promotions across digital distribution channels. We theoretically show that the policy from our proposed framework is optimal and empirically demonstrate that our framework outperforms common baselines. We find that our framework also offers large improvements over the decision-making by the experts from our partner company. Our work thus contributes to previous research by presenting a data-driven tool for optimal content promotions in settings where personalization is neither wanted nor feasible.

Funding and Competing Interests

Author 1 and Author 2 have no competing interests. Author 3 is employed at the partner company.

References

Agarwal D, Chen BC, Elango P (2009) Explore/Exploit Schemes for Web Content Optimization. *IEEE International Conference on Data Mining (ICDM)* .

Archak N, Ghose A, Ipeirotis PG (2011) Deriving the Pricing Power of Product Features by Mining Consumer Reviews. *Management Science* 57(8):1485–1509.

Ascarza E (2018) Retention Futility: Targeting High-Risk Customers Might Be Ineffective. *Journal of Marketing Research* 55(1):80–98.

Athey S, Imbens GW (2016) Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of the National Academy of Sciences* 113(27):7353–7360.

Athey S, Tibshirani J, Wager S (2019) Generalized Random Forests. *Annals of Statistics* 47(2):1179–1203.

Athey S, Wager S (2021) Policy Learning with Observational Data. *Econometrica* 89(1):133–161.

Battocchi K, Dillon E, Hei M, Lewis G, Oka P, Oprescu M, Syrgkanis V (2019) EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. Version 0.x.

Belloni A, Chernozhukov V (2013) Least Squares after Model Selection in High-Dimensional Sparse Models. *Bernoulli* 19(2):521–547.

Belloni A, Chernozhukov V, Hansen C (2014) Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies* 81(2):608–650.

Berger J, Humphreys A, Ludwig S, Moe WW, Netzer O, Schweidel DA (2020) Uniting the Tribes: Using Text for Marketing Insight. *Journal of Marketing* 84(1):1–25.

Berger J, Milkman KL (2012) What Makes Online Content Viral? *Journal of Marketing Research* 49(2):192–205.

Besbes O, Gur Y, Zeevi A (2016) Optimization in Online Content Recommendation Services: Beyond Click-Through Rates. *Manufacturing and Service Operations Management* 18(1):15–33.

Bodapati AV (2008) Recommendation Systems with Purchase Data. *Journal of Marketing Research* 45(1):77–93.

Breiman L (2001) Random Forests. *Machine Learning* 45(1):5–32.

Center PR (2021) Share of newspaper advertising revenue coming from digital advertising . URL https://www.pewresearch.org/journalism/chart/sotnm-newspapers-percentage-of-newspaper-advertising-revenue-coming-from-digital/.

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018a) Double/Debiased Machine Learning for Treatment and Structural Parameters. *Econometrics Journal* 21(1):C1–C68.

Chernozhukov V, Demirer M, Duflo E, Fernández-Val I, Athey S, Buchinsky M, Chetverikov D, Imbens G, Lehrer S, Luo S, Kasy M, Murphy S, Newey W, Power P (2018b) Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, With an Application To Immunization in India.

Devriendt F, Moldovan D, Verbeke W (2018) A Literature Survey and Experimental Evaluation of the State-of-the-Art in Uplift Modeling: A Stepping Stone Toward the Development of Prescriptive Analytics. *Big Data* 6(1):13–41.

DigiDay (2018) How Swiss news publisher NZZ built a flexible paywall using machine learning. URL https://digiday.com/media/swiss-news-publisher-nzz-built-flexible-paywall-using-machine-learning/.

DigiDay (2023) Less than half of The Independent's revenue came from advertising in 2022. URL https://digiday.com/media/less-than-half-of-the-independents-revenue-came-from-advertising-in-2022/.

Dimakopoulou M, Vlassis N, Jebara T (2019) Marginal Posterior Sampling for Slate Bandits. *International Joint Conference on Artificial Intelligence*.

Dirick YMI, Jennergren LP (1975) On the Optimality of Myopic Policies in Sequential Decision Problems. *Management Science* 21(5):550–556.

Dudík M, Erhan D, Langford J, Li L (2014) Doubly Robust Policy Evaluation and Optimization. *Statistical Science* 29(4):485–511.

Ellickson PB, Kar W, Reeder III JC (2022) Estimating Marketing Component Effects: Double Machine Learning from Targeted Digital Promotions. *Marketing Science* .

EMarketer (2020a) Marketers Expect Content-Driven Campaigns to Increase in 2020. URL https://www.emarketer.com/content/marketers-expect-content-driven-campaigns-to-increase-in-2020.

EMarketer (2020b) US Content Marketing Spending, by Channel, 2019 & 2020. URL https://www.emarketer.com/chart/241366/us-content-marketing-spending-by-channel-2019-2020-billions-of-total.

Ertefaie A, Strawderman RL (2018) Constructing Dynamic Treatment Regimes over Indefinite Time Horizons. *Biometrika* 105(4):963–977.

Foster DJ, Syrgkanis V (2019) Statistical Learning with a Nuisance Component. *Conference on Learning Theory*.

Friedman JH (2001) Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29(5):1189–1232.

Garcin F, Dimitrakakis C, Faltings B (2013) Personalized News Recommendation with Context Trees. *ACM Conference on Recommender Systems* 105–112.

Guardian (2006) How often do readers return to newspaper websites. URL https://www.theguardian.com/media/greenslade/2006/jul/18/howoftendoreadersreturnto.

Guhr O (2022) German Sentiment Classification with BERT. URL https://github.com/oliverguhr/german-sentiment-lib.

Guhr O, Schumann AK, Bahrmann F, Böhme HJ (2020) Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems. *Language Resources and Evaluation Conference*.

Hauser JR, Urban GL, Liberali G, Braun M (2009) Website Morphing. *Marketing Science* 28(2):202–223.

Hernán M, Robins JM (2020) *Causal Inference: What If* (Boca Raton: Chapman & Hall/CRC).

Hitsch GJ, Misra S (2018) Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation. *Available at SSRN* 1–64.

Holland P (1986) Statistics and Causal Inference. *Journal of the American Statistical Association* 81(396):945–960.

Huang TW, Ascarza E (2023) Doing More with Less: Overcoming Ineffective Long-term Targeting Using Short-Term Signals. *SSRN Electronic Journal* .

Imbens GW, Rubin DB (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Cambridge University Press, New York, NY).

Johar M, Mookerjee V, Sarkar S (2014) Selling vs. Profiling: Optimizing the Offer Set in Web-Based Personalization. *Information Systems Research* 25(2):285–306.

Kale S, Reyzin L, Schapirey RE (2010) Non-Stochastic Bandit Slate Problems. *Advances in Neural Information Processing Systems*.

Kennedy EH (2020) Towards Optimal Doubly Robust Estimation of Heterogeneous Causal Effects. *arXiv preprint arXiv:2004.14497* .

Kenton JDMWC, Toutanova LK (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)*.

Lada A, Peysakhovich A, Aparicio D, Bailey M (2019) Observational Data for Heterogeneous Treatment Effects with Application to Recommender Systems. *ACM Conference on Economics and Computation* .

Lai TL, Robbins H, et al. (1985) Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Mathematics* 6(1):4–22.

Lemmens A, Gupta S (2020) Managing Churn to Maximize Profits. *Marketing Science* 39(5):956–973.

Li L, Chu W, Langford J, Schapire RE (2010) A Contextual-Bandit Approach to Personalized News Article Recommendation. *International Conference on World Wide Web* .

Liao P, Klasnja P, Murphy S (2021) Off-Policy Estimation of Long-Term Average Outcomes with Applications to Mobile Health. *Journal of the American Statistical Association* 116(533):382–391.

Liu Q, Li L, Tang Z, Zhou D (2018) Breaking the Curse of Horizon: Infinite-Horizon Off-Policy Estimation. *Advances in Neural Information Processing Systems* 5356–5366.

Liu X (2022) Dynamic Coupon Targeting Using Batch Deep Reinforcement Learning: An Application to Livestream Shopping. *Marketing Science* .

Lundberg SM, Lee SI (2017) A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* .

Murphy SA, Van der Laan MJ, Robins JM, Bierman KL, Coie JD, Greenberg MT, Lochman JE, McMahon RJ, Pinderhughes E (2001) Marginal Mean Models for Dynamic Regimes. *Journal of the American Statistical Association* 96(456):1410–1423.

Oprescu M, Syrgkanis V, Wu ZS (2019) Orthogonal Random Forest for Causal Inference. *International Conference on Machine Learning*.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Pelger M, Zou J (2022) Inference for Large Panel Data with Many Covariates. *arXiv preprint arXiv:2301.00292* .

Reiss PC (2011) Structural Workshop Paper – Descriptive, Structural, and Experimental Empirical Methods in Marketing Research. *Marketing Science* 30(6):950–964.

Robertson CE, Pröllochs N, Schwarzenegger K, Parnamets P, Van Bavel JJ, Feuerriegel S (2023) Negativity Drives Online News Consumption. *Nature Human Behaviour* 7(5):812–822.

Robins JM (1999) Robust Estimation in Sequentially Ignorable Missing Data and Causal Inference Models. *American Statistical Association Section on Bayesian Statistical Science*.

Robins JM, Rotnitzky A (1995) Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association* 90(429):122.

Robins JM, Rotnitzky A, Zhao LP (1994) Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association* 89(427):846–866.

Rockwell N (2019) News in the Age of Algorithmic Recommendation. URL `https://www.datacouncil.ai/talks/news-in-the-age-of-algorithmic-recommendation`.

Rubin DB (1974) Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal Of Educational Psychology* 66(5):688–701.

Schwartz EM, Bradlow ET, Fader PS (2017) Customer Acquisition via Display Advertising Using Multi-Armed Bandit Experiments. *Marketing Science* 36(4):500–522.

Semenova V, Goldman M, Chernozhukov V, Taddy M (2023) Estimation and Inference on Heterogeneous Treatment Effects in High-Dimensional Dynamic Panels under Weak Dependence. *Quantitative Economics (conditionally accepted)* .

Simester D, Timoshenko A, Zoumpoulis SI (2019) Targeting Prospective Customers: Robustness of Machine-Learning Methods to Typical Data Challenges. *Management Science* (June):2495–2522, URL `http://dx.doi.org/10.1287/mnsc.2019.3308`.

Simester D, Timoshenko A, Zoumpoulis SI (2020) Efficiently Evaluating Targeting Policies: Improving on Champion vs. Challenger Experiments. *Management Science* 66(8):3412–3424.

Smith AN, Seiler S, Aggarwal I (2022) Optimal price targeting. *Marketing Science* 42(3):476–499.

Song Y, Sahoo N, Ofek E (2019) When and How to Diversify A Multi-Category Utility Model of Consumer Response to Content Recommendations. *Management Science* 65(8):3737–3757.

Statista (2022) Umsatz der NZZ-Mediengruppe von 2011 bis 2022. URL `https://de.statista.com/statistik/daten/studie/413228/umfrage/umsatz-der-nzz-mediengruppe/`.

Sutton RS, Barto AG (2018) *Reinforcement Learning: An Introduction* (Cambridge, Massachusetts: MIT press), 2 edition.

Swaminathan A, Krishnamurthy A, Agarwal A, Dudik M, Langford J, Jose D, Zitouni I (2017) Off-Policy Evaluation for Slate Recommendation. *Advances in Neural Information Processing Systems* .

Upworthy (2012) How To Make That One Thing Go Viral. URL `https://www.slideshare.net/Upworthy/how-to-make-that-one-thing-go-viral-just-kidding/25(2012)`.

Urban GL, Liberali GG, MacDonald E, Bordley R, Hauser JR (2014) Morphing Banner Advertising. *Marketing Science* 33(1):27–46.

van der Laan MJ, Rose S (2011) *Targeted Learning: Causal Inference for Observational and Experimental Data*, volume 4 (New York: Springer).

Wager S, Athey S (2018) Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association* 113(523):1228–1242.

Wasserman L, Roeder K (2009) High Dimensional Variable Selection. *Annals of Statistics* 37(5A):2178.

Yang J, Eckles D, Dhillon P, Aral S (2023) Targeting for Long-Term Outcomes. *Management Science* .

Yoganarasimhan H, Barzegary E, Pani A (2022) Design and Evaluation of Optimal Free Trials. *Management Science* .

Zhang Y, Li B, Luo X, Wang X (2019) Personalized Mobile Targeting with User Engagement Stages: Combining a Structural Hidden Markov Model and Field Experiment. *Information Systems Research* 30(3):787–804.

Zhao S, Witten D, Shojaie A (2021) In Defense of the Indefensible: A Very Naive Approach to High-Dimensional Inference. *Statistical Science* 36(4):562–577.

Online Appendix

Appendix A: Front pages

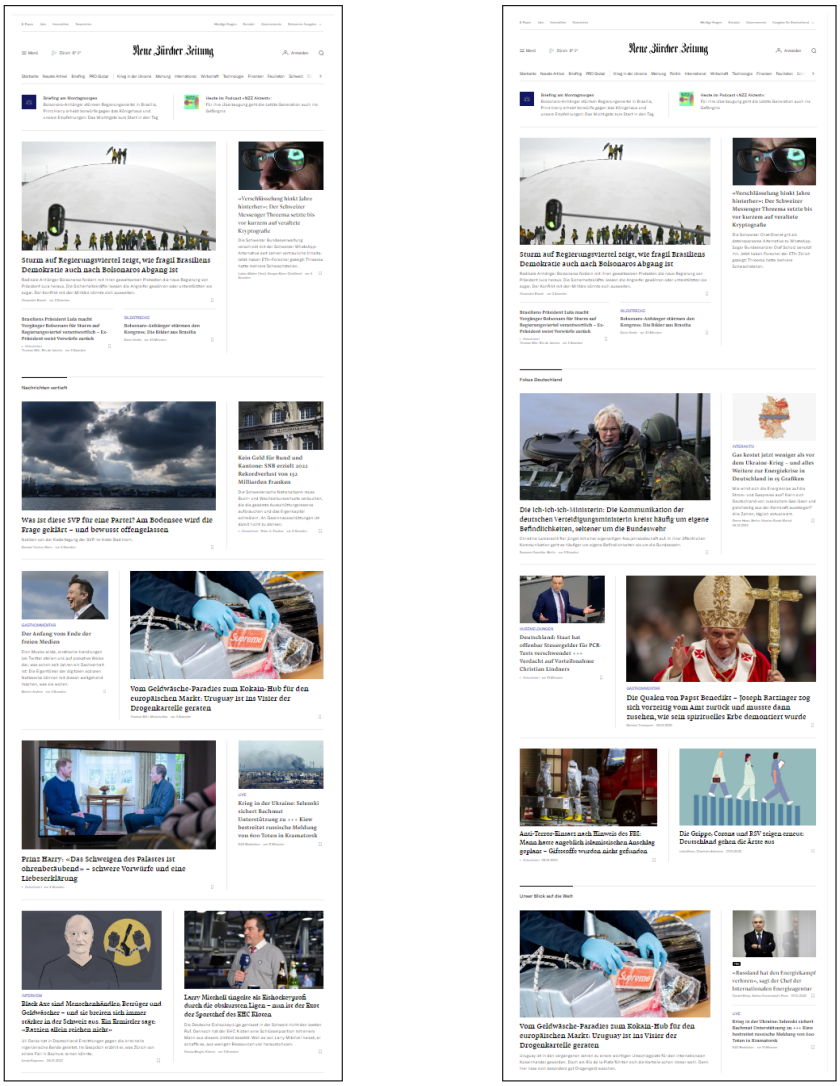


Figure 9 Screenshots of the front page of Neue Züricher Zeitung (due to space, only a few promoted items at the top are shown while the actual list of promoted articles is much longer). Shown: Swiss front page (left) and German front page (right).

Appendix B: Proofs

B.1. Proof of Proposition 1

We now show that Eq. (5) can be written as a function of the observed data distribution. To do so, we use that the density (or, equivalently, the likelihood of a data realization) can be written as a function of both the behavior policy and an arbitrary counterfactual policy.

Let P_π be the joint distribution of the historical data under the behavior policies π_m, $m = 1, \ldots, M$. Under Assumptions 5, 6, and 7, the law of total probability gives that the density of the data into a time period factorizes as

$$p(\mathbf{S}_t, \mathbf{A}_t, Y_t; \pi_1, \ldots, \pi_M) = \frac{\mathrm{d}P_\pi}{\mathrm{d}\mathbb{P}} = p(\mathbf{S}_t) \prod_{m=1}^{M} \pi_m(A_{tm} \mid \mathbf{S}_t)\, p(Y_t \mid \mathbf{S}_t, \mathbf{A}_t), \tag{23}$$

where we used that a probability density function can be defined as the Radon-Nikodým derivative, $\mathrm{d}P_\pi / \mathrm{d}\mathbb{P}$, of its distribution with respect to a reference probability distribution (i.e., measure) $\mathbb{P}$. This definition requires that P_π is absolutely continuous with respect to $\mathbb{P}$, which holds under Assumption 4 (Murphy et al. 2001).[12] Under the same assumptions as above, the density of the data under the joint distribution P_{d_m} under a counterfactual policy d_m is given by

$$p(\mathbf{S}_t, \mathbf{A}_t, Y_t; d_m, \pi_k \colon k \neq m) = \frac{\mathrm{d}P_{d_m}}{\mathrm{d}\mathbb{P}} = p(\mathbf{S}_t)\, \mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\} \prod_{k \neq m} \pi_k(A_{tk} \mid \mathbf{S}_t)\, p(Y_t \mid \mathbf{S}_t, \mathbf{A}_t), \tag{24}$$

where we again used that P_{d_m} is absolutely continuous with respect to $\mathbb{P}$. This may be interpreted as that any sequence of decisions that could occur under the counterfactual policy d_m could also have been observed in the data under the behavior policy of the same channel π_m (Murphy et al. 2001). A change of measure now gives that Eq. (3) can be written as

$$V(d_m) = \frac{1}{T} \sum_{t=1}^{T} \int \left(y(\mathbf{S}_t, A_{tm}, \mathbf{A}_t^{-m}) \frac{\mathrm{d}P_{d_m}}{\mathrm{d}P_\pi} \right) \mathrm{d}P_\pi = \frac{1}{T} \sum_{t=1}^{T} \int y_t \frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\}}{\pi_m(A_{tm} \mid \mathbf{S}_t)} \, \mathrm{d}P_\pi, \tag{25}$$

where the last equality follows from that

$$\frac{\mathrm{d}P_{d_m}}{\mathrm{d}P_\pi} = \frac{\mathrm{d}P_{d_m} / \mathrm{d}\mathbb{P}}{\mathrm{d}P_\pi / \mathrm{d}\mathbb{P}} = \frac{p(\mathbf{S}_t)\, \mathbb{1}\{a_{tm} = d_m(\mathbf{S}_t)\} \prod_{k \neq m} \pi_k(a_{tk} \mid \mathbf{S}_t) p(y_t \mid \mathbf{S}_t, \mathbf{A}_t)}{p(\mathbf{S}_t) \prod_{m=1}^{M} \pi_m(a_{tm} \mid \mathbf{S}_t) p(y_t \mid \mathbf{S}_t, \mathbf{A}_t)} = \frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\}}{\pi_m(A_{tm} \mid \mathbf{S}_t)}. \tag{26}$$

This is the likelihood ratio (also called the importance weight) between Eq. (24) and Eq. (23), where we used Assumption 1 to replace the potential outcome with the observed outcome. Eq. (25) defines the value of the counterfactual policy d_m as an expectation with respect to the observed distribution P_π. This shows that the value of an arbitrary counterfactual policy is identified from historical data. Moreover, the identification is non-parametric, as no parametric assumptions have been required to arrive at Eq. (25). What remains to show is that the value function can be written in the AIPW form of Eq. (5). For this, we let

$$\mu(\mathbf{S}_t, d(\mathbf{S}_t), \mathbf{A}_t^{-m}) \coloneqq \mathbb{E}[Y(\mathbf{S}_t, d(\mathbf{S}_t), \mathbf{A}_t^{-m}) \mid \mathbf{S}_t] = \mathbb{E}[Y_t \mid \mathbf{S}_t, d(\mathbf{S}_t), \mathbf{A}_t^{-m}] = \int y_t \, \mathrm{d}P_{Y_t \mid \mathbf{S}_t, d(\mathbf{S}_t), \mathbf{A}_t^{-m}}(y_t), \tag{27}$$

[12] We view the items in the historical data as independent and identically distributed draws from the distribution P_π and thus perform inference on the superpopulation of trajectories that could have been realized under the behavior policy. The reason for drawing inference on the superpopulation instead of the sample itself is that we wish to learn causal relationships that generalize beyond the data in the sample.

where the third equality follows from Assumption 1. We then note that

$$\mu(\mathbf{S}_t, d_m(\mathbf{S}_t), \mathbf{A}_t^{-m}) = \mathbb{E}\left[Y_t \frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\}}{\pi_m(A_{tm} \mid \mathbf{S}_t)} \mid \mathbf{S}_t\right] \tag{28}$$

$$= \mathbb{E}\left[Y(\mathbf{S}_t, A_{tm}, \mathbf{A}_t^{-m}) \frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\}}{\pi_m(A_{tm} \mid \mathbf{S}_t)} \mid \mathbf{S}_t\right] \tag{29}$$

$$= \mathbb{E}\left[Y(\mathbf{S}_t, A_{tm}, \mathbf{A}_t^{-m}) \mid \mathbf{S}_t\right] \mathbb{E}\left[\frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\}}{\pi_m(A_{tm} \mid \mathbf{S}_t)} \mid \mathbf{S}_t\right] \tag{30}$$

$$= \mu(\mathbf{S}_t, A_{tm}, \mathbf{A}_t^{-m}) \frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\}}{\pi_m(A_{tm} \mid \mathbf{S}_t)}, \tag{31}$$

where the first equality holds under Assumption 4, the second under Assumption 1, and the third by Assumption 3, and the last follows by definition of Eq. (27) and basic probability theory. Thus,

$$V(d_m) = \frac{1}{T} \sum_{t=1}^{T} \int y_t \frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\}}{\pi_m(A_{tm} \mid \mathbf{S}_t)} \, \mathrm{d}P_\pi \tag{32}$$

$$= \frac{1}{T} \sum_{t=1}^{T} \int \left(y_t \frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\}}{\pi_m(A_{im} \mid \mathbf{S}_t)} + \underbrace{\mu(\mathbf{S}_t, d_m(\mathbf{S}_t), \mathbf{A}_t^{-m}) - \mu(\mathbf{S}_t, d_m(\mathbf{S}_t), \mathbf{A}_t^{-m})}_{=0} \right) \mathrm{d}P_\pi \tag{33}$$

$$= \frac{1}{T} \sum_{t=1}^{T} \int \left(y_t \frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\}}{\pi_m(A_{tm} \mid \mathbf{S}_t)} + \mu(\mathbf{S}_t, d_m(\mathbf{S}_t), \mathbf{A}_t^{-m}) - \mu(\mathbf{S}_t, A_{tm}, \mathbf{A}_t^{-m}) \frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\}}{\pi_m(A_{tm} \mid \mathbf{S}_t)} \right) \mathrm{d}P_\pi \tag{34}$$

$$= \frac{1}{T} \sum_{t=1}^{T} \int \mu(\mathbf{S}_t, d_m(\mathbf{S}_t), \mathbf{A}_t^{-m}) + \frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\}}{\pi_m(A_{tm} \mid \mathbf{S}_t)} (y_t - \mu(\mathbf{S}_t, A_{tm}, \mathbf{A}_t^{-m}) \, \mathrm{d}P_\pi, \tag{35}$$

where the third equality follows from our derivations in Eq. (28)–(31). We now see that Eq. (35) is the oracle AIPW formula in Proposition 1. This concludes the proof. □

B.2. Proof of Theorem 1

We now show that the policy in Eq. (7) is a solution to Eq. (4) and is therefore optimal.

Due to the stationarity of the policies and the myopic nature of the decision problem, the long-run average value is maximized by maximizing the value per time period. That is, for all $t \in \mathcal{V}$,

$$d_m^* = \arg\max_{d_m \in \mathcal{D}} V_t(d_m) \tag{36a}$$

$$\text{s.t.} \sum_{i \in \mathcal{I}_t} d_m(\mathbf{s}_{it}) \leq C_{tm} \tag{36b}$$

with $V_t(d_m)$ given by Eq. (2). Thus, if we can show that the solution to Eq. (36) equals Eq. (7), we are done.

Since $d_m(\mathbf{S}_{it}) \in \{0, 1\}$, we can always decompose the expected outcome of an item under a policy decision given a state as the conditional expectation of potential outcome if the item would not be promoted plus the CATE if the policy decision is to promote the item, i. e.,

$$\mathbb{E}[Y(\mathbf{S}_{it}, d_m(\mathbf{S}_{it}), \mathbf{A}_{it}^{-m}) \mid \mathbf{S}_{it}] = \mathbb{E}[Y(\mathbf{S}_{it}, A_{itm} = 0, \mathbf{A}_{it}^{-m}) + d_m(\mathbf{S}_{it}) \cdot \tau_m(\mathbf{S}_{it}) \mid \mathbf{S}_{it}], \tag{37}$$

where $\tau_m(\mathbf{S}_{it})$ is given by Eq. (6). The value function for any time period t can thus be written as

$$V_t(d_m) = \mathbb{E}[Y(\mathbf{S}_{it}, A_{itm} = 0, \mathbf{A}_{it}^{-m}) + d_m(\mathbf{S}_{it}) \cdot \tau_m(\mathbf{S}_{it})], \tag{38}$$

where the expectation is over the states. Given the capacity constraint in Eq. (36b), it follows that

$$\max V_t(d_m) = \mathbb{E}[Y(\mathbf{S}_{it}, A_{itm} = 0, \mathbf{A}_{it}^{-m})] + \mathbb{1}\{\tau_m(\mathbf{S}_{it}) \geq \tau_{tm}^{(C_{tm})}\}] \cdot \tau_m(\mathbf{S}_{it}). \tag{39}$$

This directly implies that

$$\arg\max_{d_m \in \mathcal{D}} V_t(d_m) = \mathbb{1}\{\tau_m(\mathbf{S}_{it}) \geq \tau_{tm}^{(C_{tm})}\}], \tag{40}$$

which equals d_m^* in Eq. (7) in Theorem 1. This concludes the proof. □

B.3. Proof of Proposition 2

We now show that an estimated optimal policy is an unbiased estimate of the oracle-optimal policy.

We first note that the optimal policy is fully determined by the rank ordering of items in terms of the CATEs. Thus, a sufficient condition for the unbiasedness of an estimate of the policy is also called rank unbiasedness (Lada et al. 2019), that is, the ranks of items in terms of their estimated CATEs preserve their ranks with respect to their true CATEs. That is, for any two items $i, j \in \mathcal{I}_t$ in time period t with CATE estimates $\widehat{\tau}(\mathbf{s}_{it})$ and $\widehat{\tau}(\mathbf{s}_{jt})$, we have $\mathbb{E}[\widehat{\tau}(\mathbf{s}_{it})] > \mathbb{E}[\widehat{\tau}(\mathbf{s}_{jt})]$ if $\tau(\mathbf{s}_{it}) > \tau(\mathbf{s}_{jt})$. Note that this condition can hold even if $\widehat{\tau}(\mathbf{s}_{it})$ and $\widehat{\tau}(\mathbf{s}_{jt})$ are not unbiased estimates of the true CATEs $\tau(\mathbf{s}_{it})$ and $\tau(\mathbf{s}_{jt})$.

A sufficient condition for rank unbiasedness is that confounding in the CATE estimate $\widehat{\tau}(\mathbf{s}_{it})$ due to omitted variables is strictly larger than confounding in $\widehat{\tau}(\mathbf{s}_{jt})$ (Lada et al. 2019). By Assumption 3, there are no unmeasured confounders, and, hence, no omitted variable bias in any estimated CATE. This holds for any CATE function. It follows that an estimated optimal policy is unbiased. □

B.4. Proof of Theorem 2

We now show that the AIPW estimator of policy value given by Eq. (21) is doubly robust. We prove this for an arbitrary policy d_m. Note that, because our procedure for estimating an optimal policy is also unbiased (cf. Proposition 2), the proof implies that the estimated policy value estimate of the optimal policy is doubly robust unbiased of the true value of the true optimal policy, i.e., $V(\widehat{d_m^*})] = V(d_m^*)$. To simplify notation, we remove the arguments for the state and remaining channels such that $\mu(a_m) = \mu(\mathbf{S}_t, a_m, A_t^{-m})$, $Y_t(a_m) = Y_t(\mathbf{S}_t, a_m, \mathbf{A}_t^{-m})$ for $A = \{A_{tm}, d_m(\mathbf{S}_t)\}$.

First, we first rewrite the expression for the AIPW estimator in Eq. (21):

$$\widehat{V}(d_m) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\left[\mu(d_m(\mathbf{S}_t)) + \Big(Y_t - \mu(A_{tm})\Big)\frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\}}{\pi_m(A_{tm} \mid \mathbf{S}_t)}\right] \tag{41}$$

$$= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\left[\mu(d_m(\mathbf{S}_t)) + Y_t\frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\}}{\pi_m(A_{tm} \mid \mathbf{S}_t)} - \mu(A_{tm})\frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\}}{\pi_m(A_{tm} \mid \mathbf{S}_t)}\right] \tag{42}$$

$$= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\left[\mu(d_m(\mathbf{S}_t)) + Y_t(d_m(\mathbf{S}_t))\frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\}}{\pi_m(A_{tm} \mid \mathbf{S}_t)} - \mu(d_m(\mathbf{S}_t))\frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\}}{\pi_m(A_{tm} \mid \mathbf{S}_t)}\right] \tag{43}$$

$$= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\Bigg[\underbrace{\mu(d_m(\mathbf{S}_t))\frac{\pi_m(A_{tm} \mid \mathbf{S}_t)}{\pi_m(A_{tm} \mid \mathbf{S}_t)}}_{=\mu(d_m(\mathbf{S}_t))} + Y_t(d_m(\mathbf{S}_t))\frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\}}{\pi_m(A_{tm} \mid \mathbf{S}_t)} \tag{44}$$

$$- \mu(d_m(\mathbf{S}_t))\frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\}}{\pi_m(A_{tm} \mid \mathbf{S}_t)} + \underbrace{Y_t(d_m(\mathbf{S}_t)) - Y_t(d_m(\mathbf{S}_t))\frac{\pi_m(A_{tm} \mid \mathbf{S}_t)}{\pi_m(A_{tm} \mid \mathbf{S}_t)}}_{=0}\Bigg] \tag{45}$$

$$= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\left[Y_t(d_m(\mathbf{S}_t)) + Y_t(d_m(\mathbf{S}_t))\left(\frac{\mathbb{1}\{A_{tm}=d_m(\mathbf{S}_t)\}-\pi_m(A_{tm}\mid\mathbf{S}_t)}{\pi_m(A_{tm}\mid\mathbf{S}_t)}\right)\right. \tag{46}$$

$$\left. - \mu(d_m(\mathbf{S}_t))\left(\frac{\mathbb{1}\{A_{tm}=d_m(\mathbf{S}_t)\}-\pi_m(A_{tm}\mid\mathbf{S}_t)}{\pi_m(A_{tm}\mid\mathbf{S}_t)}\right)\right] \tag{47}$$

$$= \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}Y_t(d_m(\mathbf{S}_t)) + \frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\left[\Big(Y_t(d_m(\mathbf{S}_t)) - \mu(\mathbf{S}_t, d_m(\mathbf{S}_t)\Big)\cdot\right. \tag{48}$$

$$\left.\left(\frac{\mathbb{1}\{A_{tm}=d_m(\mathbf{S}_t)\}-\pi_m(A_{tm}\mid\mathbf{S}_t)}{\pi_m(A_{tm}\mid\mathbf{S}_t)}\right)\right]\right\}, \tag{49}$$

where division by the propensity score is allowed under Assumption 4. The third equality follows from that, if $A_{tm} = d_m(\mathbf{S}_t)$, then $Y_t = Y_t(A_{tm}) = Y_t(d_m(\mathbf{S}_t))$ and $\mu(A_{tm}) = \mu(d_m(\mathbf{S}_t)$ by Assumption 1. The last equality follows by linearity of expectations and the rest by algebra. We now note that the estimator is unbiased if the expectation of the second term in Eq. (48) is zero, i.e.,

$$\mathbb{E}[\widehat{V}(d_m)] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}Y(d_m(\mathbf{S}_{itm}))\right] \tag{50}$$

$$= \mathbb{E}_{P_{d_m}}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}Y(d_m(\mathbf{S}_{itm}))\right] \tag{51}$$

$$= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}Y(d_m(\mathbf{S}_{itm})) \tag{52}$$

$$= V(d_m). \tag{53}$$

Here, the second equality follows from that the expected potential outcome given the policy decision equals the expectation of the outcome with respect to the distribution under the policy, the third from that the sample average is a constant, and the fourth from the definition of the value.

Now, double robustness means that the estimator is unbiased if (1) $\mu(d_m(\mathbf{S}_t))$ is correctly specified but $\pi_m(A_{tm}\mid\mathbf{S}_t)$ is not, or (2) $\pi_m(A_{tm}\mid\mathbf{S}_t)$ is correctly specified but $\mu(d_m(\mathbf{S}_t))$ is not. We have shown that the first term in Eq. (48) is by itself an unbiased estimator. In the following, we thus prove doubly robust unbiasedness by showing that, in either of these cases, the expectation of the second term in Eq. (48) is zero.

Case (1): Assume that $\mu(d_m(\mathbf{S}_t))$ is correctly specified but $\pi_m(A_{tm}\mid\mathbf{S}_t)$ is not. We then have

$$\mathbb{E}\left[(Y_t(d_m(\mathbf{S}_t)) - \mu(d_m(\mathbf{S}_t)))\left(\frac{\mathbb{1}\{A_{tm}=d_m(\mathbf{S}_t)\}-\pi_m(A_{tm}\mid\mathbf{S}_t)}{\pi_m(A_{tm}\mid\mathbf{S}_t)}\right)\right] \tag{54}$$

$$= \mathbb{E}\left[\Big(Y_t(d_m(\mathbf{S}_t)) - \mathbb{E}[Y_t\mid\mathbf{S}_t, d_m(\mathbf{S}_t)]\Big)\left(\frac{\mathbb{1}\{A_{tm}=d_m(\mathbf{S}_t)\}-\pi_m(A_{tm}\mid\mathbf{S}_t)}{\pi_m(A_{tm}\mid\mathbf{S}_t)}\right)\right] \tag{55}$$

$$= \mathbb{E}\left[\mathbb{E}\left(\Big(Y_t(d_m(\mathbf{S}_t)) - \mathbb{E}[Y_t\mid\mathbf{S}_t, d_m(\mathbf{S}_t)]\Big)\left(\frac{\mathbb{1}\{A_{tm}=d_m(\mathbf{S}_t)\}-\pi_m(A_{tm}\mid\mathbf{S}_t)}{\pi_m(A_{tm}\mid\mathbf{S}_t)}\right)\,\middle|\,\mathbf{S}_t, A_{tm}\right)\right] \tag{56}$$

$$= \mathbb{E}\left[\mathbb{E}\Big[Y_t(d_m(\mathbf{S}_t)) - \mathbb{E}[Y_t\mid\mathbf{S}_t, d_m(\mathbf{S}_t)]\,\Big|\,\mathbf{S}_t, A_{tm}\Big] \mid \left(\frac{\mathbb{1}\{A_{tm}=d_m(\mathbf{S}_t)\}-\pi_m(A_{tm}\mid\mathbf{S}_t)}{\pi_m(A_{tm}\mid\mathbf{S}_t)}\right)\right] \tag{57}$$

$$= \mathbb{E}\left[\Big(\mathbb{E}[Y_t(d_m(\mathbf{S}_t))\mid\mathbf{S}_t, A_{tm}] - \mathbb{E}[Y_t\mid\mathbf{S}_t, d_m(\mathbf{S}_t)]\Big)\left(\frac{\mathbb{1}\{A_{tm}=d_m(\mathbf{S}_t)\}-\pi_m(A_{tm}\mid\mathbf{S}_t)}{\pi_m(A_{tm}\mid\mathbf{S}_t)}\right)\right] \tag{58}$$

$$= \mathbb{E}\left[\Big(\mathbb{E}[Y_t(d_m(\mathbf{S}_t))\mid\mathbf{S}_t] - \mathbb{E}[Y_t(d_m(\mathbf{S}_t))\mid\mathbf{S}_t]\Big)\left(\frac{\mathbb{1}\{A_{tm}=d_m(\mathbf{S}_t)\}-\pi_m(A_{tm}\mid\mathbf{S}_t)}{\pi_m(A_{tm}\mid\mathbf{S}_t)}\right)\right] \tag{59}$$

$$= 0. \tag{60}$$

The first equality used that $\mu(d_m(\mathbf{S}_t)) = \mathbb{E}[Y_t \mid \mathbf{S}_t, d_m(\mathbf{S}_t)]$ when the outcome model is correctly specified, the second follows from iterated expectations, the third and fourth from basic probability theory, and the fifth follows from that $\mathbb{E}[Y_t(d_m(\mathbf{S}_t)) \mid \mathbf{S}_t, \mathbf{A}_t] = \mathbb{E}[Y_t(d_m(\mathbf{S}_t)) \mid \mathbf{S}_t]$ by the unconfoundedness assumption and $\mathbb{E}[Y_t \mid \mathbf{S}_t, d_m(\mathbf{S}_t)] = \mathbb{E}[Y_t(d_m(\mathbf{S}_t)) \mid \mathbf{S}_t]$ by the consistency assumption. Thus, the two inner expectations cancel each other. The outer empirical expectation and the averaging over stages are omitted from the derivation as they have no effect on the result and the order of expectations can be changed without affecting the result.

Case (2): Now, instead, assume that $\pi_m(A_{tm} \mid \mathbf{S}_t)$ is correctly specified but $\mu(d_m(\mathbf{S}_t))$ is not. Then, we yield

$$
\begin{aligned}
&\mathbb{E}\left[\Big(Y_t(d_m(\mathbf{S}_t)) - \mu(d_m(\mathbf{S}_t))\Big)\left(\frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\} - \pi_m(A_{tm} \mid \mathbf{S}_t)}{\pi_m(A_{tm} \mid \mathbf{S}_t)}\right)\right] && (61)\\
=&\mathbb{E}\left[\mathbb{E}\left(\Big(Y_t(d_m(\mathbf{S}_t)) - \mu(d_m(\mathbf{S}_t))\Big)\left(\frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\} - \pi_m(A_{tm} \mid \mathbf{S}_t)}{\pi_m(A_{tm} \mid \mathbf{S}_t)}\right)\Big| Y_t(d_m(\mathbf{S}_t)), \mathbf{S}_t\right)\right] && (62)\\
=&\mathbb{E}\left[\Big(Y_t(d_m(\mathbf{S}_t)) - \mu(d_m(\mathbf{S}_t))\Big)\,\mathbb{E}\left(\frac{\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\} - \pi_m(A_{tm} \mid \mathbf{S}_t)}{\pi_m(A_{tm} \mid \mathbf{S}_t)}\Big| Y_t(d_m(\mathbf{S}_t)), \mathbf{S}_t\right)\right] && (63)\\
=&\mathbb{E}\left[\Big(Y_t(d_m(\mathbf{S}_t)) - \mu(d_m(\mathbf{S}_t))\Big)\,\frac{\mathbb{E}[\mathbb{1}\{A_{tm} = d_m(\mathbf{S}_t)\} \mid Y_t(d_m(\mathbf{S}_t)), \mathbf{S}_t] - \pi_m(A_{tm} \mid \mathbf{S}_t)}{\pi_m(A_{tm} \mid \mathbf{S}_t)}\right] && (64)\\
=&\mathbb{E}\left[\Big(Y_t(d_m(\mathbf{S}_t)) - \mu(d_m(\mathbf{S}_t))\Big)\,\frac{\{\mathbb{E}[\mathbb{1}\{A_{mt} = d_m(\mathbf{S}_t)\} \mid \mathbf{S}_t] - \pi_m(A_{mt} \mid \mathbf{S}_t)\}}{\pi_m(A_{mt} \mid \mathbf{S}_t)}\right] && (65)\\
=&\mathbb{E}\left[\Big(Y_t(d_m(\mathbf{S}_t)) - \mu(d_m(\mathbf{S}_t))\Big)\,\frac{\{\pi_m(A_{mt} \mid \mathbf{S}_t) - \pi_m(A_{mt} \mid \mathbf{S}_t)\}}{\pi_m(A_{mt} \mid \mathbf{S}_t)}\right] && (66)\\
=&0, && (67)
\end{aligned}
$$

where the first equality follows by iterated expectations, the second and third by basic probability theory, the fourth by Assumption 3, and the fifth by the correct specification of each propensity score model so that the terms in the numerator cancel. This concludes the proof.

□

Appendix C: Descriptive Statistics

Tables 4–5 show summary statistics of all variables, grouped by the training set and evaluation set. Recall that non-categorical variables are scaled to zero mean and unit variance. The high outliers are from articles on COVID-19, which received above attention traction among the user base.

	Training set				*Evaluation set*			
Covariate	Mean	Std	Min	Max	Mean	Std	Min	Max
Article views	200.07	525.66	0.00	50124.00	211.43	485.98	0.00	21716.00
Average engagement time	1.04	1.90	0.00	374.80	0.94	0.83	0.00	48.64
Average scroll depth	63.87	26.81	0.00	100.00	63.70	26.92	0.00	100.00
Recirculation	30.96	22.60	0.00	100.00	30.67	22.61	0.00	100.00
Cumulative registrations	0.32	1.27	0.00	44.00	0.11	0.56	0.00	29.00
Number of times on the subscription path	0.01	0.06	0.00	1.80	0.01	0.05	0.00	2.20
Content length	1145.82	1704.14	0.00	41173.00	1179.12	1912.07	0.00	41173.00
Hours online on Swiss front page	339.26	4955.98	0.00	182219.00	227.83	1404.67	0.00	15825.00
Hours online on German frontpage	312.37	4944.18	0.00	182220.00	200.04	1357.71	0.00	15836.00
Hours online since last publication	33.12	20.81	-0.52	71.94	32.40	20.74	-0.25	71.92
Count of email newsletter promotions	8.99	52.50	0.00	502.00	9.54	54.20	0.00	515.00
Count of push notification promotions	1.96	14.22	0.00	192.00	2.23	15.71	0.00	202.00
Count of Twitter promotions	4.96	28.69	0.00	333.00	5.13	29.24	0.00	343.00
Swiss front page promotion	0.17	0.37	0.00	1.00	0.17	0.37	0.00	1.00
German front page promotion	0.17	0.37	0.00	1.00	0.17	0.38	0.00	1.00
Lag of Swiss front page promotion	0.17	0.38	0.00	1.00	0.17	0.38	0.00	1.00
Lag of German front page promotion	0.17	0.37	0.00	1.00	0.17	0.38	0.00	1.00
Lag of Email newsletter promotion	0.01	0.09	0.00	1.00	0.01	0.10	0.00	1.00
Lag of Push notification promotion	0.00	0.06	0.00	1.00	0.00	0.06	0.00	1.00
Lag of Twitter promotion	0.01	0.10	0.00	1.00	0.01	0.11	0.00	1.00

Table 4 Summary statistics. Note: non-categorical variables were standardized.

Category	Training	Evaluation
Hour: 06:00	0.05	0.06
Hour: 07:00	0.05	0.06
Hour: 08:00	0.05	0.06
Hour: 09:00	0.06	0.06
Hour: 10:00	0.06	0.06
Hour: 11:00	0.06	0.06
Hour: 12:00	0.06	0.06
Hour: 13:00	0.06	0.06
Hour: 14:00	0.06	0.06
Hour: 15:00	0.06	0.06
Hour: 16:00	0.06	0.05
Hour: 17:00	0.05	0.06
Hour: 18:00	0.06	0.05
Hour: 19:00	0.06	0.06
Hour: 20:00	0.06	0.06
Hour: 21:00	0.06	0.06
Hour: 22:00	0.06	0.05
Hour: 23:00	0.05	0.04
Weekday: Monday	0.10	0.08
Weekday: Tuesday	0.09	0.09
Weekday: Wednesday	0.09	0.14
Weekday: Thursday	0.19	0.24
Weekday: Friday	0.20	0.25
Weekday: Saturday	0.18	0.11
Weekday: Sunday	0.14	0.09

Table 5 Share of items belonging to each categorical covariate of time information

Category	Training	Evaluation
Section: Culture	0.01	0.00
Section: Education	0.09	0.09
Section: Society	0.01	0.01
Section: International	0.22	0.25
Section: Opinion	0.09	0.08
Section: Mobility	0.01	0.01
Section: NZZ in English	0.01	0.01
Section: Other	0.01	0.01
Section: International politics	0.02	0.02
Section: Celebrities & events	0.08	0.08
Section: Domestic	0.08	0.10
Section: Sport	0.08	0.08
Section: Visuals	0.01	0.01
Section: Business & finance	0.15	0.13
Section: Science & technology	0.05	0.04
Section: Zurich	0.08	0.07
Type: None	0.71	0.73
Type: Editor-in-Chief	0.01	0.01
Type: Breaking news	0.08	0.09
Type: English	0.01	0.01
Type: Explained	0.01	0.00
Type: Guest comment	0.03	0.02
Type: Interview	0.03	0.02
Type: Column	0.01	0.01
Type: Comment	0.08	0.06
Type: News in brief	0.03	0.03
Type: Other	0.01	0.01
Format: Long-form standard	0.10	0.08
Format: Long-form visual	0.01	0.01
Format: Opinion	0.11	0.10
Format: Regular	0.78	0.81
Format: Video	0.00	0.00
Sentiment: Negative lead_text	0.08	0.08
Sentiment: Neutral lead_text	0.92	0.92
Sentiment: Positive lead_text	0.00	0.00
Sentiment: Negative SEO title	0.09	0.06
Sentiment: Neutral SEO title	0.89	0.93
Sentiment: Positive SEO title	0.02	0.01
Sentiment: Negative title	0.07	0.05
Sentiment: Neutral title	0.91	0.94
Sentiment: Positive title	0.02	0.01

Table 6 Share of items belonging to each categorical covariate of content characteristics.

Appendix D: Estimation and Selection of Machine Learning Models

D.1. Nuisance Models

The nuisance models are not of interest in themselves but simply are means to predict potential outcomes. Hence, we select nuisance models based on predictive performance. As candidate models, we considered random forests (Breiman 2001) and gradient-boosted trees (Friedman 2001). Both are ensembles of non-parametric decision trees, but differ in their construction. Eventually, we found that gradient-boosted trees perform better. Nevertheless, repeating the analyses from our main paper with random forests leads to qualitatively similar findings.

We perform hyperparameter tuning and model selection as follows. First, we perform 5-fold cross-validation based on an exhaustive grid search to find the hyperparameter values that minimize the loss function of each nuisance model on the training data. For the propensity score model, we search for the hyperparameters that minimize the logistic loss, given by

$$\text{Logistic loss} = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c}\sum_{i\in c}\Big(A_{im}^{(c)}\log\pi_{im}^{(c)} + (1-A_{im}^{(c)})\log(1-\pi_{im}^{(c)})\Big), \tag{68}$$

where $\pi_{im}^{(c)} = \pi_m(A_{im}^{(c)} \mid \mathbf{S}_i^{(c)})$ is short-hand notation for the propensity score of promotion on channel m on validation observation i in cross-validation fold $c = 1\ldots, C$ to channel $m = 1, \ldots, M$. For the outcome model, we search for the hyperparameter values that minimize the mean squared error (MSE) loss, given by

$$\text{MSE loss} = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c}\sum_{i\in c}\Big(\varepsilon_{im}^{(c)}\Big)^2, \tag{69}$$

where $\varepsilon_{im}^{(c)} = Y_i^{(c)} - \mu(\mathbf{S}_i^{(c)}, A_{im}^{(c)}, \mathbf{A}_i^{(c),-m)})$ is the error of the outcome model. Second, we compare the hyperparameter-tuned nuisance models in terms of their predictive ability. For this, we take the average loss across the validation splits in the cross-validation. Based on this we select one propensity score model and one outcome model. Details on the hyperparameter tuning for the propensity score models and the outcome models are shown in Table 7 and 8, respectively.

		Tuned hyperparameter values			
		Random Forest		Gradient Boosted Trees	
Hyperparameter	Grid-searched values	Switzerland	Germany	Switzerland	Germany
`n_estimators`	$[100, 200, 300, 400, 500]$	500	300	200	500
`max_features`	$[sqrt, log2, None]$	None	sqrt	log2	sqrt
`max_depth` (RF)	$[20, 40, 60, 80, 100, None]$	20	40		
`max_depth` (GB)	$[3, 5, 7, 9]$			9	9
`min_samples_split`	$[2, 4, 8, 16]$	2	16	4	2
`min_samples_leaf`	$[1, 2, 4, 8]$	8	4	1	8
`learning_rate`	$[0.01, 0.05, 0.1, 0.2]$			0.05	0.01

Table 7 Details on hyperparameter tuning for the propensity score models. The hyperparameters were tuned to minimize the loss function via 5-fold cross-validated grid search on the training data. See the documentation for scikit-learn for descriptions of the hyperparameter. Empty cells mean that the hyperparameter does not apply to the model class. Abbreviations: RF = Random Forest. GBT = Gradient Boosted Trees.

		Tuned hyperparameter values	
Hyperparameter	Grid-searched values	Random Forest	Gradient Boosted Trees
`n_estimators`	$[100, 200, 300, 400, 500]$	500	300
`max_features`	$[sqrt, log2, None]$	None	sqrt
`max_depth` (RF)	$[20, 40, 60, 80, 100, None]$	20	
`max_depth` (GB)	$[3, 5, 7, 9]$		7
`min_samples_split`	$[2, 4, 8, 16]$	2	2
`min_samples_leaf`	$[1, 2, 4, 8]$	4	1
`learning_rate`	$[0.01, 0.05, 0.1, 0.2]$		0.05

Table 8 Details on hyperparameter tuning for the outcome models. The hyperparameters were tuned to minimize the loss function via 5-fold cross-validated grid search on the training data. See the documentation for scikit-learn (Pedregosa et al. 2011) for descriptions of the hyperparameter. Empty cells mean that the hyperparameter does not apply to the model class.

D.2. CATE Function

Our procedure to construct and fit the orthogonal random model of the CATE is as follows. We first fit a random forest (Breiman 2001) to predict the nuisance model CATE estimates $\widehat{\tau}_{itm}$ from the state variables $\mathbf{S}_{it}$. We construct the random forest from non-parametric regression trees using the sub-sampled "honest" procedure of Athey and Imbens (2016), Wager and Athey (2018), where we use half of the observations for creating the tree structures and the other half for estimation. The orthogonal random forest approach is then to estimate the CATE function by solving the following local regression moment equation

$$\widehat{\tau}_m^0(\mathbf{s}) = \arg\min_{\tau_m^0} \frac{1}{|\mathcal{T}_2|} \sum_{t \in \mathcal{T}} \frac{1}{N_t} \sum_{i=1}^{N_t} K_m\Big(\mathbf{s}, \mathbf{S}_{it}\Big) \Big(\widehat{\tau}_m(\mathbf{S}_{it}) - \tau_m^0\Big)^2, \tag{70}$$

where $\mathcal{T}_2 = \bigcup_{l=1}^{L} \mathcal{T}_2^{(l)}$ is the union of held-out data from the cross-validation folds part of the cross-fitting procedure (see Appendix D.3 for details), $\widehat{\tau}_m(\mathbf{S}_{it})$ is the doubly robust nuisance CATE estimate given by Eq. (19), and $K_m(\mathbf{s}, \mathbf{S}_{it})$ is a data-driven kernel function (or, similarity metric) that calculates how often content with state $\mathbf{s}$ fall into the same leaf of the regression trees as content with state $\mathbf{S}_{it}$. The resulting CATE function estimates $\widehat{\tau}_m^0(\mathbf{s})$ are asymptotically Gaussian distributed and thereby allows for constructing asymptotically valid confidence intervals via a suitable bootstrap approach. We thereby use "bootstrap of little bags" (Athey et al. 2019) to obtain standard errors of the CATE predictions required for the construction of the confidence bounds in the LCB and UCB policies. We run the whole estimation procedure on the training data separately for the CATE functions of either front page.

We tune the hyperparameters of the orthogonal random forest CATE function as follows. We set the total number of trees to 500. We leave the maximum depth of trees unrestricted such that nodes are expanded until leaves are pure or until all leaves contain less than `min_samples_split` samples. Then, we fix the minimum number of splitting samples required to split an internal node to two and the minimum number of samples required to be at a leaf node to one. We specify the model to consider all features when looking for the best split. The kernel is tuned as part of the 5-fold cross-validated estimation.

D.3. Cross-validated Sample-Split Cross-Fitting of Nuisance Models and CATE Function.

We follow best-practice for estimating hyperparameter-tuned nuisance models and CATE function (Battocchi et al. 2019). The steps are as follows:

1. For cross-validation fold $l = \ldots, L$:
   (a) Randomly split the training data $\mathcal{T}$ into non-overlapping sets $\mathcal{T}_1^{(l)}$ and $\mathcal{T}_2^{(l)}$ of equal size $|\mathcal{T}|/2$ such that $\mathcal{T}_1^{(l)} \bigcup \mathcal{T}_2^{(l)} = \mathcal{T}$
   (b) On $\mathcal{T}_1^{(l)}$, fit the hyperparameter-tuned nuisance models in Eqs. (16)–(17)
2. For cross-validation fold $l = \ldots, L$:

(a) For each item $i \in \mathcal{T}_2^{(l)}$, predict the doubly robust scores by evaluating Eq. (18) for $a \in \{0,1\}$ using an outcome model and a nuisance model that were fitted on different folds $k, j \in \{1,\dots,L\}$, $k \neq j$, not yet paired.
(b) Use Eq. (19) to obtain CATE estimates on fold l;

3. On the union of nuisance evaluation sets $\bigcup_{l=1}^{L} \mathcal{T}_2^{(l)}$, fit the CATE function using Eq. (20).

We run the procedure using $L = 5$ cross-validation folds.

D.4. Best Linear Predictor of CATE Function with Post-Lasso Inference

Our procedure consists of the following steps.

1. We randomly partition the training data $\mathcal{T}$ into two mutually exclusive splits $\mathcal{T}_1$ and $\mathcal{T}_2$.
2. On split $\mathcal{T}_1$, we fit a Lasso regularized linear model where the CATE function estimates are the dependent variable and the state covariates are the predictors. For estimation, we use a suitable optimization algorithm (i.e., coordinate descent) to minimize the mean square error between the CATE estimates and the linear model subject to the Lasso regularization penalty,

$$(\widehat{\alpha}_m, \widehat{\boldsymbol{\beta}}_m^0) = \underset{\alpha_m, \tau_m}{\arg\min} \frac{1}{|\mathcal{T}_1|} \sum_{t \in \mathcal{T}_1} \frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} \left(\widehat{\tau}_m(\mathbf{s}_{it}) - \alpha_m - \boldsymbol{\beta}_m^\top \mathbf{S}_{it} \right)^2 + \lambda \sum_{k=1}^{p} |\beta_k|. \tag{71}$$

   Here, $\lambda \in [0, \infty]$ is a tuning parameter for the degree of penalization by the regularization function and $\widehat{\boldsymbol{\beta}}_m^0 \in \mathbb{R}^l$, $l \leq \dim(\mathbf{S}) = p$ are the non-zero parameter estimates for the selected covariates $\mathbf{S}^0 \coloneqq \{S_k \colon \widehat{\beta}_k \neq 0\} = \{S_k \in \mathbf{S} \colon \widehat{\tau}_m(\mathbf{S}) - \widehat{\tau}_m(\{\mathbf{S} \setminus S_k\}) \neq 0\}$.
3. On the other split $\mathcal{T}_2$, regress the CATE estimates on the selected covariates,

$$\widehat{\tau}_m(\mathbf{S}_{itm}^0) = \alpha_m + \boldsymbol{\xi}_m^\top \mathbf{S}_{itm}^0 + \epsilon_{itm}. \tag{72}$$

   We estimate its parameters using ordinary least squares, $\widehat{\boldsymbol{\xi}}_m = (\mathbf{S}^{0\top}\mathbf{S}^0)^{-1}\mathbf{S}^{0\top}\widehat{\boldsymbol{\tau}}_m$, and use HC2 robust standard errors for inference.

The least squares parameters of the final model are interpreted analogously to linear models but with respect to treatment effect heterogeneity. If the Lasso-regularization selects correctly selects the true covariate set, then the least squares parameter estimates equal those of the oracle estimator, and, if the Lasso-regularization does not select the true covariate subset, then the estimates are still less biased and converging at a faster rate than standard Lasso estimates. This holds in finite-samples with high-dimensional controls, under non-Gaussian and heteroskedastic errors, and even if the underlying true CATE is non-parametric (Belloni and Chernozhukov 2013, Belloni et al. 2014). We use these properties to conductasymptotically valid inference on the discovered marginal heterogeneity in the CATE by performing the following two-sided hypothesis test

$$H_0^j \colon \xi_{jm} = 0 \quad \text{vs.} \quad H_A^j \colon \xi_{jm} \neq 0. \tag{73}$$

The discovered set of covariates that significantly contribute to CATE heterogeneity is then

$$\mathbf{S}_m^0 \coloneqq \{S_j \subseteq \mathbf{S}_m^0 \ : \ H_0^j \text{ is rejected at some significance level } \alpha/l\}, \tag{74}$$

where α/l is the significance level after we have applied a Bonferroni correction for running l hypothesis tests. This procedure of combining post-Lasso inference with a correction for multiple hypothesis testing is similar to the recently proposed method of Pelger and Zou (2022). Whereas we use the procedure to select a sufficient set of covariates that explain heterogeneity in the CATE, they consider the problem of selecting a sparse set of covariates that explain a high-dimensional linear panel data regression model. In our setting, the CATE is high-dimensional and the BLP of the CATE that we estimate to perform post-Lasso inference is a linear panel data regression model. Hence, the method of Pelger and Zou (2022) may be viewed as providing the theoretical basis of our approach.

To run the method, we set the significance level α to 0.05 and find that the Lasso-regularization selects $l = 48$ state covariates for the CATE for the Swiss front page and $l = 45$ state covariates for the CATE for the German front page (see Tab. 3). Hence, $\mathbf{S}_m^0$ comprise the covariates in Tab. 3 whose p-value is at most $\alpha/l \approx 0.001$ for the respective front page CATE. The procedure returns 36 out of the 48 covariates for the CATE for Swiss front page, and 27 out of the 45 covariates for the CATE for German front page.

D.5. Estimation of the Lasso-Regularized BLP of the CATE

We perform the two-stage procedure from the post-Lasso separately for the Swiss and German front page policies using the CATE estimates of either as the dependent variable. We tune the Lasso regularized model using 5-fold cross-validation. For each cross-validation run, we run coordinate descent optimization for 1000 iterations with 500 values of the penalty λ along different regularization values (here, we use the default values from scikit-learn). The final model is selected as the respective model that best predicts the BLP of the CATE among the corresponding 10 cross-validated models. For inference, we compute HC2 robust standard errors to guard against heteroskedasticity in the second-stage error terms and obtain p-values based on those. We implement the method using the python package scikit-learn (Pedregosa et al. 2011).

D.6. Software Implementation

We perform all analyses in Python. We implement the nuisance models using the package scikit-learn (Pedregosa et al. 2011), the orthogonal random forest in our CATE function via the package EconML (Battocchi et al. 2019), and the post-Lasso using the packages scikit-learn and statsmodels.

Appendix E: Pseudocode for Off-Policy Framework

Algorithm 1 lists the step-by-step procedure for learning and evaluating the optimal policy or any of the baselines.

Algorithm 1: Off-Policy Learning and Evaluation

Input: Historical data $\mathcal{H} = \{(\mathbf{S}_{it}, \mathbf{A}_{it}, Y_{it}, C_{tm})\colon i \in \mathcal{I}_t,\ t \leq T,\ m \leq M\}$, number of folds cross-validation L
Output: $\widehat{V}(\widehat{d}_m)$, $d_m \in \mathcal{D}$, $m = 1 \ldots, M$
Split $\mathcal{H}$ into training set $\mathcal{T}$ and evaluation set $\mathcal{V}$ such that $|\mathcal{T}| + |\mathcal{V}| = |\mathcal{H}|$, where $\mathcal{T} \cup \mathcal{V} = \emptyset$, $s < t\ \forall s \in \mathcal{T}$, $t \in \mathcal{V}$
for channel $m = 1, \ldots, M$ **do**
    `/* Cross-validated cross-fitting of nuisance models and CATE model */`
    for cross-validation fold $l = 1 \ldots, L$ **do**
        Randomly split training data $\mathcal{T}$ into two sets $\mathcal{T}_1^{(l)}$, $\mathcal{T}_2^{(l)}$ of equal size $|\mathcal{T}|/2$ such that $\mathcal{T}_1^{(l)} \bigcup \mathcal{T}_2^{(l)} = \mathcal{T}$
        On $\mathcal{T}_1^{(l)}$, fit the nuisance models in Eqs. (16)–(17)
    for cross-validation fold $l = 1 \ldots, L$ **do**
        for $i \in \mathcal{T}_2^{(l)}$ **do**
            Get the doubly robust scores by evaluating Eq. (18) for $a = 1$ and $a = 0$ using an outcome model fitted on $\mathcal{T}_1^{(k)}$ and a propensity score model fitted on $\mathcal{T}_1^{(j)}$ for a pair of folds $k, j \in \{1, \ldots, L\}$, $k \neq j$, not yet considered
            Evaluate Eq. (19) to get the nuisance CATE estimate $\widehat{\tau}_{itm}$
    if *Policy* = Optimal, LCB, or UCB **then**
        On $\bigcup_{l=1}^{L} \mathcal{T}_2^{(l)}$, fit the CATE function in Eq. (20) with $\widehat{\tau}_m$ as dependent variable
    `/* Off-Policy Evaluation */`
    forall time periods $t \in \mathcal{V}$ **do**
        forall *items* $i \in \mathcal{I}_t$ **do**
            if Policy = Optimal, LCB, or UCB **then**
                Predict the CATE by evaluating Eq. (20) at the state $\mathbf{S}_{itm}$
                Set $Q_i \leftarrow \widehat{\tau}_m(\mathbf{S}_{itm})$
                if Policy = LCB or UCB **then**
                    Obtain the confidence bound $\text{CB}_i = 1.96 \times \text{SE}_{itm}$ of the CATE prediction
                    if Policy = UCB **then**
                        Set $Q_i \leftarrow Q_i + \text{CB}_i$
                    if Policy = LCB **then**
                        Set $Q_i \leftarrow Q_i - \text{CB}_i$
            if Policy = Naïve **then**
                Using Eq. (18), obtain the doubly robust score for $A_{itm} = 1$
                Set $Q_i \leftarrow \widehat{Y}(\mathbf{S}_{it}, A_{itm} = 1, \mathbf{A}_{it}^{-m})$
            if Policy = Autoregressive **then**
                Set $Q_i \leftarrow Y_{i,t-1}$
            Sort items $i \in \mathcal{I}_t$ in descending order with respect to Q_i **if** *rank*$(i) \geq$ *rank*(C_{tm}) **then**
                $\widehat{d}_{itm} = 1$
            else
                $\widehat{d}_{itm} = 0$
            Predict counterfactual outcome under policy d_m by evaluating Eq. (18) at $\widehat{d}_{itm}$
    Obtain the policy value estimate $\widehat{V}(\widehat{d}_m)$ using Eq. (21)

Appendix F: Additional Results for Reduced Optimal Policy

We empirically check if covariates selected by the post-Lasso inference procedure on the training data are sufficient for the optimal policy on the evaluation data. We thus first select a covariate set according to the procedure in Appendix D.4. Following our fitting procedure (see Appendix D.3, we re-estimate the orthogonal random forest models of the CATE function of either front page using only the selected covariates. We then estimate the value of the reduced optimal policy on the evaluation data according to estimation step 3 (Sec. 4.6.4, where the policy is given by $\widehat{d}^*_m(\mathbf{S}^0_{itm})$ with $\widehat{d}^*_m(\mathbf{S}^0_{itm}) = \mathbb{1}\{\widehat{\tau}_m(\mathbf{S}^0_{itm}) \geq \widehat{\boldsymbol{\tau}}^{(C_{tm})}_{tm}\}$ and where $\mathbf{S}^0_{itm}$ is given by Eq. (74).

Table 9 reports the performance of the reduced optimal policy. First, in line with our expectations, the reduced optimal policy outperforms the behavior policy (i.e., current practice) by a large margin. Second, by comparing the reduced optimal policy with the results from the main paper, we find that the reduced optimal policy substantially outperforms all of the baselines. Third, the reduced optimal policy and the optimal policy are in par in the sense that the 95% confidence intervals of their values overlap. This is confirmed by Fig. 10, which shows that, on average, the 95% confidence intervals of their values overlap at all hours of the day and all days of the week. This shows that there is no significant loss in the value when an optimal policy based on the CATE only uses the covariates that have statistical significant effects on the heterogeneity in the CATE. Altogether, the above analysis demonstrates that even a subset of the data collected by our partner company is sufficient to make optimal decisions. Taking our partner company as an example, the managerial implication is that the dashboard would only need to show data on at most three-quarters of all covariates. This could support the editors' decision-making by reducing the amount of information they need to process during their decisions.

Table 9 Policy value estimates on the evaluation set.

	Swiss front page				*German front page*			
Policy	Value	SE	Regret	Gain	Value	SE	Regret	Gain
Optimal policy	1.354	0.0076	0.015%	5.21%	1.316	0.0075	0.007%	2.25%
Reduced optimal policy	1.374	0.0078	0.00%	6.76%	1.325	0.0075	0.00%	2.95%

Notes. Value = average performance score across content and time periods. Larger is better. SE: standard error. Regret: decrease in value relative to the best policy for the front page. Gain: increase in value relative to behavior policy. Optimal policy: Our optimal policy when the predicted CATEs are based on all state covariates. Reduced optimal policy: The optimal policy when the predicted CATEs are based only on the state covariates selected by our post-Lasso inference procedure.

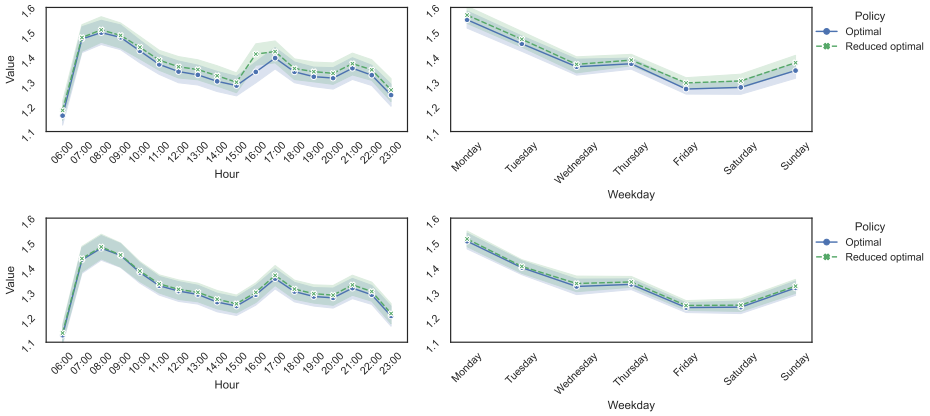


Figure 10 **Value of the optimal policy and the reduced optimal policy over time. Shaded regions are 95% confidence intervals across 1000 bootstrap runs. Top: Swiss front page. Bottom: German front page.**