Targeting Treatment Assignment by combining Behavioral Measurement and Machine Learning

*

KEVIN BAUER[†], ANDREAS GRUNEWALD[‡],

FLORIAN HETT[§], JOHANNA JAGOW, MAXIMILIAN SPEICHER[¶]

March 26, 2026

We study how behavioral economics and machine learning can be combined to construct effective treatment targeting rules. In a large field experiment at a major German online fashion retailer ($N \approx 500,000$ visitors), a loss-framed discount message has essentially zero average effect on purchases and returns. We implement a nested incentivized behavioral measurement experiment ($N = 582$) to elicit individual loss aversion, a moderator for loss framing proposed by economic theory, and train a prediction model to impute loss-aversion categories from commonly observed digital footprints at scale. Targeting exposure to visitors with high predicted loss aversion yields statistically significant revenue gains. A causal machine learning approach using causal forests based on large RCT data identifies heterogeneity but delivers smaller, less robust improvements. The results illustrate how combining behavioral economics and machine learning can improve both the performance and interpretability of algorithmic treatment assignment.

Keywords: Treatment targeting; Heterogeneous treatment effects; Behavioral measurement; Machine learning; Field experiments; Loss aversion

JEL Codes: C93, C55, D91, M31, L81

*Financial support by the DFG, grant FOR 5392 project number 462020252, is gratefully acknowledged.

[†]Goethe University, Department of Information Systems, 60629 Frankfurt am Main, E-Mail: Bauer@wiwi.uni-frankfurt.de

[‡] Frankfurt School of Finance and Management, Department of Economics, Adickesallee 32-34, 60322 Frankfurt am Main, E-Mail: a.grunewald@fs.de

[§]University of Mainz, Department of Business and Economics, Jakob-Welder-Weg 4, 55128 Mainz, E-Mail: florian.hett@uni-mainz.de

[¶] Jagow Speicher Consulting, 08037 Barcelona, E-Mails: info@johannajagow.com, max@maxspeicher.com

1. Introduction

A central challenge in implementing behavioral interventions is not only to determine whether a treatment works on average, but also to decide how it should be allocated in practice. In many settings, treatment effects are likely to vary systematically across individuals. Targeting treatments should therefore improve the effectiveness of an intervention by assigning it to those individuals for whom it is most beneficial. This shifts attention from average treatment effects to treatment assignment rules: the practical object of interest is a policy that maps available information into treatment decisions and can be deployed at scale (Manski, 2004; Kitagawa and Tetenov, 2018; Athey and Wager, 2021).

How to derive such targeting rules, however, remains an open practical question. For many policies, economic theory makes meaningful predictions about treatment heterogeneity and suggests specific moderator variables that should determine who responds to an intervention, how they respond, and who does not respond at all. Experimental and behavioral methods even have made these variables empirically accessible at the individual level (Holt and Laury, 2002; Niederle and Vesterlund, 2007; Andreoni and Sprenger, 2012; Augenblick, Niederle and Sprenger, 2015). This bears potential for treatment targeting: theory can point to a moderator, and behavioral measurement can make this moderator observable. However, while recent papers have successfully adopted behavioral measurement in the field (Meier and Sprenger, 2010; Buser, Niederle and Oosterbeek, 2014; Falk et al., 2018; Andreoni et al., 2023; Schneider and Sutter, 2026), running a behavioral elicitation for every individual and only then decide whom to treat is typically infeasible.

We illustrate how combining behavioral measurement with machine learning can scale empirical findings on heterogeneous treatment effects from lab-like measurement settings to treatment targeting in the field. The idea is simple. Theory identifies a behavioral construct that should moderate treatment effects. This construct is then measured in a comparatively small *behavioral measurement sample*, that is, a sample in which individual behavior can be elicited directly. A machine-learning model is then trained to predict this construct from variables that are observed both in the behavioral measurement sample and in the *field sample*, that is, in the setting in which treatment decisions are to be made. This overlap in observables allows to link the two samples and to transport treatment heterogeneity defined over a behaviorally measured economic construct into the full field sample and to use it for treatment targeting, even tough direct behavioral measurement is not feasible.

We apply this approach to the targeting of a loss-framed discount message in e-commerce. We combine a large field study at a major German online fashion retailer with an incentivized behavioral measurement experiment with a subsample of website visitors. In the field setting, the intervention varies the framing of a discount message while holding fixed its economic content. In the baseline condition, visitors see a neutral discount message; in the treatment condition, the same discount is presented in a loss frame. The field sample covers close to 500,000 website visits and records *digital footprints* – variables such as referral source, entry page, browser, device,

operating system, and IP-based information. In the behavioral measurement sample, participants are endowed with a voucher and decide whether to keep it or stake it in a lottery that involves a potential loss, yielding a coarse but theory-based measure of individual loss aversion. We then use the digital footprints observed in both datasets to predict this measure for visitors in the field sample and assign treatment accordingly.

The results demonstrate the effectiveness of our approach. First, the loss-framing intervention is ineffective on average. A firm that rolled out the treatment to all visitors would not improve performance. Second, the behavioral measurement experiment shows that there is in fact meaningful heterogeneity to exploit: when loss aversion is measured directly, treatment effects vary systematically with this theoretically motivated construct. Individuals with a higher measured degree of loss aversion respond more favorably to the loss-framed message, whereas individuals with lower levels of loss aversion do not. This establishes that the relevant moderator is economically meaningful, but also highlights the practical problem that it is not directly observed in the field. We therefore turn to prediction. Using the behavioral measurement sample, we train a model that maps digital footprints to the loss-aversion measure and apply it to the full field sample. In the field sample, predicted loss aversion again explains treatment heterogeneity in the expected direction. A policy that treats visitors with a high predicted level of loss aversion yields a statistically significant increase in revenues. Thus, an intervention that is ineffective on average becomes profitable when deployed selectively using a behavioral construct that can be predicted at scale.

A natural alternative, in principle, would be to rely on recent developments in causal machine learning. If one has access to a large randomized trial before implementing treatment targeting, methods such as causal forests can be used to estimate heterogeneous treatment effects directly from experimental data and to derive treatment rules on that basis (Wager and Athey, 2018). In many applications, however, generating such data may be difficult. Running a large randomized trial before targeting may be operationally infeasible, too risky because individuals must be deliberately mistreated for learning purposes, or simply too costly given the nature of the intervention. Our approach is useful precisely because it does not require such large-scale experimental training data.

As in our application randomized field data are available for evaluation, we can compare the approaches directly to assess their relative performance. We find that treatment targeting based on causal-forest estimates does not generate robust, statistically significant gains, whereas the loss-aversion-based targeting rule does. While this performance advantage of our approach may not generalize to other settings, the comparison illustrates the central idea of our approach: a theory-guided procedure combining behavioral measurement and machine learning prediction can be valuable precisely in environments in which large-scale experimentation is infeasible or too costly, and it can perform well even when compared to methods that do rely on such data.

We contribute to several strands of literature. First, we relate to the broader rise of randomized field experiments and large-scale online A/B testing as central tools of data-driven decision-making in the social sciences, particularly economics, and increasingly in business and policy (Harrison

and List, 2004; Duflo, Glennerster and Kremer, 2007; Einav and Levin, 2014; Ascarza, 2018). Rather than only asking whether a treatment works on average, researchers focus on for whom and how to apply it, deriving assignment rules that map observables into interventions. The literature on policy-learning formalizes this shift and shows how assignment rules can be learned from experimental or observational data (Manski, 2004; Kitagawa and Tetenov, 2018; Athey and Wager, 2021). However, learning such rules in controlled environments is only the first step. Recent work on scaling experiments to policy emphasizes that evidence on efficacy must be complemented by evidence on implementability under real-world constraints (List, 2022, 2024; Rasul, 2026). In practice, constraints like limited overlap, operational costs, and noisy signals make large-scale RCTs difficult or expensive (Lewis and Rao, 2015). Our paper speaks directly to this concern. The central challenge in our setting is not only to identify a relevant source of heterogeneity, but also to make it usable in the field when direct measurement at scale is infeasible. By combining behavioral measurement in a small sample with prediction in routine field data, the paper illustrates one concrete way to design treatment targeting with implementation and scaling in mind from the outset.

Second, we contribute to the growing literature on the relation between economics and machine learning and more specifically, whether the two are substitutes or complements. Recent work has shown that machine learning can help assess how complete existing economic models are, how restrictive they are relative to more flexible alternatives, where their predictions break down, and how they might be improved. Predictive algorithms have also been shown to diagnose human decision systems or to generate anomalies that reveal the limits of established theories (Fudenberg and Liang, 2019; Fudenberg et al., 2022; Andrews et al., 2022; Mullainathan and Obermeyer, 2022; Mullainathan and Rambachan, 2024, 2025; Ludwig and Mullainathan, 2024; Liang, 2025; Fudenberg, Gao and Liang, 2025). Our approach follows the same general logic. We do not use machine learning as a purely agnostic search procedure for treatment heterogeneity. Instead, theory disciplines the choice of the relevant behavioral construct, while machine learning is used to recover that construct from variables observed at scale in the field. In this sense, the paper provides a concrete example of how economic theory and modern machine learning methods can be combined to produce an implementable treatment targeting rule.

Third, we add to the literature on behavioral measurement and experimental economics. A large literature has developed incentivized methods to elicit economically meaningful preferences such as risk attitudes, patience, and social preferences, and has shown that these measures capture substantively important individual heterogeneity (Holt and Laury, 2002; Dohmen et al., 2011; Falk et al., 2018). Our contribution is to shift the role of such measurements from description and validation to implementation. We show how an experimentally elicited preference measure can serve not only as an object of behavioral analysis, but also as an intermediate input into a scalable targeting rule. This links the measurement literature more directly to questions of policy design and field implementation.

Fourth, the paper contributes to the applied literature on loss aversion, reference dependence,

and loss framing. Our focus on loss aversion builds on the foundational literature on reference-dependent preferences (Kahneman and Tversky, 1979; Tversky and Kahneman, 1991) and on established approaches to measuring loss aversion experimentally (Abdellaoui, Bleichrodt and Paraschiv, 2007). It also relates to applied work showing that loss framing can shape behavior in field settings such as firms, health, and education (Hossain and List, 2012; Fryer Jr et al., 2022). Our contribution to this literature is not primarily to ask whether loss framing works on average, but to study how it can be targeted more effectively.

2. Research Design and Methods

We utilize two nested experiments to collect data for our study: a large-scale randomized controlled trial (RCT) and a behavioral measurement experiment.[1] The RCT, conducted in collaboration with a major European fashion retailer in their online store, is detailed in Section 2.1. The behavioral measurement experiment which we ran for a subsample of RCT participants to elicit their loss attitudes, is described in Section 2.2. Section A.5 shows how we apply a causal forest algorithm to estimate individual treatment effects.

2.1. The RCT with a Fashion Retailer

We ran an RCT together with an online fashion store over several weeks in autumn 2020. Although the online store is globally accessible with a country-specific setup, we only consider its German version and the associated visitors. The corresponding website traffic amounts to tens of thousands visits per day.

The company maintains an infrastructure to conduct and analyze RCTs to continuously improve their website in terms of user friendliness, acquired revenues, and conversion rates. Through this infrastructure it is possible to randomly allocate online store visitors to different versions of the website (A/B testing) and to collect basic user data for each consumer. We refer to this data as "digital footprints" (see also Berg et al., 2020). These footprints include only data which is typically generated by default when visiting *any* online website, e.g., the IP address, the browser type, or the operating system of the visitor. Hence, the data are not specific to our setting but typically universally available for website hosts. For a full description of the observed variables see Table 8 in the Appendix. In addition, we also observe whether or not a purchase was made and if so, the number of products ultimately bought and overall price of the shopping basket. We also observe returns of orders. Next to revenues we also consider additional measures such as the conversion rate and the value of the returned products as secondary outcome variables.

[1] Both experiments and their nested structure were preregistered under AEARCTR-0006291. Note that the preregistration also contains three more treatment arms of the RCT which are not part of the nested structure and are therefore not included in this paper.

2.1.1. Treatment: loss framing of rebates

As is the standard in the fashion retail industry, our partner frequently uses discounts to increase revenues and conversion rates for website users. To understand the optimal display of such discounts, they agreed to collaborate on an RCT that exogenously manipulated the framing of discounts that consumers could receive if they finalized their purchase. All consumers were informed about the discount via an advertising banner displayed at the top of their screen when searching for suitable products.

While the placement and the design of the banner was constant across conditions, its exact wording differed between CONTROL and TREATMENT.[2] In the CONTROL condition the discount was framed as a gain at the end of the purchase:

"Secure your discount of up to 70% on items on sale."

In TREATMENT, in contrast the framing emphasized that the rebate would soon be lost if the consumer did not purchase:

"Don't lose your discount of up to 70% on items on sale."

The intervention follows classical framing experiments that manipulate whether individuals perceive a given product or monetary outcome either as a gain or as their endowment that might be lost (Kahneman, Knetsch and Thaler, 1990, 1991). Although the treatment only changes the framing but not the characteristics of the received good, some papers document substantial effects of such loss frames and considerable treatment heterogeneity, for instance, in the areas of education or precautionary savings (Ibanez and Schneider, 2021; Karle, Engelmann and Peitz, 2022). Importantly, the extent to which such a loss frame changes consumer behavior is typically hypothesized to be immediately associated with individuals' loss attitudes. If individuals are loss averse, their tendency to avoid losses should make them particularly susceptible to the treatment manipulation. Applied to our setting, this notion motivates the hypothesis that loss averse participants are more likely to redeem a rebate if they perceive not doing so as losing the rebate.

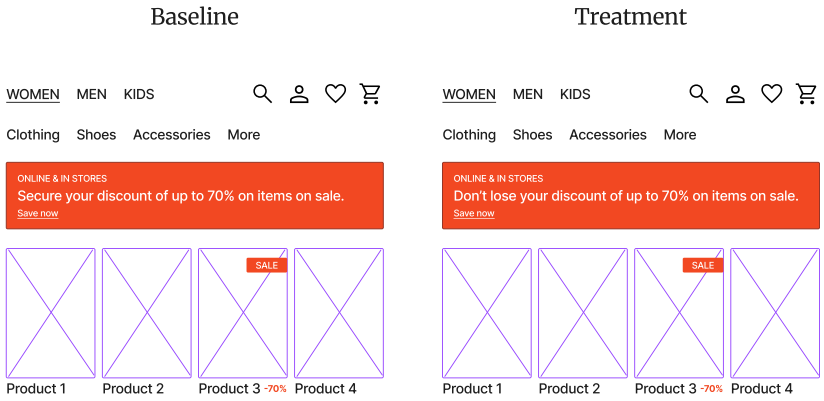
2.1.2. Procedures and implementation

We conducted the intervention in autumn 2020 in the German online store of the retailer.[3] An automated script in the backend randomly assigned visitors to either CONTROL or TREATMENT. This assignment was implemented as soon as a visitor first entered the website, regardless of their entry point into the store. To avoid contamination across treatments, the company used cookies to

[2] We report translations from the original German text.

[3] We accessed a certain share of the entire website traffic.

Figure 1: Schematic depiction of the treatment.



Notes: The figure shows a schematic depiction of the treatment to maintain anonymity of the cooperation partner.

track visitors and to assign them to the same treatment in case they revisited the website during the time of the campaign.[4]

During the campaign we collected a total of 242.228 observations in the CONTROL condition and 242.108 observations in the TREATMENT condition, leaving us with an overall sample of almost 500.000 website visits.

2.2. The Behavioral Measurement experiment

With a subsample of website visitors (N=651), we ran an additional behavioral measurement experiment with the objective of eliciting individuals' degree of loss aversion.[5] As the treatment represents a change in the framing of the rebate from gains to losses, individual loss attitudes should moderate the treatment effect.

While precisely eliciting loss aversion often relies on rather extensive choice lists (see for examples, Fehr and Goette, 2007; Booij, Van Praag and Van De Kuilen, 2010; Gächter, Johnson and Herrmann, 2022), the online setting precludes the use of such time intensive measurement protocols. We therefore elicit only a partial ordering of loss aversion between subjects rather than specific individual preference parameters. Given the nature of our study as intended to provide a general proof-of-concept, this very coarse measurement approach renders our results rather conservative. A more elaborate and longer-lasting measurement protocol likely provides more precise assessments of individual heterogeneity. Such improved assessments should then also lead to even better predictions of treatment effect heterogeneity.

[4]The company used cookies to ensure that an individual using a specific device would not be recorded more than once. It is possible, however, that visitors cleared their cookies such that the same person is included more than once on different occasions.

[5]Despite technical measures aimed at preventing individuals to participate multiple times, we remove 60 observations from our overall sample of 711 records due to repeated participation, leaving us with a sample of 651 distinct observations.

2.2.1. Implementation of the Behavioral Measurement Experiment

An important feature of our experiment is that we measure loss attitudes of consumers on our industry partner' web site. As this goal and restrictions by the industry partner precluded a detailed elicitation protocol, we designed a measurement that consists of only a single choice task. This task was incentivized by vouchers for future purchases on the retailer's web page and varied across participants in order to derive an ordering of loss preferences between subjects. We first endowed each participant with a 15% voucher for their next purchase. To elicit loss aversion it is important to induce an observable reference point against which subjects evaluate the outcome of the experiment and can actually perceive losses. We made sure that participants understood that they can keep this initial voucher and additionally showed them a picture of the respective voucher before they took any further decisions to make it the relevant reference point.

Afterwards, individuals chose whether they want to stake their voucher in a potential gamble. There were three versions of this potential gamble that varied between subjects: in version 1 the individual decided between keeping the 15% voucher or playing a binary lottery that yields a 5% voucher and a 20% voucher with equal probabilities (N=229), in version 2 subjects could keep the 15% voucher or gamble between a 5 % voucher and a 25% voucher (N=210), and in version 3 they could stake the voucher for a gamble between a 5% voucher and a 30% voucher (N=212). Hence, we kept the potentially occurring loss identical across all subjects while varying the potential gain of the lottery. Each participant was randomly assigned to one of the above conditions with equal probability.

Website visitors participating in the RCT received invitations to join the behavioral measurement experiment through a pop-up banner that appeared at a random point in time during their online store visit. This banner invited visitors to participate in a customer survey by offering a link to another website where they could complete the survey. This method of conducting customer surveys follows the typical approach in the online store. We distributed the vouchers provided by our industry partner (750 total) to incentivize the experiment evenly across the six weeks of the RCT. In every given week, we thus randomly invited participants of the RCT to the experiment until we reached 125 submissions. If we did not reach this amount, we carried over the remaining vouchers to the next week.

2.2.2. Classification of Loss Aversion Based on Behavioral Measurement Data

To derive a partial ordering of individual loss attitudes, we assume that the preference of each participant $i \in \{1, \cdots, 651\}$ over binary risks can be described by Cumulative Prospect Theory (Kahneman and Tversky, 1979; Quiggin, 1982) and use an elicitation approach in line with the arguments in Köbberling and Wakker (2005). Suppose that i faces the binary gamble $L = (p, x_1; x_2)$, which yields a consumption utility x_1 with probability p and a consumption utility x_2 otherwise with $x_2 > x_1$. Then, a prospect theory agent evaluates the binary lottery relative to a reference

point $r_i \in \mathbb{R}$ according to:

$$CPT_i(L(p, x_1, x_2)) = \begin{cases} (1-\omega_i^+(p))u_i(x_1) + \omega_i^+(p)u_i(x_2) & \text{if} \quad r_i \leq x_1 \\ \omega_i^-(1-p)u_i(x_1) + \omega_i^+(p)u_i(x_2) & \text{if} \quad x_1 < r_i \leq x_2 \\ \omega_i^-(1-p)u_i(x_1) + (1-\omega_i^-(1-p))u_i(x_2) & \text{if} \quad x_2 < r_i, \end{cases}$$

with non decreasing weighting functions $\omega_i^+, \omega_i^- : [0,1] \to [0,1]$ satisfying $\omega_i^+(1) = \omega_i^-(1) = 1$, and a continuous value function $u_i(\cdot) : \mathbb{R} \to \mathbb{R}$, such that $u_i(r_i) = 0$. To derive the ranking, we impose three assumptions on these objects. First, we assume that the weighting of probabilities does not depend on the position of the respective outcome in relation to the reference point, i.e., $\omega_i^+(p) = \omega_i^-(p) \ \ \forall p, i$. Second, we assume that u_i is a piecewise linear function with

$$u_i(x) = \begin{cases} m_i(x - r_i) & \text{if} \quad x \geq r_i \\ -\lambda_i m_i(x - r_i) & \text{if} \quad x < r_i, \end{cases}$$

where m_i captures differences across subjects in how much they value vouchers on the retailer's webpage, λ_i captures different degrees of loss aversion, and x describes the consumption value of the voucher. The assumption of risk neutrality for small stake gambles is in line with the well established idea that risk aversion can not plausibly explain choice behavior in such small stake gambles (see, Rabin, 2000; Köbberling and Wakker, 2005) and the experimental literature on loss aversion (e.g., Fehr and Goette, 2007; Karle, Engelmann and Peitz, 2022; Gächter, Johnson and Herrmann, 2022). Third, embracing the notion that an individual's reference point is shaped by the endowment at the beginning of the behavioral measurement experiment, we finally assume that $r_i = r$ is the consumption value of a voucher with a 15% rebate for all i.

To derive a ranking over individual's loss aversion, we first classify the different choices that we observe in the experiment. Since there are overall three different versions of the gamble, we observe six different types of individuals. We denote the types by VJ, where $V \in \{1, 2, 3\}$ describes the version of the gamble that the subject faces and $J \in \{A, D\}$, whether the subject accepted (A) or declined (D) the gamble. Hence, type $1A$, for example, faces version one of the gamble and decided to accept the gamble. Denoting the average loss aversion among the subjects of observed type VJ by $\bar{\lambda}(VJ)$, we get the following partial ordering of individual loss aversion:[6]

Measurement 1. *For any distribution of parameters m_i, λ_i in the subject pool with full support over $\mathbb{R}^2$, the expected loss aversion of observed types has the following partial ranking:*
(i) $\bar{\lambda}(1D) \leq \bar{\lambda}(2D) \leq \bar{\lambda}(3D)$,
(ii) $\bar{\lambda}(1A) \leq \bar{\lambda}(1D)$, $\bar{\lambda}(2A) \leq \bar{\lambda}(2D)$, $\bar{\lambda}(3A) \leq \bar{\lambda}(3D)$

By conceptualizing the idea that declining a more favorable gamble implies that subjects are more loss averse, the result shows that the different conditions in our experiment allow us to clas-

[6]The formal arguments to derive Measurement 1 are relegated to Appendix A.6.

sify subjects into loss aversion categories. There are three points worth discussing with respect to our classification. First, it does not allow to distinguish loss aversion at the individual level. In particular, consider a subject that would decline all gambles, because it is extremely loss averse, but happens to face version 2 of the experiment. Our procedure will classify her level of loss aversion as only intermediate (observed type $2D$). While such (partial) mis-classifications will introduce some noise in the estimated ranking, individuals that are classified as $2D$ are still *on average* less loss averse than those being classified as type $3D$. Second, the classification does not provide a complete ordering of types across all individuals in the subject pool. In the later analysis, we will therefore rely on a classification of loss aversion such that observed types 3D and 2D are classified to be loss averse while all other types are classified as not loss averse. In particular, in order to reduce noise in the labelling, we do not label any subject as loss averse that accepts the gamble. Third, the measurement allows for heterogeneity in the subject pool in terms of loss aversion, the utility from the voucher (m_i) and the probability weighting functions, while not imposing any restrictions on their distribution across the population.

We classify all participants that did not accept the two most extreme lotteries as loss averse and all other participants as not loss averse. In our data approximately 10% of subjects are accordingly classified as being loss averse. Chapman et al. (2022) obtain a measure for loss aversion in risky choices on the individual level by conducting experiments in a representative US sample. They find that approximately 50% of participants are loss averse. Our loss aversion indicator is therefore conservative in the sense that relatively few subjects are being classified as loss averse. We opted for this conservative approach, because our procedure is inherently prone to measurement error due to the small number of choices.

Notably, to preserve the integrity of our predictive analysis and avoid over-fitting, these participants are subsequently excluded from any future analysis testing the ML-based targeting.

2.3. Predicting Individual Loss Aversion based on Digital Footprints

We model the prediction of an individual's degree of loss aversion as a multi-class prediction problem because our experimental measure of loss aversion cannot be assumed to be cardinal. That is, we predict to which of the six possible classes of loss aversion we measured in the experiment an individual belongs. As input features, we rely on traces that website visitors' leave whenever they go into the online shop. We leverage these *digital footprints* for two reasons. First, we follow previous work in Finance (e.g., Berg et al., 2020) and Marketing (e.g., von Zahn et al., 2024) that has shown the relevance of digital footprints in the prediction of relevant heterogeneity in indivdiuals' behavior. Second, these footprints are readily available to any web-site host at that moment: the referring website, entry page, IP region, device type, operating system, IP provider, browser type, and time of the week (see Table 8 for details). We treat these predictors as categorical variables and apply one-hot encoding during data preparation. Overall, we rely on $|X| = 33$ features.

We train a supervised multiclass classifier that maps each individual's digital footprint to one of six loss-aversion classes. The training sample is constructed from the experimental dataset after removing duplicate observations at the visitor level. The target variable is recoded to zero-based labels to match the requirements of the learning algorithm. As base learner we use gradient-boosted decision trees implemented in XGBoost (Chen and Guestrin, 2016), which yields estimated class-membership probabilities $\widehat{\Pr}(Y_i = k \mid X_i)$ for $k \in \{1, \ldots, 6\}$. Model selection is performed via randomized hyperparameter search over a predefined grid space (number of trees, maximum depth, learning rate, row subsampling, column subsampling, and regularization-related parameters), using 40 random draws and 3-fold cross-validation. We optimize the macro-averaged $F1$-score to weight all loss-aversion classes equally. Finally, the tuned model is refit on the full training sample. We use this refitted model to predict loss-aversion classes where predicted labels are mapped back to the original 1–6 class coding for interpretation and subsequent analysis.

3. Results

This section reports the results of our empirical analyses. We first estimate the effect of non-targeted loss framing and then test whether individual loss aversion explains systematic heterogeneity in this effect within our behavioral measurement sample. We next turn to scalability: we train a machine-learning model to predict loss aversion from digital footprints and use these predictions to evaluate targeted treatment assignment in the field sample. Finally, we compare this behaviorally motivated targeting policy against purely data-driven alternatives.

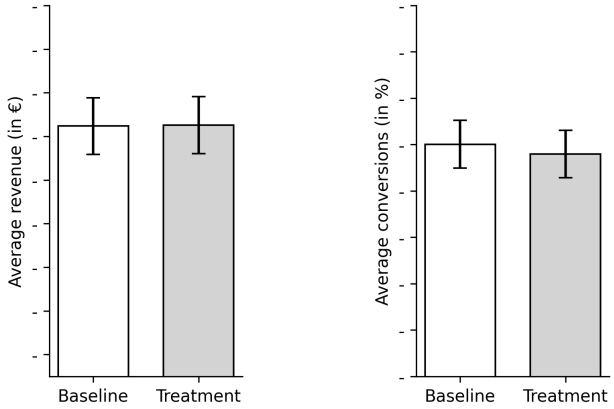
Owing to a non-disclosure agreement with the field partner, we report all business metrics as relative changes against the baseline rather than absolute figures. For example, an increase from a baseline average purchase volume of €1.00 to a treatment average of €1.10 is reported as a 10% increase.

3.1. Effectiveness of Loss Framing without Targeting

Figure 2 shows the effectiveness of an indiscriminatory loss framing policy via a simple treatment-control comparison. The average purchase volume in the TREATMENT increased by 0.63% (95% CI: [–2.7%, +2.8%]) relative to the CONTROL. Despite the substantial sample size, this average treatment effect (ATE) lacks statistical significance ($p = 0.964$, F-test;). Similarly, the loss-framing intervention induced only a marginal decrease in the conversion rate by 0.64% (95% CI: [–2.8%, +1.6%]). This decrease is statistically indistinguishable from zero as well ($p = 0.57$, F-test). Taken together, these results demonstrate that the non-targeted loss-framing intervention failed to generate statistically or economically significant changes across either the intensive or extensive margin of purchasing decisions.

Result 1. *Indiscriminately exposing website visitors to the loss framing has neither an economically nor statistically significant effect.*

Figure 2: Overview RCT.



Notes: Purchasing behavior of online store visitors with 95% confidence intervals. The scale of the y-axis is concealed due to confidentiality. The left panel depicts average purchasing volume across visitors in Euro. The right panel depicts conversion rates in %.

This overall ineffectiveness of loss framing is consistent with several other findings from the literature. For example, Apostolova-Mihaylova et al. (2015) and Krawczyk (2011) find that loss frames do not change average behavior in educational settings. For the purpose of our study, Result 1 establishes our benchmark for the effectiveness of the treatment intervention: Indiscriminate exposure of online store visitors to a loss framing does not foster business outcomes of interest.

Given the small set of covariates observed when users enter the website, there is only limited scope to search for moderators that could reveal treatment effect heterogeneity and, in turn, be used for targeting. Unsurprisingly, exploratory one-dimensional interaction analyses based on these available features indeed fails to uncover latent heterogeneity (see Table 9 in the Appendix). In the following, we therefore investigate whether a targeted approach combining behavioral measurement and machine learning can render the policy effective.

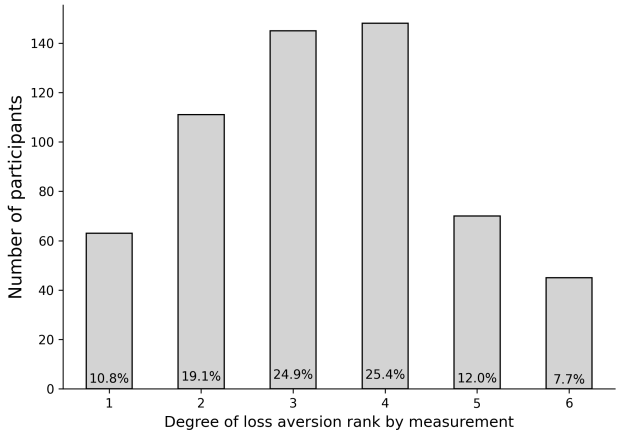
3.2. Treatment Targeting by Behavioral Measurement

This section addresses our central research questions: can we combine behavioral measurement and machine learning to predict individual loss aversion from digital footprints, and does this provide a scalable basis for personalized treatment targeting?

Before developing the predictions, we use our behavioral measurment sample to validate the capacity of the measured loss-aversion levels to capture treatment heterogeneity. Consistent with the analysis of the field sample, we find that the average treatment effect on purchasing behavior is statistically indistinguishable from zero. More specifically, revenues for the treatment group exhibit a non-significant decrease of 6.8% compared to the control group ($p = 0.56, F$-test) and no significant impact on conversion rates (–8.5%, $p = 0.41, F$-test). These results confirm that applying loss framing indiscriminately does not affect revenues.

Our primary objective is, however, to explore whether a behavioral measure of loss aversion explains treatment effect heterogeneity. The protocol described in Section 2.2 classifies participants

Figure 3: Measured loss aversion levels.



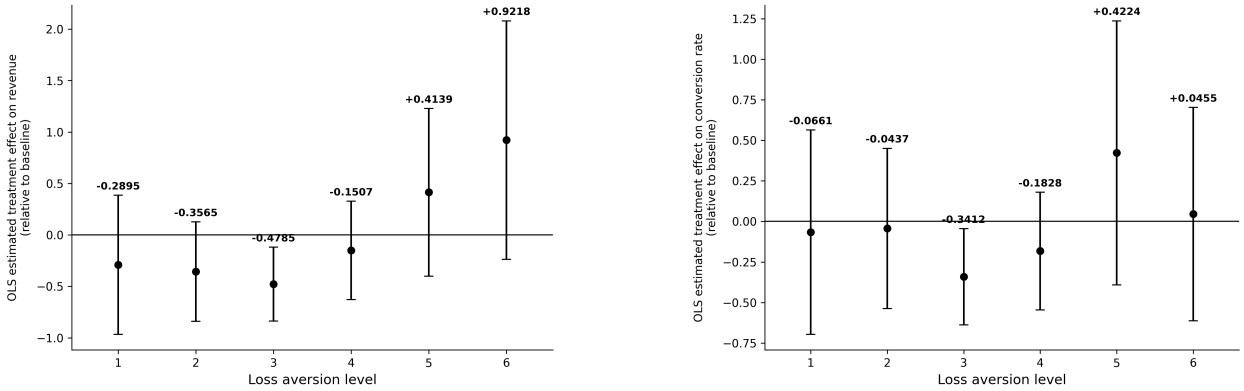
Notes: Average purchasing behaviors of online store visitors with 95% confidence intervals. Actual business figures are concealed due to confidentiality. The left panel depicts average purchasing volume across visitors in Euro. The right panel depicts average conversion rate across visitors in %.

into six distinct categories based on their elicited loss attitudes. Figure 3 illustrates the distribution of individuals across these categories, revealing substantial variation in loss attitudes among store visitors (for additional summary statistics see Table 2).

Figure 4: Treatment heterogeneities for experiment participants

(a) Purchase volume GATE

(b) Conversion rate GATE



Notes: Panel (a) shows for each category of loss aversion separately the estimated ATE for purchasing volume. Panel (b) shows for each category of loss aversion separately the estimated ATE for the conversion rate.

Figure 4 presents the Group Average Treatment Effects (GATE), demonstrating how treatment responses vary across loss aversion categories for both purchase volume (Panel (a)) and conversion rate (Panel (b)). For purchase volumes, we find that consumers with high levels of loss aversion (Categories 5 and 6) exhibited higher purchase volumes relative to the control (+41.4% and +92.2%, respectively). Regression analyses, reported in columns (1)–(3) of Table 1, provide further support for this result: As shown in column (2), participants with above-median loss aversion (Categories 4–6) show a statistically larger treatment effect than those with below-median loss aversion (Categories 1–3). This interaction remains statistically significant (column (3)) if we use

the loss aversion category rather than a median split as moderator. The systematic heterogeneity indicates that behavioral measures of loss aversion can be effective for treatment targeting.

Table 1: Treatment heterogeneity in purchasing behaviors.

	Purchase volume		Conversion rate	
DV:	(1)	(2)	(4)	(5)
Treatment	-0.0898	-0.3844***	-0.0845	-0.1877
	(0.126)	(0.135)	(0.102)	(0.131)
Loss Averse (binary)		0.0516		-0.0526
		(0.164)		(0.124)
Treatment		0.6612**		0.2298
× Loss Averse (binary)		(0.259)		(0.206)
Observations	582	582	582	582
R-squared	0.001	0.028	0.001	0.04

*Notes: We show estimates from OLS regressions with robust standard errors in parentheses. *, **, ***, denote significance levels at p<.01, p<0.05, and p<0.01, respectively. Treatment is a dummy variable indicating whether a participant was exposed to the loss framing.*

In contrast, the results for the conversion rate (the extensive margin) are more nuanced. While Panel (b) in Figure 4 shows a directional trend consistent with the purchase volume findings (with positive effects restricted to Categories 5 and 6), the regression analyses indicate that these interactions are not statistically significant (see Table 1, columns (4)–(6)). This finding suggests that the influence of loss aversion on the treatment effect with respect to the extensive margin (probability of purchasing) is less pronounced or does not even exist.

Embracing the notion that measured loss-aversion is an important predictor of treatment effects, we next proceed with the approach detailed in Section 2.3. In particular, we train a machine learning model to predict individual loss aversion categories based on digital footprints in the field sample. Given that our experimental results identified loss aversion as a robust moderator primarily for the intensive margin, the subsequent analyses focuses exclusively on purchase volume (analyses for conversion rate can be found in the appendix).

Table 2: Loss aversion distribution.

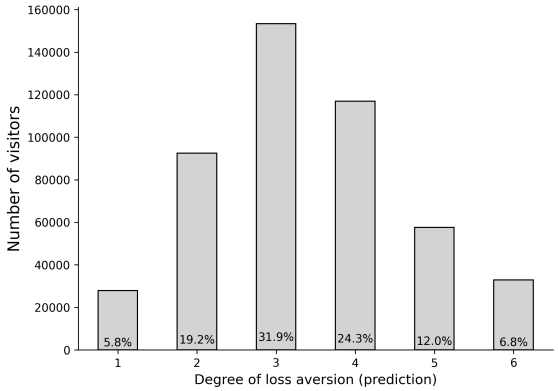
	Summary statistics				
	N	Mean	Median	Variance	Skew
Measured	582	3.319	3	1.9562	0.1258
Predicted	481005	3.379	3	1.617	0.24117

We report summary statistics of the loss aversion distribution.

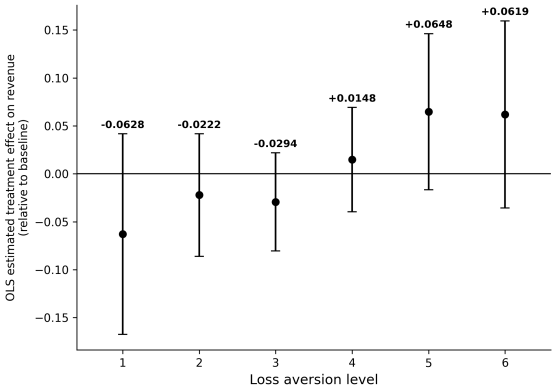
Panel (a) of Figure 5 illustrates the distribution of predicted loss aversion categories in the field sample. Comparing this to the distribution of in the behavioral measurement sample (Figure 3)

Figure 5: Treatment heterogeneities for field sample

(a) Distribution of predicted loss aversion ranks



(b) Purchase volume GATE



Notes: Panel (a) depicts the distribution of predicted loss attitudes of website visitors. Panel (b) shows for each category of loss aversion separately the differences between treatment and baseline.

reveals a high degree of alignment (Table 2): The two distributions share identical medians (4.0) and feature comparable means ($\mu_{pred} = 3.6$ vs. $\mu_{exp} = 3.7$) and variances ($\sigma^2_{pred} = 1.62$ vs. $\sigma^2_{exp} = 1.96$). A Mann–Whitney-U test shows that the distributions are not statistically distinct ($p = 0.375$).

We now examine whether *predicted* loss aversion is able to explain treatment effect heterogeneity. Panel (b) of Figure 5 presents the GATEs for predicted loss aversion categories, comparing average purchase volumes between treatment and control groups within each category. The results demonstrate that predicted loss-aversion categories effectively segment the population with respect to the size and sign of treatment effects: For visitors with predicted loss aversion above the median (Categories 4–6), loss framing increases purchase volumes by 1.5% to 6.5%. By contrast, visitors with below median predicted loss-aversion exhibit a decreases in purchase volumes of about -2.2% to –6.3% in response to the treatment.

Table 6 provides additional regressions confirming the predictive power of predicted loss-aversion categories. Column (2) shows that the interaction between the treatment and a dummy for a median split in predicted loss aversion is statistically significant ($p < 0.02, F$-test). Hence, the treatment effect is 6.9 percentage points larger for individuals with above-median loss aversion. The interaction remains statistically significant if we use the loss-aversion category as the moderator in Column 3 ($p < 0.01, F$-test). Here, a one level increase in predicted loss aversion leads to an increase of the purchase volume by 2.96 percentage points. This corresponds to a standardized effect of XX.

Result 2. *Individual levels of loss aversion predicted based on digital footprints explain heterogeneity in the treatment effect of loss framing.*

Table 3: Treatment heterogeneity

DV: Purchase volume	(1)	(2)
Treatment	0.0006	-0.029*
	(0.014)	(0.018)
Predicted Loss Aversion (binary)		0.1481***
		(0.020)
Treatment		0.0687**
× Predicted Loss Aversion (binary)		(0.029)
Observations	481,005	481,005
R-squared	0.000	0.000

Notes: OLS regressions with robust standard errors in parentheses.

Result 2 shows that machine learning models can effectively scale behavioral measurements to broader populations, thereby facilitating targeted interventions even if no pilot data is available. To quantify the economic value of such a strategy in our setting, we employ an off-policy evaluation (Hitsch, Misra and Zhang, 2023), an econometric framework that can be used to estimate the causal effect of a proposed targeting policy using historical data from a randomized controlled trial. The methodology relies on the orthogonality of the treatment assignment in the RCT to the predicted levels of loss aversion and, therefore, the proposed targeting rule. As the treatment in the RCT sometimes aligns with the treatment recommended by our predictive model, the procedure allows to simulate outcomes for the counterfactual situation in which the treatment had been implemented.

To operationalize the off-policy evaluation, we employ the Inverse Propensity Score (IPS) estimator. A common challenge in evaluating a new policy with historical data is that the target policy (our behavioral targeting rule) would assign treatment at a different frequency than the logging policy (the original RCT) did. IPS addresses this by reweighting each observation by the inverse of the probability that the observed treatment was assigned under the logging policy. Specifically, the IPS estimator rescales the observed outcomes to create a synthetic population that reflects the distribution of the target policy. For an individual i with digital footprints x_i, if our model recommends treatment but the individual was in the RCT's control group, that observation is zero-weighted for the treatment outcome. Conversely, if the recommendation and actual assignment align, the observation is weighted by the inverse of the propensity score—the probability $P(T_i|x_i)$ to account for the fact that it represents a broader segment of the population that would have been treated under the new regime. This reweighting ensures that the resulting purchase volume estimate is a consistent and unbiased representation of the counterfactual, effectively neutralizing any distributional shift between the experimental data and the proposed deployment.

We evaluate a policy that restricts the loss-framing intervention to visitors whose predicted loss attitudes fall within Categories 4 to 6. This policy results in a target cohort comprising 43.1% of all website visitors.[7] We refer to this targeted subgroup as the "Targeted for Treatment" (TT) group,

[7]Our results are robust to alternative policy specifications, such as targeting only Categories 5 and 6, or Category 6 exclusively.

while all other individuals constitute the "Non-Targeted" (NT) group.

We first analyze the average treatment effects (ATE) separately for the TT and NT groups (see also Chernozhukov et al., 2018). Note that we observe four different combinations of actual treatment assignment in the RCT — Control (C) or Treatment (T) — and hypothetical policy assignment if treatment had been targeted according to our approach — TT and NT. Let R_J denote the observed average revenues for groups J with $J \in \{(TT,T),(TT,C),(NT,T),(NT,C)\}$, representing the four combinations of actual and hypothetical treatment assignment. The treatment effect for the policy-aligned group is $R_{TT,T} - R_{TT,C}$, while the effect for those the policy would exclude is $R_{NT,T} - R_{NT,C}$.

We find a statistically significant, positive average treatment effect for the TT-group. In contrast, the average treatment effect for the NT-group is negative. Specifically, the revenue for TT-group individuals increases by +3.98% ($p < 0.08, F$-test). For the NT-group, the average treatment effect equals -2.89% ($p = 0.105, F$-test). This heterogeneity in individual responses to the treatment offers an explanation for the low average treatment effect when assigning treatment indiscriminatory: For the majority of visitors (NT-group; 56.9%) loss framing has a negative impact, which counteracts the positive revenue-increasing effect for another, smaller group of visitors (TT-group; 43.1%).

To quantify the aggregate revenue impact of the targeting strategy, we calculate the counterfactual average revenue, $\hat{R}$, as:

$$\hat{R} = \alpha R_{TT,T} + (1-\alpha) R_{NT,C}$$

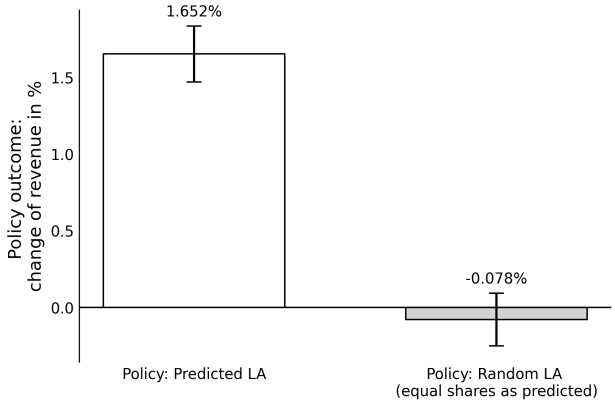
where α denotes the proportion of consumers designated for treatment by the policy (see also Yoganarasimhan, Barzegary and Pani, 2023). This weighting accounts for potential minor deviations between the share of individuals targeted and those actually treated during the RCT. The causal effect of the targeting policy is then the difference between counterfactual revenue $\hat{R}$ and the observed average revenue in the control group.

We benchmark our approach against a "random targeting" policy, which would treat a random share of the population equal in size to our TT group (43.1%). We run Monte Carlo simulations, randomly drawing 50% of the field experiment data over 100 iterations, to bootstrap confidence intervals for the IPS estimates.

Figure 6 shows that combining behavioral measurement and machine learning constitutes an effective targeting approach. The policy based on predicted loss aversion ranks yields a statistically significant revenue increase of 1.65% (95% CI: [1.4%, 1.8%], Wilcoxon signed-rank test: $p < 0.01$). In contrast, the random targeting benchmark yields a negligible and statistically insignificant dencrease of -0.08% (95% CI: [–0.18%, 0.08%], Wilcoxon signed-rank test: $p = 0.18$).

Result 3. *Combining behavioral measurement and machine learning to implement targeting based on predicted individual loss attitudes leads to a statistically significant increase in revenue.*

Figure 6: Off-policy evaluation.



Notes: The Figure shows inverse propensity score estimates of the policy-induced change in revenue for two targeting policies: (i) the loss-aversion–based policy that treats visitors predicted to be loss averse ("Predicted LA") and (ii) a placebo benchmark that assigns treatment at random while matching the treated share of the Predicted LA policy ("Random LA; equal shares as predicted"). Bars show the mean treatment effect of a policy bootstrapped across 100 simulation iterations. Error bars denote 95% confidence intervals around the mean. Figures on top of bars represent means.

3.3. Treatment Targeting by Data Driven approach

Up to this point, we have shown that combining behavioral measurement with machine learning can yield effective treatment targeting even when no pilot data is available. In many applied contexts, however, pilot data either exists or can be collected. Next, we therefore compare our pilot-free targeting approach with a purely data-driven approach from causal machine learning (CML) that learns assignment policies from experimental pilot outcomes. In this scenario, we assume that the organization has conducted a randomized pilot study, thereby generating the labeled data required for a data-driven approach. In contrast to our approach based on measured loss attitudes, CML is agnostic to the underlying mechanism, rather focusing on maximizing the predictive accuracy of the treatment response.

CML represents a class of algorithms designed to predict heterogeneous treatment effects by leveraging pilot data from randomized controlled trials (Athey and Imbens, 2016; Athey, 2017). The predicted individual-level treatment effects constitute a plausible foundation for treatment targeting. We consider a specific CML algorithm — a Causal Forest (Wager and Athey, 2018) — and test if it can effectively uncover latent heterogeneity in consumer responses to loss framing in order to target treatment. Intuitively, the algorithm partitions the covariate space to identify the subgroups with the largest variance in treatment response, thereby learning the optimal targeting policy directly from the observed outcomes in the pilot study.[8]

We train the Causal Forest to predict the marginal change in purchase volume for each online store visitor when exposed to loss framing, utilizing the same digital footprint data as beforehand. Given that the majority of observations are zero we employ a hurdle model approach to refine

[8]Technical details regarding the algorithm's architecture and the implementation of the Causal Forest are provided in Appendix A.5.

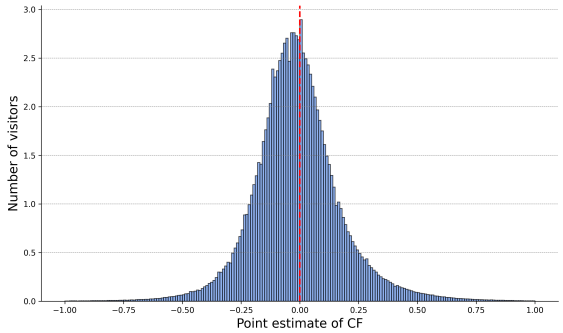
the outcome prediction component of the forest.[9] By integrating this structure into the Causal Forest, we ensure that the estimated treatment effects are not confounded by the high frequency of non-purchasers, thus providing a more accurate prediction of consumer responsiveness.

To assess the effectiveness of targeting based on CML, we again employ an off-policy evaluation framework. We run 50 iterations, where in each iteration the RCT data is randomly partitioned into a training set (70%) and a hold-out evaluation set (30%). The Causal Forest is trained on the 70% partition to estimate individual treatment effects (ITEs). We then apply the trained model to implement a targeting policy on the 30% evaluation set treating only those visitors for whom the model predicts a strictly positive treatment effect. To ensure a reasonable comparison, we simultaneously apply our behavioral measurement targeting approach to the same 30% evaluation set. As before, we implement the policy of treating only those visitors with a predicted loss aversion rank of 4 or higher.

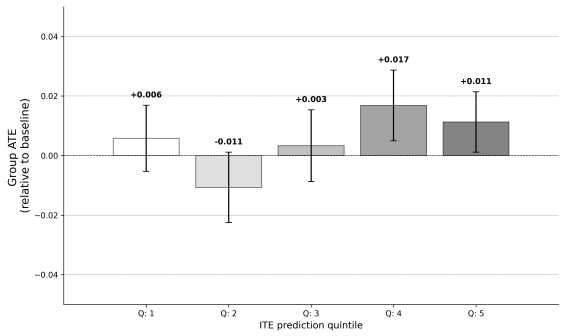
By evaluating both the CML-based policy and the loss-aversion-based policy on the identical hold-out datasets, we obtain a sample of 50 inverse propensity score (IPS) estimates for each approach. This procedure allows us to perform a direct comparison of the two policies.

Figure 7: Causal Forest Results

(a) Distribution of predicted ITEs



(b) GATE per ITE quintile



Notes: Panel (a) depicts the distribution of predicted individual treatment effects (ITE) of website visitors across simulations. Panel (b) shows the mean and 95% confidence intervals of group average treatment effects per quintile of ITE across simulations.

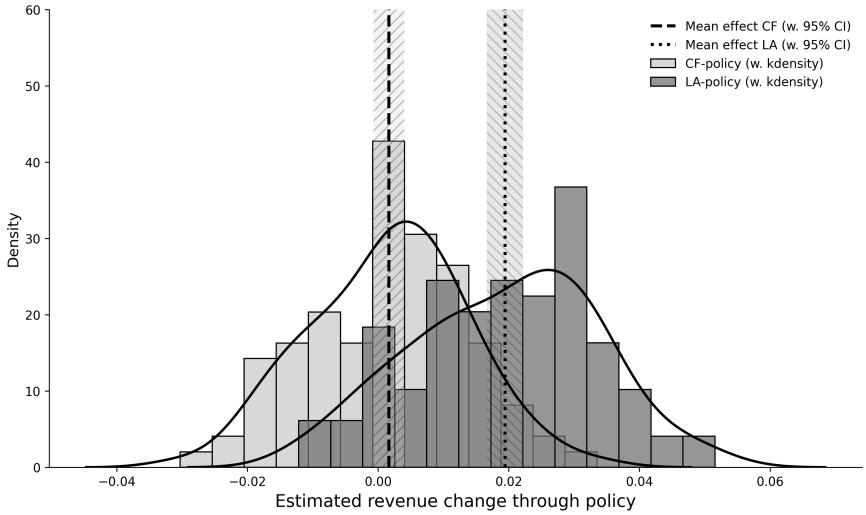
Figure 7 presents the distribution of ITEs across all 50 iterations (Panel (a)) and the mean GATEs partitioned by quintiles (Panel (b)). On average, the model predicts a mean ITE of –0.01% in revenue. However, the GATEs across quintiles indicate that the Causal Forest captures distinct heterogeneity in treatment effects, with mean effects ranging from –0.01% to +0.017%. Notably, positive treatment effects are concentrated in the two highest quintiles (Q4: +0.052%; Q5: +0.242%), whereas the lower quintiles exhibit negative responses.[10] These results suggest that a targeting policy based on strictly positive Causal Forest predictions should outperform an approach with

[9] A hurdle regression is a two-stage modeling framework designed to handle zero-inflated data by separately estimating (i) the probability of an individual crossing the "hurdle" to make a purchase and (ii) the expected purchase volume conditional on a transaction occurring.

[10] Mean ITEs for the remaining quintiles are Q1: –0.236%; Q2: –0.09%; and Q3: –0.023%.

indiscriminatory treatment.

Figure 8: Distribution of estimated policy impact (Off-policy simulation).



Notes: The figure compares the estimated purchase volume impact of two targeting policies in off-oolicy evaluations: Causal Forest (CF) and the loss aversion prediction (LA) policy. The estimated changes reflect inverse propensity scoring (IPS) estimates. The distribution is bootstrapped.

Figure 8 compares the performance of the Causal Forest policy with our behavioral measurement policy. We present the distributions and density estimates of the IPS estimator for revenue changes across all 50 iterations. The empirical results indicate that the CF policy achieves a marginal revenue increase of 0.16% (95% CI: [–0.01%, 0.4%]) by treating 44.4% of the visitor population. Hence, the increase in revenues is not statistically significant at conventional levels. In contrast, the loss aversion prediction policy—which targets a comparable 43.1% of visitors—yields a substantially higher revenue increase of 1.94% (95% CI: [1.66%, 2.21%]). Consequently, the LA-based policy significantly outperforms the data-driven Causal Forest approach within our framework.

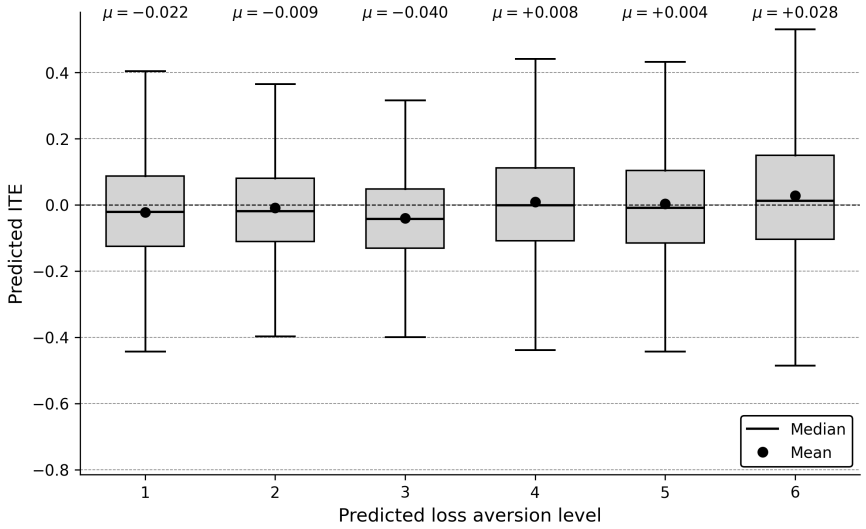
Result 4. *Targeting based on estimated loss attitudes outperforms targeting based on a Causal Forest (revenue increase of approx. +1.9% vs. approx. +0.2%).*

3.4. Association of Predictions on the Individual Level

Finally, our data allow us to compare, at the individual level, the predictions generated by the behavioral measurement policy with those from the Causal Forest policy. Examining the correlation between these two sets of individual-level predictions provides a direct way to assess whether the heterogeneity detected by the Causal Forest aligns with the theoretically grounded construct of loss aversion. A strong alignment between the two sets of predictions would be consistent with the interpretation that the Causal Forest captures meaningful preference heterogeneity, whereas weaker alignment suggests that it the data driven predictions reflect heterogeneity unrelated to loss aversion. This comparison is informative in light of ongoing debates about whether the limited structure of machine learning approaches constrain their ability to generalize beyond randomly

partitioned training and holdout samples drawn from the same underlying population Andrews et al. (2022). In particular, given the low signal-to-noise ratio inherent in online footprint data, the absence of economic structure may increase the risk of overfitting, leading to spurious individual-level predictions.

Figure 9: Box plot of predicted ITE conditional on predicted loss aversion.



Notes: *This figure shows the distribution of predicted individual treatment effects (ITEs) by predicted loss aversion level (1–6). Boxes indicate the interquartile range (25th–75th percentiles). Whiskers extend to the most extreme observations within 1.5 times IQR; outliers are suppressed for readability.*

To explore the association of predictions on the individual level, Figure 9 presents the distribution of predicted ITEs generated by the Causal Forest across each level of predicted loss aversion. The depicted results suggest a weak association between the behavioral measurement prediction and the causal forrest's ITE prediction. Mean predicted ITEs are negative for lower levels of predicted loss aversion (Levels 1–3: –0.022, –0.009, and –0.040, respectively) and positive for higher levels of loss aversion (Levels 4–6: +0.008, +0.004, and +0.028). To quantify this relationship, we estimated a linear regression, incorporating fixed effects for the 50 iterations and clustering standard errors at the iteration level. The results confirm a significant but weak correlation ($p < 0.01$, F-test). Specifically, a one-unit increase in the predicted loss aversion level is associated with a 0.01 Euro increase in the predicted ITE. Due to the weak correlation, the model only explains a modest portion of the variance ($R^2 = 0.108$) in ITE estimates, and the variance of the ITE predictions within each predicted behavioral category remains notable.

These patterns suggests that a substantial part of the heterogeneity identified by the Causal Forrest is not aligned with the structured measure of loss aversion. In line with the idea that the limited structural assumptions of machine learning methods may lead to spurious predictions in settings characterized by low signal-to-noise-ratios, this finding could explain why targeting based on the more structured approach leads to larger revenue gains. On the other hand, it also indicates that there is additional variation not captured by our coarse measure of loss-aversion that is predictive of individual level treatment effects.

4. Discussion and Conclusion

Our findings suggest several broader implications for treatment targeting in practice. The central result of the paper is that a treatment that is ineffective on average can still be valuable when deployed selectively, and that a useful targeting rule can be constructed even when the relevant source of heterogeneity is not directly observed in the field. More broadly, the paper highlights how behavioral measurement and machine learning can be combined to bridge the gap between theory-driven heterogeneity identified in a small behavioral measurement sample and implementation in a large-scale field sample.

Targeting without pilot outcome data. A first implication is that the approach expands the set of settings in which effective targeting is feasible. In the absence of pilot outcome data, researchers can in principle still derive targeting rules from variables observed in the field, but in many applications those observables are only distant proxies for the constructs that economic theory suggests should matter. In practice, this often pushes targeting toward readily observed outcomes rather than toward the underlying mechanisms that theory identifies. As Ascarza (2018) shows, for example, targeting customers who are ex ante most likely to churn need not identify those for whom a retention intervention is most effective. Our approach offers a middle ground. It does not require large pilot RCT data to learn heterogeneous treatment effects directly, yet it also does not restrict theory-driven targeting to the raw covariates available in the field. By measuring a behaviorally meaningful construct in a small sample and linking it to routinely observed field data, it becomes possible to construct theory-guided targeting rules even when the field sample itself contains little information beyond standard operational variables.

Theory-guided targeting, scaling, and transferability. A second implication concerns scaling. Recent work emphasizes that the challenge in moving from experimental evidence to large-scale implementation is not merely to know whether an intervention can work, but whether the logic used to deploy and target it remains valid under real-world conditions and across new environments (List, 2024; Rasul, 2026). A targeting rule based on a purely data-driven, a-theoretical "black-box" approach may be fragile if the underlying relationships are sample-specific or change across populations. By contrast, when the targeted heterogeneity is anchored in a theoretically motivated mechanism and measured behaviorally, this might lead to greater robustness and portability. Our results do not establish such transferability directly, but they point to a plausible advantage of theory-guided targeting: when the rule is built around a stable theory of change rather than an unrestricted predictive fit, it may extrapolate more reliably across settings, participants, or target audiences. In that sense, the value of the approach is not only that it allows targeting without large pilot RCTs, but also that it may generate targeting rules that are better suited for scaling.

When to run pilots and what to measure. A third implication concerns the relative costs and benefits of different forms of data collection before targeting. When experimentation is cheap, operationally simple, and low-risk, running large A/B tests and using causal machine learning to estimate heterogeneous treatment effects may be a sensible strategy. In other environments, however, such pilots may be infeasible, too costly, or too risky because mistreatment is expensive, because implementation is slow, or because treatment effects are too noisy to estimate precisely (Lewis and Rao, 2015). In such settings, a small behavioral measurement exercise combined with theory-guided prediction may be a more attractive alternative. Our comparison with the causal-machine-learning benchmark can be interpreted in this light. We do not claim that theory-guided targeting will generally dominate more agnostic approaches. But the comparison suggests that the choice between large-scale experimental learning and smaller-scale behavioral measurement should depend on the economics of experimentation itself: on how costly it is to randomize, how informative the resulting data are likely to be, and whether there is a clear theoretically motivated moderator that can be approximated from available field data.

Economic theory, machine learning, and feedback to model building. Finally, we also speak to the broader relation between economic theory and machine learning. In our setting, machine learning is not only a tool for scaling a behaviorally grounded construct, but also a tool for learning where existing theory is informative and where it leaves residual variation unexplained. This connects to recent work that uses machine learning to assess the completeness and restrictiveness of economic models, to diagnose where their predictions break down, and to improve them in light of richer empirical patterns (Fudenberg et al., 2022; Fudenberg, Gao and Liang, 2025; Liang, 2025; Mullainathan and Rambachan, 2024). It also relates to the idea, emphasized by Ludwig and Mullainathan (2024), that machine learning can serve as a systematic tool for hypothesis generation by surfacing novel and interpretable regularities in high-dimensional data that standard researcher-driven approaches may miss. Our approach fits naturally into this broader perspective. It uses machine learning not as a replacement for economic reasoning, but as a way to operationalize a theory-motivated behavioral construct in the field and, in turn, to evaluate how far that construct helps explain real-world heterogeneity. In this sense, the approach does not only transport behaviorally measured heterogeneity into the field; it also creates feedback from field evidence back to theory, helping identify which behaviorally grounded mechanisms are most useful for understanding and targeting treatment responses in practice.

References

Abdellaoui, Mohammed, Han Bleichrodt, and Corina Paraschiv. 2007. "Loss Aversion Under Prospect Theory: A Parameter-Free Measurement." *Management Science*, 53(10): 1659–1674.

Andreoni, James, and Charles Sprenger. 2012. "Estimating Time Preferences from Convex Budgets." *American Economic Review*, 102(7): 3333–3356.

Andreoni, James, Michael Callen, Karrar Hussain, Muhammad Yasir Khan, and Charles Sprenger. 2023. "Using preference estimates to customize incentives: an application to polio vaccination drives in Pakistan." *Journal of the European Economic Association*, 21(4): 1428–1477.

Andrews, Isaiah, Drew Fudenberg, Annie Liang, and Chaofeng Wu. 2022. "The Transfer Performance of Economic Models." *arXiv preprint arXiv:2202.04796*.

Apostolova-Mihaylova, Maria, William Cooper, Gail Hoyt, and Emily C Marshall. 2015. "Heterogeneous gender effects under loss aversion in the economics classroom: A field experiment." *Southern Economic Journal*, 81(4): 980–994.

Ascarza, Eva. 2018. "Retention futility: Targeting high-risk customers might be ineffective." *Journal of marketing Research*, 55(1): 80–98.

Athey, Susan. 2017. "Beyond prediction: Using big data for policy problems." *Science*, 355(6324): 483–485.

Athey, Susan, and Guido Imbens. 2016. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences*, 113(27): 7353–7360.

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

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized random forests." *The Annals of Statistics*, 47(2): 1148–1178.

Augenblick, Ned, Muriel Niederle, and Charles Sprenger. 2015. "Working over time: Dynamic inconsistency in real effort tasks." *The Quarterly Journal of Economics*, 130(3): 1067–1115.

Battocchi, Keith, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. 2023. "EconML user guide: Forest based estimators." `https://econml.azurewebsites.net/spec/estimation/forest.html`, Accessed: 2023-11-04.

Berg, Tobias, Valentin Burg, Ana Gombović, and Manju Puri. 2020. "On the rise of fintechs: Credit scoring using digital footprints." *Review of Financial Studies*, 33(7): 2845–2897.

Booij, Adam S, Bernard Van Praag, and Gijs Van De Kuilen. 2010. "A parametric analysis of prospect theory's functionals for the general population." *Theory and Decision*, 68(1): 115–148.

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek. 2014. "Gender, competitiveness, and career choices." *The quarterly journal of economics*, 129(3): 1409–1447.

Chapman, Jonathan, Erik Snowberg, Stephanie W Wang, and Colin Camerer. 2022. "Looming large or seeming small? Attitudes towards losses in a representative sample." National Bureau of Economic Research.

Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A scalable tree boosting system." 785–794.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters." *Econometrics Journal*, 21(1): C1–C68.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 2011. "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences." *Journal of the European Economic Association*, 9(3): 522–550.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. "Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics*. Vol. 4, , ed. T. Paul Schultz and John A. Strauss, Chapter 61, 3895–3962. Amsterdam:Elsevier.

Einav, Liran, and Jonathan Levin. 2014. "Economics in the age of big data." *Science*, 346(6210): 1243089.

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 2018. "Global Evidence on Economic Preferences." *The Quarterly Journal of Economics*, 133(4): 1645–1692.

Fehr, Ernst, and Lorenz Goette. 2007. "Do workers work more if wages are high? Evidence from a randomized field experiment." *American Economic Review*, 97(1): 298–317.

Fryer Jr, Roland G, Steven D Levitt, John List, and Sally Sadoff. 2022. "Enhancing the efficacy of teacher incentives through framing: A field experiment." *American Economic Journal: Economic Policy*, 14(4): 269–299.

Fudenberg, Drew, and Annie Liang. 2019. "Predicting and understanding initial play." *American Economic Review*, 109(12): 4112–4141.

Fudenberg, Drew, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. 2022. "Measuring the completeness of economic models." *Journal of Political Economy*, 130(4): 956–990.

Fudenberg, Drew, Wayne Gao, and Annie Liang. 2025. "How flexible is that functional form? Quantifying the restrictiveness of theories." *Review of Economics and Statistics*, 1–16.

Gächter, Simon, Eric J Johnson, and Andreas Herrmann. 2022. "Individual-level loss aversion in riskless and risky choices." *Theory and Decision*, 92(3): 599–624.

Guo, Ruocheng, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. 2020. "A survey of learning causality with data: Problems and methods." *ACM Computing Surveys (CSUR)*, 53(4): 1–37.

Harrison, Glenn W., and John A. List. 2004. "Field Experiments." *Journal of Economic Literature*, 42(4): 1009–1055.

Hitsch, Günter J, Sanjog Misra, and Walter Zhang. 2023. "Heterogeneous treatment effects and optimal targeting policy evaluation." *Working Paper*.

Holt, Charles A., and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review*, 92(5): 1644–1655.

Hossain, Tanjim, and John A List. 2012. "The behavioralist visits the factory: Increasing productivity using simple framing manipulations." *Management Science*, 58(12): 2151–2167.

Ibanez, Marcela, and Sebastian O Schneider. 2021. "Income Risk, Precautionary Saving, and Loss Aversion–An Empirical Test." *MPI Collective Goods Discussion Paper*, , (2021/6).

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, 47(2): 263–292.

Kahneman, Daniel, Jack L Knetsch, and Richard H Thaler. 1990. "Experimental tests of the endowment effect and the Coase theorem." *Journal of Political Economy*, 98(6): 1325–1348.

Kahneman, Daniel, Jack L Knetsch, and Richard H Thaler. 1991. "Anomalies: The endowment effect, loss aversion, and status quo bias." *Journal of Economic perspectives*, 5(1): 193–206.

Karle, Heiko, Dirk Engelmann, and Martin Peitz. 2022. "Student performance and loss aversion." *Scandinavian Journal of Economics*, 124(2): 420–456.

Kitagawa, Toru, and Aleksey Tetenov. 2018. "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice." *Econometrica*, 86(2): 591–616.

Köbberling, Veronika, and Peter P Wakker. 2005. "An index of loss aversion." *Journal of Economic Theory*, 122(1): 119–131.

Krawczyk, Michał. 2011. "Framing in the field. A simple experiment on the reflection effect." *University of Warsaw Faculty of Economic Science Working Paper*, 14: 54.

Lewis, Randall A., and Justin M. Rao. 2015. "The Unfavorable Economics of Measuring the Returns to Advertising." *The Quarterly Journal of Economics*, 130(4): 1941–1973.

Liang, Annie. 2025. "Using Machine Learning to Generate, Clarify, and Improve Economic Models." Working paper, August 27, 2025.

List, John A. 2022. *The voltage effect: How to make good ideas great and great ideas scale.* Crown Currency.

List, John A. 2024. "Optimally generate policy-based evidence before scaling." *Nature*, 626(7999): 491–499.

Ludwig, Jens, and Sendhil Mullainathan. 2024. "Machine learning as a tool for hypothesis generation." *The Quarterly Journal of Economics*, 139(2): 751–827.

Manski, Charles F. 2004. "Statistical Treatment Rules for Heterogeneous Populations." *Econometrica*, 72(4): 1221–1246.

Meier, Stephan, and Charles Sprenger. 2010. "Present-biased preferences and credit card borrowing." *American Economic Journal: Applied Economics*, 2(1): 193–210.

Mullainathan, Sendhil, and Ashesh Rambachan. 2024. "From predictive algorithms to automatic generation of anomalies." National Bureau of Economic Research.

Mullainathan, Sendhil, and Ashesh Rambachan. 2025. "Science in the Age of Algorithms." *The Economics of Transformative AI*.

Mullainathan, Sendhil, and Ziad Obermeyer. 2022. "Diagnosing physician error: A machine learning approach to low-value health care." *The Quarterly Journal of Economics*, 137(2): 679–727.

Niederle, Muriel, and Lise Vesterlund. 2007. "Do Women Shy Away from Competition? Do Men Compete Too Much?" *Quarterly Journal of Economics*, 122(3): 1067–1101.

Oprescu, Miruna, Vasilis Syrgkanis, and Zhiwei Steven Wu. 2019. "Orthogonal random forest for causal inference."

Quiggin, John. 1982. "A theory of anticipated utility." *Journal of Economic Behavior & Organization*, 3(4): 323–343.

Rabin, Matthew. 2000. "Risk aversion and expected-utility theory: a calibration theorem." *Econometrica*, 68: 1281–1292.

Rasul, Imran. 2026. "From field experiments to policy interventions at scale."

Schneider, Sebastian O, and Matthias Sutter. 2026. "Risk preferences and field behavior: The relevance of higher-order risk preferences." *American Economic Review*, 116(1): 88–118.

Schuler, Alejandro, Michael Baiocchi, Robert Tibshirani, and Nigam Shah. 2018. "A comparison of methods for model selection when estimating individual treatment effects." *arXiv preprint arXiv:1804.05146*.

Syrgkanis, Vasilis, Greg Lewis, Miruna Oprescu, Maggie Hei, Keith Battocchi, Eleanor Dillon, Jing Pan, Yifeng Wu, Paul Lo, Huigang Chen, et al. 2021. "Causal inference and machine learning in practice with EconML and CausalML: Industrial use cases at Microsoft, Tripadvisor, Uber." 4072–4073.

Tversky, Amos, and Daniel Kahneman. 1991. "Loss Aversion in Riskless Choice: A Reference-Dependent Model." *The Quarterly Journal of Economics*, 106(4): 1039–1061.

von Zahn, Moritz, Kevin Bauer, Cristina Mihale-Wilson, Johanna Jagow, Maximilian Speicher, and Oliver Hinz. 2024. "Smart Green Nudging: Reducing Product Returns through Digital Footprints and Causal Machine Learning." *Marketing Science*, forthcoming.

Wager, Stefan, and Susan Athey. 2018. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association*, 113(523): 1228–1242.

Yoganarasimhan, Hema, Ebrahim Barzegary, and Abhishek Pani. 2023. "Design and evaluation of optimal free trials." *Management Science*, 69(6): 3220–3240.

A. Appendix

A.1. One-Hot Regression Specification

Table 4: Treatment heterogeneity in revenue and conversions (Measurement data).

	Revenue	Conversions
DV:	(1)	(2)
Treatment	-0.2474 (0.293)	-0.0639 (0.309)
Loss Type Level: 2	0.2463 (0.281)	0.0511 (0.255)
Loss Type Level: 3	0.0836 (0.232)	0.0855 (0.233)
Loss Type Level: 4	0.1033 (0.248)	-0.0373 (0.226)
Loss Type Level: 5	0.1429 (0.393)	-0.0639 (0.286)
Loss Type Level: 6	0.4707 (0.465)	0.2502 (0.395)
Treatment × Loss Type Level: 2	-0.1450 (0.400)	0.0194 (0.402)
Treatment × Loss Type Level: 3	-0.2014 (0.340)	-0.2952 (0.348)
Treatment × Loss Type Level: 4	0.1031 (0.375)	-0.1060 (0.354)
Treatment × Loss Type Level: 5	0.6603 (0.506)	0.4452 (0.485)
Treatment × Loss Type Level: 6	1.4691* (0.827)	0.1192 (0.508)
Constant	0.8546*** (0.202)	0.9668*** (0.204)
Observations	582	582
R-squared	0.070	0.019
Adj. R-squared	0.052	0.000
F-statistic	2.372	1.071
Prob (F-statistic)	0.00716	0.383

Notes: We show estimates from OLS regressions with heteroskedasticity-robust (HC1) standard errors in parentheses. The omitted category for the loss-type indicators is Loss Type 1.
** p<0.10, ** p<0.05, *** p<0.01.*

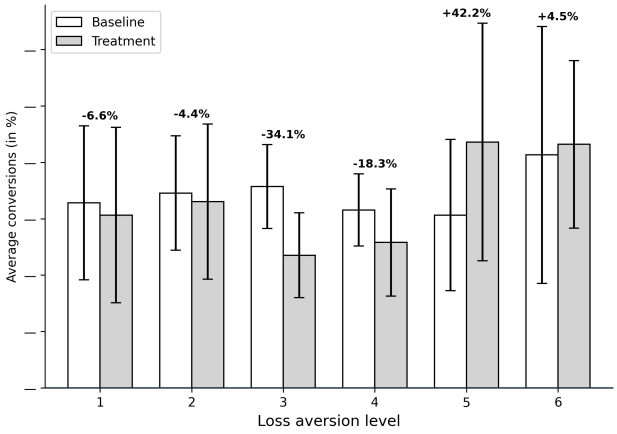
Table 5: Treatment heterogeneity in revenue and conversions (field study).

	Revenue	Conversions
DV:	(1)	(2)
Treatment	-0.0694 (0.059)	-0.0434 (0.048)
Loss Type Level: 2	-0.1256** (0.049)	-0.1326*** (0.039)
Loss Type Level: 3	-0.2263*** (0.046)	-0.2285*** (0.037)
Loss Type Level: 4	-0.0422 (0.048)	-0.0735* (0.038)
Loss Type Level: 5	-0.0838 (0.052)	-0.0889** (0.042)
Loss Type Level: 6	0.1644*** (0.062)	0.1141** (0.052)
Treatment × Loss Type Level: 2	0.0477 (0.067)	0.0114 (0.054)
Treatment × Loss Type Level: 3	0.0436 (0.063)	0.0214 (0.051)
Treatment × Loss Type Level: 4	0.0851 (0.066)	0.0614 (0.053)
Treatment × Loss Type Level: 5	0.1356* (0.073)	0.0711 (0.058)
Treatment × Loss Type Level: 6	0.1480* (0.086)	0.0681 (0.071)
Constant	1.1053*** (0.043)	1.1190*** (0.035)
Observations	481005	481005
R-squared	0.001	0.001
Adj. R-squared	0.001	0.001
F-statistic	24.77	28.41
Prob (F-statistic)	5.53e-52	2.11e-60

Notes: We show estimates from OLS regressions with heteroskedasticity-robust (HC1) standard errors in parentheses. The omitted category for the loss-type indicators is Loss Type 1. * *p<0.10,* ** *p<0.05,* *** *p<0.01.*

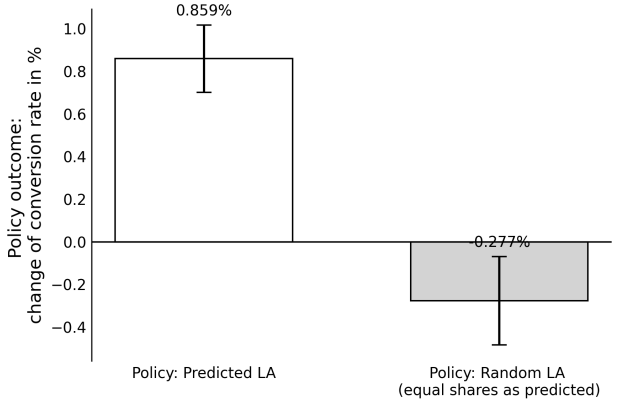
A.2. Conversion rate with predicted LA.

Figure 10: GATEs for conversion rate.



Notes: .

Figure 11: Off-policy evaluation (conversion rate).



Notes: .

Table 6: Treatment heterogeneity

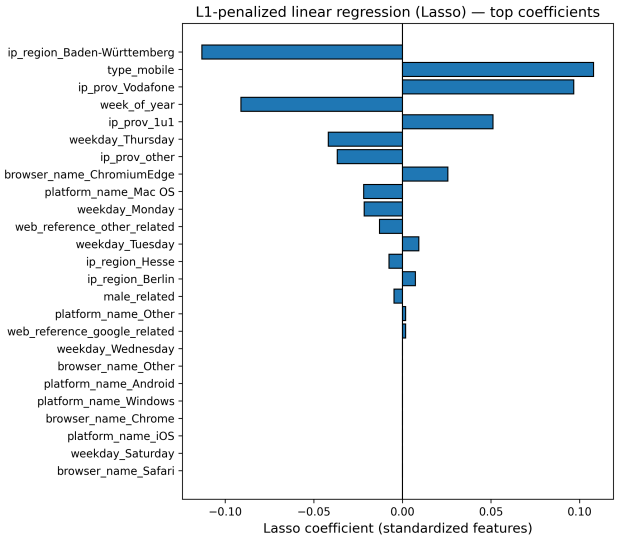
DV: Purchase volume	(1)	(2)
Treatment	-0.0064	-0.0276**
	(0.011)	(0.014)
Predicted Loss Aversion (binary)		
		(0.016)
Treatment		0.0495**
× Predicted Loss Aversion (binary)		(0.023)
Predicted Loss Aversion (level)		
Treatment		
× Predicted Loss Aversion (level)		
Observations	481,005	481,005
R-squared	0.000	0.000

Notes: OLS regressions with robust standard errors in parentheses.

A.3. What the model has learned

A natural question is which part of the digital footprints has explanatory power for loss aversion. Although this ignores the role of non-linear interaction of different elements of footprints, Figure **??** presents Lasso OLS regressions with participants' loss aversion rank as the dependent variable and individual features of their digital footprint as independent variables. The results reveal that certain footprint features are significantly correlated with loss aversion. For example, individuals using a mobile device are significantly less, while those from the Baden-Württemberg region are significantly more loss averse. We also provide SHAP plots for our non-linear XGboost model in Figure 13.

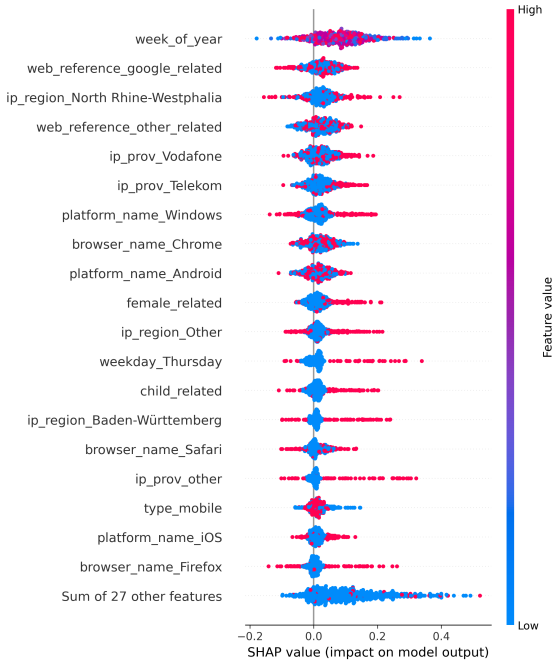
Figure 12: Lasso OLS coefficients.



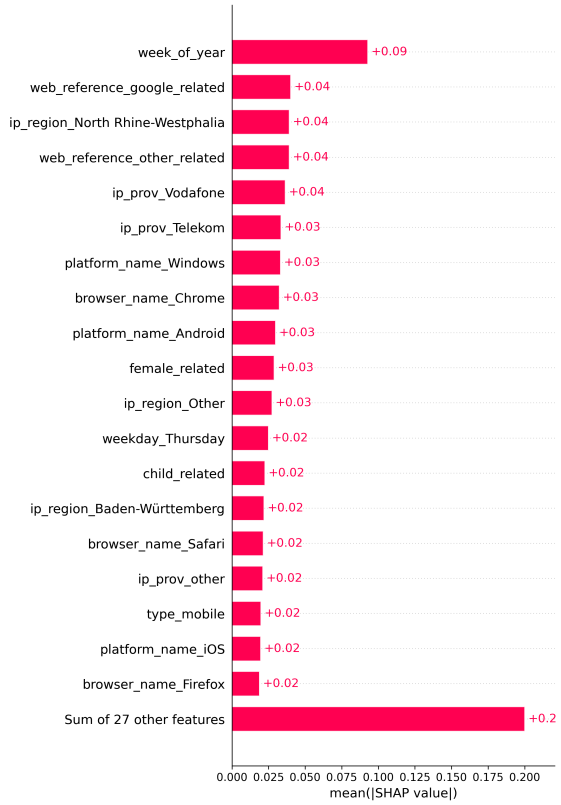
Notes: .

Figure 13: SHAP plots

(a) Distribution of predicted loss aversion ranks



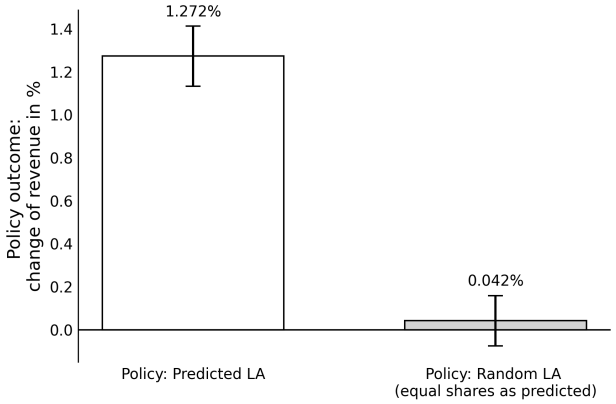
(b) ATE per loss aversion rank



Notes:

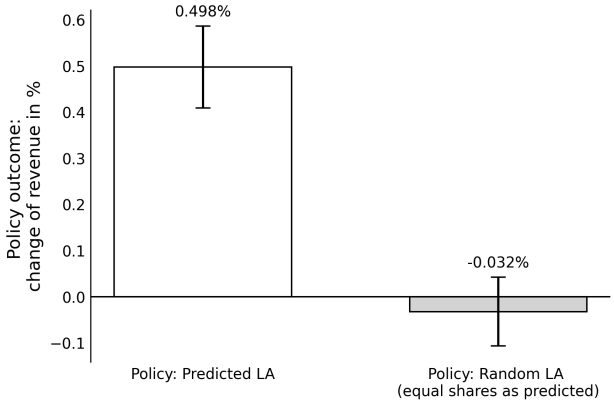
A.4. Robustness tests off-policy evaluations

Figure 14: Off-policy evaluation (treating only loss averse 5 & 6).



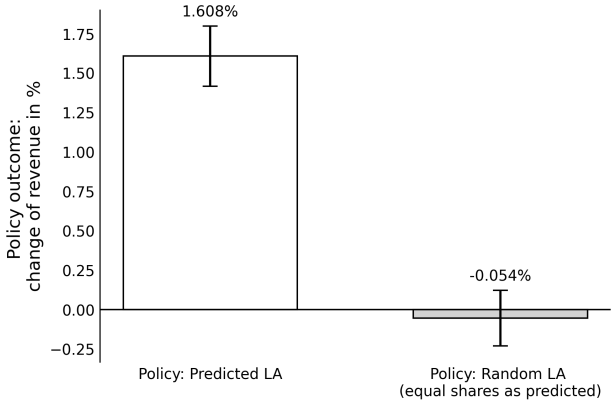
Notes: .

Figure 15: Off-policy evaluation (treating only loss averse 6).



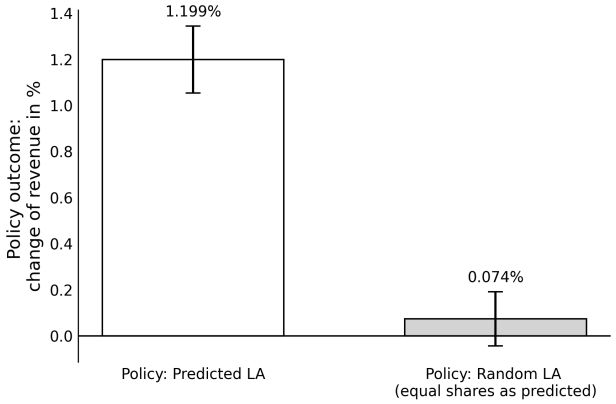
Notes: .

Figure 16: Off-policy evaluation with IPS score.



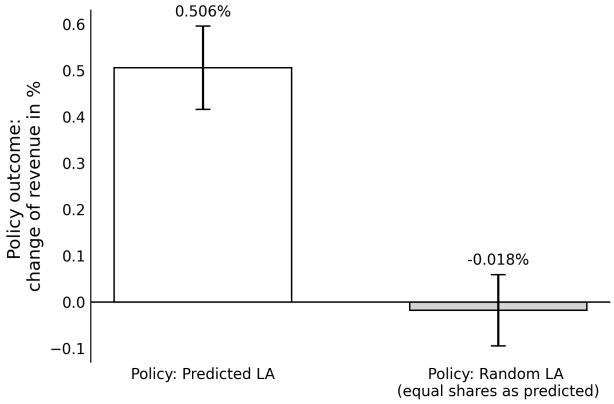
Notes: The Figure shows inverse propensity score estimates of the policy-induced change in revenue for two targeting policies: (i) the loss-aversion–based policy that treats visitors predicted to be loss averse ("Predicted LA") and (ii) a placebo benchmark that assigns treatment at random while matching the treated share of the Predicted LA policy ("Random LA; equal shares as predicted"). Bars show the mean treatment effect of a policy bootstrapped across 100 simulation iterations. Error bars denote 95% confidence intervals around the mean. Figures on top of bars represent means.

Figure 17: Off-policy evaluation with IPS score (treating only loss averse 5 & 6).



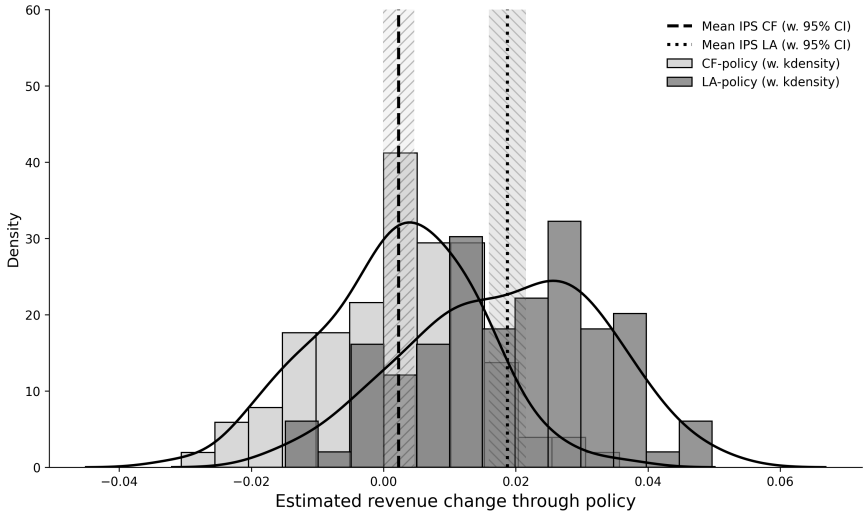
Notes: The Figure shows inverse propensity score estimates of the policy-induced change in revenue for two targeting policies: (i) the loss-aversion–based policy that treats visitors predicted to be loss averse ("Predicted LA") and (ii) a placebo benchmark that assigns treatment at random while matching the treated share of the Predicted LA policy ("Random LA; equal shares as predicted"). Bars show the mean treatment effect of a policy bootstrapped across 100 simulation iterations. Error bars denote 95% confidence intervals around the mean. Figures on top of bars represent means.

Figure 18: Off-policy evaluation with IPS score (treating only loss averse 6).



Notes: The Figure shows inverse propensity score estimates of the policy-induced change in revenue for two targeting policies: (i) the loss-aversion–based policy that treats visitors predicted to be loss averse ("Predicted LA") and (ii) a placebo benchmark that assigns treatment at random while matching the treated share of the Predicted LA policy ("Random LA; equal shares as predicted"). Bars show the mean treatment effect of a policy bootstrapped across 100 simulation iterations. Error bars denote 95% confidence intervals around the mean. Figures on top of bars represent means.

Figure 19: Distribution of estimated policy impact (Off-policy simulation IPS).



Notes: The figure compares the estimated purchase volume impact of two targeting policies in off-oolicy evaluations: Causal Forest (CF) and the loss aversion prediction (LA) policy. The estimated changes reflect inverse propensity scoring (IPS) estimates. The distribution is bootstrapped.

Table 7: Overview of treatment effects.

	Average purchase behaviors	
Variable:	In Euro	Conversion rate
%-Change	+0.63%	-0.64%
p-value (F-test)	0.964	0.57

We report non-parametric differences in %. Associated p-values originate from OLS regression with robust standard errors.

Table 8: Summary statistics and randomization check.

	RCT	Measurement experiment	p-value
Weekday	0.683	0.835	< 0.01
Mobile device	0.762	0.698	< 0.01
Web referrer:			
Newsletter	0.031	0.06	< 0.01
Facebook	0.062	0.031	< 0.01
Instagram	0.011	0.003	0.12
Google	0.568	0.486	< 0.01
Other	0.287	0.402	< 0.01
Entry page:			
Children categ.	0.163	0.174	0.53
Female categ.	0.292	0.225	$p < 0.01$
Male categ.	0.048	0.052	0.79
IP-region:			
Baden-Württ.	0.11	0.125	0.28
Bavaria	0.136	0.112	< 0.1
Berlin	0.059	0.072	0.223
Hesse	0.097	0.091	0.662
North Rhine-West.	0.237	0.256	0.31

Other	0.23	0.266	< 0.05
IP-provider:			
Telekom	0.384	0.326	< 0.1
Vodafone	0.3	0.369	< 0.1
Telefonica	0.106	0.113	0.62
1&1	0.054	0.07	0.11
Other	0.155	0.12	< 0.05
Web browser:			
Chrome	0.522	0.515	0.82
Safari	0.373	0.311	< 0.01
Firefox	0.067	0.143	< 0.01
ChromiumEdge	0.024	0.015	0.24
Other	0.015	0.015	0.999
Operating sys:			
Android	0.45	0.421	0.18
Mac	0.068	0.045	< 0.05
IoS	0.313	0.278	< 0.1
Windows	0.161	0.254	< 0.01
Other	0.009	0.002	0.11

Notes: The table presents summary statistics for the variables used to train the causal forest (one-hot encoded). Th variables are defined as follows: "Weekday" indicates whether the online shop visit occurred on a weekday, "Mobi device" whether the shopping was conducted on a phone or tablet, and "Web referrer" specifies the source throug which the customer accessed the website. "Entry page" refers to the first page visited in the online shop, "IP regio identifies the general location based on the customer's IP address, and "IP provider" indicates the network operat providing internet access. Additionally, "Web browser" and "Operating system" represent the type of browser and th operating system used by the customer, respectively. "Operating system" indicates the operating system of the devic customers used. We further report p-values from ^{2}-tests as randomization checks.

A.5. Estimating individual treatment effects using Causal Forests

We developed a CML system based on a causal forest (CF) (Athey, Tibshirani and Wager, 2019; Wager and Athey, 2018) that predicts the conditional average treatment effect (CATE) $E(Y_{i,1} - Y_{i,0}|X = x_i)$ measuring the expected change in the outcome variable $Y_{i,j}$ for individual i with certain characteristics $X = x_i$ when she is exposed to the treatment intervention $j = 1$ (compared to not being exposed $j = 0$).

A.5.1. Causal Forests

CFs build on causal trees – regression trees adapted to predicting treatment effects. Conceptually, a regression tree sorts observations according to binary questions – *Is the visitor on a mobile device?, Did she arrive from the Google website?, Is the visit on a weekend?* – and, at each split, sends observations with different answers down different branches, ultimately forming final leaves. At each split, the algorithm chooses the respective feature which maximizes the heterogeneity of treatment effects across the resulting leaves, subject to a penalty that prevents over-fitting.[11] A tree trained with this criterion can predict the CATE for an individual i with characteristics x_i: Characteristics x_i determine the leaf L individual i ends up in. The tree's CATE estimate is then the simple difference in mean outcomes between treated and control visitors in that leaf, $\overline{Y}_{1,L} - \overline{Y}_{0,L}$. Critically, we adopt the honest variant: Deciding splits and therefore building a tree is done on one subsample of the data, while the calculation of corresponding treatment effects is done on a disjoint subsample and hence "out-of-sample". (add the corresponding Athey reference here?)

For a single tree, the potential questions splitting the observations are selected from a randomly chosen subsample of the full set of features. To form a causal forest, one therefore grows several individual trees, each based on a different bootstrap sample of the data and a random subset of features at each split. The forest then averages the CATE over all individual trees.

A.5.2. Bootstrapping Procedure

As a random forest consists of several individual trees and every individual tree is random due to variation in the split of the data as well as the feature space, also an individual random forest always entails sampling variation. To make sure that our results are not an artifact of a particularly lucky draw, we employ a bootstrapping strategy where we estimate 25 separate random forests, each based on randomly splitting the data into train-test sets (50%/50%). This procedure ensures that treatment effects are always estimated on data not available for growing the underlying forest and hence constitutes an "out-of-sample" calculation not prone to overfitting. Further, the resulting distribution of estimates from the 25 separate random forests allows us to calculate p-values and confidence intervals in the spirit of a Monte-Carlo simulation. The resulting confidence intervals thus provide a measure for how robust our findings are with respect to the underlying split of the data in the train and test sets as well as to the randomness that is implied by the construction of the random forest.

A.5.3. Implementation details

We employed the CausalForestDML estimator, a tool recommended for integrating a Causal Forest into a double machine learning framework (Oprescu, Syrgkanis and Wu, 2019). In a first step, this procedure partials out the variation in the outcome variable as well as the treatment indicator that

[11] This approach is mathematically equivalent to minimizing the expected mean squared error of hypothetical treatment effects, as demonstrated by Athey and Imbens (2016).

can be explained by other covariates. In a second step, treatment effects are estimated using only the residual data, which is orthogonal to other covariates by construction.

We implemented the CF in Python, utilizing the popular EconML and scikit-learn libraries[12] and adhered to well-established procedures, primarily using recommendations and default settings from EconML, a widely-used and publicly accessible library (Syrgkanis et al., 2021; Battocchi et al., 2023). We fine-tuned the CF's hyperparameters (number of trees, size of leaves, ...) through a grid search and 5-fold cross-validation on the training data, focusing on optimizing out-of-sample R-score performance (Schuler et al., 2018).

A.5.4. Features

Our CF utilizes the digital footprints of website visitors as predictor variables to estimate the CATE of our loss framing on revenue (in Euro). Similar to traditional machine learning models, the accuracy of the CF's predictions hinges on the richness of the data available during the training process (Guo et al., 2020). Since the goal is to assign visitors to a treatment upon their entry into the online store, we limit ourselves to data that is typically available to any web-site host at that moment: the referring website, entry page, IP region, device type, operating system, IP provider, browser type, and time of the week (see Table 8 for details). We treat these predictors as categorical variables and apply one-hot encoding during data preparation. Overall, the CF uses $|X| = 33$ features.

A.6. Derivation of the Behavioral Measurement

To derive Measurement 1, consider the decision of an arbitrary subject i when facing a safe 15% voucher or a Lottery $L(y, x, \frac{1}{2})$, where r is the consumption value of a 15% rebate, y is the consumption value of a 5% rebate and x is the consumption value of a voucher with more than 15% such that $y < r < x$. He will take the gamble if and only if:

$$\begin{aligned} &\omega_i\left(\frac{1}{2}\right)\lambda_i m_i(y-r) + \omega_i\left(\frac{1}{2}\right) m_i(x-r) > 0 \\ &\Leftrightarrow \lambda_i < \frac{x-r}{r-y} \end{aligned}$$

Hence, subjects decide for a gamble if and only if their loss aversion is below the threshold $\tilde{\lambda} = \frac{x-r}{r-y}$, which is independent of m_i. This insight immediately implies part (ii) of Measurement 1. Part (i) follows because the threshold $\tilde{\lambda}$ is increasing in x.

[12] See `https://github.com/microsoft/EconML` and `https://scikit-learn.org/stable/`.

A.7. One-dimensional heterogeneity analyses

Table 9: Simple heterogeneities.

Variable:	Purchase volume (in Euro) relative to baseline			
LA framing	0.0566	0.0017	-0.0006	-0.0087
	(0.035)	(0.015)	(0.014)	(0.015)
Mobile device	-0.6014***			
	(0.027)			
Mobile device	-0.073*			
× LA framing	(0.038)			
Article on sales		-0.0601		
		(0.046)		
Article on sales		-0.0244		
× LA framing		(0.063)		
Sunday			0.4111***	
			(0.046)	
Sunday			0.0240	
× LA framing			(0.066)	
Apple user				0.0930***
				(0.028)
Apple user				0.0548
× LA framing				(0.040)

Notes: We present estimated coefficients and standard errors from OLS regressions. The dependent variable represents the visitors' normalized purchase volume (in Euros). Results are displayed for regressions both without (w/o) and with (w.) the inclusion of controls. We include one-dimensional interactions to examine potential treatment heterogeneities. Specifically, we consider dummies for device type (tablet/phone vs. computer), weekday vs. weekend visits, and whether the device was an Apple product. Control variables include the web referral source, the first page viewed in the online shop, the IP region, network operator, and operating system of the device in use.

Table 9 presents an example of one-dimensional treatment heterogeneity analysis using OLS regression with treatment interaction terms. We explore potential sources of heterogeneity by considering (i) whether visitors accessed the store via a mobile device, (ii) whether the visit occurred on a weekday, and (iii) whether the device used was an Apple product. The results from these regressions consistently indicate no significant treatment interactions for the variables examined. These findings are indicative of our broader inability to identify substantial treatment heterogeneities using a one-dimensional exploratory approach.