DICE

DISCUSSION PAPER

NO 412

Human-Machine Interactions in Pricing: Evidence from Two Large-Scale Field Experiments

Tobias Huelden

Vitalijs Jascisens

Lars Roemheld

Tobias Werner

March 2024

IMPRINT

DICE DISCUSSION PAPER

Published by:
Heinrich-Heine-University Düsseldorf,
Düsseldorf Institute for Competition Economics (DICE),
Universitätsstraße 1, 40225 Düsseldorf, Germany
www.dice.hhu.de

Editor:
Prof. Dr. Hans-Theo Normann
Düsseldorf Institute for Competition Economics (DICE)
Tel +49 (0) 211-81-15125, E-Mail normann@dice.hhu.de

ISSN 2190-9938 (online) / ISBN 978-3-86304-411-4
The working papers published in the series constitute work in progress circulated to stimulate discussion and critical comments. Views expressed represent exclusively the authors' own opinions and do not necessarily reflect those of the editor.

Human-Machine Interactions in Pricing: Evidence from Two Large-Scale Field Experiments

Tobias Huelden* Vitalijs Jascisens† Lars Roemheld‡

Tobias Werner§

March 2024

Abstract

While many companies use algorithms to optimize their pricing, additional human oversight and price interventions are widespread. Human intervention can correct algorithmic flaws and introduce private information into the pricing process, but it may also be based on less sophisticated pricing strategies or suffer from behavioral biases. Using fine-grained data from one of Europe's largest e-commerce companies, we examine the impact of human intervention on the company's commercial performance in two field experiments with around 700,000 products. We show that sizeable heterogeneity exists and present evidence of interventions that harmed commercial performance and interventions that improved firm outcomes. We show that the quality of human interventions can be predicted with algorithmic tools, which allows us to exploit expert knowledge while blocking inefficient interventions.

Keywords: Artificial Intelligence, Human-Computer-Interaction, Uniform pricing

JEL-Codes: C93, D22, L2, L81

*tobias.huelden@zalando.de. Zalando SE

†vitalijs.jascisens@zalando.de. Zalando SE

‡lars.roemheld@zalando.de. Zalando SE

§Corresponding author. werner@mpib-berlin.mpg.de. Center for Humans and Machines at the Max Planck Institute for Human Development & Düsseldorf Institute for Competition Economics (DICE)

1 Introduction

Algorithmic pricing is becoming more and more prevalent in firms across all industries. Such systems optimize product-level prices using specific inputs, like demand forecasts and optimization targets. Importantly, however, these systems are typically not fully autonomous but often require human oversight and allow for "human interventions". Hence, managers or other decision-makers can often manually adjust or override the algorithms' prices.

While there has been extensive discussion of the impact of algorithmic pricing in its own right, the impact of human-machine interaction within the firm has often been neglected.[1] This provides an incomplete picture of the pricing process used by firms in practice, and thus may lead to incorrect conclusions about the origins of market outcomes when discussing the implications of algorithmic pricing. With the rapid advent of artificial intelligence (AI), this issue will become even more relevant given that even traditional industries are increasingly moving to automated-with-supervision systems.

This paper examines how humans interact with pricing algorithms at Zalando, a leading e-commerce company in Europe. At the company, algorithms price the vast majority of products. However, managers monitor the decision and can override the algorithmic decision, deeply integrating the interaction between humans and machines into the pricing process. We quantify the impact of those interventions on the firm's commercial performance using two field experiments and investigate how human-machine interactions in pricing can be improved. We find that human interventions can reduce firm profits, especially when those interventions are predicted to perform poorly from an ex-ante perspective. However, there is substantial heterogeneity in the quality of human interventions.

From an ex-ante perspective, it is ambiguous how human intervention affects firm performance. Human interventions may, on the one hand, be used to insert human expert domain knowledge not available to the algorithm into the pricing process. Additionally, they may help to prevent adverse consequences following a flawed algorithmic optimization.[2] On the other hand, humans may be overconfident about their pricing strategy or have other behavioral biases when making pricing decisions. In addition,

[1] Aparicio and Misra (2023) provide a recent survey on pricing algorithms. Furthermore, there have been discussions by policy makers on their implications, for instance, by the European Commission (2017) or The White House (2015).

[2] The involvement of humans in the algorithmic decision process is generally perceived as desirable to prevent any adverse effects algorithms can have. For instance, a group of industry experts and academics set up by the European Commission declares in their ethics guidelines for trustworthy AI that "human oversight helps ensure that an AI system does not undermine human autonomy or causes other adverse effects". See European Commission (2019) for further details.

human pricing may be overly uniform and fail to take into account product or market attributes compared to algorithmic tools. Similar to recent evidence from U.S. retail chains (see, for instance, DellaVigna and Gentzkow, 2019), managers in our setup often rely on uniform pricing strategies across markets. Those factors may worsen the firm's performance in the aggregate compared to a fully automated pricing solution. Thus, given these opposing potentials, the trade-off is whether the benefits of human insight and flexibility outweigh the risks associated with human bias and less sophisticated pricing strategies.

To investigate this trade-off, we conduct two large-scale field experiments that exogenously vary whether proposed human price interventions are realized. To assess the effects of human intervention in pricing, our first field experiment with 32,000 products in 12 countries selectively blocks selected human-suggested prices. For business risk mitigation reasons, we focus only on human price interventions that are expected to underperform relatively to other interventions following an algorithmic prediction metric.[3] We block proposed human interventions in the treatment group within this subset of products. We show that, on average, blocking human interventions significantly increases profits. Importantly, however, we find heterogeneity in outcomes: products predicted to exhibit particularly poor performance following a proposed human intervention indeed perform worse, and the adverse effects diminish as we move up the predicted quality ranking. The results suggest that human intervention can worsen the firm's commercial performance and that preventing human intervention is desirable for products where the algorithms are expected to perform particularly well.

Motivated by the heterogeneity in our findings in the first experiment, we conduct a second, more extensive field experiment with 650,000 products in eight countries. The experiment extends the analysis to a broader set of products, covering the entire assortment range offered by Zalando, thus enabling a more representative assessment of the effects of human price interventions. Furthermore, we include all products, regardless of whether they have been subject to human interventions, to align the experiment with the broader context of Zalando's business operations. This allows us to study the overall effect of human intervention on commercial outcomes. On average, blocking all human interventions independently of their expected quality reduces revenues and has no significant effect on profits. However, in this second experiment, we also find evidence that interventions expected to perform well did indeed increase profits. Similar to the initial field experiment, blocking human interventions expected to underperform increases profits, but the effect is not statistically significant, likely due to lower statistical power in this sub-sample.

[3] The metric employs the predicted profit margin difference between the proposed algorithmic and human price. This risk assessment is based on a measure of pricing efficiency frequently used in the company. Notably, it has not previously been applied to evaluate the performance of human pricing decisions.

Our results highlight the importance of human-machine interaction in pricing. Human interventions can help improve algorithmic pricing solutions, highlighting that human expertise is indeed valuable. Notably, however, we detect substantial heterogeneity in the quality of human interventions. The ones that are expected to perform poorly reduce profits, while interventions expected to be relatively efficient have a substantial positive effect on profits. This suggests that human-machine interactions can be further refined and firm performance improved by blocking specific interventions. Implementing a system that guardrails and optimizes the balance between human knowledge and algorithmic sophistication could help maximize overall profitability. Our results suggest that the predicted profit margin difference between algorithmic and proposed human prices can be valuable for this purpose.

Related Literature Our paper concerns the use of algorithmic pricing tools in online markets. These tools can leverage vast amounts of data and often use complex automation routines to find a price that is optimal given the firm's objective. There exists widespread evidence that these tools are becoming more prevalent (see, for example, Baker et al., 2014; Chen et al., 2016; Aparicio and Misra, 2023). Our paper contributes to this literature by documenting the use of algorithmic pricing tools with data from a large e-commerce company. Furthermore, our research shows that pricing algorithms often do not act completely autonomously, but that there is an interplay between human and algorithmic decision makers.

Aparicio et al. (2021) highlight how algorithmic pricing tools have led to increased price discrimination over time, location, and among different sellers. In contrast, offline markets are often characterized by uniform pricing across markets (see, for instance, DellaVigna and Gentzkow, 2019; Adams and Williams, 2019; Hitsch et al., 2021). Our research is closely related to this branch of the literature, as Zalando's pricing algorithm can choose different prices across time and countries.[4] However, human pricing decisions in our context often follow a uniform pricing strategy. This pricing behavior results from the organizational structure in which we study human-machine interaction, as well as the sheer scale of Zalando's pricing systems. In doing so, we can relate to recent discussions of the relationship between the organizational structure of firms and how it relates to seemingly behavioral biases of firms (see, for instance, Hortaçsu et al., 2023; Cho and Rust, 2010; Huang, 2022). Our setup allows us to explore the interplay between the uniform pricing strategy of humans and algorithmic pricing, offering novel insights into the complexities of their interactions in pricing decisions within a modern e-commerce setting.

Our work centers around the interaction between humans and algorithms. It has

[4] When we refer to targeted pricing in the context of Zalando, we are referring to price differences across time or country. Zalando does not engage in personalized pricing.

been studied in abstract experimental scenarios (Crandall et al., 2018; Kasberger et al., 2023) as well as more applied scenarios. Often, the focus is on algorithms providing a recommendation to human decision markers and evaluating its effect on outcomes. In those scenarios, the human decision-maker usually receives some advice from prediction algorithms and can decide to follow it or overrule it. For instance, algorithmic advice has been studied in the context of medical (Agarwal et al., 2023; Tschandl et al., 2020) or legal decision-making (Kleinberg et al., 2018; Angelova et al., 2023), as well as online news curation (Peukert et al., 2023). We extend this research by considering a novel environment by focusing on pricing. Furthermore, we consider a setting where algorithms do not provide recommendations but are more deeply embedded in the system than in previous studies.

Few papers consider the interaction between humans and algorithms in pricing directly. Garcia et al. (2023) consider algorithmic price recommendations for hotel room pricing. Based on observational data, they focus on the aspect that human decisions can be slow and costly. They show that while managers may possess or perceive themselves to have private information, it does not necessarily lead to superior pricing decisions than the recommendation. Karlinsky-Shichor and Netzer (2023) discuss the interaction between pricing algorithms and managers in a business-to-business setting. They build an automated version of salespeople based on their past pricing behavior and show that recommendations from this algorithm increase profits in a field experiment. Our study extends this research on human-AI interactions in pricing. To the best of our knowledge, our study is the first to conduct a fully randomized experiment within the existing business operations of a firm, focusing on this interaction and the implications for firm outcomes. It allows for robust identification while providing a high degree of external validity. Furthermore, unlike scenarios where algorithms merely provide recommendations, algorithms actively set prices by default in our setup, with human decision-makers retaining the power to override these algorithmic decisions. It allows us to uniquely identify and quantify the benefits and limitations of human and algorithmic pricing in a different scenario than studied before.

The remainder of this paper is organized as follows. In Section 2, we introduce the institutional setting that we consider in this paper. Then, in Section 3, we discuss the designs of the two field experiments and present the main results. Section 4 concludes and discusses the implications of our findings.

2 Algorithmic and Human Pricing at Zalando

Zalando is one of Europe's leading e-commerce companies, focusing on fashion retail. The publicly traded company serves more than 50 million customers annually. In

2022, the company generated €15bn in gross merchandise volume, with more than 260 million orders in 25 countries.[5]

At Zalando, human-machine interactions in pricing are of significant commercial relevance. By default, algorithms price products, but humans can intervene and overwrite algorithmic decisions. In the time period studied, these interventions accounted for about 7% of all daily price observations and accounted for about 15% of revenues and 7% of profits. Hence, understanding the mechanisms and effects of these human pricing interventions is essential for the firm. It can help generate further business value by streamlining and rationalizing current pricing processes and inform the design and implementation of new pricing algorithms and systems.

In the following, we introduce the institutional and commercial setting in which we study the interaction between humans and machines in pricing. We start by explaining the general algorithmic pricing process at Zalando. We then outline how human pricing is conducted and how humans can intervene in the algorithmic processes and potentially override algorithmic decisions. Finally, we discuss the potential implications of the interplay between human-machine pricing and our conjecture regarding the overall effectiveness of this interaction.

2.1 Algorithmic Pricing

Every product enters Zalando at the beginning of its lifecycle at a non-discounted price. At some point, a re-pricing process is initiated that determines a discount level between 0% and 70% for the product for the coming weeks. The discount adjustment is performed by a pricing algorithm that aims to maximize profits given a revenue target. This optimization takes place continuously to account for market environment changes.

It is important to highlight that the data landscape that Zalando operates in, typical but not exclusive to e-commerce firms, presents unique scientific challenges for algorithmic pricing due to sparse data on pricing and sales. While many products are offered online, only a much smaller subset contribute substantial sales in any given time period. Data such as information about prices and sales, and thus also demand forecasts, are sparse in the product cross-section. Moreover, historical time series are relatively short due to the seasonality and fast dynamics of the e-commerce and fashion business.

The weekly price optimization process starts with a demand forecast, which is

[5]Zalando also operates an off-price "shopping club" and a marketplace business. This paper focuses on the retail business of the main online fashion store, where Zalando sells its own inventory. See Zalando (2022) for more details on the company's financial performance.

explained in detail by Kunz et al. (2023).[6] The forecaster predicts weekly demand for each product within a specific country, at 15 possible discount levels.

The demand forecast is the primary input to the (inventory-constrained) price optimization algorithm. This algorithm is described by Li et al. (2021). For the given forecast, it uses a mixed-integer optimization approach to derive for each product a discount level that is expected to maximize the firm's profits.

The price optimization algorithm optimizes the assortment with respect to certain constraints. Some constraints represent business preferences and commercial strategy; others constitute "safety guardrails" designed to prevent the algorithm from making sudden or drastic changes. Such guardrails include a maximum week-on-week price change or a maximum discount rate, both of which reduce the likelihood of extrapolation bias or accidental over-optimization using high-uncertainty forecasts.

The price optimizer yields a discount level for each product for a given week in a given country. After the algorithmic price is computed, human decision-makers observe it and may decide to override it. Our study focuses on these interventions, and we explain the decision process behind human overrides in the sections below.

2.2 Human Interventions

Price interventions are proposed by category managers who oversee a smaller part of the assortment. After category managers propose new discounts, pricing managers review and approve these changes at the country level. Pricing managers do not consider the exact pricing decisions of the category managers but only control whether the implied overall discount level is compatible with the country-specific business constraints. Furthermore, there are no approval decisions for specific products, but groups of products are approved or rejected in batches. These groups can contain thousands of products and are usually grouped by some proximity in the product space.[7] In the following, we will use the term "human price intervention" to refer to the actual discount decision made by category managers and not the approval by pricing managers.

The human repricing process is performed regularly at Zalando. Typically, human price interventions occur once a week, but more frequent ad-hoc interventions are possible. While it is not encouraged, humans could change prices on a daily basis. Any possible discount change by humans is weakly decreasing the price.

Category managers, who propose all human pricing interventions and are the critical decision-makers, are specialized individuals with a deep understanding of the fash-

[6]While the main forecaster treats the problem of finding an appropriate price as a prediction problem, there have been recent efforts to introduce notions of causality into the process at Zalando. For details, see Schultz et al. (2023).

[7]The average product category contains approximately 39,000 products. Examples of product categories include "Men's Tall Boots" and "Women's Textile Dresses".

ion industry. They focus on specific categories in the fashion industry, which they monitor closely. For example, a category manager may only be responsible for men's outdoor sportswear. It allows them to identify emerging trends and developments. As a result, they have a good understanding of the latest shifts in consumer trends and styles.[8] Often, this type of soft information about fashion trends is complex to encode into data that pricing algorithms can use. In addition, fashion trends can evolve quickly and, therefore, cannot be accurately tracked by the algorithm. Human price interventions could be a way to bring this expertise and information, which the algorithm may not be able to observe, into the pricing process and thereby improve it.[9] In addition, they do not price every product but rather cherry-pick products where they see sufficient room to improve on the already optimized algorithmic price. As a result, human intervention could improve pricing decisions and increase the firm's profits compared to algorithmic pricing.

Notably, category managers always propose a uniform discount for a specific product across all countries in which Zalando operates. Thus, human discounts do not factor in preferences or other demand differences across markets compared to algorithmic pricing. While being an obvious shortcoming of human pricing, this reduces the operational complexity given the large number of markets and products that category managers have to manage. This uniform pricing strategy by category managers is similar to patterns observed for retail chains in the U.S. (see DellaVigna and Gentzkow, 2019).[10] Category managers may underestimate the impact of their pricing strategy across countries. Humans may focus on the effectiveness of their pricing decisions in a particular market while neglecting the overall impact of that intervention. As a result of the uniform pricing strategy, human interventions could lead to poorer company performance and lower profits compared to algorithmic pricing.

Furthermore, while the ultimate goal of category managers is to maximize the company's profits, other intermediate factors may drive overriding decisions that potentially lead to inefficiencies. It is well documented that humans can exhibit algorithmic aversion or distrust of algorithms (see, for instance, Dietvorst et al., 2015). Category managers may distrust algorithms to make the right pricing decision for a given prod-

[8] Supply-side considerations often drive the decision to price a product at a higher discount. For example, new fashion trends not captured by the pricing algorithm may cause certain products to perform worse than expected. Because category managers are domain experts, they can respond to specific changes in demand and try to correct possible overstocking problems by proposing a different discount. We confirmed this through qualitative interviews with stakeholders and category managers at Zalando.

[9] As highlighted by Angelova et al. (2023), who study algorithmic support for bail decisions, humans may overwrite the algorithms to induce their private expert knowledge into the process. However, it may also reintroduce human mistakes and biases.

[10] Note that uniform pricing refers only to the category manager's decision. Algorithmic pricing decisions are made at the country level.

uct, for example due to a lack of algorithmic transparency.[11] While this assessment could be correct if the category manager has private information, it is also possible that overconfidence is driving this decision.[12]

A-priori, it is unclear how human intervention affects profits relative to algorithmic pricing. On the one hand, if human interventions are mostly driven by overconfidence or limited foresight concerning their uniform pricing approach, they may worsen firm outcomes. On the other hand, humans could add expert knowledge to the pricing process and thereby help improve the pricing system. In the following section, we quantify the effect of human pricing interventions on algorithmic prices using two large-scale field experiments. The results will help to understand which effect dominates and whether there is scope for further improvement in human-machine interaction in pricing.

3 Experimental Evidence

This section describes the experimental design and presents the results of the field experiments. We first present the design and analysis of the initial experiment. It studies human interventions that are predicted to be particularly inefficient. Then, we discuss the large-scale field experiment where we mimic the "business-as-usual" scenario as closely as possible.

3.1 First Field Experiment: Low-Risk Products

3.1.1 Sample and Experimental Design

The first experiment serves as a pilot for the second large-scale field experiment. To mitigate business risk in the first field experiment, we focused on products where human interventions are predicted to perform particularly poorly. This prediction ranks product-level interventions by the expected profit margin difference between the algorithmic and human price. We then use this difference to rank the human interventions.

The intuition behind using the predicted margin difference for commercial risk mitigation is that products predicted to have a comparatively smaller margin difference are unlikely to be priced better by a human. Any additional information that humans might add to the process by adjusting prices will unlikely improve the margin enough

[11] Qualitative interviews with category managers confirm that this is indeed the case at Zalando.

[12] Managers overestimating their own abilities is well documented for CEOs (see, for example, Malmendier and Tate, 2005, 2015). The possible overconfidence we are referring to concerns the lower level category managers for very specific pricing decisions. For a general discussion on overconfidence in psychology see Moore and Healy (2008).

to be better than under the algorithmic pricing regime. They are more likely to misperceive the correct price or not sufficiently account for the impact of their uniform pricing strategy across countries. Consequently, it was deemed less risky to conduct an initial experiment using only human interventions ranked lowest by predicted margin difference.

The resulting sample for the initial field experiment consists of the bottom 50% of human interventions by the predicted margin difference. Furthermore, the set is restricted to the 12 countries selected by the pricing managers to participate in the field experiment. We are left with 32,897 unique products and 124,923 product-country pairs.[13] We also refer to these product-country pairs as "articles" in the following.

We use a clustered-stratified experiment design to allocate articles into treatment and control groups, reducing concerns that substitution effects lead to interference bias.[14] To achieve this, we first sort articles into higher-level fashion clusters. These in-house clusters were created using domain knowledge to minimize the substitution across clusters. In the initial experiment, the clustering combines the in-house cluster and the country. Next, we calculate cluster-level average revenue using the data two weeks before the experiment and allocate clusters into strata of size 32 based on the calculated average revenue. Finally, we assign 16 clusters to the treatment group and 16 to the control group within each stratum. In this way, we allocate the 124,923 articles into 5,593 clusters distributed across 174 strata.

In the control group, products were priced through the usual process: After the algorithm decided on a price, possible proposed price changes from category managers were applied. In the treatment group, any proposed human price intervention was blocked. Importantly, the category managers proposing an additional price change were unaware of the experiment, so that they could not adjust their pricing behavior due to the treatment. The experiment lasted for 15 days.

3.1.2 Results

We analyze both experiments in a difference-in-differences framework, accounting for the clustered and stratified randomization.[15] To estimate the causal effect of human price interventions on economic outcomes, we use the following specification:

[13]Note that not all products are available in every country. As a result, the number of product-country pairs is less than the number of countries times the number of unique products.

[14]Randomization at the article level is common in e-commerce because treatment assignment at the customer level would lead to personalized price discrimination, which is considered unethical by most e-commerce companies (see, for example, Cooprider and Nassiri, 2023).

[15]Note that pre-treatment outcomes in the treatment and control groups are balanced (see Table A.1 in Appendix A). We use the difference-in-differences specification as a variance reduction technique. Results coming from a simple differences specification are qualitatively similar (albeit with more variance) and are presented in Appendix A.

$$Y_{it} = \beta_0 + \beta_1 D_i + \beta_2 1\{t \geq T_0\} + \beta_3 D_i 1\{t \geq T_0\} + \epsilon_{it} \tag{3.1}$$

where Y_{it} denotes the outcome for article i on day t, D_i equals one for treated articles and zero otherwise, and the intervention starts at time T_0.

Since the clusters used for randomization can vary in size substantially, an equal number of treated and control clusters within each stratum does not translate into evenly sized groups at the article level. As a result, the treatment probability at the article level varies across blocks. To overcome this challenge, we use inverse probability weighting in all regressions. This ensures that β_3 can be interpreted as an average treatment effect (ATE).

Notice that in the difference-in-differences specification, $ATE = \beta_3$ and the counterfactual is given by $\beta_0 + \beta_1 + \beta_2$. To obtain the relative average treatment effect, we use the following expression:

$$\text{ATE}_{rel} = \frac{\beta_3}{|\beta_0 + \beta_1 + \beta_2|} \tag{3.2}$$

In the following sections, we will focus on the sample analogue of ATE_{rel} and do inference using the delta method.

Table 1 presents estimated absolute ATEs for log-prices and relative ATEs for all other outcome variables (as they may be 0 or negative). We conclude that preventing human pricing interventions that are anticipated to perform particularly poorly generated a 30% increase in prices, which led to a drop in sales and revenue of 65% and 60%, respectively. Furthermore, it increased profit by 237%.

Next, we examine the time evolution of obtained treatment effects. Towards this end, we use the following specification:

$$Y_{it} = \beta_0 + \beta_1 D_i + \sum_{\substack{k=-14 \\ k \neq -1}}^{14} \gamma_k 1\{t = k\} + \sum_{\substack{k=-14 \\ k \neq -1}}^{14} \xi_k D_i 1\{t = k\} + \epsilon_{it} \tag{3.3}$$

Figure 2 displays estimated $\hat{\xi}_k$'s. As expected, we do not see significant differences between treatment and control groups before the treatment.[16] Once the treatment is switched on, we see large changes in prices, sales, revenue, and profits.

We take a closer look at the treatment effect on price dispersion to highlight the potential mechanisms behind these observed effects. Figure 1 shows the distribution of the standard deviation of prices (in terms of chosen discount) for each product across all countries before and after the start of the experiment. Before the start of the exper-

[16]Implicitly, the absence of effects before the start of the experiment supports the parallel trends assumption.

iment, the standard deviation is virtually the same across the treatment and control group. After the experiment begins, we see that the standard deviation in the control group is smaller, and the spread of the deviation becomes narrower. It provides clear evidence that the treatment induces more price variation across countries. Notably, this aligns with our ex-ante expectations since we are blocking human interventions that rely on uniform pricing in the treatment group. The exercise highlights one of the mechanisms discussed in Section 2 by which human pricing may not be sufficiently targeted to specific markets compared to pricing algorithms.

Finally, we examine the treatment effect heterogeneity with respect to the predicted quality of human price interventions. For this, we again use the ranking of the predicted profit margin difference before the experiment. We use the ranking to divide all product-country pairs into two groups: above and below median quality. We expect articles below the median to perform worse than those above the median.[17]

To estimate the treatment effect heterogeneity we use the following specification:

$$\begin{aligned} Y_{it} = \beta_0 + \beta_1 D_i + \beta_2 1\{t \geq T_0\} + \beta_3 Above_i + \\ \beta_4 D_i 1\{t \geq T_0\} + \beta_5 D_i Above_i + \\ \beta_6 Above_i 1\{t \geq T_0\} + \beta_7 D_i Above_i 1\{t \geq T_0\} + \epsilon_{it} \end{aligned} \tag{3.4}$$

where all notation corresponds to previous specifications and $Above_i$, equals one for articles where human discounts were predicted to generate above median performance and zero otherwise. Notice that β_4 and $\beta_4 + \beta_7$ correspond to ATEs in the $Above = 0$ and $Above = 1$ categories. We denote these ATEs by ATE_B and ATE_A. The difference in ATEs between $Above = 1$ and $Above = 0$ categories is represented by $\beta_7 = \text{ATE}_A - \text{ATE}_B$. Table 2 presents sample counterparts of $ATE_{B,rel} = \frac{\beta_4}{|\beta_0+\beta_1+\beta_2|}$, $ATE_{A,rel} = \frac{\beta_4+\beta_7}{|\beta_0+\beta_1+\beta_2+\beta_3+\beta_5+\beta_6|}$ and of $\frac{ATE_A - ATE_B}{|ATE_B|} = \frac{\beta_7}{|\beta_4|}$.

We find that the effect of human pricing interventions on revenue is not statistically significantly different between the above and below-median quality interventions. Regarding profits, the positive treatment effect of lifting human interventions is 58% smaller for articles with an above-median intervention quality.[18] Those results offer two interesting insights: First, they highlight that there are indeed quality differences between human interventions, and some are less harmful than others. Secondly, it suggests that we can predict the ex-post effectiveness of a human intervention using

[17] Note that concerning all human interventions, products below the median in this heterogeneity analysis are part of the bottom 25 % of all products. It follows from the sample restriction to the bottom 50% products for this initial experiment.

[18] It is important to exercise caution when comparing the relative average treatment effects, $ATE_{B,rel}$ and $ATE_{A,rel}$. This is the only instance in two experiments when counterfactuals were of the opposite signs. Therefore, we obtain a somewhat counterintuitive result: $ATE_{B,rel}$ is smaller than $ATE_{A,rel}$, but ATE_B is larger than ATE_A.

the ex-ante prediction of the profit margin difference between the human and the algorithmic price. We will further investigate those aspects in the following section.

In our experimental design, we did not stratify on the $Above_i$ dummy. To understand whether previously presented heterogeneity results are not driven by time-varying differences between the above and below categories, we utilize the following specification:

$$\begin{aligned} Y_{it} = \beta_0 + \beta_1 D_i + \beta_2 Above_i + \beta_3 D_i Above_i + \\ \sum_{\substack{k=-14 \\ k \neq -1}}^{14} \gamma_k 1\{t = k\} + \sum_{\substack{k=-14 \\ k \neq -1}}^{14} \xi_k D_i 1\{t = k\} + \\ \sum_{\substack{k=-14 \\ k \neq -1}}^{14} \mu_k Above_i 1\{t = k\} + \\ \sum_{\substack{k=-14 \\ k \neq -1}}^{14} \lambda_k D_i Above_i 1\{t = k\} + \epsilon_{it} \end{aligned} \tag{3.5}$$

Notice that in this specification:

$$\lambda_k = ATE(k)_{Above} - ATE(k)_{Below} \tag{3.6}$$

in words, λ_k equals the difference between average treatment effects in the above and below category at time k. Figure 3 displays the evolution of this difference for all outcomes of interest. We conclude that treatment effects for prices, sales, and revenue do not vary between the two groups. On the other hand, for profit, we see stark treatment effect heterogeneity appearing at the beginning of the experiment. It further supports our findings from Table 2.

We next turn our attention to the second, large-scale field experiment.

3.2 Second Field Experiment: Whole Assortment

The results of the initial field experiment indicate that algorithms outperform human price interventions. Moreover, we show treatment effect heterogeneity with respect to the quality of price interventions. Crucially, the initial experiment restricted the set of products to those interventions that are predicted to underperform, given our algorithmic forecast.

To understand whether findings from the previous experiment generalize to the whole assortment of Zalando, we decided to run a large-scale field experiment that tests blocking all human interventions in the "business-as-usual" process. In this setup, human discount allocation follows the process described in Section 2. First, category managers propose interventions, and then pricing managers choose to either upload those discounts or not. In what follows, we label the sub-sample where human discounts were uploaded as articles eligible for human discounts. Notably, at the time of the experimental design, we did not know which discounts managers would upload.

3.2.1 Sample and Experimental Design

To mimic this business reality as closely as possible, we randomized around five million product-country pairs (in eight countries participating in the second experiment) into treatment and control. Similarly, as before, we use in-house clusters and clustered randomization. Importantly, this time, we do randomization country by country. Because the randomization was done country by country, the treatment probability slightly varies across the countries. As in the first experiment, we use inverse probability weights to account for this imbalance. In the treatment group, any proposed human price intervention was blocked. The experiment lasted for 13 days. Human interventions were proposed and approved for around 7% of product-country pairs in the entire sample. In contrast to the initial experiment, we do not account for the predicted quality of an intervention when blocking it in the treatment group. We thus measure the treatment effect of cutting all human pricing interventions instead of cutting only the low-quality interventions in the initial experiment.

3.2.2 Results

In the following, we present the results for the whole sample and the sub-sample of articles that received a human intervention. Tables A.3 and A.4 in the Appendix show that in both samples, there are no statistically significant differences between treated articles and control articles at the baseline. Furthermore, similar to the initial field experiment, we find that the treatment caused an increase in price variation on the product level across countries compared to the control group (see Figure A.1 in Appendix A).

To show the main effect of the treatment on the firms' outcomes, we again use the specification 3.1 as in the analysis of the initial field experiment. Tables 3 and 4 show sample counterparts of ATE and ATE_{rel} obtained for both samples. In the whole sample, we find small negative effects on prices and sales and no statistically significant effects on revenue and profit. In the sub-sample of articles with human interventions, we find that the experiment caused the price to increase by around 14%.[19] This increase

[19]The treatment of blocking human interventions increased prices; we believe that the very small but significant negative effect on prices for the overall assortment was a function of the pricing algorithm balancing overall price level in the treatment group.

led to a decrease in sales and revenue of 39% and 38%, respectively. Estimated ATE_{rel} for profit is negative but not statistically different from zero at conventional significance levels. Hence, we find no evidence that blocking all human price interventions, rather than just low-quality ones as in the initial experiment, enhances profitability.

We proceed by showing event study plots for both samples. We again use specification 3.3 to accomplish this. Event study plots presented in Figures 4 and 5 reconfirm the findings from Tables 3 and 4. We do not see differences between treatment and control groups before the experiment in both samples. Even after the start of the experiment we do not see substantial changes in the whole sample. On the other hand, in the subset of articles that received a human intervention, we see a sharp reaction to prices, sales, and revenue and slightly underpowered effects on profit, likely due to the smaller number of observations in this sub-sample.

In the following, we investigate the heterogeneity in human intervention quality. While the treatment effects on profits in the whole sample are not statistically significant at conventional levels, we suspect heterogeneity drives this. We again leverage variations in the predicted quality of human price interventions as in the initial field experiments. We use the ranking of quality of human interventions discussed in Section 3.1 to group all articles with a human intervention into two groups: above and below median quality interventions. Hence, we investigated the treatment effect heterogeneity below and above the median intervention quality among all approved human pricing interventions.[20]

We run specifications 3.4 and 3.5 on the sub-sample of articles with a human intervention. Results presented in Table 5 and Figure 6 provide a holistic picture of the treatment effect heterogeneity. We find that there is no treatment effect heterogeneity in price but large negative effects on sales, revenue and profits. More precisely, the treatment effects of blocking human pricing interventions on sales and revenues in the above median quality category are around 170% and 260% lower as compared to below median quality interventions category. The effect on profit is around $\widehat{ATE}_{A,rel} - \widehat{ATE}_{B,rel} = 43$ percentage-points lower in the above median category.[21] Moreover, we find that blocking human interventions that are above the median actually reduces profits.

[20]Note that in the initial experiment, our analysis focused on heterogeneity across interventions ranked below and above the 25% quality quantile. In this experiment, the median split refers to the actual median across the entire population of interventions, offering a broader understanding of the heterogeneity. Hence, articles below the median in this second experiment would be those eligible for the entire sample in the initial experiment.

[21]The relative treatment effect for profit in the below-median category is not statistically different from zero at conventional significance levels. Therefore, the previously reported difference in percent is not very meaningful. To have certain consistency throughout the paper we decided not to report statistical differences between $ATE_{A,rel}$ and $ATE_{B,rel}$.

Together with the findings from the initial experiment, the results indicate substantial heterogeneity in the quality and impact of human pricing interventions. While some human interventions reduce profits, others benefit the firms' performance. Our findings align with the postulated trade-off that humans have private information but rely on uniform pricing strategies and suggest that completely blocking or accepting all human interventions may not be the most profitable approach for the company. The interplay between humans and pricing algorithms is essential to obtain the best outcomes, and as such, it is vital to optimize it further. Our findings indicate a clear possible way to improve the interaction in pricing. The predicted profit margin differences between the algorithmic and human prices allow us to forecast the quality of human intervention. Our field experiments show that this metric can be used to block inefficient human interventions while leveraging those that improve the firms' performance. Notably, the metric has the advantage of building on existing tools at Zalando and does not induce further engineering work. Furthermore, its design is simple and universal, making our approach applicable and transferable to other companies and situations.

4 Discussion and Concluding Remarks

While algorithmic pricing is widely used in many industries, humans often oversee the automated pricing process and can override algorithmically selected prices. In many cases, humans are domain experts and can bring their knowledge to the pricing process. However, humans may also be less sophisticated in their pricing strategies than algorithms. This creates a trade-off, and it is unclear whether human intervention increases or decreases firm profits.

We investigate this trade-off in two field experiments at Zalando, one of Europe's largest e-commerce companies. In the first experiment, we block proposed human price interventions, focusing on a subset of the assortment for which we expect human interventions to be particularly inefficient. We utilize algorithmic forecasts to derive this efficiency measure and confirm that blocking the expected-damaging human intervention increases profits. In the second experiment, we remove those sample restrictions. Here, we find that, on average, blocking human interventions, regardless of their expected quality, reduces revenues and has no significant effect on profits. However, we find significant heterogeneity similar to the first experiment: Human interventions predicted to perform poorly reduce profits, while interventions that our algorithmic tools expect to perform well increase profits.

Importantly, our results indicate strong heterogeneity in the quality of human interventions and suggest that the effect of interventions is broadly predictable using algorithmic heuristics: that is, it is possible to *a priori* predict which manual pricing interventions will perform poorly, and which ones will be profitable—our heterogeneity results suggest that (conditional) ATEs are in line with the predicted expectation.

Overall, the findings indicate the superiority of a well calibrated human-machine collaboration in pricing over one dimensional pricing strategies. Completely unguarded human pricing interventions as well as fully blocked interventions are not beneficial for firm outcomes. However, the careful identification and implementation of high quality human pricing interventions can improve algorithmic pricing outcomes.

This highlights an important mechanism for human-machine collaboration in pricing: typical industry pricing systems implement safety guardrails to avoid algorithmic failure. When human domain experts override the algorithmic price, they may also violate such guardrails. We believe that this mechanism explains the heterogeneity we found: Human pricing interventions can suffer from overconfidence and especially from uniformity across markets, since human decision-making cannot optimize at the article level considering the scale of the optimization problem. Our algorithmic predictions identify such cases, and can prevent damaging interventions. However, we also show some domain expert interventions to be highly profitable. In such cases, the algorithmic prediction confirms the human intervention to be a good idea, and the collaboration between humans and algorithms provides the safety required to violate standard system guardrails.

Interactions between humans and machines are widespread, and they allow for improvements whenever carefully implemented: When algorithmic optimization underlies binding constraints designed to prevent damage from extrapolation bias or excessive uncertainty, human domain experts can instill confidence and private knowledge into the overall process, yielding profit improvements even to mature algorithmic systems. It suggests a framework for optimizing human-computer collaboration: algorithms should provide strong, reliable, and safe baseline solutions within well-defined guardrails. Aided by decision-support systems, humans can then improve on this baseline by selectively going beyond the guardrail constraints.

Further work is needed to uncover the mechanisms behind the heterogeneity in the quality of human interventions. Qualitative interviews with managers indicate that it is, in fact, algorithmic (in)transparency that can drive human algorithmic distrust and, thus, interventions. More work is needed to quantify the relevance of this mechanism. It would help to improve the guardrail approach we present in this paper further.

References

Adams, Brian and Kevin R Williams, "Zone pricing in retail oligopoly," *American Economic Journal: Microeconomics*, 2019, *11* (1), 124–156.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz, "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology," Working Paper, National Bureau of Economic Research 2023.

Angelova, Victoria, Will S Dobbie, and Crystal Yang, "Algorithmic recommendations and human discretion," Working Paper, National Bureau of Economic Research 2023.

Aparicio, Diego and Kanishka Misra, *Artificial intelligence and pricing*, Vol. 20, Emerald Publishing Limited, 2023.

_, Zachary Metzman, and Roberto Rigobon, "The pricing strategies of online grocery retailers," Working Paper, National Bureau of Economic Research 2021.

Baker, Walter, Dieter Kiewell, and Georg Winkler, "Using Big Data to Make Better Pricing Decisions," McKinsey & Company 2014. Available at: https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/using-big-data-to-make-better-pricing-decisions Accessed: 2023-02-21.

Chen, Le, Alan Mislove, and Christo Wilson, "An empirical analysis of algorithmic pricing on amazon marketplace," in "Proceedings of the 25th international conference on World Wide Web" 2016, pp. 1339–1349.

Cho, Sungjin and John Rust, "The flat rental puzzle," *The Review of Economic Studies*, 2010, *77* (2), 560–594.

Cooprider, Joe and Shima Nassiri, "Science of price experimentation at Amazon," *Business Economics*, 2023, *58* (1), 34–41.

Crandall, Jacob W, Mayada Oudah, Tennom, Fatimah Ishowo-Oloko, Sherief Abdallah, Jean-François Bonnefon, Manuel Cebrian, Azim Shariff, Michael A Goodrich, and Iyad Rahwan, "Cooperating with machines," *Nature communications*, 2018, *9* (1), 233.

DellaVigna, Stefano and Matthew Gentzkow, "Uniform pricing in us retail chains," *The Quarterly Journal of Economics*, 2019, *134* (4), 2011–2084.

Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey, "Algorithm aversion: people erroneously avoid algorithms after seeing them err.," *Journal of Experimental Psychology: General*, 2015, *144* (1), 114.

European Commission, "Final Report on the E-commerce Sector Inquiry," Technical Report, European Commission 2017.

__ , "Ethics guidelines for trustworthy AI," 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai Accessed: 2023-02-21.

Garcia, Daniel, Juha Tolvanen, and Alexander K Wagner, "Strategic Responses to Algorithmic Recommendations: Evidence from Hotel Pricing," Working Paper 2023.

Hitsch, Günter J, Ali Hortacsu, and Xiliang Lin, "Prices and promotions in US retail markets," *Quantitative Marketing and Economics*, 2021, pp. 1–80.

Hortaçsu, Ali, Olivia R Natan, Hayden Parsley, Timothy Schwieg, and Kevin R Williams, "Organizational Structure and Pricing: Evidence from a Large U.S. Airline," *The Quarterly Journal of Economics*, 2023.

Huang, Yufeng, "Pricing frictions and platform remedies: the case of Airbnb," *Available at SSRN 3767103*, 2022.

Karlinsky-Shichor, Yael and Oded Netzer, "Automating the b2b salesperson pricing decisions: A human-machine hybrid approach," *Marketing Science*, 2023.

Kasberger, Bernhard, Simon Martin, Hans-Theo Normann, and Tobias Werner, "Algorithmic Cooperation," *Available at SSRN 4389647*, 2023.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human decisions and machine predictions," *The Quarterly Journal of Economics*, 2018, *133* (1), 237–293.

Kunz, Manuel, Stefan Birr, Mones Raslan, Lei Ma, Zhen Li, Adele Gouttes, Mateusz Koren, Tofigh Naghibi, Johannes Stephan, Mariia Bulycheva et al., "Deep Learning based Forecasting: a case study from the online fashion industry," *arXiv preprint arXiv:2305.14406*, 2023.

Li, Hanwei, David Simchi-Levi, Rui Sun, Michelle Xiao Wu, Vladimir Fux, Torsten Gellert, Thorsten Greiner, and Andrea Taverna, "Large-scale price optimization for an online fashion retailer," in "Innovative Technology at the Interface of Finance and Operations: Volume II," Springer, 2021, pp. 191–224.

Malmendier, Ulrike and Geoffrey Tate, "CEO overconfidence and corporate investment," *The Journal of Finance*, 2005, *60* (6), 2661–2700.

_ and _ , "Behavioral CEOs: The role of managerial overconfidence," *Journal of Economic Perspectives*, 2015, *29* (4), 37–60.

Moore, Don A and Paul J Healy, "The trouble with overconfidence.," *Psychological review*, 2008, *115* (2), 502.

Peukert, Christian, Ananya Sen, and Jörg Claussen, "The editor and the algorithm: Recommendation technology in online news," *Management science*, 2023.

Schultz, Douglas, Johannes Stephan, Julian Sieber, Trudie Yeh, Manuel Kunz, Patrick Doupe, and Tim Januschowski, "Causal Forecasting for Pricing," *arXiv preprint arXiv:2312.15282*, 2023.

The White House, "Big Data and Differential Pricing," Technical Report, Council of Economic Advisors, Washington, D.C. 2015. Available at: https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf Accessed: 2023-02-21.

Tschandl, Philipp, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy et al., "Human–computer collaboration for skin cancer recognition," *Nature Medicine*, 2020, *26* (8), 1229–1234.

Zalando, "Annual Report 2022," 2022. Available at: https://corporate.zalando.com/en/investor-relations/annual-report-2022 Accessed: 2023-02-21.

Tables

Table 1: Estimates of ATE, Difference in Differences Estimator: First Experiment

	Log Price	Sales	Revenue	Profit
$\widehat{ATE}$	0.305			
	(0.004)			
$\widehat{ATE}_{rel}$		-0.652	-0.603	2.365
		(0.018)	(0.019)	(0.298)
N	3,622,383	3,622,383	3,622,383	3,622,383

Notes: This table presents estimates of ATE for various outcomes of interest. To obtain estimates of ATE we use difference-in-differences estimator. $\widehat{ATE}_{rel}$ denotes an estimate of the relative treatment effect. For relative quantities inference is done using delta method. Standard errors in parentheses, robust and clustered at the cluster level.

Table 2: Heterogeneity Analysis, Difference in Differences Estimator: First Experiment

	Log Price	Sales	Revenue	Profit
$\widehat{ATE}_B$	0.311			
	(0.004)			
$\widehat{ATE}_A$	0.300			
	(0.004)			
$\widehat{ATE}_A - \widehat{ATE}_B$	-0.011			
	(0.003)			
$\widehat{ATE}_{B,rel}$		-0.668	-0.606	1.441
		(0.020)	(0.024)	(0.068)
$\widehat{ATE}_{A,rel}$		-0.635	-0.600	4.771
		(0.019)	(0.020)	(3.766)
$\frac{\widehat{ATE}_A - \widehat{ATE}_B}{\lvert\widehat{ATE}_B\rvert}$		0.099	-0.097	-0.575
		(0.048)	(0.076)	(0.055)
N	3,622,383	3,622,383	3,622,383	3,622,383

Notes: This table presents heterogeneity analysis. To obtain estimates of ATE we use difference in differences estimator. $\widehat{ATE}_A$ and $\widehat{ATE}_B$ denote estimated ATEs in the above and below categories. $\widehat{ATE}_{A,rel}$ and $\widehat{ATE}_{B,rel}$ denote estimated relative ATEs in the above and below categories. Inference on ratios and relative treatment effects is done using delta method. Standard errors in parentheses, robust and clustered at the cluster level.

Table 3: Estimates of ATE, Difference in Differences Estimator: Second Experiment, Full Sample

	Log Price	Sales	Revenue	Profit
$\widehat{ATE}$	-0.003			
	(0.001)			
$\widehat{ATE}_{rel}$		-0.055	-0.037	-0.036
		(0.026)	(0.024)	(0.028)
N	125,426,232	126,530,641	126,530,641	126,530,641

Notes: This table presents estimates of ATE for various outcomes of interest. To obtain estimates of ATE we use difference-in-differences estimator. $\widehat{ATE}_{rel}$ denotes an estimate of the relative treatment effect. For relative quantities inference is done using delta method. Standard errors in parentheses, robust and clustered at the cluster level.

Table 4: ATEs, Difference in Differences Estimator, Articles Eligible for Human Discounts: Second Experiment

	Log Price	Sales	Revenue	Profit
$\widehat{ATE}$	0.143			
	(0.005)			
$\widehat{ATE}_{rel}$		-0.394	-0.381	-0.127
		(0.032)	(0.042)	(0.085)
N	2,720,051	2,720,051	2,720,051	2,720,051

Notes: This table presents estimates of ATE for various outcomes of interest. The sample of consists of articles that were eligible for human discounts. To obtain estimates of ATE we use difference-in-differences estimator. Standard errors in parentheses, robust and clustered at the cluster level.

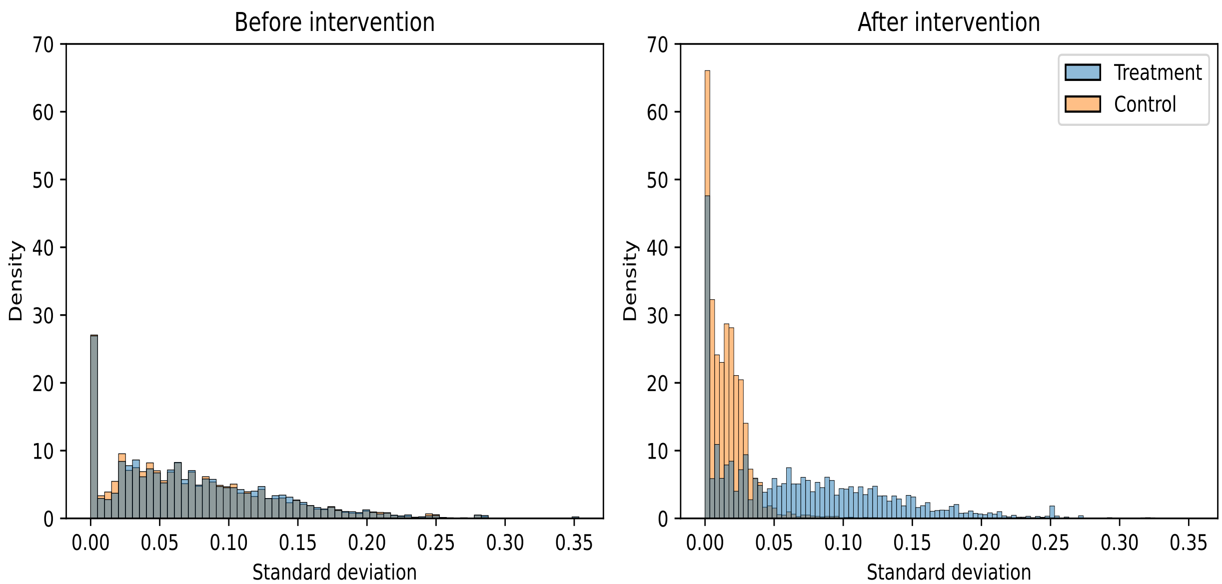
Table 5: Heterogeneity Analysis, Difference in Differences Estimator: Second Experiment, Articles Eligible For Human Discounts

	Log Price	Sales	Revenue	Profit
$\widehat{ATE}_B$	0.139			
	(0.007)			
$\widehat{ATE}_A$	0.144			
	(0.008)			
$\widehat{ATE}_A - \widehat{ATE}_B$	0.004			
	(0.009)			
$\widehat{ATE}_{B,rel}$		-0.269	-0.209	0.105
		(0.041)	(0.035)	(0.099)
$\widehat{ATE}_{A,rel}$		-0.496	-0.512	-0.321
		(0.038)	(0.054)	(0.110)
$\frac{\widehat{ATE}_A - \widehat{ATE}_B}{\lvert\widehat{ATE}_B\rvert}$		-1.698	-2.628	-4.955
		(0.645)	(0.988)	(4.110)
N	2,720,051	2,720,051	2,720,051	2,720,051

Notes: This table presents heterogeneity analysis. To obtain estimates of ATE we use difference in differences estimator. $\widehat{ATE}_A$ and $\widehat{ATE}_B$ denote estimated ATEs in the above and below categories. $\widehat{ATE}_{A,rel}$ and $\widehat{ATE}_{B,rel}$ denote estimated relative ATEs in the above and below categories. Inference on ratios and relative treatment effects is done using delta method. Standard errors in parentheses, robust and clustered at the cluster level.

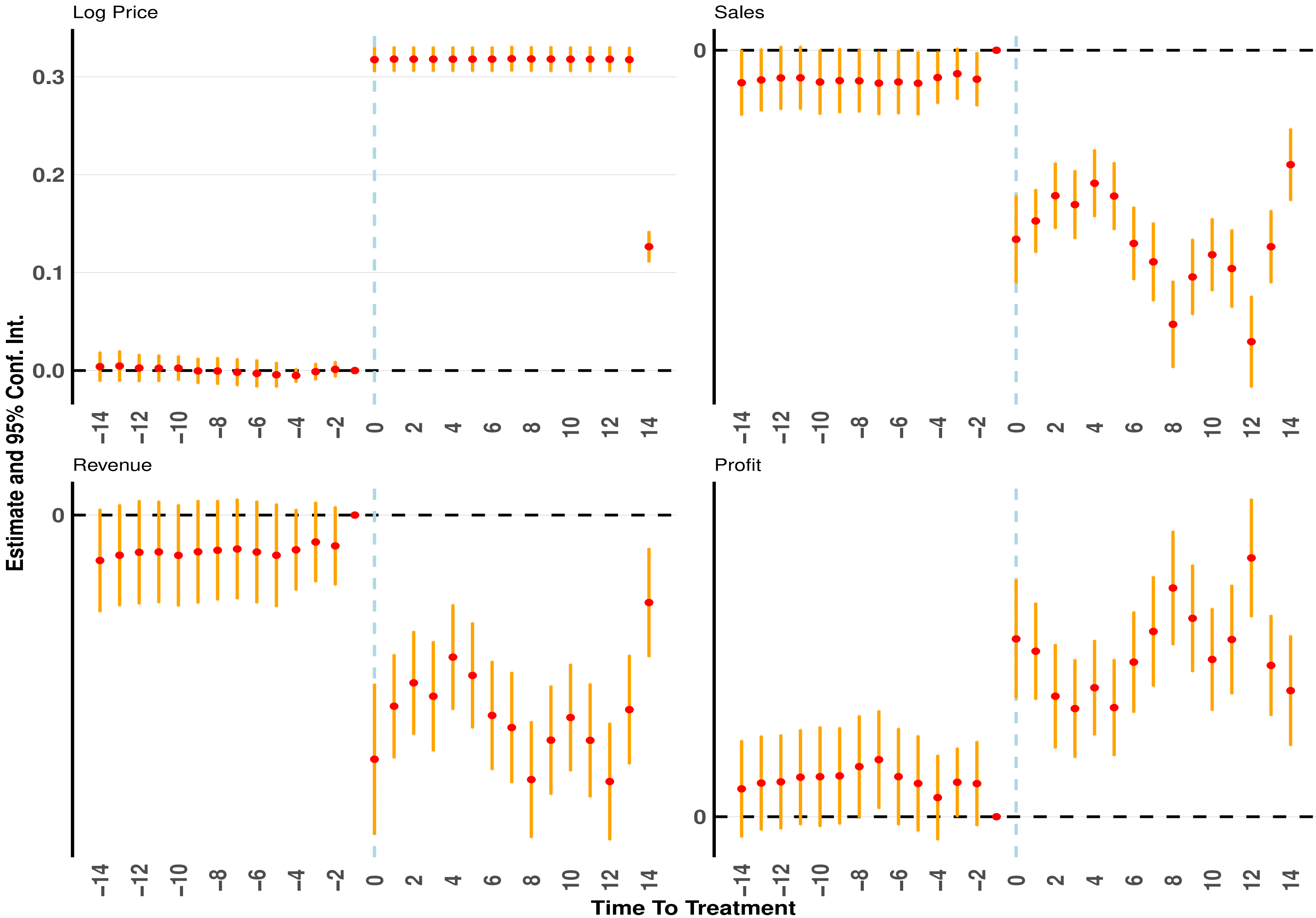
Figures

Figure 1: Variability in Product Discounts Across Countries Before and After the Start of the Initial Experiment



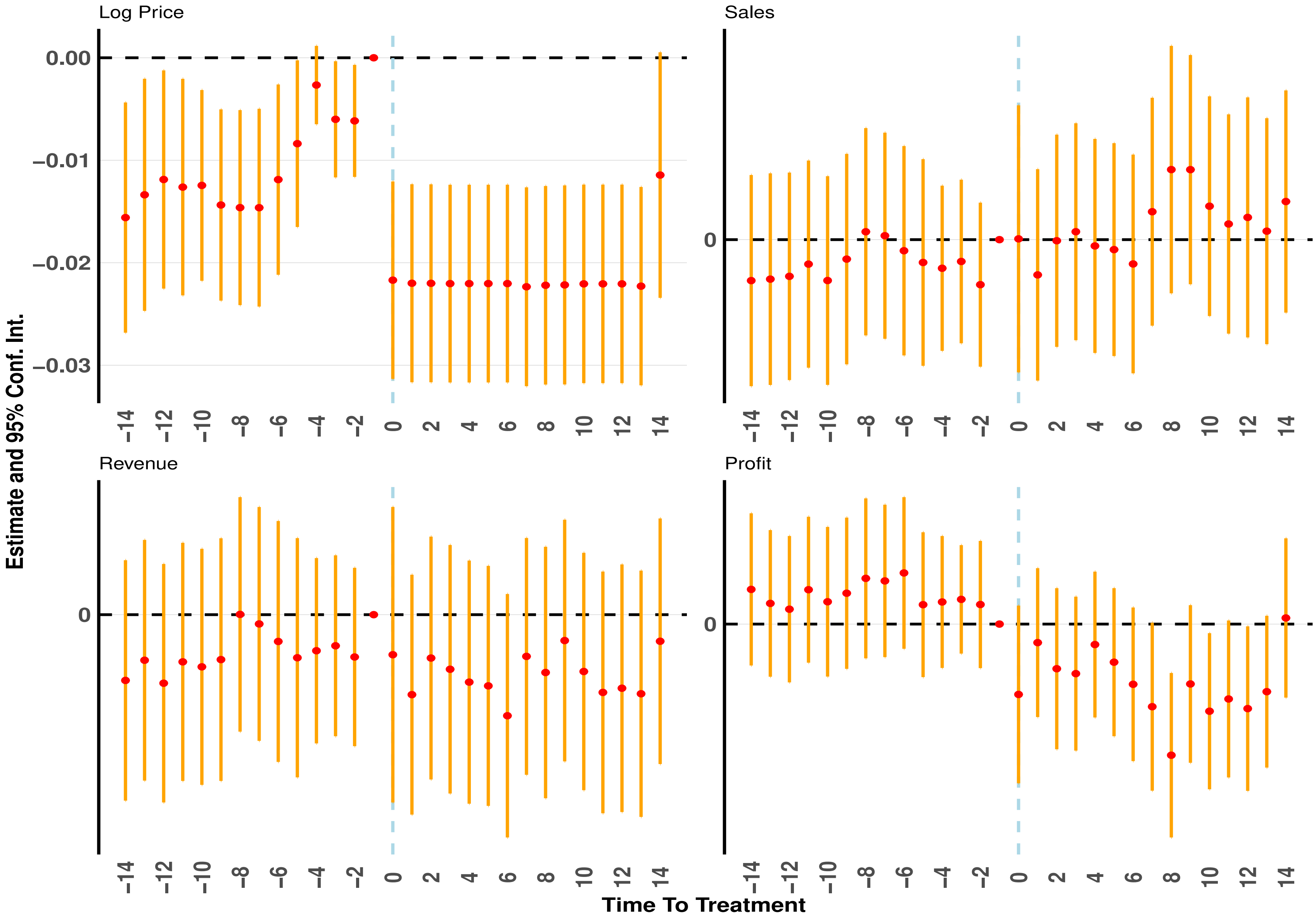
Notes: This figure illustrates the standard deviation in average discounts for each product across different countries, calculated before and after the start of the initial experiment. The displayed histogram shows that human interventions lead to less price variation than the algorithmic status quo.

Figure 2: Difference in Differences Event Study Estimates: First Experiment



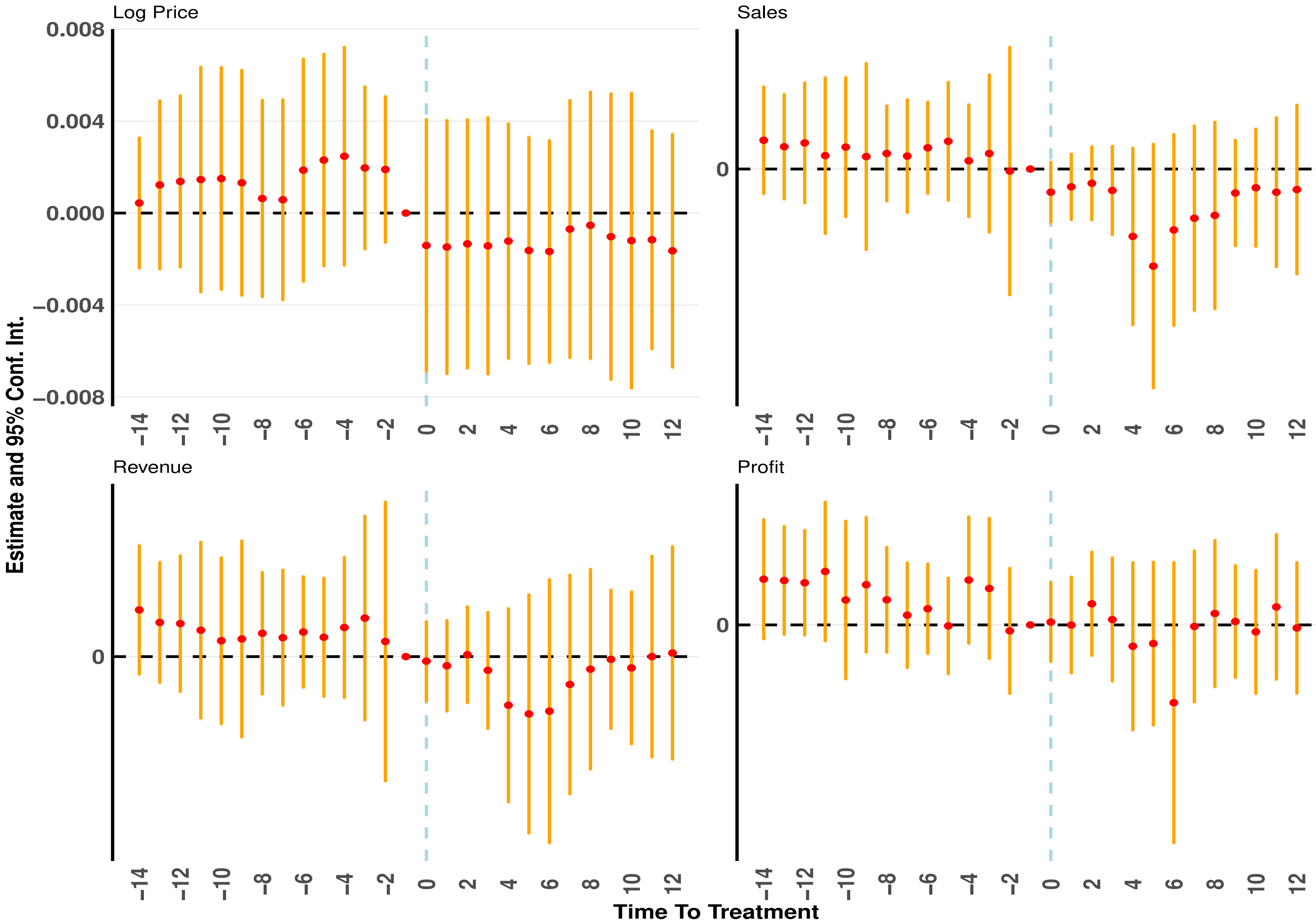
Notes: This figure displays event study coefficients for the first experiment. One day before the start of the experiment serves as a baseline in this specification.

Figure 3: Difference in Estimated ATEs for Above and Below Category: Initial Experiment



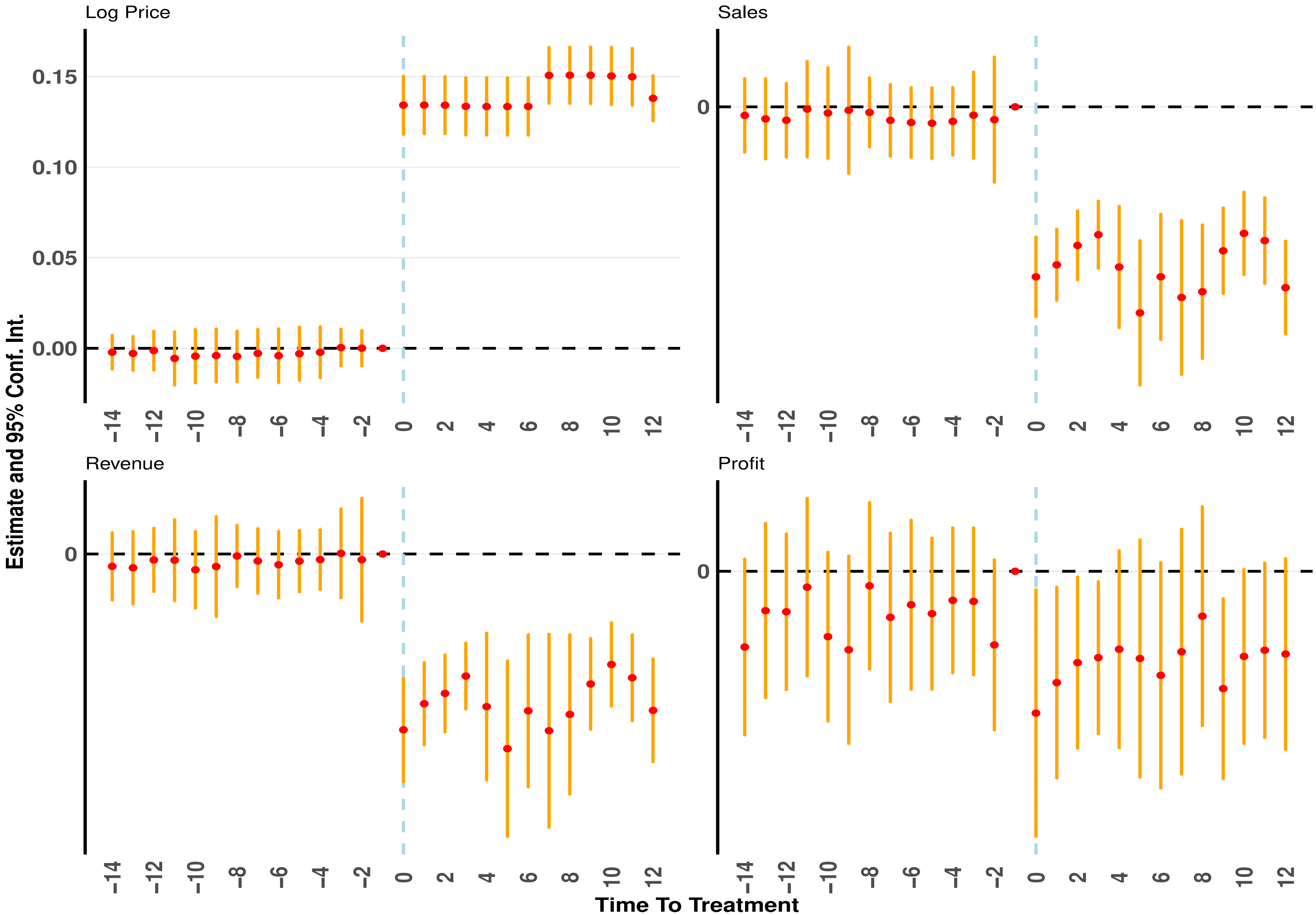
Notes: This figure displays difference in estimated ATEs for above and below categories. One day before the start of the experiment serves as a baseline in this specification.

Figure 4: Difference in Differences Event Study Estimates, Full Sample: Second Experiment



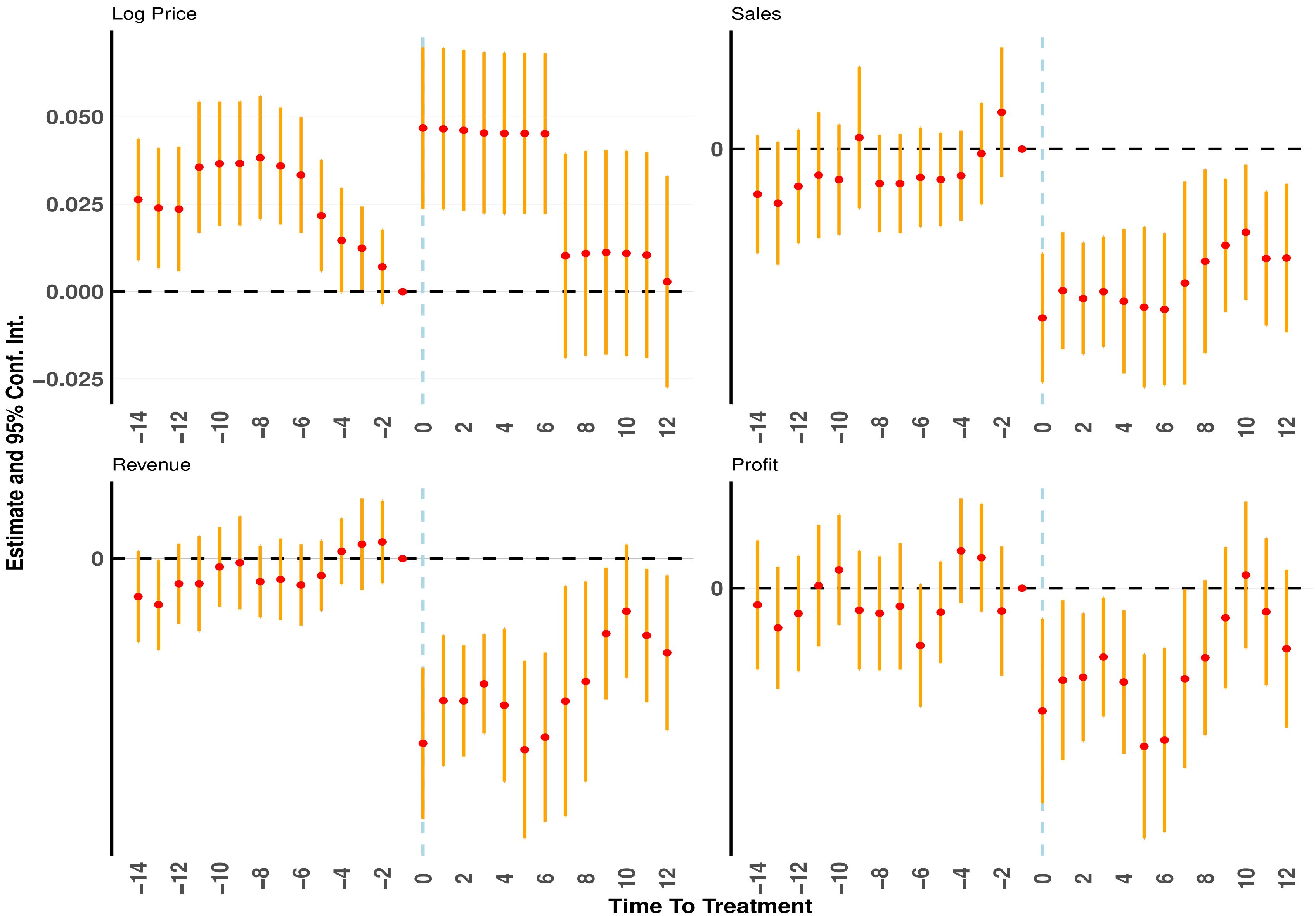
Notes: This figure displays event study estimates. One day before the start of the experiment serves as a baseline in this specification.

Figure 5: Difference in Differences Event Study Estimates, Sample of Articles Eligible for Human Discounts: Second Experiment



Notes: This figure displays event study estimates. One day before the start of the experiment serves as a baseline in this specification.

Figure 6: Difference in Estimated ATEs for Above and Below Category, Sample of Articles Eligible for Human Discounts: Second Experiment



Notes: This figure displays difference in estimated ATEs for above and below categories. One day before the start of the experiment serves as a baseline in this specification.

A Additional Tables and Figures

Table A.1: Balance Table: First Experiment

	Log Price	Sales	Revenue	Profit
Treat	0.005			
	(0.045)			
Treat_{rel}		0.027	0.027	-0.035
		(0.079)	(0.066)	(0.111)
N	1,748,538	1,748,538	1,748,538	1,748,538

Notes: This table tests for statistically significant differences in pre-treatment outcomes between treatment and control articles, using 2 weeks of pre-treatment data and a regression like $Y_{it} = \beta_0 + \beta_1 D_i + \epsilon_{it}$. Treat equals one for treated articles and zero otherwise. Treat_{rel} denotes an indicator for re-scaling. For scaled quantities inference is done using the delta method. Standard errors in parentheses, robust and clustered at the cluster level.

Table A.2: Estimates of ATE, Difference Estimator: First Experiment

	Log Price	Sales	Revenue	Profit
$\widehat{ATE}$	0.310			
	(0.045)			
$\widehat{ATE}_{rel}$		-0.648	-0.598	2.426
		(0.027)	(0.028)	(0.270)
N	1,873,845	1,873,845	1,873,845	1,873,845

Notes: This table presents estimates of ATE for various outcomes of interest. To obtain estimates of ATE we use difference estimator. $\widehat{ATE}_{rel}$ denotes an estimate of the relative treatment effect. For relative quantities inference is done using delta method. Standard errors in parentheses, robust and clustered at the cluster level.

Table A.3: Balance Table: Second Experiment, Full Sample

	Log Price	Sales	Revenue	Profit
Treat	0.005			
	(0.045)			
Treat_{rel}		-0.022	-0.010	-0.020
		(0.079)	(0.095)	(0.115)
N	64,985,084	65,557,636	65,557,636	65,557,636

Notes: This table tests for statistically significant differences in pre-treatment outcomes between treatment and control articles. Treat equals one for treated articles and zero otherwise. Treat_{rel} denotes rescaled indicator. For scaled quantities inference is done using delta method. Standard errors in parentheses, robust and clustered at the cluster level.

Table A.4: Balance Table, Articles Eligible for Human Discounts: Second Experiment

	Log Price	Sales	Revenue	Profit
Treat	-0.024			
	(0.041)			
Treat_{rel}		-0.026	-0.082	-0.222
		(0.100)	(0.118)	(0.149)
N	1,409,885	1,409,885	1,409,885	1,409,885

Notes: This table tests for statistically significant differences in pre-treatment outcomes between treatment and control articles. The sample of consists of articles that were eligible for human discounts. Standard errors in parentheses, robust and clustered at the cluster level.

Table A.5: Estimates of ATE, Difference Estimator: Second Experiment, Full Sample

	Log Price	Sales	Revenue	Profit
$\widehat{ATE}$	0.003 (0.045)			
$\widehat{ATE}_{rel}$		-0.075 (0.080)	-0.046 (0.092)	-0.053 (0.110)
N	60,441,148	60,973,005	60,973,005	60,973,005

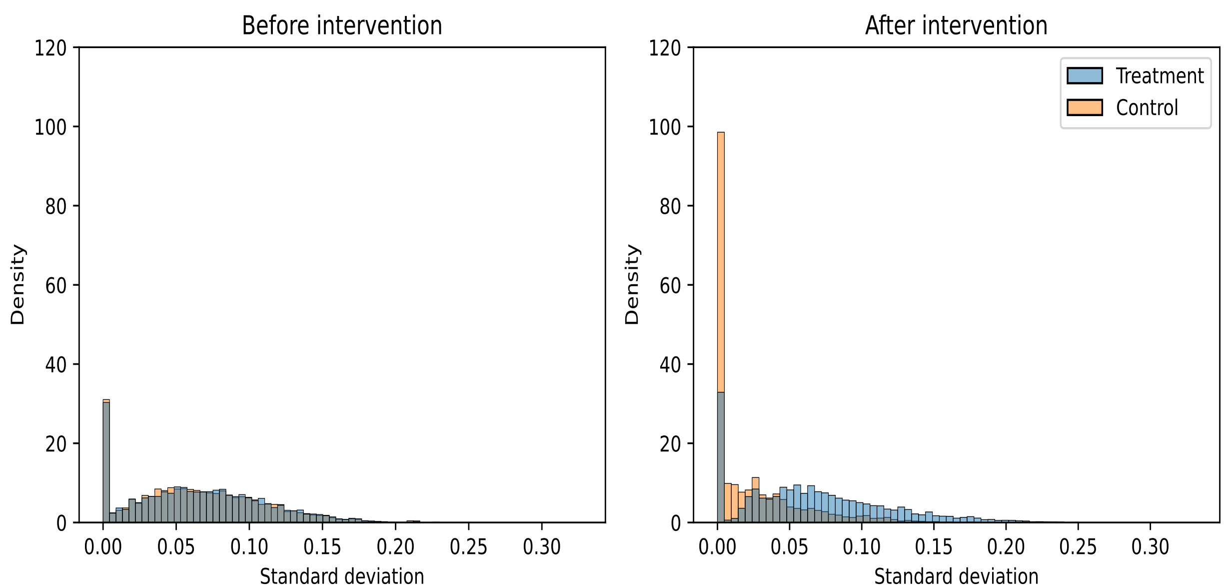
Notes: This table presents estimates of ATE for various outcomes of interest. To obtain estimates of ATE we use difference estimator. $\widehat{ATE}_{rel}$ denotes an estimate of the relative treatment effect. For relative quantities inference is done using delta method. Standard errors in parentheses, robust and clustered at the cluster level.

Table A.6: ATEs, Difference Estimator, Articles Eligible for Human Discounts: Second Experiment

	Log Price	Sales	Revenue	Profit
$\widehat{ATE}$	0.120 (0.040)			
$\widehat{ATE}_{rel}$		-0.407 (0.066)	-0.419 (0.085)	-0.297 (0.151)
N	1,310,166	1,310,166	1,310,166	1,310,166

Notes: This table presents estimates of ATE for various outcomes of interest. To obtain estimates of ATE we use difference estimator. The sample of consists of articles that were eligible for human discounts. Standard errors in parentheses, robust and clustered at the cluster level.

Figure A.1: Variability in Product Discounts Across Countries Before and After the Start of the Second Experiment



Notes: This figure illustrates the standard deviation in average discounts for each product across different countries, calculated before and after the start of the second experiment and only for the article assortment that was eligible for human pricing interventions.

PREVIOUS DISCUSSION PAPERS

412 Huelden, Tobias, Jascisens, Vitalijs, Roemheld, Lars and Werner, Tobias, Human-Machine Interactions in Pricing: Evidence from Two Large-Scale Field Experiments, March 2024.

411 Hermes, Henning, Lergetporer, Philipp, Mierisch, Fabian, Schwerdt, Guido and Wiederhold, Simon, Does Information about Inequality and Discrimination in Early Child Care Affect Policy Preferences? January 2024.

410 Hunold, Matthias and Werner, Tobias, Algorithmic Price Recommendations and Collusion: Experimental Evidence, December 2023.

409 Herzog, Sabrina, Schildberg-Hörisch, Hannah, Trieu, Chi and Willrodt, Jana, Who is in Favor of Affirmative Action? Representative Evidence from an Experiment and a Survey, November 2023.

408 Stadelmann, David, Thomas, Tobias and Zakharov, Nikita, Too Hot to Play it Cool? Temperature and Media Bias, November 2023.

407 Dertwinkel-Kalt, Markus and Wey, Christian, Why "Energy Price Brakes" Encourage Moral Hazard, Raise Energy Prices, and Reinforce Energy Savings, September 2023.
Forthcoming in: RAND Journal of Economics.

406 Bertermann, Alexander, Kamhöfer, Daniel A. and Schildberg-Hörisch, Hannah, More Education Does Make You Happier – Unless You Are Unemployed, September 2023.

405 Trieu, Chi, Who's Who: How Uncertainty About the Favored Group Affects Outcomes of Affirmative Action, August 2023.
Published in: Journal of the Economic Science Association, 9 (2023), pp. 252-292.

404 Fischer, Kai, Martin, Simon and Schmidt-Dengler, Philipp, The Heterogeneous Effects of Entry on Prices, July 2023.

403 García-Vega, María, Gupta, Apoorva and Kneller, Richard, Is Acquisition-FDI During an Economic Crisis Detrimental for Domestic Innovation?, July 2023.

402 Haucap, Justus and Stiebale, Joel, Non-price Effects of Mergers and Acquisitions, July 2023.

401 Simion, Ștefania and Sulka, Tomasz, Multidimensional Cognitive Ability, Intermediate Channels, and Financial Outcomes, May 2023.

400 Bartling, Björn, Cappelen, Alexander W., Hermes, Henning, Skivenes, Marit and Tungodden, Bertil, Free to Fail? Paternalistic Preferences in the United States, May 2023.

399 Kandelhardt, Johannes, Flexible Estimation of Random Coefficient Logit Models of Differentiated Product Demand, May 2023.

398 Hermes, Henning, Lergetporer, Philipp, Mierisch, Fabian, Peter, Frauke and Wiederhold, Simon, Discrimination on the Child Care Market: A Nationwide Field Experiment, April 2023.

397 Schmal, W. Benedikt, Haucap, Justus and Knoke, Leon, The Role of Gender and Coauthors in Academic Publication Behavior, March 2023.
Published in: Research Policy, 52 (2023), 104874.

396 Magin, Jana Anjali, Neyer, Ulrike and Stempel, Daniel, The Macroeconomic Effects of Different CBDC Regimes in an Economy with a Heterogeneous Household Sector, March 2023.

395 Dertwinkel-Kalt, Markus and Wey, Christian, Resale Price Maintenance in a Successive Monopoly Model, February 2023.
Forthcoming in: Journal of Industrial Economics.

394 Hermes, Henning, Krauß, Marina, Lergetporer, Philipp, Peter, Frauke and Wiederhold, Simon, Early Child Care and Labor Supply of Lower-SES Mothers: A Randomized Controlled Trial, December 2022.

393 Chowdbury, Shyamal, Schildberg-Hörisch, Hannah, Schneider, Sebastian O., and Sutter, Matthias, Information Provision Over the Phone Saves Lives: An RCT to Contain COVID-19 in Rural Bangladesh at the Pandemic's Onset, November 2022.

392 Normann, Hans-Theo and Sternberg, Martin, Human-Algorithm Interaction: Algorithmic Pricing in Hybrid Laboratory Markets, October 2022.
Published in: European Economic Review, 152 (2023), 104347.

391 Hunold, Matthias and Petrishcheva, Vasilisa, Foreclosure and Tunneling with Partial Vertical Ownership, September 2022.

390 Haucap, Justus and Heldman, Christina, The Sociology of Cartels, August 2022.
Published in: European Journal of Law and Economics, 56 (2023), pp. 289-323.

389 Döpper, Hendrik, Sapi, Geza and Wey, Christian, A Bargaining Perspective on Vertical Integration, May 2022.
Published in: Canadian Journal of Economics, 57 (2024), pp. 199-224.

388 Bachmann, Ronald, Gonschor, Myrielle, Lewandowski, Piotr and Madoń, Karol, The Impact of Robots on Labour Market Transitions in Europe, May 2022.

387 Fremerey, Melinda, Hörnig, Lukas and Schaffner, Sandra, Becoming Neighbors with Refugees and Voting for the Far-Right? The Impact of Refugee Inflows at the Small-Scale Level, April 2022.
Published in: Labour Economics, 86 (2024), 102467.

386 Fischer, Kai, Alcohol Prohibition and Pricing at the Pump, March 2022.

385 Cobb-Clark, Deborah A., Dahmann, Sarah C., Kamhöfer, Daniel A. and Schildberg-Hörisch, Hannah, The Determinants of Population Self-Control, March 2022.
Forthcoming in: Economic Journal under the title "Surveillance and Self-Control".

384 Sulka, Tomasz, Planning and Saving for Retirement, March 2022.
Published in: European Economic Review, 160 (2023), 104609.

383 Cattan, Sarah, Kamhöfer, Daniel A., Karlsson, Martin and Nilsson, Therese, The Long-term Effects of Student Absence: Evidence from Sweden, March 2022.
Published in: Economic Journal, 133 (2023), pp. 888-903.

382 Martin, Simon and Rasch, Alexander, Collusion by Algorithm: The Role of Unobserved Actions, March 2022.
Published in: International Journal of Industrial Organization, 92 (2024), 103036 under the title "Demand Forecasting, Signal Precision, and Collusion with Hidden Actions".

381 Haucap, Justus, Nedic, Radivoje and Şimşek, Talha, An Empirical Analysis of German Casino Locations, March 2022.
Published in: European Journal of Law and Economics, 55 (2023) pp. 291-311.

380 Haucap, Justus, Heldman, Christina and Rau, Holger A., Gender and Collusion, March 2022.

379 Schain, Jan Philip, Foreign Institutional Investors and the Great Productivity Slowdown, November 2022 (First Version February 2022).

378 Neyer, Ulrike and Stempel, Daniel, How Should Central Banks React to Household Inflation Heterogeneity?, January 2022.

Older discussion papers can be found online at:
http://ideas.repec.org/s/zbw/dicedp.html

Heinrich-Heine-Universität Düsseldorf

Düsseldorfer Institut für Wettbewerbsökonomie (DICE)

Universitätsstraße 1, 40225 Düsseldorf

ISSN 2190-992X (online)
ISBN 978-3-86304-411-4