EUR Research Information Portal

Personalization and Targeting

Published in:
International Journal of Research in Marketing

Publication status and date:
Published: 01/06/2026

DOI (link to publisher):
10.1016/j.ijresmar.2025.07.004

Document Version
Publisher's PDF, also known as Version of record

Document License/Available under:
CC BY

Citation for the published version (APA):
Lemmens, A., Roos, J., Gabel, S., Ascarza, E., Bruno, H., Gordon, B., Israeli, A., McDonnell Feit, E., Mela, C., & Netzer, O. (2026). Personalization and Targeting: how to experiment, learn & optimize. *International Journal of Research in Marketing*, *43*(2), 537-559. https://doi.org/10.1016/j.ijresmar.2025.07.004

Link to publication on the EUR Research Information Portal

Terms and Conditions of Use
Except as permitted by the applicable copyright law, you may not reproduce or make this material available to any third party without the prior written permission from the copyright holder(s). Copyright law allows the following uses of this material without prior permission:

- you may download, save and print a copy of this material for your personal use only;
- you may share the EUR portal link to this material.

In case the material is published with an open access license (e.g. a Creative Commons (CC) license), other uses may be allowed. Please check the terms and conditions of the specific license.

Take-down policy
If you believe that this material infringes your copyright and/or any other intellectual property rights, you may request its removal by contacting us at the following email address: openaccess.library@eur.nl. Please provide us with all the relevant information, including the reasons why you believe any of your rights have been infringed. In case of a legitimate complaint, we will make the material inaccessible and/or remove it from the website.

Erasmus University Rotterdam
Making Minds Matter

International Journal of Research in Marketing xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Research in Marketing

journal homepage: www.elsevier.com/locate/ijresmar

Personalization and targeting: how to experiment, learn & optimize ⋆

Aurélie Lemmens [a,*], Jason M.T. Roos [a], Sebastian Gabel [a], Eva Ascarza [b], Hernán A. Bruno [c], Brett R. Gordon [d], Ayelet Israeli [b], Elea McDonnell Feit [e], Carl F. Mela [f], Oded Netzer [g]

[a] Rotterdam School of Management, Erasmus University, the Netherlands
[b] Harvard Business School, USA
[c] University of Cologne, Germany
[d] Kellogg School of Management, Northwestern University, USA
[e] Drexel University's LeBow College of Business, USA
[f] The Fuqua School of Business, Duke University, USA
[g] Columbia Business School, USA

ARTICLE INFO

Keywords:
Personalization
Test and learn
Experiments

ABSTRACT

Personalization has become the heartbeat of modern marketing. The rapid expansion of individual-level data, the proliferation of personalized communication channels, and advancements in experimentation have fundamentally reshaped how firms tailor their marketing strategies. Furthermore, causal inference and machine learning enable companies to understand how the same marketing action can impact the choices of individual customers differently. This article provides an academic overview of these developments. We formalize personalization as a causal inference problem embedded in the test and learn framework. We review key challenges and solutions that arise when personalization is approached through causal inference, including data limitations, treatment effect heterogeneity, policy evaluation, and ethical considerations. Finally, we identify emerging research trends stemming from new methodologies such as generic and double machine learning, direct policy learning, foundation models, and generative AI.

1. A new era of personalization

Personalization has become the heartbeat of modern marketing. By leveraging consumer data, firms craft tailored offers, content, and communications that match consumers' individual preferences and needs. Seventeen years ago, Arora et al. (2008) published an influential article in the 7th Tri-Annual Choice Symposium Special Issue about what was then an emerging concept: *personalization*. They defined it as the process by which "the firm decides, usually based on previously collected customer data, what marketing mix is suitable for the individual." Personalization thus refers to tailoring firm decisions to individual consumers or groups based on observed characteristics, rather than implementing a uniform policy for the entire customer base.

⋆ This paper is the result of discussions that took place during the 12th Triennial Invitational Choice Symposium held from August 9–12, 2023 at the INSEAD Campus, in Fontainebleau, France. Aurélie Lemmens, Jason Roos, and Sebastian Gabel led the writing and editing process. All other authors contributed equally and are listed alphabetically.

* Corresponding author.
E-mail address: lemmens@rsm.nl (A. Lemmens).

https://doi.org/10.1016/j.ijresmar.2025.07.004
Received 27 June 2024
Available online 25 July 2025
0167-8116/© 2025 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Please cite this article as: Aurélie Lemmens et al., *International Journal of Research in Marketing*, https://doi.org/10.1016/j.ijresmar.2025.07.004

Table 1
Examples of personalized marketing according to the customer lifecycle stage.

	Customer acquisition	Customer development	Customer retention
Offline	**BMW:** Personalized direct mail campaigns for test drives	**Nordstrom:** In-store stylists recommend clothing based on purchase history	**Ritz-Carlton:** Customized experiences based on preferences
	Timberland: Location-based targeting to encourage store visit	**Whole Foods:** Tailored in-store promotions based on local demographics	**Local boutique:** Personalized thank-you notes with offers
Online	**Nike:** Personalized ads tailored to their browsing history and location	**Netflix:** Dynamic content previews tailored to viewing history (sizzles)	**Spotify:** AI-curated playlists to keep users engaged
	Shopify: Personalized product landing pages for first-time visitors	**Amazon:** Upselling via personalized email recommendations	**Chewy:** Reminders for replenishable products

Today, personalization shapes all stages of the customer life cycle, transforming the way businesses attract, engage, and retain customers. Table 1 provides examples of personalization focused on each stage of the customer life cycle. In the customer acquisition phase, luxury car manufacturers like BMW use personalized direct mail to invite potential customer targets[1] for test drives of specific models based on demographic and lifestyle data (Pereira, 2024), while online platforms like Nike deliver micro-targeted ads tailored to individual browsing behaviors and interests (Barseghian, 2022). In customer development, Nordstrom enhances engagement through personalized in-store experiences, with stylists offering tailored recommendations (Patov, 2024), while Netflix employs personalized previews to dynamically showcase content based on a subscriber's viewing history (Wobbe & Kwok, 2023). Finally, in customer retention, hotels like The Ritz-Carlton remember repeat guests' preferences for room setup and amenities to create a bespoke experience (Bare, 2023), just as Spotify retains subscribers by curating personalized playlists that align with their unique musical tastes (Spotify, 2024). These examples demonstrate how personalization—whether online or offline—enables firms to build profitable, long-term relationships with their customers, making it a cornerstone of contemporary marketing strategy (Rust & Huang, 2014).

Since Arora et al. (2008), the landscape of personalization has undergone a profound transformation. This evolution has been driven by the emergence of new data sources (e.g., social media and mobile data), significant technological advancements (e.g., cloud computing, SaaS, Generative AI), methodological innovations (e.g., causal inference and machine learning), the proliferation of new marketing channels (e.g., mobile marketing), and evolving business priorities that increasingly emphasize societal considerations such as fairness and privacy. These disruptions call for a fresh perspective on the personalization landscape.

Accordingly, this paper examines the current state of the field of personalization in marketing, with two main objectives. Our first goal is to provide a framework to formalize the personalization problem through the lens of causal inference, which aligns with the way the field has been evolving over the last decade (Rafieian & Yoganarasimhan, 2023). In this context, we describe how "test and learn" has emerged as a recognized approach to personalization. Our second objective is to explore the key challenges that arise when personalization is approached through causal inference. This perspective introduces challenges such as estimating heterogeneous treatment effects in the presence of evolving markets, platforms, regulations, businesses, and society. The article offers concrete guidance for marketing researchers on addressing these challenges and outlines what we believe are emerging research trends in personalization. Together, these objectives align closely with several of the 2024 MSI Priorities that call for better guidance on how to utilize algorithmic decision support in interactions with customers (Marketing Science Institute, 2024, pages 4–5).

The structure of the paper is as follows. Section 2 examines the technological advances that have transformed personalization. Section 3 reviews the recent literature on personalization, while Section 4 provides a mathematical foundation. Section 5 describes the test and learn approach. Finally, Section 6 identifies challenges in four domains, and Section 7 outlines emerging research directions in personalization.

2. Technological advances enabling personalization

The success of personalized marketing has been made possible by technological advancements that have transformed how companies collect, analyze, and act on customer data. We categorize these advances into three key areas: (1) enhanced access to individual-level data, (2) personal communication channels enabling personalized marketing, and (3) technological breakthroughs that facilitate continuous experimentation and prediction of individual customer responses.

2.1. Enhanced access to individual-level data

The first technological advance is related to the availability of new data sources. Firms can now track interactions across a wide range of touch points, including website visits, in-store purchases, call center engagements, and even customer complaints.

[1] In this paper, we refer to targeting specifically in the context of targeting an individual with a personalized offer.

The unprecedented access to granular, individual-level customer data enables businesses to move beyond understanding just what customers buy, delving into how, when, and why they engage with brands (Rust & Huang, 2014).

Insights from individual-level behavioral data allow marketers to identify customer preferences and anticipate needs with a previously unattainable level of precision. Smartphone apps collect personal information, behavioral data, and contextual data to enable personalization. Snapchat, for example, leverages user data, such as location from Snap Map, browsing habits in Discover, and engagement with features like filters and lenses, to deliver personalized content and ads (Malik, 2024). By analyzing responses to push notifications and in-app purchases, Snap optimizes interventions to maximize user engagement and profitability.

Cross-app usage insights offer an even broader view of user behavior across multiple firms. Allstate collects user behavior data from driving apps like Life360 and GasBuddy to generate driving scores, which are then used to set insurance rates tailored to individual driving patterns (Hill, 2024). These personalized rates represent a form of price differentiation, leveraging individual-level data to adjust pricing based on risk profiles and behaviors. This type of sharing of customer data is becoming a key motivation for strategic link-ups between firms.

Social media platforms such as Facebook and Instagram generate large amounts of data that provide valuable insights into user behaviors and preferences. Detailed information on how users engage with content, including likes, shares, comments, and watch time, reflects users' preferences and state of mind. Additionally, data on the formation and dynamics of social networks reveal how individuals connect with peers, influencers, and brands. Meta's advertising products allow marketers to leverage these data for personalized advertising.

Finally, Internet of Things (IoT) technologies facilitate the collection of rich, real-time data from devices embedded in users' daily lives. Data from voice assistants like Alexa generate insights into users' daily routines and contextual needs. Similarly, companies like Samba TV collect detailed usage behavior from smart TVs by tracking watch histories and engagement to inform personalized content recommendations and advertising strategies (Maheshwari, 2018). Wearable devices, including smartwatches and fitness trackers, extend this personalization by capturing real-time movement patterns, health metrics, and even sleep data.

Firms increasingly integrate diverse data sources into unified customer data platforms (data lakes), combining customer information from in-store interactions, online activity, and social media into a single, actionable profile. However, growing consumer awareness and tightening privacy regulations, such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), are beginning to constrain firms' ability to collect and use granular individual-level data. We discuss these challenges in Section 6.

2.2. Personalized communication channels

The second technological advance is modern communication channels that have revolutionized firms' ability to connect with customers individually. Unlike traditional mass media, such as billboards or television ads that broadcast uniform messages to broad audiences, today's platforms facilitate personalized interactions. Companies have always had the ability to address their customers individually, e.g., via regular mail (Esterzon et al., 2023) or catalogs (Zantedeschi et al., 2017), but the rise of online channels has facilitated personalized interactions. Companies can now send personalized emails, display targeted ads on social media (Haenlein & Libai, 2013; Libai et al., 2013), and deliver context-specific push notifications through mobile apps (Andrews et al., 2016) to a select set of customers.

Mobile marketing stands out for its ability to leverage hyper-contextual data (Tong et al., 2020). Mobile devices provide access to rich, real-time information, including a consumer's location, the time of an activity, environmental conditions, and even whether the consumer is in the company of others. This hyper-contextual information, obtained through built-in sensors such as GPS, accelerometers, and gyroscopes, enables companies to advance beyond personalization to what is termed hyper-personalization. For example, Burger King's "Whopper Detour" utilized geo-targeted promotions to attract customers near McDonald's locations (Black, 2019).

Communication channels can also provide data and insights to service personnel, which can be used to personalize interactions. For example, Delta Airlines uses an application to provide flight attendants with relevant customer data in real time (Bachman, 2018). This tool provides access to passenger information, such as travel history, preferences, and past issues, empowering them to resolve concerns efficiently and deliver personalized assistance.

These examples illustrate that the unprecedented immediacy and context-awareness empower businesses to deliver highly personalized and impactful customer experiences. Edge computing-data processing directly on users' devices-allows firms to instantly react to incoming data. This seamless integration of real-time data, hyper-contextual insights, and one-to-one communication channels illustrates the trend toward hyper-personalized marketing interventions.

2.3. Continuous experimentation and predictions of customer responses

The third technological advance is continuous experimentation, which has become a cornerstone of personalized marketing by enabling firms to test a wide array of interventions, such as tailored offers or promotions, across randomized samples of their customer base. Through experiments, companies can obtain precise predictions of how specific customers are likely to react to particular offers, considering not only their characteristics, such as past behaviors and demographic data, but also hyper-contextual factors like location and time of day.

At Amazon, Booking.com, and Google, experimentation is deeply embedded in the organizational culture. Amazon employs a sophisticated experimentation infrastructure that supports testing hypotheses around personalized promotions, pricing strategies, and

dynamic bundling options, tailoring its marketing to the unique preferences and needs of individual customers. Similarly, Booking.com runs thousands of experiments annually to fine-tune customer interactions, from website design and search filters to personalized messaging and promotional offers (Thomke, 2020). This iterative testing allows the company to dynamically optimize each customer's experience, using real-time insights. Similarly, Google's Ad Experiments platform further illustrates the power of continuous testing. Such platforms have been increasingly used by marketing scholars to study consumer response in a real-world environment, even though Boegershausen et al. (2025) warns that such platforms suffer from inherent confoundedness.

Experimentation is complemented by a methodological revolution that has made it possible to leverage the wealth of experimental data generated by continuous testing. When Arora et al. (2008) considered personalization, heterogeneous customer responses were typically modeled using a random-coefficients approach. For instance, Rossi et al. (1996) uses individual consumer panel data to model household-specific sensitivity to marketing mix variables through a multinomial brand choice model, in which heterogeneity distributions were linked to observable demographics. Similarly, Ansari and Mela (2003) leverages user-level click stream data to estimate the probability of clicking on links within emails, enabling personalized email design and content.

In modern randomized controlled trials, the large number of experiments, customers, and features creates challenges related to statistical power and multiple testing. Causal machine learning, a methodological breakthrough that integrates causal inference techniques with machine learning (ML) algorithms, provides a promising solution to these challenges (Wager & Athey, 2018). By offering a highly flexible approach to estimating heterogeneous treatment effects, causal ML enables marketers to uncover nuanced customer responses to interventions. Among these recent advances, double and generic machine learning enable the estimation of treatment effects based on high-dimensional observational or experimental data (Chernozhukov et al., 2018, 2025). While the technological advances discussed in this section enable personalization, novel data sources, channels, and methodologies also lead to new challenges, which we will discuss in Section 6. Next, we provide an overview of the recent literature on personalization in view of these recent changes.

3. Personalization in marketing research

Since the publication of Arora et al. (2008), personalization has received considerable attention in the marketing literature, driving the development of novel approaches to attract, engage, and retain customers. Table 2 provides an overview of a representative set of studies in this domain over the past two decades, highlighting the focal marketing instrument being personalized in each case. Personalization spans a wide range of marketing-mix instruments, including ads (Andrews et al., 2016; Gordon et al., 2023b; Urban et al., 2014), free trials (Datta et al., 2015; Yoganarasimhan et al., 2023), loyalty programs (Daljord et al., 2023), pricing and promotions (Gabel & Timoshenko, 2022; Liu, 2023; Zhang & Krishnamurthi, 2004), product recommendations (Chung et al., 2009; Padilla et al., 2024; Xu et al., 2023), and website design (Hauser et al., 2009; Liberali & Ferecatu, 2022). While some studies implement personalization offline, such as through regular mail (Esterzon et al., 2023; Hitsch et al., 2024), the majority takes place online, e.g., on social media platforms (Gordon et al., 2019). Delivery formats include banners (Bleier & Eisenbeiss, 2015) and push messages (Luo et al., 2014). Recently, several studies have also explored how to combine personalization across multiple channels, offline and online (Valenti et al., 2024).

Designing personalized policies involves multiple decisions (Ascarza, 2018; Tong et al., 2020): *whom* to target, *what* to offer and communicate, *when* to engage, and *where* to reach out (e.g., location, channel). Table 2 highlights the specific question(s) addressed by each study. Identifying *whom* to target is crucial to ensure that marketing resources are allocated efficiently to audiences most likely to respond positively. For instance, Ascarza (2018) and Lemmens and Gupta (2020) demonstrate that retention offers should not be targeted at potential churners, but instead, firms should identify customers with high expected (profit) lifts. Likewise, Blake et al. (2015) reports that paid search ads are most effective if targeted at new and infrequent users, and should not be shown to loyal ones, similar to Sahni et al. (2017) which finds that discounts are more effective for dormant customers. Finally, von Zahn et al. (2024) emphasizes the varying effectiveness of green nudges in reducing product returns, noting that green interventions can even backfire.

The literature has also focused on optimizing the *what* decision. For instance, Ellickson et al. (2023) estimates the heterogeneous treatment effects of 13 different promotional emails and finds that discounts framed as clearance events sharply outperform those tied to particular products. Similarly, Zantedeschi et al. (2017) shows that catalogs have longer-lasting effects on purchases than emails. More broadly, personalizing what to offer can extend to decisions on personalized assortments, recommendations, and product design. For instance, Chung et al. (2009) designs a real-time hyper-personalization approach for mobile music recommendation using listening data and music attributes as predictor variables. Hauser et al. (2009) develops a system for website design personalization called "website morphing" that matches the content, look, and "feel" of the website to visitors' cognitive styles.

Other research has focused on temporal (Hui et al., 2013) or geographical personalization (Ghose et al., 2013a). For instance, Zhang and Krishnamurthi (2004) develops a decision support system to personalize promotion timing in order to maximize profits. Similarly, Sahni et al. (2019) personalizes retargeting timing by comparing the reactivation of lapsed customers directly after the website visit to reactivation four weeks later, while Valenti et al. (2025) shows that retargeting can be effectively combined with contextual targeting when users are prompted to customize the product. Hauser et al. (2014) also focuses on the *when*, specifically in terms of the optimal timing to deploy website morphing. Morphing earlier generates more customer clicks, while morphing later reveals more about the customer's latent segment. Finally, in terms of the *where*, Molitor et al. (2020) studies the effectiveness of location-based coupons while Andrews et al. (2016) explores how features of locations (i.e., how crowded they are) influence promotional effectiveness.

Table 2
Representative literature overview on personalization with focus on whom, what, when, and/or where?.

Paper	Channel	Instrument	Whom	What	When	Where	Contextual factors
Andrews et al. (2016)	Online (mobile)	Ads	✓		✓	✓	**Whom:** Based on contextual data; **When:** Predefined times that differ in crowdedness; **Where:** Target commuters in a subway train
Ascarza (2018)	Offline (mail)/ Online (email)	Free credits, gifts	✓				**Whom:** Select based on their sensitivity to retention interventions
Blake et al. (2015)	Online (search)	Ads	✓				**Whom:** New and infrequent users
Bleier and Eisenbeiss (2015)	Online (banner)	Ads		✓	✓	✓	**What:** Different degrees of content personalization; **When:** Elapsed time since last store visit; **Where:** Banner-site congruence
Chung et al. (2009)	Online (audio device)	Recommendations	✓	✓	✓		**Whom:** Tailor playlists to users' preferences; **What:** Choosing songs by tempo, genre, relevance; **When:** Recent listening data
Daljord et al. (2023)	Offline (in-store)	Loyalty program	✓	✓			**Whom:** Compliers (behavior changes due to promotion); **What:** Reward levels and spending hurdles
Datta et al. (2015)	Offline (various)	Ads, direct marketing	✓	✓			**Whom:** Free-trial customers; **What:** Direct marketing and mass advertising
Ellickson et al. (2023)	Online (email)	Price promotions	✓	✓		✓	**Whom:** RFM, engagement with past promotions; **What:** Clearance vs. product-specific discounts; **Where:** Open vs. purchase stage
Esterzon et al. (2023)	Offline (mail)	Donation request	✓	✓			**Whom:** Frequent, generous, and long-tenured donors; **What:** Targeting-via-options and targeting-via-amounts
Gabel and Timoshenko (2022)	Offline (in-store)	Coupons	✓	✓	✓		**Whom:** Which customers to target; **What:** Which product to promote; **When:** Purchase sequence
Ghose et al. (2013b)	Online (mobile/ website)	Rankings	✓	✓	✓	✓	**Whom:** Store proximity; **What:** Brand post ranking position; **When:** Information recency; **Where:** Posts for nearby stores
Gordon et al. (2019)	Online (social media)	Ads	✓	✓			**Whom:** Likely converters based on activity; **What:** Ad types and creatives
Gordon et al. (2023b)	Offline (TV)	Ads		✓	✓	✓	**What:** Positive and negative ad tones; **When:** Election cycle phases; **Where:** Battleground states and media markets
Hauser et al. (2009)	Online (website)	Website design	✓	✓			**Whom:** Cognitive styles from clickstream data; **What:** Customizing website layouts
Hauser et al. (2009)	Online (website)	Website design	✓	✓	✓		**Whom:** Cognitive style from clickstream data; **What:** Website designs; **When:** Morph updates based on behavior
Hitsch et al. (2024)	Offline (catalog)	Catalogs	✓				**Whom:** Identify customers with high predicted incremental spending (CATEs)
Huang and Ascarza (2023)	Online (mobile)	Coupon	✓				**Whom:** Identify customers with high long-term spending potential
Huang et al. (2024)	Online (app)	Coupon	✓	✓			**Whom:** Target high incremental spending customers; **What:** Tailored promotions
Hui et al. (2013)	Online (mobile)	Coupon	✓	✓		✓	**Whom:** Travel patterns within a store; **What:** Coupons customized to path; **Where:** In-store
Lemmens and Gupta (2020)	Offline (mail)	Promotions	✓				**Whom:** Targets customers with highest profit lift

Table 2
Continue

Paper	Channel	Instrument	Whom	What	When	Where	Contextual factors
Liberali and Ferecatu (2022)	Online (website)	Website design	✓	✓	✓		**Whom:** Position in purchase funnel (clickstream); **What:** Abstract vs. concrete morphs; **When:** Adjust morph by to funnel position
Liu (2023)	Online (mobile)	Coupons	✓	✓	✓		**Whom:** Spending habits, engagement with hosts; **What:** Target discount depth; **When:** Adjusting discounts over time
Luo et al. (2014)	Online (mobile)	Ads	✓		✓	✓	**Whom:** Users near cinemas and mobile activity; **When:** Same-day, one-day prior, two-day prior; **Where:** Geographical proximity
Ma et al. (2025)	Online (game)	Product features	✓	✓	✓		**Whom:** Playing habits, past activity; **What:** Adjust difficulty; **When:** Based on in-game progress
Molitor et al. (2020)	Online (mobile)	Coupons	✓	✓		✓	**Whom:** Based on location; **What:** With or without distance information, distance-based ranking; **Where:** Proximity to retailers
Padilla et al. (2024)	Online (website)	Recommendations	✓				**Whom:** Target based on journey attributes, such as trip type, number of travelers, and flight preferences
Sahni et al. (2017)	Online (email)	Promotions	✓		✓		**Whom:** Based on RFM metrics; **When:** Customer inactivity
Sahni et al. (2019)	Online (banner)	Ads	✓		✓		**Whom:** Browsing users or cart abandoners; **When:** Frequency caps and timing
Urban et al. (2014)	Online (banner)	Ads	✓	✓			**Whom:** Cognitive styles from clickstream data; **What:** Dynamically optimizing banner ad size, color, content
Valenti et al. (2024)	Offline (mail)/ Online (email)	Ads	✓			✓	**Whom:** New vs. existing customers; **Where:** Online vs. offline
Valenti et al. (2025)	Online (website)	Ads	✓	✓	✓		**Whom:** Prospective vs. retargeted users; **What:** Prompt to customize or not; **When:** Combined with contextual ad
von Zahn et al. (2024)	Online (e-commerce)	Nudges	✓				**Whom:** Customers prone to product returns
Xu et al. (2023)	Online (e-commerce)	Recommendations	✓	✓			**Whom:** Targets users by demographics and behaviors; **What:** Product recommendations based on attributes
Yoganarasimhan et al. (2023)	Online (email)	Free trials	✓	✓			**Whom:** Users based on demographics and skill level; **What:** Trial duration
Zantedeschi et al. (2017)	Offline (catalog)/ Online (email)	Ads	✓			✓	**Whom:** Individual differences in purchase propensity and marketing response; **Where:** Email vs. catalog
Zhan et al. (2024)	Offline (mail)	Coupons	✓				**Whom:** RFM data (recency, frequency, monetary value), spending during holiday season
Zhang and Krishnamurthi (2004)	Online (e-store)	Price promotions	✓	✓	✓		**Whom:** Price sensitivity and variety-seeking; **What:** Optimal price discount; **When:** Dynamically update promotions

Notably, several studies address multiple decisions (whom, what, when, where) simultaneously, demonstrating the interdependencies between these factors. For instance, Bleier and Eisenbeiss (2015) finds that the effectiveness of personalized banners depends on timing and placement factors. Specifically, banners with a high degree of content personalization are most effective when displayed shortly after a consumer visits the advertiser's website but quickly lose their impact over time. Together, these examples illustrate the breadth of (hyper-)personalization in marketing.

4. Causal foundations of personalization in marketing

A major development in academic studies of personalization in recent years has been the formalization of personalization as a decision problem rooted in causal inference (Ascarza, 2018; Daljord et al., 2023; Ellickson et al., 2023; Gordon et al., 2023a, 2019; Hitsch et al., 2024; Lemmens & Gupta, 2020; Yoganarasimhan et al., 2023). This shift has transformed the field from primarily descriptive analyses that focus on observing differences between customers to normative approaches that guide how interventions should be personalized. By explicitly modeling and predicting how customers respond to marketing actions, this framework enables firms to *optimize* the intended objectives of personalization, such as profits. In this section, we formally define the personalization problem using the notation and concepts from causal inference that have become standard in the personalization literature.

Personalization involves selecting an *action* (a.k.a. *treatment* or *intervention*) W_i for each individual $i = 1, \ldots, N$. The action W_i is expected to have a causal effect on an outcome of interest, Y_i, which the firm aims to optimize. For instance, a company may personalize coupon offers, providing different discount levels to customers based on their predicted redemption rates or net margin. The company determines the optimal action based on the individual's feature vector X_i, which includes all relevant characteristics associated with that individual, such as demographics or past behavior, as well as other contextual variables. The action $W_i = w$ is chosen from a predefined action set $w \in \mathcal{W}$. In the coupon example, each w might represent a different discount level for a coupon. Associated with each action w is a *potential outcome*, $Y_i(w)$, which represents the level of Y_i that would arise if W_i were set to $W_i = w$.

Formally, a *personalized policy*, p, is defined as a mapping from the company's current state of information, $X \in \mathcal{X}$, to the set of actions $\mathcal{W}$, so that each X_i is associated with exactly one action W_i (Rafieian & Yoganarasimhan, 2023),

$$p : \mathcal{X} \rightarrow \mathcal{W}. \tag{1}$$

Whichever policy p maximizes the relevant objective (conditional on X) is said to be *optimal* and is denoted by p^*. The effectiveness of personalized policies depends on the degree of heterogeneity in individuals' responses to actions and the firm's ability to identify this heterogeneity.

A common approach to determining a personalized policy relies on estimation of *conditional average treatment effects* (*CATEs*). The CATE of taking the action $W_i = w$ against an individual i characterized by $X_i = x$ is defined as

$$\tau_w(x) = \mathbb{E}\left[Y(w) - Y(w_0) \mid X_i = x\right], \tag{2}$$

where w_0 represents the *baseline action*, typically representing the status quo, or "business as usual." For example, in the coupon example, w_0 could represent the action "no coupon is sent." The CATE then represents the difference in expected outcomes when the customer receives a coupon offering a discount $W_i = w$, versus not receiving a coupon offer, $W_i = 0$.

The total expected value or *return* of a policy p can be expressed as a value function that sums the CATEs across all individuals assuming their values of W_i are determined by the policy p,

$$R(p) = \sum_{i=1}^{N} \tau_{p_i}(X_i = x), \tag{3}$$

where p_i denotes individual i's assigned action W_i under policy p. The optimal policy p^* maximizes this value function:

$$p^* = \arg\max_p R(p). \tag{4}$$

Note that the value function $R(p)$ must also incorporate the expected cost associated with executing the policy. Costs that can be attributed to specific actions $W_i = w$ are typically incorporated directly into $\tau_w(x)$, whereas more complex cost structures may be more parsimoniously included in the definition of $R(p)$.

In practice, personalized policies can be designed by predicting CATEs, using either experimental (randomized controlled trials) or observational data that capture X_i, W_i, and Y_i for a large number of customers. The next section focuses on the experimental approach to personalization.

5. Test and learn

Even though personalized policies can be developed using observational data (Ellickson et al., 2023), our focus in this article is primarily on using experiments to improve personalization. Whether using observational or experimental data, a fundamental requirement within the test and learn cycle is being able to causally predict how a particular policy (e.g., personalized discounts) will impact an outcome of interest (e.g., individual sales). Observational and experimental approaches to causal prediction differ substantially in their implementation, and each has its own strengths and weaknesses. For example, an ineptly designed experiment might have a devastating impact on customers; by comparison, passively analyzing observational data is less risky.

Here we emphasize experimentation for a few reasons. First, observational data may simply be unavailable. Second, even if available, observational data may lack sufficient variation to make causal predictions, especially if new customer interventions are being considered. Third, and perhaps most importantly, the causes and effects in observational data may be jointly determined by confounding factors in the environment that cannot be adequately controlled for during the analysis. Experiments may overcome these limitations by controlling the policies to be evaluated and substituting randomization for an otherwise unknown and uncontrolled data generating process.

With that said, we do not want to downplay the importance of observational causal inference for personalization. It is not always feasible (or ethical) to conduct a randomized experiment. In such cases, observational studies may be the only viable options, although

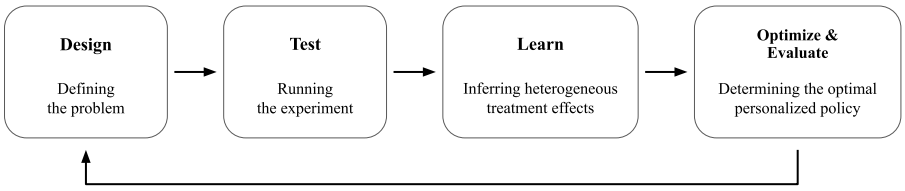


Fig. 1. Test and Learn.

quasi-experimental approaches might also be possible alternatives (Goldfarb et al., 2022). Hence, in Sections 6.1.3 and 7, we discuss adaptations for learning from observational data, including double machine learning and direct policy learning. Moreover, approaches to causal inference that use pre-experiment data can be invaluable tools for designing more efficient experiments.

As shown in Fig. 1, a typical test and learn cycle comprises four stages: design, test, learn, and optimize/evaluate. We briefly introduce each of these steps to connect them with the formal model of personalization (see Section 4), and to provide an organizing framework to introduce practical considerations related to personalization in Section 6. Below, we illustrate each step of the cycle with a running example of a *food delivery service* seeking to optimize revenue through promotional offers in the form of coupons. We also provide a code tutorial (R and Python) in which we implement each step of test and learn.[2]

5.1. Design

The goal of the design phase is to identify the relevant customer population, personalization decisions (i.e., treatments), outcomes, value function, data, and business rules or constraints. In the earlier coupon example, say that the company is a food delivery service, and the goal is to optimize discount offers. The set of personalization decisions might then be the discount level, expressed as a percent of the total order value: $W_i \in \{0\,\%, 5\,\%, 10\,\%, 15\,\%\}$, with $W_i = 0\,\%$ representing the baseline treatment of "no discount." A business rule requires discounts to be redeemed within 24 hours of the offer, and in this context, there is a belief that discounts may lead to forward-purchasing. The outcome of interest Y_i is therefore defined as the total margin from customer i, net of the discount, for all orders placed within seven days of the discount offer.

The relevant population comprises all users of the company's mobile app. Each customer has a rich set of data, X_i, including their history of interactions with the service's mobile app, past orders, and previous responses to promotional offers. Twice per day—once at 11 a.m. and again at 5 p.m.—a *random* subsample of customers receives a mobile app notification, such that after three weeks, every customer has received exactly one app notification. For each customer, a discount level is chosen at random; customers with $W_i = 0\,\%$ receive the app notification, but are not offered a discount. To improve the precision of the CATE estimates, stratified sampling is employed during random assignment (Imbens & Rubin, 2015). This entails dividing customers into discrete subgroups (strata) based on values of X_i, and randomizing W_i within each stratum.

The experimental design should anticipate later stages in the test and learn cycle, especially the causal assumptions necessary for CATE estimation. For example, one of these assumptions is the stable unit treatment value assumption (*SUTVA*), which says that an individual's potential outcomes, $Y_i(w)$, are unchanged by the act of assigning a particular treatment (to customer i or any other individuals). In the running example, SUTVA would require the amount spent by a customer offered a $w\%$ discount to be unaffected by their spouse receiving their own discount offer. Anticipating the need for SUTVA, the experiment's randomization procedure is constrained so that individuals from the same household never receive discounts within one week of each other.

Beyond SUTVA, the estimation of causal effects also relies on two other key assumptions, which are satisfied by design in properly randomized experiments, but must be carefully considered in observational studies. First, *unconfoundedness* requires that treatment assignment—the collection of randomized values of W_i—is independent of potential outcomes (conditional on observed covariates X_i). Unconfoundedness ensures that any difference in observed outcomes can be attributed to the action W_i, rather than to pre-existing, and possibly unobserved, customer heterogeneity. In the running example, it may be that in the past, discounts were assigned to individuals expected to spend the most with the discount (independent of how much more the discount caused them to spend). The offer of a discount would thus be confounded with an individual's treated potential outcomes. By replacing the decision to offer a discount with a random number generator, the unconfoundedness assumption is likely to hold.

Second, the *overlap* assumption requires that every individual has a non-zero probability of receiving each possible treatment. Without overlap, there exist subgroups for which certain values of $w \in \mathcal{W}$ are never observed in data, making it impossible to estimate causal effects for those groups of customers. In the food delivery example, imagine that it is decided that any customer who has not placed an order within two months will be automatically offered a discount of $W = 10\,\%$. In that case, the company will be able to estimate, for this subset of dormant customers, the average margin after receiving a 10 % discount. But it will not be able to estimate the average margin for this group in the absence of a discount offer. Without an estimate for the average margin when $W = 0\,\%$, the company cannot infer how the 10 % discount affected customer margin. Failure to satisfy the overlap assumption can arise in both experimental (if randomization is not properly executed) and observational contexts (if some groups always or never receive a treatment).

[2] https://github.com/sbstn-gbl/cs23-personalization

5.2. Test

The test phase involves running the experiment; in our example, it means randomizing $W_i \in \{0\,\%, 5\,\%, 10\,\%, 15\,\%\}$ among the customer population defined in the design phase. However, before implementing the actual experiment, it is advisable to run a placebo test, often referred to as an "A/A" test, in which all assigned treatments are identical (Kohavi et al., 2020). For example, randomization to different values of W_i would take place, but all randomized values of W_i are then replaced with the same discount level (e.g., $W_i = 0$). If any significant differences in outcomes across treatment groups are then observed, it suggests potential flaws in the experimental setup, such as incorrect randomization or software bugs.

In the food delivery service example, the company runs the experiment for four weeks (randomized offers for three weeks, plus a fourth week of data collection). During the experiment, continuous monitoring ensures that app notifications are being sent correctly, that randomization is carried out successfully, and that there are no other problems. Monitoring is particularly critical in complex environments, such as multi-channel campaigns, where deviations from the planned interventions can bias estimates and reduce the reliability of the findings, or overlapping promotions or experiments can introduce spillover effects or unintended treatment exposures. These interfering efforts can potentially bias or invalidate experimental results.

At the conclusion of the experiment, each customer has an observed value of X_i, W_i and Y_i. For each customer, their observed outcome is equal to the potential outcome corresponding with their treatment, $Y_i = Y_i(W_i)$.

5.3. Learn

In the learning phase, experimental data are leveraged to infer heterogeneous treatment effects (CATEs) so that they can inform the design of an optimal personalized policy. In the food delivery service example, the company might want to personalize discounts based on two potential X_i variables: household size and average order value. Accordingly, data from customers sharing identical values of X_i are pooled to estimate conditional-on-X_i CATEs.

One of the simplest approaches for estimating CATEs is using simple averages,

$$\hat{\tau}_w(x) = \hat{Y}_{w,x} - \hat{Y}_{0,x}, \tag{5}$$

where $\hat{Y}_{w,x}$ is the average observed outcome among individuals with $X_i = x$ under treatment $W_i = w$, and serves as an estimator of $E[Y_i(w) \mid X_i = x]$. Similarly, $\hat{Y}_{0,x}$ is the average outcome among those in the baseline group (e.g., those who did not receive an offer, $W_i = 0$), and serves to estimate $E[Y_i(0) \mid X_i = x]$. Because the treatment variable W_i is randomized conditional on X_i, the difference in the two group averages is a consistent CATE estimator. The aforementioned overlap assumption is especially important for establishing the validity of CATE estimates. If all customers with $X_i = x'$ (where x' indicates no recent purchases) are assigned $W = 10\,\%$, then the company can estimate $\hat{Y}_{10\,\%,x'}$, but not $\hat{Y}_{0\,\%,x'}$ or $\hat{\tau}_{10\,\%}(x')$.

Although conceptually simple, the simple difference-in-means estimator in Eq. (5) can produce noisy estimates for $\hat{\tau}_w(x)$, for example when the proportion of treated to untreated observations is very close to zero or one, or when the variance of the outcome variable is especially high. An alternative CATE estimator is the *causal forest* (Athey et al., 2019b), which has been used extensively in marketing (Chen et al., 2020; Esterzon et al., 2023; Gelper et al., 2024; Habel et al., 2024; Huang & Ascarza, 2024; Rafieian et al., 2025; Turjeman & Feinberg, 2024; von Zahn et al., 2024). Like traditional random forests, causal forests are ensembles of decision trees. In a typical random forest, individual trees are built by recursively splitting the data based on values of X with the goal of minimizing prediction error for Y. Causal forests, however, are optimized for CATE prediction. As such, the individual trees in a causal forest are constructed in order to maximize the variance in estimates of $\tau_w(x)$. This produces a model that minimizes prediction error for $\tau_w(x)$.[3] The successful adoption of causal forests in marketing is undoubtedly due to its known statistical properties, which contrasts with the machine learning approaches previously used in marketing. Causal forests achieve $\sqrt{N}$ consistency for the CATEs, and their residuals are asymptotically normal, allowing for the construction of confidence intervals around estimated CATEs. This is particularly useful for researchers because true treatment effects can never be observed.

5.4. Optimize and evaluate

In the last stage of the test and learn cycle, CATE predictions are used to derive the personalized policy. In the running example, a customer with a particular household size and average order value $X_i = x$ would be offered a discount w only if $\hat{\tau}_w(x) > 0$—meaning the expected net value to the company from the discount offer is positive. As there could be many discount levels that are profitable for a customer with $X_i = x$, offers should be set to the discount level w with the largest estimated CATE. Hence, under SUTVA, the optimal policy for this promotion, p^*, would be a function that takes customers' information X as its input, and returns discount levels that maximize $\hat{\tau}_w(X_i)$ for each customer i. That is, $p_i^* = \arg_w \max \hat{\tau}_w(X_i)$. It is extremely common that for the majority of customers, $p_i^* = 0$, meaning *not* offering a discount is optimal.

Even after an optimal policy is derived based on estimated treatment effects, it is essential to evaluate how well the policy performs in practice. This is because the estimated treatment effects, $\hat{\tau}_w(X_i)$, are only approximations of the true effects. They are

[3] Because the true CATEs are unobserved in data, they must be approximated. This is accomplished by using a subset of observations to construct estimators for $\tau_w(x)$, and the remaining observations to calculate $\hat{\tau}_w(x)$ for the purpose of splitting the decision tree. Causal forests are thus a type of direct estimator, which we discuss in Section 6.1.3.

subject to sampling variability, model misspecification, and limitations in the available data. As a result, the policy derived from these estimates may not achieve the desired outcomes when implemented. To assess the effectiveness of a proposed policy, researchers and practitioners must carry out a policy evaluation step. There are two main approaches to doing so: *online* policy evaluation and *offline* policy evaluation (also called on-policy and off-policy). Online policy evaluation typically entails implementing the new policy and the business-as-usual policy separately against two randomly selected customer samples (Kumar et al., 2008). In the running example, this would entail running an A/B test comparing the profitability of offering discounts according to the new policy p^* versus some pre-existing policy p^0 (e.g., not offering any discounts). Note that the comparison is between *policies*—rules for which discounts are offered to which individuals—rather than particular *treatments*.

Online evaluation can be impracticable. Convincing a company to actually implement a policy without a guarantee of success is difficult, especially when it involves implementing multiple policies and comparing them, or when the intervention is costly (e.g., promotions) or conveys the risk of negative treatment effects (e.g., reactance). In such cases, offline evaluation is an attractive alternative. Offline evaluation entails evaluating a new policy using historical data, typically the experimental data used to derive the new policy. Treatments that correspond with a policy's output—for example, observations for which the randomized discount level turned out to also be the optimal discount—can be used to estimate the value generated by the policy. Because this approach does not make full use of the experimental data, the comparison may be inefficient, in which case a popular alternative is to use an inverse propensity weighted (IPW) or Horvitz-Thompson estimator (Horvitz & Thompson, 1952) for offline evaluation. These estimators re-weight observations by the inverse of their propensity to receive the observed treatment. Several variants of this estimator have been used for offline policy evaluation in marketing (Ascarza, 2018; Ascarza & Israeli, 2022; Esterzon et al., 2023; Hitsch et al., 2024; Rafieian & Yoganarasimhan, 2021; Simester et al., 2020).[4]

As Fig. 1 shows, the cycle can now repeat, with the information gained from the experiment becoming part of an updated X in the next iteration. The test and learn cycle presented here is intentionally simple, but can be made more complex in order to improve efficiency. In particular, estimates of $\hat{Y}_{0,x}$ or $\hat{\tau}_w(x)$ obtained in the current cycle can be used for stratified randomization in the next. Even more complex approaches allow for adaptive experimentation that balances profit maximization with information gain during the experiment (Rafieian, 2023; Wang et al., 2023).

Having outlined how personalization aligns with the test and learn framework, we next turn to the main academic challenges that arise from personalization.

6. Challenges arising in the new era of personalization

Personalization is a complex task that presents many challenges and opportunities for academic research. Here, we discuss four key areas addressed in the recent literature: (1) predicting customer responses with limited data, (2) identifying causality amid numerous influencing factors, (3) evaluating policies in line with business objectives, and (4) ensuring ethical and responsible personalization (see Table 3). We outline real-life challenges, the research problems they raise, and recent advances in academic literature that offer potential solutions. We then consider avenues for future research in Section 7.

6.1. Predicting customer responses with limited data

To develop personalized policies, firms depend on data describing individual customers. Decades of marketing research have demonstrated the value of behavioral and preference data over demographic data (e.g., Matz & Netzer, 2017), access to which has been greatly expanded by the digital revolution. Variables such as customers' past responses to marketing treatments, purchase histories, and product reviews are both highly informative and increasingly commonplace.

In spite of this, predicting customer responses to personalized marketing actions often meets with severe limitations. Here we highlight three particularly challenging instances: the cold-start problem (6.1.1), enhanced data regulation and data loss (6.1.2), and efficient CATE estimation (6.1.3). Although not exhaustive, these challenges exemplify broader issues faced by marketers when accessing and utilizing data effectively, as well as showcase recent advances in the literature.

6.1.1. Cold-start problem when targeting new customers

Personalization is inherently challenging when a company has little or no data about its customers, a situation commonly referred to as the *cold-start problem*. This issue arises in customer acquisition contexts, where prospective customers lack historical interaction data (Ahn, 2008), but also when targeting newly acquired customers in a customer development context (Padilla & Ascarza, 2021). The data sparsity challenge becomes even more pronounced in personalized online recommendations, especially when new items—such as news, music, or movies—are frequently added to the catalog. In these cases, companies face a dual data gap: limited information on new customers, and uncertainty about their responses to novel items (Xu et al., 2023).

Probabilistic machine learning offers a compelling approach to addressing the cold-start problem. Whereas the statistical models that have been standard in marketing for many decades often depend on latent variable representations of data, machine learning techniques potentially allow for considerably more flexibility in estimation and prediction, potentially capturing complex behavioral

[4] Although simple IPW estimators improve efficiency, they do not fully mitigate the high variance associated with offline policy evaluation; the development of better estimators for offline policy evaluation is an active area of research in causal inference and computer science (Sachdeva et al., 2024; Uehara et al., 2022).

Table 3
Personalization challenges and state-of-the-art solutions.

Type of challenge	Challenge	State-of-the-art solutions
Predicting Customer Responses with Limited Data (6.1)	Cold-Start Problem When Targeting New Customers	- Probabilistic machine learning - Reinforcement learning
	Enhanced Data Regulation and Data Loss	- New zero-, first-, and second-party -data sources
	Efficient CATE Estimation	- Causal machine learning - Use of surrogates
Identifying Causality Amid Numerous Influencing Factors (6.2)	Treatment Spillovers and Interference	- Cluster randomization - Randomized clustering - Variance-minimization methods
	Non-Compliance in Marketing Interventions	- Local average treatment effects - Marginal treatment effects
	Personalization in Changing Market Environments	- Adaptive experiments - Meta-analyses
Evaluating Policies in Line with the Business Objectives (6.3)	Aligning Evaluation with Business Objectives	- Off-policy evaluation - Rank-weighted average treatment effects
	Measuring Long-Term Effect of Personalized Policies	- Surrogate models
Ensuring Ethical and Responsible Personalization (6.4)	Fair Personalization	- Fair treatment effect estimation - Fair active learning
	Privacy-Aware Personalization	- Differentially private CATE estimation
	Transparent Personalization	- Explainable AI - Features of the CATE

patterns with minimal assumptions. Machine learning algorithms can also address data scarcity by enabling data fusion, allowing the model to infer information about new customers from related data sources, such as the behavior of similar customers.

A number of recent approaches based on probabilistic machine learning have been proposed to address the cold-start problem. Padilla and Ascarza (2021) introduces the First-Impression Model, which leverages data collected at the moment of acquisition (e.g., the first transaction) to infer latent customer traits predictive of future behavior. This approach identifies high-value customers and pinpoints those most responsive to marketing actions immediately after their first purchase.

Other approaches draw more directly on the reinforcement learning literature to efficiently learn about (new) user preferences. Xu et al. (2023) proposes a method for personalized recommendations that combines collaborative filtering with multi-armed bandits, enabling systems to make initial recommendations and iteratively refine them by learning user preferences over time. The approach integrates user demographics and item attributes into the collaborative filtering process, enhancing prediction accuracy by utilizing auxiliary information in settings where historical user-item interaction data is limited. Liberali and Ferecatu (2022) develops a framework that combines multi-armed bandits with Hidden Markov Models to personalize website content in real time. The model learns latent states of visitor engagement in order to serve content expected to maximize user engagement and conversion rates.

6.1.2. Enhanced data regulation and data loss

Recent regulatory changes have significantly restricted companies' access to customer data collected by other firms. Such *third-party data* is often aggregated by external entities, such as ad networks and data brokers, and then sold to firms for personalization purposes. Apple now requires explicit user consent for cross-app tracking; Google offers users greater control over third-party cookies and has announced plans for a Chrome feature allowing users to mask their IP addresses (Joseph & Peterson, 2025). Enhanced privacy laws, including the General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), and California Privacy Rights Act (CPRA), have further tightened data-sharing practices by requiring user consent and restricting collection of sensitive information.

While these privacy protections may benefit consumers, they also limit marketers' ability to track customer behavior across platforms, making it harder to deliver personalized ads and promotions. Moreover, they may disproportionately affect smaller firms, as well as companies that sell through external partners, as they tend to have relatively less customer data on consumer interactions to work with. By contrast, major platforms and retailers, such as Google, Meta, TikTok, and Walmart, engage directly with large numbers of consumers and can build a formidable data advantage. Coping with data loss is now a major priority for both companies and marketing scholars (Marketing Science Institute, 2024, pages 4–5).

Data collection strategies are changing in response to new privacy constraints. One approach is to collect so-called *zero-party* data by setting up private customer panels or reward-based data-sharing programs that encourage users to *voluntarily* disclose their preferences, interests, and purchase intentions. Amazon's Shopper Panel offers monetary incentives to customers who upload offline purchase receipts, giving Amazon a view into customer share-of-wallet. Cashback apps like Dosh and Fetch collect purchase data from consumers in return for rewards. Another strategy is to strengthen the collection of *first-party data*, data generated passively through

normal customer interactions with the firm, by shifting toward direct-to-consumer (DTC) models. Such companies gain access to a rich array of data from website visits, purchase histories, and other customer interactions. Yet another tactic is to expand data-sharing agreements with trusted partners. These agreements, such as one between supermarket chain Albert Heijn and online retailer bol.com in the Netherlands, allow partner companies to leverage each other's *second-party data* to improve customer insights and refine personalization efforts.

In addition to finding new ways to collect data, companies can also enhance the value of data already collected. Padilla and Ascarza (2021) shows how to infer preferences by dynamically integrating first-party click stream data, past customer journeys, and the behavior of similar customers. By capturing evolving preferences as the customer journey unfolds, the model facilitates personalization in contexts with limited prior data.

A type of first-party data, which for many companies remains an underutilized resource, is unstructured data, such as text and images (Berger et al., 2020; Compiani et al., 2023). Recent advances in off-the-shelf procedures for converting unstructured text and images into high-dimensional vectors called *embeddings* have paved the way for using such data in personalization. For instance, Veitch et al. (2020) adapts deep language models (e.g., BERT) to learn low-dimensional embeddings from long Reddit posts and estimate the causal effect of their linguistic properties. Embeddings reduce the dimensionality of unstructured data while maintaining much of their richness. Their approach can potentially be extended to estimate heterogeneous treatment effects of personalized communications in a marketing context.

6.1.3. Efficient CATE estimation

Marketing data is inherently noisy due to the numerous unobservable factors that influence customer behavior. This noise complicates the estimation of CATEs, especially when dealing with outcomes that are difficult to predict or accurately measure. A growing body of research spanning marketing, economics, statistics, and computer science seeks to improve CATE estimation in both experimental and observational settings. Given the importance of the test and learn cycle for personalization, we focus on the challenge of estimating CATEs from experimental data. Challenges related to missing data, rare events, endogenous treatments, and massive data sets (regarding rows or columns) also affect personalization; here, we emphasize issues with the strongest affinity with the test and learn cycle.

Indirect approaches to CATE estimation start with a model of the outcome Y as a function of a treatment variable W and covariates X. This is almost always an additive error model, $Y = \mu(W, X) + \epsilon$, where $\mu(W, X)$ is the conditional mean of Y. After model training, a CATE estimate can be obtained by predicting outcomes for customers sharing the same covariates $X = x$ under i) a treatment of interest, $\hat{\mu}(w, x)$, and ii) a baseline condition, $\hat{\mu}(w_0, x)$. Under appropriate causal assumptions, the difference in the two predicted levels of Y corresponds with the CATE for that group of customers: $\hat{\tau}_w(x) = \hat{\mu}(w, x) - \hat{\mu}(w_0, x)$.

If the expected outcome levels, $\mu(w, x)$, are estimated in a single regression model, wherein W is treated like any other covariate, then the model is called an *S-learner* (Künzel et al., 2019, **S** denoting a **s**ingle regression model). Returning to the example from Section 5, a food delivery service uses coupons (the treatment) to increase revenues (the outcome). For CATE estimation, an S-Learner models revenues as a function of the provided discount level W and other covariates X. For inference, the company predicts revenues at different discount levels (e.g., a discount of 15 % and a discount of 0 %), conditional on X, and obtains the CATE as the difference in the two predicted outcome levels. A drawback of S-learners is that they might ignore contributions from the treatment variable when the treatment effect is weak relative to the correlation between the outcome and the covariates, biasing predicted treatment effects towards zero (Künzel et al., 2019).

An alternative is to train a set of *T-learners* on subsets of data corresponding with each (discrete) treatment, $\mu_0(x), \mu_1(x), \mu_2(x), \ldots$ (**T** denoting **t**wo separate regression models when treatment is binary-valued). In our running example, the company splits the data according to the observed discount levels and trains separate models on the covariates X in each data set. During inference, the company can then use the models corresponding to the discount levels of interest to derive the corresponding conditional-on-X revenue predictions and derive CATEs as the differences in the two predicted outcome levels. Although T-learners help to alleviate problems associated with S-learners, T-learners may also introduce their own biases if the outcome variable is nonlinear and sample sizes differ across treatment conditions (Künzel et al., 2019). Nevertheless, indirect learners are straightforward to implement and can range in complexity from the simple difference-in-means estimator in Eq. (5) to deep neural networks.

Given that the goal of indirect methods is CATE estimation, there is a discrepancy between how these models are trained and what they need to accomplish. Indirect estimators are trained by minimizing differences (i.e., MSE) between observed outcomes Y and model predictions $\hat{\mu}(w, x)$. Ideally, a single model would be trained to minimize the difference between true CATEs, $\tau_w(x)$, and model predictions, $\hat{\tau}_w(x)$, but the fundamental unobservability of CATEs in data prevents this. Hence, indirect estimators are inefficient because they are trained to perform the wrong task.

Direct approaches to CATE estimation seek to address this problem. The causal forest (Wager & Athey, 2018) described in Section 5.3 is a well-known estimator that uses a direct approach. Another direct approach is implemented in a two-stage estimator called an *X-learner* (Künzel et al., 2019). In the first stage, an S- or T-learner is trained to estimate $\mu(w, x)$. These predictions are then used to generate proxy outcomes, $\tilde{\tau}_i$, which stand in for the unobserved CATEs in a second-stage regression of $\tilde{\tau}_i$ on X_i. By splitting the training sample, some observations are used to train a model that generates $\tilde{\tau}_i$, and the remainder are used to train a model that predicts τ_i; no observations are used twice (Chernozhukov et al., 2018; Nie & Wager, 2021).[5] The proxy outcomes, $\tilde{\tau}_i$, are derived

[5] This approach is called *cross-fitting*. Causal forests require a similar sample split, although this is provided automatically by out-of-bag observations—the observations that were not bootstrap sampled to train a given base regression tree.

by reweighting the residuals from the first stage in proportion to their inverse propensity scores, and the second-stage model may further transform the treatment variable W so that it is orthogonal to the covariates X (Chernozhukov et al., 2018; Kennedy, 2023; Nie & Wager, 2021). Orthogonalization and reweighting improve the robustness and accuracy of CATE predictions in settings where treatment assignment is imbalanced and a propensity score is known or can be estimated. Recent marketing applications include Ellickson et al. (2023) and Gordon et al. (2023c).

Finally, much of the aforementioned work in improving CATE estimation has focused on settings lacking a significant time-series component. When outcomes are observed over time, the noisiness in the outcome variable accumulates, but the fundamental treatment effect does not. Huang and Ascarza (2024) provides a theoretical analysis demonstrating how noise in long-term outcome variables, such as repeat purchases, can significantly reduce the precision of widely used CATE estimators, ultimately diminishing the effectiveness of personalized marketing policies. This study introduces a surrogate approach that utilizes short-term outcomes as lower-noise proxies for predicting long-term behavior. Estimating CATEs based on this proxy offers more effective personalized policies than directly estimating CATEs on the noisy long-term outcome.

6.2. Identifying causality amid numerous influencing factors

Personalization strongly relies on the ability to predict heterogeneous treatment effects. In addition to the aforementioned problems stemming from limited data, additional challenges to CATE estimation arise due to the complexity of real-world marketing environments. In this section, we describe three of these: spillover effects and interference (6.2.1), non-compliance (6.2.2), and changing market environments (6.2.3).

6.2.1. Treatment spillovers and interference

As mentioned in Section 5, most experimental methods are predicated on the assumption that SUTVA is satisfied (Rubin, 2005). Yet, in many marketing settings, this assumption may not be appropriate. Firms and customers interact with each other and the broader environment, creating spillovers and interferences across different marketing actions that can bias CATE estimates. For example, customers may pass personalized offers to their friends by word of mouth (Peres et al., 2010). Likewise, customer churn is highly influenced by friends' churn decisions, such that a personalized retention campaign may affect not only targeted individuals, but also customers in their network (Ascarza et al., 2017; Godinho de Matos et al., 2018; Nitzan & Libai, 2011). Similarly, campaigns and experiments running simultaneously can interfere with one another, making it difficult to attribute treatment effects to a single marketing action.

To avoid such interferences, Huang et al. (2020) limits users' ability to see more than one ad at a time, and many larger companies require field experiments to be centrally registered and coordinated. At the design phase of an experiment, researchers can use interference-minimizing approaches such as (graph) cluster randomization (Eckles et al., 2017; Ugander et al., 2013), randomized clustering (Ugander & Yin, 2023), and variance-minimization methods (Candogan et al., 2024). After the experiment, interference can be detected using conditional randomization tests (Aronow, 2012). Finally, the risk of biased CATE estimation can be reduced by modeling network dependencies directly (Athey et al., 2018).

6.2.2. Non-compliance in marketing interventions

An ideal marketing experiment uses randomization to prevent confounding, but users do not always receive the treatment they were randomly assigned to. For example, if a user is assigned to "treatment" in a paid media campaign, that user must visit the media site, and an ad slot must be allocated to the advertiser in order for the individual to be shown the ad (Gordon et al., 2019; Johnson et al., 2017; Waisman & Gordon, 2024). The advertiser thus controls treatment *assignment*, whereas *receiving* the treatment is not fully under their control. Since media is paid for when the ad is shown, advertisers want to know the effect of *receiving* the treatment. This situation is known in the experimental literature as *one-sided non-compliance*. The key challenge under one-sided non-compliance is that the received treatment has not been fully randomized, and the non-compliance creates a potential confound. For example, customers in the treatment group who are more likely to actually receive the ad might also be less sensitive to it. This violates the *unconfoundedness* assumption required to estimate heterogeneous CATEs.

One solution to non-compliance is to estimate conditional *intent-to-treat effects* for each customer type—causal effects due to the assigned treatment, as opposed to the received treatment. Ascarza (2018) compares retention probabilities among customers who are sent thank-you emails to those customers who are not, thus estimating the ITT for *sending* (as opposed to *reading*) the email. Conditional ITTs are relevant for the personalization of direct mail when companies incur costs for sending mail, even if customers do not read it (Hitsch et al., 2024; Zantedeschi et al., 2017). In many settings, conditional ITTs are the only treatment effects that can be estimated as the company does not know whether the assigned treatment was received (e.g., whether the mail was opened).

If the company *does* observe which customers receive treatment, it may be possible to estimate causal effects of *receiving* treatment among individuals in the treatment group—a treatment effect called a conditional average treatment effect on the treated (CATT). In a standard randomized experiment where non-compliance is only possible in the treatment group (i.e., one-sided), CATTs can be estimated by dividing the conditional ITT by the proportion of customers who actually receive the treatment within the subgroup (Gordon et al., 2019). This estimation strategy is equivalent to using the randomized treatment assignment as an instrumental variable (IV) for the received treatment (Angrist et al., 1996).

Two-sided non-compliance arises when all customers are able to receive treatment, not just those assigned to the treatment group. For example, Daljord et al. (2023) considers a field experiment with an encouragement design, wherein some customers are offered

a random number of bonus loyalty reward points for staying at a predetermined number of hotels (called a *hurdle*) during the promotional period. In this case, the causal effect of most interest is the conditional local average treatment effect (CLATE), which represents the increase in hotel visits due to reaching or surpassing the spending hurdle among the subset of customers who were motivated enough by the incentive to reach or surpass the hurdle, but otherwise would have spent below it.[6] This represents the incremental effect that the program had on hotel visits, which can be compared to the cost of the program. Another example of two-sided noncompliance in marketing is Gabel et al. (2024a), which uses coupon exposure as an incentive to encourage the purchase of a brand, and estimates the causal effect of this brand purchase on subsequent (non-coupon) repurchases (i.e., structural state dependence).

By definition, CLATEs are based on the differences in discrete levels of the randomized incentive to receive treatment. When the goal is to personalize the outcome-maximizing level of the reward, it becomes desirable to evaluate treatment effects corresponding with previously unseen levels of the randomized incentive. This can be accomplished by estimating a generalized version of the CLATE, called the *marginal treatment effect* (MTE, Heckman & Vytlacil, 2007), as it provides a coherent framework for interpolation or extrapolation of treatment effects under these conditions. MTE estimation is greatly aided by having many values of the randomized incentive to receive treatment, and when designing such experiments, the company should randomize as many levels of the incentive as possible (Daljord et al., 2023).

6.2.3. Personalization in changing market environments

Predictions of heterogeneous treatment effects based on an experiment are useful for personalizing marketing campaigns only if the environment during campaign implementation is sufficiently similar to the environment during the experiment. Changing market conditions can lead to *concept drift*, threatening the generalizability of experimental findings to new contexts and undermining the effectiveness of personalized policies. Concept drift occurs when the relationship between covariates, treatments, and outcomes evolves over time, possibly due to external shocks. The recent energy crisis, for example, caused significant changes in electricity prices and likely affected consumer price elasticities and responsiveness to discount promotions (Jamissen et al., 2024). Even in the absence of external shocks, treatment effects may exhibit seasonal variation or be sensitive to broader contextual influences that are not typically observed in data (Andrews et al., 2016; Blake et al., 2015). For example, other advertisers' activities can influence treatment effects, but may not be recorded in the company's data (Waisman et al., 2024).

If these changes in the environment are observed, they can be included as covariates in the CATE estimates. Otherwise, more sophisticated methods are needed. For instance, Yang et al. (2024) proposes an adaptive experimentation approach where data are iteratively added over time to account for changes in the environment, thus enabling robust personalization.

Another approach to ensure that treatment effect predictions generalize across environments involves synthesizing insights from multiple experiments. Huang et al. (2024) introduces incrementality representation learning, a framework designed to predict CATEs for a wide variety of marketing actions. The method extracts a low-dimensional representation of marketing actions and customer segments based on past experiments. This facilitates the efficient design of personalized interventions while ensuring generalization to previously untested actions and customer segments.

Meta-analyses play an important role in evaluating the generalizability of models and findings by investigating effect distributions across many brands or categories (e.g., Gabel et al., 2024b; Kalyanam et al., 2018; Shapiro et al., 2021; Simonov et al., 2018; Wernerfelt et al., 2025). The scale of these studies allows researchers to uncover systematic patterns and differences in treatment effects by applying meta-analytic frameworks to cross-sectional and longitudinal measurements of treatment effects. For example, Meager (2019) proposes a (hierarchical) Bayesian meta-analysis model to statistically aggregate individual measurements. Gabel et al. (2024a) uses this framework as a principled evidence aggregation approach to provide generalizable conclusions about structural state dependence obtained from a coupon field experiment with 223 brands. This meta-regression also identifies covariates that explain systematic differences in treatment effects across brands and categories. An extention to the meta-analysis calculates the correlation between experimental measures of state dependence and results from structural brand choice models (Matzke et al., 2017), making it possible to evaluate the consistency of the two approaches while accounting for the effects of measurement error. This example illustrates that combining large-scale field experiments with meta-analytic modeling approaches can enable researchers to evaluate how contextual factors such as brand characteristics, customer demographics, and competitive environments moderate treatment effects, supporting the generalizability of findings.

6.3. Evaluating policies in line with the business objectives

Policy evaluation is essential for determining how well a policy performs and how it compares to the alternatives. One challenge is to articulate the goal of personalization, or *policy objective*, and then choose an appropriate numerical summary, or *evaluation metric*, that measures how well the policy performs (6.3.1). A second challenge arises from the fact that the impact of personalized policies often takes time to manifest, in particular in contexts where firms want to maximize long-term outcomes (6.3.2). These challenges raise complex research questions.

[6] The treatment effect of interest under one-sided non-compliance is also a CLATE, although in that case, the CLATE is equivalent to the CATT, and is usually referred to as such. Under two-sided non-compliance, the CLATE is *not* equal to the CATT without additional, and often strong, assumptions.

6.3.1. Aligning valuation with business objectives

The alignment of policy objectives and evaluation metrics is a fundamental challenge. It is also highly consequential, as any misalignment between these may lead to suboptimal marketing campaigns, biased assessments of policy effectiveness, and, ultimately, lower profits. For example, many recommendation systems recommend the product with the highest utility (Ansari et al., 2018; Chung & Rao, 2012), leading to policies that treat customers already likely to purchase rather than those that are most affected by treatment (Bodapati, 2008). Similarly, personalized ad campaigns are often evaluated using click-through rates, which ignore that the customer could have made a purchase anyway (Bleier & Eisenbeiss, 2015).

To address this misalignment, researchers have proposed evaluation metrics that reflect the true objectives of the policy, such as incremental sales or profit lift. For example, Lemmens and Gupta (2020) uses an off-policy IPW estimator that evaluates the cumulative profit lift generated by a proactive retention campaign as a function of the number of customers targeted. This estimator belongs to a larger family of evaluation metrics called *rank-weighted average treatment effect* (RATE) metrics (Yadlowsky et al., 2024). Such metrics capture the extent to which customers who are highly ranked by a prioritization rule (e.g., with a high predicted CATE) are more responsive to treatment than randomly selected customers. Other metrics that belong to the RATE family are the Qini coefficient (Radcliffe, 2007), and the AUTOC (area under the targeting operating curve Zhao et al., 2013). Yadlowsky et al. (2024) proposes to design RATE estimators by constructing scores using inverse-propensity weighting, potentially with cross-fitting (Chernozhukov et al., 2025, 2022), and provides a central limit theorem that applies to the wide class of RATE metrics. In case of more than two alternative treatments (i.e., multi-arm problem), Sverdrup et al. (2024) developes an efficient algorithm for computing treatment-specific Qini curves, and proposes exact bootstrap-based confidence intervals for any point on the curve. These confidence intervals can be used to conduct hypothesis tests comparing the return on investment of a policy that offers an optimal combination of treatments compared to alternative policies using a subset of treatments.

6.3.2. Measuring long-term effect of personalized policies

Personalized policies can have long-lasting positive effects (Liao et al., 2021; Theocharous et al., 2015). However, in some cases, the effect of a marketing action can take months to fully materialize. For instance, in the domain of customer retention, a personalized retention offer might gradually reshape consumer perceptions about a brand and slowly influence behavior. In the customer acquisition domain, a discount might drive immediate adoption but lead to lower customer lifetime value and increased price sensitivity in the long run (Datta et al., 2015). Although firms may care about a policy's long-term effectiveness, in practice, they often rely on short-term measures when evaluating them. For example, Yoganarasimhan et al. (2023) personalizes promotions (the length of free trials) to maximize *subscriptions*, even though the firm might ultimately care more about long-term objectives such as maximizing total revenue or customer retention, outcomes that are only observable years after the promotion.

Evaluating the long-term effects of personalized policies poses significant challenges for marketers (Marketing Science Institute, 2024, pages 4–5). While short-term metrics, such as immediate purchases or click-through rates, are easier to observe, they frequently fail to capture the broader, enduring impacts of marketing interventions. To address this gap, Athey et al. (2019a) introduces the concept of a "surrogate model." In this approach, an experiment is first conducted to estimate the effect of a treatment on a short-term metric. Then, the surrogate model is estimated from observational data to predict the long-term outcome from the short-term outcome. By combining these two models, the effect of the treatment on the long-term outcome can be inferred. Yang et al. (2024) adapts this approach to a personalization context, optimizing the three-year net revenues of a discount campaign among newspaper readers. Huang and Ascarza (2024) further enhances the method to maximize long-run retention by proposing a separate imputation strategy designed to reduce the additional variance caused by the interdependent nature of customer churn and purchase intensity, which is prevalent in marketing contexts.

To reliably infer long-term effects, the chosen surrogates must fully mediate the treatment's impact on the outcome. However, even if this condition is met, the estimated causal direction may still be incorrect if the surrogate and the outcome share an unobserved confounder. This phenomenon is referred to as the surrogate paradox. In the food delivery example, this paradox would arise if the discount increases short-term margin among customers who are already highly engaged and unlikely to churn, but reduces the short-term margin among price-sensitive customers who are at high risk of churn (e.g., because they delay purchases to wait for discounts, or switch to competitors when no new offers appear). If the short-term gains from loyal customers outweigh the small short-term losses from at-risk customers, the average short-term margin appears positive. Yet the overall long-term customer value may decline, as more churn-prone users reduce their engagement or defect entirely. To address this challenge, Bibaut et al. (2023) proposes an approach based on cross-fold procedures to derive instruments for the surrogates using historical experiments. This method not only eliminates biases stemming from confounding but also accommodates effects that are not perfectly mediated by the surrogates.

Emerging experimental designs also offer solutions for evaluating long-term effects. For instance, staggered rollout designs (Xiong et al., 2024) allow firms to observe the outcomes of personalized policies over extended periods. These designs involve selecting an initial treatment time for each unit in order to most precisely estimate both the instantaneous and cumulative effects of the treatment over time.

6.4. Ensuring ethical and responsible personalization

Personalized marketing enhances relevance and effectiveness, but its inherent differentiation among individuals introduces significant ethical concerns. These include the risk of algorithmic discrimination, leading to unfair personalized policies (6.4.1), and consumer privacy risks (6.4.2). In addition, responsible personalization also creates a need for more transparency (6.4.3). The recent academic literature has focused on the methodological challenges that these concerns create.

6.4.1. Fair personalization

Personalization can inadvertently lead to discriminatory outcomes. Facebook, for example, was accused of gender discrimination for excluding women from seeing certain job ads. The danger of algorithmic biases stands even when algorithms exclude protected attributes like gender or race. For instance, ride share companies like Uber and Lyft were found to offer higher prices in areas predominantly inhabited by non-white populations, because the ride supply was lower in these areas. In general, personalized policies have a greater risk of discriminating against protected groups when the protected attributes are correlated with the features (X) used by the model. A typical example is when the heterogeneous treatment effects are predicted based on a household location (e.g., zip code), which is known to correlate with income.

The vast literature on fairness and algorithmic bias (e.g., Cowgill, 2019; Cowgill & Tucker, 2020; Lambrecht & Tucker, 2019) has focused on uncovering biases and their causes, conceptualizing the algorithmic bias problem, and developing solutions for researchers and practitioners. These solutions include the use of auditing tools to detect bias (e.g., Ukanwa et al., 2023) and modifying personalization procedures to ensure fair outcomes. Many efforts to mitigate algorithmic bias frame the problem in terms of constrained optimization (e.g., Lu et al., 2023), with explicit constraints to ensure, for example, that the demographic composition of targeted individuals does not systematically diverge according to protected characteristics. Although effective in theory, this approach struggles in high-dimensional scenarios involving multiple protected attributes. Other techniques modify non-protected variables so they are statistically independent of protected characteristics (e.g., Feldman et al., 2015; Johndrow & Lum, 2019). However, this requires the collection and use of protected attribute data in the test and learn cycle.

To address the challenges of scalability and data collection in mitigating bias, Ascarza and Israeli (2022) introduces Bias Eliminating Adapted Trees (BEAT). This model leverages a computationally efficient approach for CATE estimation (Athey et al., 2019b; Wager & Athey, 2018), providing ease of implementation and scalability. Notably, BEAT allows for "fair personalization" without requiring protected attributes for ongoing decisions.

Beyond the methodological complexity of fair personalization, decision-makers face the critical challenge of defining fairness itself, which generally encompasses equality in outcomes, treatment allocation, and/or prediction error. Fairness can also be categorized into group fairness, which seeks equitable treatment distribution across protected and non-protected groups, and individual fairness, which seeks to ensure equal treatment among individuals with identical unprotected attributes across different groups.[7] These different views on fairness may be in conflict with each other, further complicating the decision-making process. Although existing methods predominantly achieve group fairness, they may inadvertently compromise individual fairness (Dwork et al., 2012), highlighting the need for decision-makers to carefully select the most appropriate fairness metric for their context. The BEAT (Ascarza & Israeli, 2022) algorithm attains both group and individual fairness in allocation, but does not address fairness in prediction errors. Conversely, Shi and Lemmens (2025) proposes a Fair Active Learning (FAL) model aimed at eliminating disparities in prediction errors across protected groups. This approach focuses on the experimental stage, selectively oversampling units from protected groups using an accuracy-maximizing acquisition function. The active learning algorithm can be used in combination with causal forests (Wager & Athey, 2018) if the goal is to maximize fairness in prediction errors, or in combination with BEAT (Ascarza & Israeli, 2022) when seeking a trade-off between equality in outcomes and fairness in prediction errors.

Although these methodologies enable some degree of personalization while reducing bias against protected groups, they usually result in less efficient personalized outcomes. This performance-fairness trade-off presents a significant challenge for decision-makers striving to balance marketing effectiveness against the risk of producing unequal outcomes. Despite the theoretical capacity of certain models to evaluate the cost of fairness, empirical research quantifying this trade-off remains scarce.

6.4.2. Privacy-aware personalization

Another challenge that arises under personalization is the need to protect customer privacy (Marketing Science Institute, 2024, page 6), as effective personalized marketing relies heavily on comprehensive consumer data, and thus comes with inherent privacy risks (Pauwels & Aksehirli, 2025). Firms that neglect these risks may face severe penalties, including fines of up to 4 % of their total revenue under the GDPR. These challenges compel marketing scholars to examine, quantify, and mitigate the privacy risks associated with personalization.

Differential privacy has emerged as a common framework for defining and quantifying privacy risks, and is already being adopted by major firms, including Apple, Google, and Meta (Kifer et al., 2020). Differential privacy acknowledges that any data analysis poses a privacy risk to individuals (Dwork et al., 2014). In essence, differential privacy anticipates various types of adversarial attacks, such as re-identifying customers, detecting the inclusion of a particular customer in a dataset, inferring sensitive attributes, or uncovering unknown vulnerabilities (Ponte et al., 2024). These risks are indexed by a single privacy risk parameter, denoted as ϵ, which can be specified when designing a personalized campaign. The value of ϵ determines the level of protection for a customer's privacy by introducing probabilistic changes to the personalized policy. This ϵ parameter operationalizes an inherent trade-off between lower privacy risk and more efficient use of sensitive data.

To mitigate privacy risks, scholars have recently proposed methods to estimate heterogeneous treatment effects in a privacy-preserving fashion. One common approach involves adding noise to the data to protect user privacy (e.g., Kasiviswanathan et al., 2011; Schneider et al., 2017). However, this introduces a fundamental trade-off: the noise necessary for privacy protection often reduces the data's precision, thereby diminishing the effectiveness of personalized marketing efforts (Showkatbakhsh et al., 2018).

[7] Alternative definitions include, among others appearing in the literature, ensuring equality in prediction errors, odds ratios, or equal opportunity across groups

For instance, Huang and Ascarza (2023) demonstrates that privacy-enhancing mechanisms like local differential privacy (LDP) can distort a firm's CATE predictions, ultimately weakening the impact of personalized campaigns.

Ponte et al. (2025) introduces private targeting strategies that incorporate privacy considerations into different stages of a typical targeting workflow. The framework includes two key components: differentially private conditional average treatment effect estimation (DP-CATE) and differentially private policy optimization (DP-policy). Using a novel approach called private causal neural networks, the study demonstrates how firms can design profitable personalized campaigns while maintaining low levels of privacy risk. Additionally, it quantifies the privacy elasticity of targeting profit, a measure of how a 1 % increase in privacy risk affects the profitability of targeting. This elasticity helps identify the levels of privacy risk at which targeting profitability is most sensitive to changes in privacy protection, enabling firms to make informed decisions about balancing privacy and effectiveness.

6.4.3. Transparent and explainable personalization

Effective personalization often relies on causal machine learning and deep learning models that can autonomously identify complex patterns and nonlinear relationships in large datasets. A key downside of these approaches is their lack of transparency and the difficulty in explaining their predictions. While classic ML algorithms, such as decision trees, are directly interpretable, deep neural networks and ensemble methods, such as random forests, operate as a "black box." They generate predictions without an explicit explanation of how they are related to the model inputs.

This lack of interpretability leads to three challenges. First, it complicates the identification and correction of prediction errors (Bauer et al., 2023). For instance, diagnosing mistargeting due to concept drift becomes challenging without a clear understanding of how a model generates targeting decisions. Second, a lower explainability reduces accountability. Black-box models can obfuscate algorithmic biases, potentially resulting in discrimination (Rai, 2020). Transparency is essential for regulatory compliance, particularly when addressing fairness concerns. Third, the lack of transparency can erode trust in the models and the resulting policies. Managers rolling out personalization often require a clear understanding of the factors driving the outcomes to confidently endorse and implement the strategy (von Zahn et al., 2024).

To address these challenges, the literature identifies three main types of explainable AI (XAI) approaches. First, researchers can embed interpretability constraints directly into the architecture of deep learning models. For instance, Gabel and Timoshenko (2022) explicitly structures the neural network into interpretable components that capture customer heterogeneity in purchase frequencies, purchase incidence, and coupon effects. By incorporating direct effects of coupons in the model's output layer, the architecture ensures these effects are modeled independently of the other network components. Zhang (2025) restricts personalization to a subset of policies that can be expressed in a single sentence, e.g., "Target a customer with a promotion if she has not bought in the last thirty days and lives in Chicago (p.2)."

Second, researchers can approximate black-box AI models with glass-box models, which are simpler, interpretable models designed to mimic the predictions of black-box models while providing insights into the decision logic. For instance, a firm may train a deep learning model to optimize product recommendations and subsequently approximate its predictions with a decision tree to understand which customer features, such as browsing history or recent purchases, most strongly influence the recommendations (Rai, 2020). Relatedly, Chernozhukov et al. (2025) uses doubly-robust machine learning to explore the features of the CATE predictions (e.g., GATE plots), and Esterzon et al. (2023) extends this approach to partial dependence plots.

Third, post-hoc explanation techniques for black-box models, such as Local Interpretable Model-agnostic Explanations (LIME, Ribeiro et al., 2016) and SHapley Additive exPlanations (SHAP, Lundberg, 2017), provide insights into model predictions after training without altering the model's structure. These approaches enable modelers to quantify the influence of individual features on the model predictions and provide insights into how different variables marginally contribute to the outcome (Senoner et al., 2022). In marketing, SHAP values have been applied to assess feature importance to predict intervention effectiveness (von Zahn et al., 2024; Zhou et al., 2021). In the context of personalization, SHAP values can analyze the contribution of individual features—such as customer characteristics, contextual factors, or past purchases—to a customer's likelihood of responding to a promotion.

While XAI methods are valuable tools for enhancing transparency in personalization, ongoing challenges remain that offer important opportunities for further research. A key concern is the risk of misinterpreting explanations, as it is possible for variables that have no moderating impact on the magnitude of a treatment effect to nevertheless be extremely useful for treatment effect prediction. This concern is most relevant for CATE estimators based on the indirect approach, which first predict treated and untreated levels of the outcome, and then estimate the causal effect as their difference. A variable like past purchase frequency may be strongly related to customer heterogeneity in sales, and thus useful for predicting sales, but may or may not be related to the treatment effect of interest. XAI methods may flag this variable as "important" regardless of how it moderates the treatment effect.

Failing to recognize this distinction can lead to incorrect conclusions about interventions' effectiveness (von Zahn et al., 2024). In addition, explanations might fail to generalize to unseen inputs, especially in dynamic or evolving environments. A promising new direction is to use direct policy learning to learn simple targeting rules (Zhang, 2025), which we discuss further in Section 7.2.

7. Future research directions and conclusion

The challenges outlined in the previous section highlight the complexities of personalization and recent advancements to address them. In this section, we conclude by highlighting several emerging methodological trends that we anticipate will play a central role in shaping the future of personalization research. For each trend, we propose key research questions to guide and inspire future inquiry in the domain of personalized marketing.

7.1. Advances in causal machine learning

The rise of field experiments in marketing has introduced new challenges related to CATE estimation. Although experiments have generally grown larger in terms of their sample sizes, they have grown even more rapidly in terms of the number of observed covariates X. This explosion in dimensionality introduces computational and conceptual complexities, because as the dimension of X grows, the number of ways to characterize treatment effect heterogeneity increases exponentially. Causal forests, for example, can become unstable when applied to high-dimensional data in finite samples, in spite of their attractive asymptotic properties. A recent methodological advancement with the potential to overcome this barrier is GenericML (Chernozhukov et al., 2025). This direct estimation approach is designed to handle high-dimensional covariates while ensuring valid inference for estimated CATEs. It can be implemented using any base machine learning method, and requires minimal assumptions on the first-stage estimator. We expect even more methodological advances along these lines in the coming years, and anticipate this will further stimulate the adoption of causal machine learning for personalization. As causal ML becomes more commonplace, the following open questions will become even more salient:

1. Which causal ML methods work best in different contexts? Is it possible to guide the selection of causal ML methods based on general considerations such as sample size, the dimensionality of covariates, treatment types, and propensities, and expected treatment effect heterogeneity? If so, at what level of generalization should this guidance be developed?
2. What is the best way to integrate domain expertise or learning from prior experiments into black-box CATE estimators? How does augmenting a data set with variables describing past personalization efforts compare with modifying or constraining learning algorithms to adhere to established norms and business rules?
3. How can causal ML methods be extended or augmented to be more robust to real-world challenges, such as concept drift, privacy and fairness considerations, experimental non-compliance, and multidimensional treatments and outcomes?

7.2. Directly learning the personalization policy

The test and learn approach includes CATE estimation as an intermediate step before policy optimization. This step can be inefficient when policy optimization does not fully align with model estimation. For instance, when the policy only depends on whether a treatment effect is positive, prediction errors that imply a sign switch are more consequential than errors that do not. Direct policy learning is a solution to this problem (e.g., Mnih et al., 2015). It defines the model objective function according to the campaign goal (e.g., profitability, Lemmens & Gupta, 2020) and directly learns the policy function that outputs treatment assignments. Directly learning the policy function instead of CATEs avoids some of the noise introduced during the treatment effect estimation step. Applications of this approach in the context of personalization have appeared in the literature (Liu, 2023; Ma et al., 2025; Zhang, 2025). We identify a number of important research challenges:

1. Under which circumstances is direct policy learning a more efficient approach and leads to better personalization? To what extent does it depend on the alignment of the loss function of the CATE prediction model w.r.t. the policy goal? What are the benefits of direct policy learning in large-scale field experiments where efficiency is less of an issue?
2. How can direct policy learning integrate multiple, possibly competing objectives in the same loss function, e.g., maximizing short-run and long-run profitability or balancing profitability and fairness? Does policy learning lead to more unfair personalized policies by focusing even more on high-profit targets, and to what extent does it obfuscate unfairness?
3. Can we transport insights from direct policy learning to (i) new tasks that have different goals and/or to (ii) new environments where customers' behavior has changed? Are direct policy methods more or less robust to such variations than a causal inference approach where CATE predictions can be interpreted?

7.3. Foundation models for personalization

Test and learn is typically implemented with a specific business objective in mind. Evaluating policies for new objectives requires new instances of the test and learn cycle, such as new models for treatment effect estimation and possibly even additional experiments. Standard methods for carrying out test and learn may struggle to effectively use prior knowledge. Foundation models, initially developed in the area of natural language processing (GPT is a well-known example), offer a promising research direction to address these challenges (Bommasani et al., 2022). During *pre-training*, foundation models learn generalizable patterns and structures from large datasets. During *adaptation*, pre-trained models are then applied to specific downstream tasks using smaller, task-specific datasets (e.g., through fine-tuning or continuation training). For example, Gabel and Ringel (2024) proposes a foundation model designed to learn general shopping patterns and basket composition from large market basket data. The study illustrates how foundation models can be adapted to downstream applications, such as predicting coupon redemption in a field experiment. Foundation models hold much potential for scalable personalization in marketing. One could imagine consortia of retailers and advertisers building foundation models for basic marketing responses, such as pricing and advertising. Important directions for future research include:

1. Can non-causal foundation models be successfully fine-tuned on experimental data to learn causal treatment effects without bias? Can fine-tuned foundation models trained on non-experimental data outperform causal ML models on experimental data? What can mitigate potential biases from pre-training foundation models on observational training data?

2. What foundation model architectures are most suitable for marketing data sets? What value does multi-modality (i.e., combining multiple data sources such as purchase, text, images, and click stream data) provide?
3. How do we achieve interpretability in foundation models that rely on large deep learning architectures? How do we ensure fairness and privacy preservation when training these models?

7.4. Generative AI for personalization

Generative AI is poised to significantly influence personalization in the coming years. Among its many potential applications, GenAI can assist both marketing scholars and practitioners in developing novel personalization interventions (Peres et al., 2023). For instance, tools like DALL-E can generate personalized creatives (what to offer), as recently done by Meta and Coca-Cola (Mehta, 2023; Rogers, 2024). Gartner predicts that by 2025, GenAI will generate 30 % of brands' marketing messages (Gartner, 2024). Large Language Models (LLMs) also present new opportunities for semi-automating search engine optimization content to reach specific customer segments (whom to target) (Reisenbichler et al., 2022). Additionally, AI-driven personalization extends to customer service interactions. For example, Huang and Rust (2024) proposes the "caring machine," which personalizes customer care by recognizing and managing emotions over time (when to engage). Beyond content creation and personalization, GenAI has the potential to mitigate the cold-start problem discussed in Section 6.1.1. By synthesizing plausible customer profiles from minimal data, it provides firms with a valuable starting point for personalization efforts (Argyle et al., 2023; Brand et al., 2023; Goli & Singh, 2024; Horton, 2023). As GenAI becomes increasingly integrated into personalization strategies, several open research questions arise:

1. How can firms optimize GenAI-generated content to maximize engagement and conversion? In particular, how would model parameters—such as temperature settings, prompt structures, or fine-tuning methods—affect the persuasiveness and effectiveness of personalized messages? To what extent do specific tuning strategies align with different marketing goals, from short-term sales to long-term brand loyalty?
2. Can direct policy learning be combined with GenAI to generate unstructured, high-quality marketing outputs, such as product descriptions, advertisements, and promotional offers? Can GenAI replace experimentation in some settings?
3. How efficient would GenAI be in tackling the challenges discussed in this article, in particular the cold-start problem, data loss, concept drift, and non-compliance, relative to the classical approaches to personalization?
4. Can we ensure responsible AI-driven personalization? Given that LLMs are prone to biases and hallucinations, can they be safely used for personalization, and how? What is the best way to mitigate risks, e.g., through managing inputs (e.g., prompt-tuning), modifying the models themselves (e.g., bias-aware fine-tuning), or managing outputs through human-in-the-loop intervention?

Given the rapid evolution of GenAI, addressing these questions will be critical for both advancing academic research and guiding responsible industry adoption.

Over the past decade, personalization has evolved from a promising concept into a core pillar of marketing, fueled by advances in causal inference, machine learning, and new data sources. Yet, the field continues to change rapidly. Looking ahead, two possibilities emerge: One is that within a decade, another major reassessment will be necessary, as new technologies and consumer expectations continue to redefine the boundaries of personalization. The other is that personalization will become so deeply embedded in marketing that its evolution will no longer be a distinct track of research but rather an intrinsic part of how firms engage with customers. If the latter holds true, then personalization is no longer just a strategy—it is marketing itself. We therefore close with a prediction: Personalization is the fifth P.

CRediT authorship contribution statement

Aurélie Lemmens: Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Funding acquisition, Conceptualization; **Jason M.T. Roos:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Methodology, Conceptualization; **Sebastian Gabel:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization; **Eva Ascarza:** Writing – review & editing, Writing – original draft, Conceptualization; **Hernán A. Bruno:** Writing – review & editing, Writing – original draft, Conceptualization; **Brett R. Gordon:** Writing – review & editing, Writing – original draft, Conceptualization; **Ayelet Israeli:** Writing – review & editing, Writing – original draft, Conceptualization; **Elea McDonnell Feit:** Writing – review & editing, Writing – original draft, Conceptualization; **Carl F. Mela:** Writing – review & editing, Writing – original draft, Conceptualization; **Oded Netzer:** Writing – review & editing, Writing – original draft, Conceptualization.

Data availability

No data was used for the research described in the article.

References

Ahn, H. J. (2008). A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. *Information Sciences*, *178*(1), 37–51.
Andrews, M., Luo, X., Fang, Z., & Ghose, A. (2016). Mobile ad effectiveness: Hyper-contextual targeting with crowdedness. *Marketing Science*, *35*(2), 218–233.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455.
Ansari, A., Li, Y., & Zhang, J. Z. (2018). Probabilistic topic model for hybrid recommender systems: A stochastic variational bayesian approach. *Marketing Science*, *37*(6), 987–1008.
Ansari, A., & Mela, C. F. (2003). E-customization. *Journal of Marketing Research*, *40*(2), 131–145.
Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351.
Aronow, P. M. (2012). A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research*, *41*(1), 3–16.
Arora, N., Dreze, X., Ghose, A., Hess, J. D., Iyengar, R., Jing, B., Joshi, Y., Kumar, V., Lurie, N., Neslin, S. et al. (2008). Putting one-to-one marketing to work: Personalization, customization, and choice. *Marketing Letters*, *19*, 305–321.
Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research*, *55*(1), 80–98.
Ascarza, E., Ebbes, P., Netzer, O., & Danielson, M. (2017). Beyond the target customer: Social effects of customer relationship management campaigns. *Journal of Marketing Research*, *54*(3), 347–363.
Ascarza, E., & Israeli, A. (2022). Eliminating unintended bias in personalized policies using bias-eliminating adapted trees (BEAT). *Proceedings of the National Academy of Sciences*, *119*(11), e2115293119.
Athey, S., Chetty, R., Imbens, G. W., & Kang, H. (2019a). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical Report National Bureau of Economic Research.
Athey, S., Eckles, D., & Imbens, G. W. (2018). Exact p-values for network interference. *Journal of the American Statistical Association*, *113*(521), 230–240.
Athey, S., Tibshirani, J., & Wager, S. (2019b). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.
Bachman, J. (2018). Airlines have your personal data, and they're using it. *The Denver Post*. Accessed: 2025-06-10 https://www.denverpost.com/2017/11/25/airlines-using-personal-data/.
Bare, M. (2023). Luxury hotel tactics: Elevating guest satisfaction and loyalty with simple gestures, (last accessed 2025-06-06). https://www.costar.com/article/630780836/luxury-hotel-tactics-elevating-guest-satisfaction-and-loyalty-with-simple-gestures
Barseghian, A. (2022). How nike is using analytics to personalize their customer experience. *Forbes*. Accessed: 2025-06-10 https://www.forbes.com/councils/forbestechcouncil/2019/10/07/how-nike-is-using-analytics-to-personalize-their-customer-experience/.
Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl (AI) ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research*, *34*(4), 1582–1602.
Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2020). Uniting the tribes: Using text for marketing insight. *Journal of Marketing*, *84*(1), 1–25.
Bibaut, A., Kallus, N., Ejdemyr, S., & Zhao, M. (2023). Long-term causal inference with imperfect surrogates using many weak experiments, proxies, and cross-fold moments. https://arxiv.org/abs/2311.04657.
Black, N. (2019). Campaign of the year: Burger king's whopper detour, (last accessed 2025-06-06). https://www.marketingdive.com/news/burger-king-whopper-detour-mobile-marketer-awards/566224/
Blake, T., Nosko, C., & Tadelis, S. (2015). Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica : Journal of the Econometric Society*, *83*(1), 155–174.
Bleier, A., & Eisenbeiss, M. (2015). Personalized online advertising effectiveness: The interplay of what, when, and where. *Marketing Science*, *34*(5), 669–688.
Bodapati, A. V. (2008). Recommendation systems with purchase data. *Journal of Marketing Research*, *45*(1), 77–93.
Boegershausen, J., Cornil, Y., Yi, S., & Hardisty, D. J. (2025). On the persistent mischaracterization of google and facebook a/b tests: How to conduct and report online platform studies. *International Journal of Research in Marketing*, (forthcoming).
Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E. et al. (2022). On the opportunities and risks of foundation models. arXiv preprint arXiv: 2108.07258
Brand, J., Israeli, A., & Ngwe, D. (2023). Using GPT for market research. *Harvard Business School Marketing Unit Working Paper*.
Candogan, O., Chen, C., & Niazadeh, R. (2024). Correlated cluster-based randomized experiments: Robust variance minimization. *Management Science*, *70*(6), 4069–4086.
Chen, Y., Lee, J.-Y., Sridhar, S., Mittal, V., McCallister, K., & Singal, A. G. (2020). Improving cancer outreach effectiveness through targeting and economic assessments: Insights from a randomized field experiment. *Journal of Marketing*, *84*(3), 1–27.
Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68.
Chernozhukov, V., Demirer, M., Duflo, E., & Fernandez-Val, I. (2025). Generic machine learning inference on heterogenous treatment effects in randomized experiments. *Econometrica : Journal of the Econometric Society*, (forthcoming).
Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., & Robins, J. M. (2022). Locally robust semiparametric estimation. *Econometrica : Journal of the Econometric Society*, *90*(4), 1501–1535.
Chung, J., & Rao, V. R. (2012). A general consumer preference model for experience products: Application to internet recommendation services. *Journal of Marketing Research*, *49*(3), 289–305.
Chung, T. S., Rust, R. T., & Wedel, M. (2009). My mobile music: An adaptive personalization system for digital audio players. *Marketing Science*, *28*(1), 52–68.
Compiani, G., Morozov, I., & Seiler, S. (2023). Demand estimation with text and image data. *SSRN Electronic Journal*, 4588941.
Cowgill, B. (2019). Bias and productivity in humans and machines. *SSRN Electronic Journal*, 3584916.
Cowgill, B., & Tucker, C. E. (2020). Algorithmic fairness and economics. *SSRN Electronic Journal*, 3361280.
Daljord, Ø., Mela, C. F., Roos, J. M. T., Sprigg, J., & Yao, S. (2023). The design and targeting of compliance promotions. *Marketing Science*, *42*(5), 866–891.
Datta, H., Foubert, B., & Van Heerde, H. J. (2015). The challenge of retaining customers acquired with free trials. *Journal of Marketing Research*, *52*(2), 217–234.
Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (p. 214–226). https://doi.org/10.1145/2090236.2090255
Dwork, C., Roth, A. et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, *9*(3–4), 211–407.
Eckles, D., Karrer, B., & Ugander, J. (2017). Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, *5*(1), 20150021. https://doi.org/10.1515/jci-2015-0021
Ellickson, P. B., Kar, W., & Reeder III, J. C. (2023). Estimating marketing component effects: Double machine learning from targeted digital promotions. *Marketing Science*, *42*(4), 637–837.
Esterzon, E., Lemmens, A., & Bergh, B. V. d. (2023). Enhancing donor agency to improve charitable giving: Strategies and heterogeneity. *Journal of Marketing*, *87*(4), 636–655. https://doi.org/10.1177/00222429221148969
Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 259–268).
Gabel, S., Guhl, D., Hitsch, G. J., & Kook, K. (2024a). Experimental evidence on structural state dependence in demand. Available at SSRN 4955593.
Gabel, S., & Ringel, D. (2024). The market basket transformer: A new foundation model for retail. Available at SSRN 4335141.
Gabel, S., Simester, D., & Timoshenko, A. (2024b). In-store coupon effectiveness: A large-scale field experiment. https://ssrn.com/abstract=4335525.
Gabel, S., & Timoshenko, A. (2022). Product choice with large assortments: A scalable deep-learning model. *Management Science*, *68*(3), 1808–1827.
Gartner (2024). Gartner experts answer the top generative AI questions for your enterprise, (last accessed 2025-02-14). https://www.gartner.com/en/topics/generative-ai
Gelper, S., Lovett, M. J., & Peres, R. (2024). The effect of second screening on repeat viewing: Insights from large-scale mobile diary data. *Journal of the Academy of Marketing Science*, (pp. 1–24).

Ghose, A., Goldfarb, A., & Han, S. P. (2013a). How is the mobile internet different? search costs and local activities. *Information Systems Research*, *24*(3), 499–882. https://doi.org/10.1287/isre.1120.0453
Ghose, A., Goldfarb, A., & Han, S. P. (2013b). How is the mobile internet different? search costs and local activities. *Information Systems Research*, *24*(3), 613–631.
Goldfarb, A., Tucker, C., & Wang, Y. (2022). Conducting research in marketing with quasi-experiments. *Journal of Marketing*, *86*(3), 1–20.
Goli, A., & Singh, A. (2024). Frontiers: Can large language models capture human preferences? *Marketing Science*, *43*(4), 709–722.
Gordon, B., Moakler, R., & Zettelmeyer, F. (2023a). Predictive incrementality by experimentation (PIE) for ad measurement. https://arxiv.org/abs/2304.06828.
Gordon, B. R., Lovett, M. J., Luo, B., & Reeder III, J. C. (2023b). Disentangling the effects of ad tone on voter turnout and candidate choice in presidential elections. *Management Science*, *69*(1), 220–243.
Gordon, B. R., Moakler, R., & Zettelmeyer, F. (2023c). Close enough? a large-scale exploration of non-experimental approaches to advertising measurement. *Marketing Science*, *42*(4), 768–793.
Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science*, *38*(2), 193–225.
Habel, J., Alavi, S., & Heinitz, N. (2024). Effective implementation of predictive sales analytics. *Journal of Marketing Research*, *61*(4), 718–741.
Haenlein, M., & Libai, B. (2013). Targeting revenue leaders for a new product. *Journal of Marketing*, *77*(3), 65–80.
Hauser, J. R., Liberali, G., & Urban, G. L. (2014). Website morphing 2.0: Switching costs, partial exposure, random exit, and when to morph. *Management Science*, *60*(6), 1594–1616.
Hauser, J. R., Urban, G. L., Liberali, G., & Braun, M. (2009). Website morphing. *Marketing Science*, *28*(2), 202–223.
Heckman, J. J., & Vytlacil, E. J. (2007). Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. *Handbook of Econometrics*, *6*, 4875–5143.
Hill, K. (2024). Is your driving being secretly scored? *The New York Times*. Accessed: 2025-06-10 https://www.nytimes.com/2024/06/09/technology/driver-scores-insurance-data-apps.html.
Hitsch, G. J., Misra, S., & Zhang, W. W. (2024). Heterogeneous treatment effects and optimal targeting policy evaluation. *Quantitative Marketing and Economics*, *22*(2), 115–168.
Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? Technical Report National Bureau of Economic Research.
Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663–685.
Huang, M.-H., & Rust, R. T. (2024). The caring machine: Feeling AI for customer care. *Journal of Marketing*, *88*(5), 1–23.
Huang, S., Aral, S., Hu, Y. J., & Brynjolfsson, E. (2020). Social advertising effectiveness across products: A large-scale field experiment. *Marketing Science*, *39*(6), 1142–1165.
Huang, T.-W., & Ascarza, E. (2023). Debiasing treatment effect estimation for privacy-protected data: A model audition and calibration approach. Available at SSRN 4575240.
Huang, T.-W., & Ascarza, E. (2024). Doing more with less: Overcoming ineffective long-term targeting using short-term signals. *Marketing Science*, (forthcoming).
Huang, T.-W., Ascarza, E., & Israeli, A. (2024). Incrementality representation learning: Synergizing past experiments for intervention personalization. *SSRN Electronic Journal*, 4859809.
Hui, S. K., Inman, J. J., Huang, Y., & Suher, J. (2013). The effect of in-store travel distance on unplanned spending: Applications to mobile promotion strategies. *Journal of Marketing*, *77*(2), 1–16.
Imbens, G. W., & Rubin, D. B. (2015). Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.
Jamissen, D., Vatne, J., Holz, F., & Neumann, A. (2024). The price elasticity of natural gas demand of small consumers in germany during the energy crisis 2022. *Energy Efficiency*, *17*(8), 98.
Johndrow, J. E., & Lum, K. (2019). An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *Annals of Applied Statistics*, (pp. 189–220).
Johnson, G. A., Lewis, R. A., & Reiley, D. H. (2017). When less is more: Data and power in advertising experiments. *Marketing Science*, *36*(1), 43–53.
Joseph, S., & Peterson, T. (2025). Google's latest chrome update leaves third-party cookie phase-out as unclear as ever, accessed on january 29, 2025. https://digiday.com/marketing/google-reveals-chrome-cookie-consent-plan-barely,
Kalyanam, K., McAteer, J., Marek, J., Hodges, J., & Lin, L. (2018). Cross channel effects of search engine advertising on brick & mortar retail sales: Meta analysis of large scale field experiments on google. com. *Quantitative Marketing and Economics*, *16*, 1–42.
Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., & Smith, A. (2011). What can we learn privately? *SIAM Journal on Computing*, *40*(3), 793–826.
Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, *17*(2), 3008–3049.
Kifer, D., Messing, S., Roth, A., Thakurta, A., & Zhang, D. (2020). Guidelines for implementing and auditing differentially private systems. arXiv preprint arXiv: 2002.04049
Kohavi, R., Tang, D., & Xu, Y. (2020). Trustworthy online controlled experiments: A practical guide to A/B testing. Cambridge University Press.
Kumar, V., Venkatesan, R., Bohling, T., & Beckmann, D. (2008). Practice prize report–the power of CLV: Managing customer lifetime value at IBM. *Marketing Science*, *27*(4), 585–599.
Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165.
Lambrecht, A., & Tucker, C. E. (2019). Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science*, *65*(7), 2966–2981.
Lemmens, A., & Gupta, S. (2020). Managing churn to maximize profits. *Marketing Science*, *39*(5), 956–973.
Liao, P., Klasnja, P., & Murphy, S. (2021). Off-policy estimation of long-term average outcomes with applications to mobile health. *Journal of the American Statistical Association*, *116*(533), 382–391.
Libai, B., Muller, E., & Peres, R. (2013). Decomposing the value of word-of-mouth seeding programs: Acceleration versus expansion. *Journal of Marketing Research*, *50*(2), 161–176.
Liberali, G., & Ferecatu, A. (2022). Morphing for consumer dynamics: Bandits meet hidden markov models. *Marketing Science*, *41*(4), 769–794.
Liu, X. (2023). Dynamic coupon targeting using batch deep reinforcement learning: An application to livestream shopping. *Marketing Science*, *42*(4), 637–658.
Lu, H., Simester, D., & Zhu, Y. (2023). Optimizing scalable targeted marketing policies with constraints. Available at SSRN 4668582.
Lundberg, S. (2017). A unified approach to interpreting model predictions. arXiv preprint arXiv: 1705.07874
Luo, X., Andrews, M., Fang, Z., & Phang, C. W. (2014). Mobile targeting. *Management Science*, *60*(7), 1738–1756.
Ma, L., Huang, T.-W., Ascarza, E., & Israeli, A. (2025). Dynamic personalization with multiple customer signals: Multi-response state representation in reinforcement learning. Available at SSRN 5126129.
Maheshwari, S. (2018). Two senators call for investigation of smart TV industry. *The New York Times*. Accessed: 2025-06-10 https://www.nytimes.com/2018/07/12/business/media/senators-smart-tv-investigation.html.
Malik, A. (2024). Snapchat's new footsteps feature tracks your location history, (last accessed 2025-06-06). https://techcrunch.com/2024/09/30/snapchats-new-footsteps-feature-tracks-your-location-history/
Marketing Science Institute (2024). MSI research priorities, https://www.msi.org/wp-content/uploads/2024/04/2024-RP.pdf.
Godinho de Matos, M., Pedro, F., & Rodrigo, B. (2018). Target the ego or target the group: Evidence from a randomized experiment in proactive churn management. *Marketing Science*, *37*(5), 793–811.
Matz, S. C., & Netzer, O. (2017). Using big data as a window into consumers' psychology. *Current Opinion in Behavioral Sciences*, *18*, 7–12.

Matzke, D., Ly, A., Selker, R., Weeda, W., Scheibehenne, B., Lee, M., & Wagenmakers, E.-J. (2017). Bayesian inference for correlations in the presence of measurement error and estimation uncertainty. *Collabra: Psychology*, *3*(1), 25. https://doi.org/10.1525/collabra.25

Meager, R. (2019). Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, *11*(1), 57–91.

Mehta, I. (2023). Meta wants to use generative AI to create ads, (last accessed 2025-02-14). https://techcrunch.com/2023/04/05/meta-wants-to-use-generative-ai-to-create-ads

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G. et al. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Molitor, D., Spann, M., Ghose, A., & Reichhart, P. (2020). Effectiveness of location-based advertising and the impact of interface design. *Journal of Management Information Systems*, *37*(2), 431–456.

Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, *108*(2), 299–319.

Nitzan, I., & Libai, B. (2011). Social effects on customer retention. *Journal of Marketing*, *75*(6), 24–38.

Padilla, N., & Ascarza, E. (2021). Overcoming the cold start problem of customer relationship management using a probabilistic machine learning approach. *Journal of Marketing Research*, *58*(5), 981–1006.

Padilla, N., Ascarza, E., & Netzer, O. (2024). The customer journey as a source of information. *Quantitative Marketing and Economics*, (pp. 1–40).

Patov, A. (2024). How nordstrom fosters customer loyalty with high-quality customer experience (CX), (last accessed 2025-06-06). https://www.renascence.io/journal/how-nordstrom-fosters-customer-loyalty-with-high-quality-customer-experience-cx

Pauwels, K., & Aksehirli, Z. (2025). Big data analytics democratized with clean collaboration and customer privacy choice. *Journal of Business Research*, *188*, 115112.

Pereira, D. (2024). Bmw marketing strategy, (last accessed 2025-06-06). https://businessmodelanalyst.com/bmw-marketing-strategy/

Peres, R., Muller, E., & Mahajan, V. (2010). Innovation diffusion and new product growth models: A critical review and research directions. *International Journal of Research in Marketing*, *27*(2), 91–106.

Peres, R., Schreier, M., Schweidel, D., & Sorescu, A. (2023). On chatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice. *International Journal of Research in Marketing*, *40*(2), 269–275.

Ponte, G., Wieringa, J., & Boot, T. (2025). Differentially private targeting strategies, *working paper*.

Ponte, G. R., Wieringa, J. E., Boot, T., & Verhoef, P. C. (2024). Where's waldo? A framework for quantifying the privacy-utility trade-off in marketing applications. *International Journal of Research in Marketing*, *41*(3), 529–546.

Radcliffe, N. (2007). Using control groups to target on predicted lift: Building and assessing uplift model. *Direct Marketing Analytics Journal*, (pp. 14–21).

Rafieian, O. (2023). Optimizing user engagement through adaptive ad sequencing. *Marketing Science*, *42*(5), 910–933.

Rafieian, O., Kapoor, A., & Sharma, A. (2025). Multiobjective personalization of marketing interventions. *Marketing Science*, *44*(2), 457–477.

Rafieian, O., & Yoganarasimhan, H. (2021). Targeting and privacy in mobile advertising. *Marketing Science*, *40*(2), 193–218.

Rafieian, O., & Yoganarasimhan, H. (2023). Ai and personalization. In K. Sudhir, & O. Toubia (Eds.), *Artificial intelligence in marketing* (pp. 77–102). Leeds: Emerald Publishing Limited (*20*). Review of Marketing Research. https://doi.org/10.1108/S1548-643520230000020004

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*, 137–141.

Reisenbichler, M., Reutterer, T., Schweidel, D. A., & Dan, D. (2022). Frontiers: Supporting content marketing with natural language generation. *Marketing Science*, *41*(3), 441–452.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD)* (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

Rogers, C. (2024). Coca-cola: The future is 'AI meets human ingenuity', (last accessed 2025-02-14). https://www.marketingweek.com/coca-cola-artificial-intelligence

Rossi, P. E., McCulloch, R. E., & Allenby, G. M. (1996). The value of purchase history data in target marketing. *Marketing Science*, *15*(4), 321–340.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322–331.

Rust, R. T., & Huang, M.-H. (2014). The service revolution and the transformation of marketing science. *Marketing Science*, *33*(2), 206–221.

Sachdeva, N., Wang, L., Liang, D., Kallus, N., & McAuley, J. (2024). Off-policy evaluation for large action spaces via policy convolution. In *Proceedings of the ACM on web conference 2024* (pp. 3576–3585).

Sahni, N. S., Narayanan, S., & Kalyanam, K. (2019). An experimental investigation of the effects of retargeted advertising: The role of frequency and timing. *Journal of Marketing Research*, *56*(3), 401–418.

Sahni, N. S., Zou, D., & Chintagunta, P. K. (2017). Do targeted discount offers serve as advertising? evidence from 70 field experiments. *Management Science*, *63*(8), 2688–2705.

Schneider, M. J., Jagpal, S., Gupta, S., Li, S., & Yu, Y. (2017). Protecting customer privacy when marketing with second-party data. *International Journal of Research in Marketing*, *34*(3), 593–603.

Senoner, J., Netland, T., & Feuerriegel, S. (2022). Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing. *Management Science*, *68*(8), 5704–5723.

Shapiro, B. T., Hitsch, G. J., & Tuchman, A. E. (2021). Tv advertising effectiveness and profitability: Generalizable results from 288 brands. *Econometrica : Journal of the Econometric Society*, *89*(4), 1855–1879.

Shi, Z., & Lemmens, A. (2025). Fair active learning for targeting. *Working Paper*.

Showkatbakhsh, M., Karakus, C., & Diggavi, S. (2018). Privacy-utility trade-off of linear regression under random projections and additive noise. In *2018 IEEE international symposium on information theory (ISIT)* (pp. 186–190). IEEE.

Simester, D., Timoshenko, A., & Zoumpoulis, S. I. (2020). Efficiently evaluating targeting policies: Improving on champion vs. challenger experiments. *Management Science*, *66*(8), 3412–3424.

Simonov, A., Nosko, C., & Rao, J. M. (2018). Competition and crowd-out for brand keywords in sponsored search. *Marketing Science*, *37*(2), 200–215.

Spotify (2024). Spotify premium users can now turn any idea into a personalized playlist with AI playlist in beta, (last accessed 2025-06-06). https://newsroom.spotify.com/2024-04-07/spotify-premium-users-can-now-turn-any-idea-into-a-personalized-playlist-with-ai-playlist-in-beta/

Sverdrup, E., Wu, H., Athey, S., & Wager, S. (2024). Qini curves for multi-armed treatment rules. *Journal of Computational and Graphical Statistics*, 1–13.

Theocharous, G., Thomas, P. S., & Ghavamzadeh, M. (2015). Ad recommendation systems for life-time value optimization. In *Proceedings of the 24th international conference on world wide web* (pp. 1305–1310).

Thomke, S. (2020). Building a culture of experimentation. *Harvard Business Review*, *98*(2), 40–47.

Tong, S., Luo, X., & Xu, B. (2020). Personalized mobile marketing strategies. *Journal of the Academy of Marketing Science*, *48*, 64–78.

Turjeman, D., & Feinberg, F. M. (2024). When the data are out: Measuring behavioral changes following a data breach. *Marketing Science*, *43*(2), 440–461.

Uehara, M., Shi, C., & Kallus, N. (2022). A review of off-policy evaluation in reinforcement learning. arXiv preprint arXiv: 2212.06355

Ugander, J., Karrer, B., Backstrom, L., & Kleinberg, J. (2013). Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining* (p. 329–337). https://doi.org/10.1145/2487575.2487695

Ugander, J., & Yin, H. (2023). Randomized graph cluster randomization. *Journal of Causal Inference*, *11*(1), 20220014. https://doi.org/10.1515/jci-2022-0014

Ukanwa, K., Rand, W., & Zubcsek, P. P. (2023). Why firms should want algorithmic accountability. *Available at SSRN 4148214*.

Urban, G. L., Liberali, G., MacDonald, E., Bordley, R., & Hauser, J. R. (2014). Morphing banner advertising. *Marketing Science*, *33*(1), 27–46.

Valenti, A., Miller, C. J., & Tucker, C. E. (2025). Combining ad targeting techniques: Evidence from a field experiment in the auto industry. *Management Science*, *0*(0).

Valenti, A., Srinivasan, S., Yildirim, G., & Pauwels, K. (2024). Direct mail to prospects and email to current customers? modeling and field-testing multichannel marketing. *Journal of the Academy of Marketing Science*, *52*(3), 815–834.

Veitch, V., Sridhar, D., & Blei, D. (2020). Adapting text embeddings for causal inference. In *Conference on uncertainty in artificial intelligence* (pp. 919–928). PMLR.

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

Waisman, C., & Gordon, B. R. (2024). Multi-cell experiments for marginal treatment effect estimation of digital ads. arXiv preprint arxiv: 2302.13857

Waisman, C., Sahni, N. S., Nair, H. S., & Lin, X. (2024). Parallel experimentation on advertising platforms, arXiv preprint arxiv: 1903.11198.

Wang, W., Li, B., Luo, X., & Wang, X. (2023). Deep reinforcement learning for sequential targeting. *Management Science*, *69*(9), 5439–5460.

Wernerfelt, N., Tuchman, A., Shapiro, B. T., & Moakler, R. (2025). Estimating the value of offsite tracking data to advertisers: Evidence from meta. *Marketing Science*, *44*(2), 268–286.

Wobbe, B., & Kwok, L. (2023). The next step in personalization: Dynamic sizzles, (last accessed 2025-06-06). https://https://netflixtechblog.com/the-next-step-in-personalization-dynamic-sizzles-4dc4ce2011ef

Xiong, R., Athey, S., Bayati, M., & Imbens, G. (2024). Optimal experimental design for staggered rollouts. *Management Science*, *70*(8), 5317–5336.

Xu, B., Deng, Y., & Mela, C. (2023). A scalable recommendation engine for new users and items. arXiv preprint: arxiv: 2209.06128

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., & Wager, S. (2024). Evaluating treatment prioritization rules via rank-weighted average treatment effects. *Journal of the American Statistical Association*, (pp. 1–14).

Yang, J., Eckles, D., Dhillon, P., & Aral, S. (2024). Targeting for long-term outcomes. *Management Science*, *70*(6), 3841–3855.

Yoganarasimhan, H., Barzegary, E., & Pani, A. (2023). Design and evaluation of optimal free trials. *Management Science*, *69*(6), 3220–3240.

von Zahn, M., Bauer, K., Mihale-Wilson, C., Jagow, J., Speicher, M., & Hinz, O. (2024). Smart green nudging: Reducing product returns through digital footprints and causal machine learning. *Marketing Science*.

Zantedeschi, D., Feit, E. M., & Bradlow, E. T. (2017). Measuring multichannel advertising response. *Management Science*, *63*(8), 2706–2728.

Zhan, R., Ren, Z., Athey, S., & Zhou, Z. (2024). Policy learning with adaptively collected data. *Management Science*, *70*(8), 5270–5297.

Zhang, J., & Krishnamurthi, L. (2004). Customizing promotions in online stores. *Marketing Science*, *23*(4), 561–578.

Zhang, W. W. (2025). Optimal comprehensible targeting, available at: https://marketing.wharton.upenn.edu/wp-content/uploads/2023/09/10.05.2023-Zhang-Walter-PAPER-JMP.pdf.

Zhao, L., Tian, L., Cai, T., Claggett, B., & Wei, L.-J. (2013). Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association*, *108*(502), 527–539.

Zhou, M., Chen, G. H., Ferreira, P., & Smith, M. D. (2021). Consumer behavior in the online classroom: Using video analytics and machine learning to understand the consumption of video courseware. *Journal of Marketing Research*, *58*(6), 1079–1100.