big data and cognitive computing

MDPI

Review

Big Data Analytics and AI for Consumer Behavior in Digital Marketing: Applications, Synthetic and Dark Data, and Future Directions

Leonidas Theodorakopoulos *, Alexandra Theodoropoulou and Christos Klavdianos

Department of Management Science and Technology, University of Patras, 26334 Patras, Greece; theodoropouloua@upatras.gr (A.T.); cklavdianos@ac.upatras.gr (C.K.)
* Correspondence: theodleo@upatras.gr

Abstract

In the big data era, understanding and influencing consumer behavior in digital marketing increasingly relies on large-scale data and AI-driven analytics. This narrative, concept-driven review examines how big data technologies and machine learning reshape consumer behavior analysis across key decision-making areas. After outlining the theoretical foundations of consumer behavior in digital settings and the main data and AI capabilities available to marketers, this paper discusses five application domains: personalized marketing and recommender systems, dynamic pricing, customer relationship management, data-driven product development and fraud detection. For each domain, it highlights how algorithmic models affect targeting, prediction, consumer experience and perceived fairness. This review then turns to synthetic data as a privacy-oriented way to support model development, experimentation and scenario analysis, and to dark data as a largely underused source of behavioral insight in the form of logs, service interactions and other unstructured records. A discussion section integrates these strands, outlines implications for digital marketing practice and identifies research needs related to validation, governance and consumer trust. Finally, this paper sketches future directions, including deeper integration of AI in real-time decision systems, increased use of edge computing, stronger consumer participation in data use, clearer ethical frameworks and exploratory work on quantum methods.

Keywords: consumer insights; behavioral analytics; machine learning; personalized advertising; recommender systems; real-time decision-making; customer journey; customer relationship management (CRM); data governance

Academic Editor: Yin Zhang

Received: 10 December 2025
Revised: 13 January 2026
Accepted: 26 January 2026
Published: 2 February 2026

1. Introduction

Digital technology has spread at high speed, which has allowed businesses to monitor consumer activities through traceable data that can also be manipulated. The combination of online platforms with mobile channels, loyalty programs and connected devices produces endless streams of behavioral information, which includes search data and click patterns, application logs and user locations, transaction history and service usage records. The available data enable marketing decision-makers to better understand user preferences, and offer immediate signal response and synchronized channel operations [1–3]. Research studies show that big data analytics enables digital marketing to achieve better targeting results, improved campaign performance and enhanced customer interactions [4–6]. However, the implementation of algorithm-based marketing systems brings forth new privacy

concerns for consumers alongside issues of fair treatment, deceptive practices and trust problems because these systems control what products appear, at what prices and when messages are delivered [7,8].

Artificial intelligence and machine learning sit at the center of this transformation. In many organizations, they already underpin recommendation engines, dynamic pricing policies, churn and response models, fraud detection systems and A/B testing platforms [9]. These tools do more than forecast outcomes. They structure the decision environment by selecting which options are visible, how information is ordered and which transactions are blocked or challenged. As a result, they reshape not only what firms can observe and predict about consumers, but also how consumers themselves search, decide, form habits and evaluate fairness. Research on AI in marketing and consumer psychology has grown rapidly in recent years, but it often treats data availability as given, focusing on specific applications or technologies rather than on the evolving data landscape that supports them [10].

Two developments illustrate this gap. First, synthetic data techniques are moving from technical niches into mainstream marketing and market research practice. Brands and research providers now use synthetic panels, "digital twins" and other artificial datasets to train and test models, explore scenarios and accelerate insight generation under tightening privacy and access constraints [11,12]. Second, the notion of dark data has gained traction in analytics and consulting circles: organizations sit on large volumes of unexploited information in the form of free-text complaints, call-center recordings, technical logs, sensor streams and legacy CRM files that are rarely analyzed, even though they contain rich behavioral signals [13]. Both synthetic data and dark data are starting to influence how firms study and model consumer behavior, yet their implications for marketing decisions and consumer experience remain under-explored in the academic literature.

The present paper employs a concept-based narrative review to analyze how big data and AI enable digital marketers to understand consumer behavior patterns. It has three main aims. The first objective involves explaining how big data and AI methods operate in fundamental decision-making processes, including personalization and recommender systems, dynamic pricing, customer relationship management, product and service development, and fraud detection (Section 4). The second is to examine how synthetic data practices are beginning to reshape model development, experimentation and scenario analysis in consumer analytics (Section 5). The third objective involves studying dark data as an underutilized resource which could deliver important consumer journey understanding, pain point identification and risk detection (Section 6) while examining its impact on privacy, transparency and governance requirements.

This article is positioned as a narrative review rather than a systematic review or meta-analysis. Section 2 outlines the theoretical foundations, drawing on established work in consumer behavior, digital marketing and technology acceptance to frame how data-driven interventions affect decision processes, perceived intrusiveness and trust. Section 3 briefly describes the methodological approach and scope of the literature considered. Sections 4–6 present the core of this review, moving from operational applications of big data and AI to the emerging roles of synthetic and dark data. Section 7 discusses the main insights, implications for digital marketing practice, research directions and limitations, and sketches several future trajectories for consumer behavior analytics, including deeper integration of AI into real-time decision systems, the growth of edge computing and new forms of consumer participation in data use. Section 8 concludes by reflecting on the opportunities and risks of an environment where consumer behavior is increasingly inferred and acted upon through continuous digital traces, synthetic records and previously unused data sources.

Unique Contribution of This Review

Prior reviews in digital marketing analytics typically treat personalization, pricing, CRM, product innovation, and security as separate application silos, or discuss algorithms without tracing how they reshape consumer decision environments. This review contributes a unified consumer-behavior lens: it treats AI-enabled marketing systems as intervention mechanisms that modify option visibility, choice architecture, and perceived risk in real time, and it links these mechanisms to constructs such as perceived autonomy, fairness, trust, cognitive load, and behavioral adaptation over time.

A second contribution is the joint treatment of (i) AI applications, (ii) synthetic data, and (iii) dark data. Considering these streams together makes visible a recurring tension that is easy to miss when they are reviewed separately: firms increasingly act on consumers via automated decisions built from behavioral traces that may be incomplete (dark data), partially simulated (synthetic data), or inferred beyond what consumers expect. This review therefore enables a research agenda that connects marketing performance questions (conversion, retention, CLV) with legitimacy questions (transparency, proportionality, privacy expectations) in the same analytical frame.

2. Theoretical Foundations

2.1. Consumer Behavior in Digital Marketing

Consumer behavior in digital environments is shaped by a sequence of micro-decisions that occur along the customer journey: problem recognition, information search, evaluation of alternatives, choice, post-purchase evaluation and, in recurring contexts, habit formation [14]. Digital channels change both the amount and the structure of information available at each stage, and this has direct consequences for how people search, compare and decide. Recent work on online shopping shows that information overload and interface complexity increase stress, perceived risk and frustration, and can ultimately reduce purchase likelihood when consumers feel unable to process available options [15,16].

Several behavioral frameworks are particularly relevant for AI-driven digital marketing. Stimulus–Organism–Response (SOR) models are often used to explain how platform features (e.g., personalization, interface design, recommendation cues) act as stimuli that influence internal states such as perceived relevance, trust or privacy concern, which then drive responses such as click-through, purchase and word-of-mouth [17]. Technology Acceptance Model (TAM) constructs—perceived usefulness and perceived ease of use—remain central for understanding adoption of AI-enabled services, including personalized recommenders and conversational agents [18]. Consumers tend to accept these technologies when they see a clear reduction in effort and a meaningful improvement in decision quality, provided that the interface remains understandable and controllable [19].

In the domain of personalization, recent empirical studies highlight a tension between perceived relevance and perceived intrusiveness. AI-driven recommendations and targeted ads can increase purchase intentions when users experience them as accurate and helpful, but the same mechanisms can harm trust when consumers feel over-tracked, manipulated or profiled in opaque ways [20]. Privacy calculus theory and related work on self-disclosure help explain this trade-off: individuals weigh expected benefits (better offers, less search effort, more appropriate content) against perceived costs (loss of control, data misuse, profiling). When the calculus shifts (because of a data breach, perceived unfairness in pricing, or exposure to "dark patterns" that exploit cognitive biases) consumers may reduce information sharing, adopt ad-blocking tools, or switch providers [21,22].

Trust plays a structuring role across these mechanisms. In digital markets, consumers must often judge the trustworthiness of algorithms and platforms rather than individual salespeople. Trust is influenced by transparency of data practices, consistency of prior

experiences, perceived competence of the provider and the extent to which AI systems are seen as aligned with users' interests [23]. Recent research shows that AI-personalized recommendations increase online purchase intentions mainly when they enhance perceived relevance and do not undermine trust [24,25]. This perspective is important for the rest of this paper: applications in Sections 4.1–4.5 (personalization, pricing, CRM, product innovation and fraud detection) do not only change prediction accuracy, they also reconfigure the psychological environment in which consumers make decisions.

2.2. Big Data and AI Capabilities in Marketing

Big data in marketing typically refers to large, fast and heterogeneous data sources that describe consumers, interactions and contexts. These include transactional records, clickstream and app events, search queries, social media activity, text reviews, location traces, sensor data from connected devices and, increasingly, log-level advertising data [3]. The value of this data for marketing lies in their ability to represent behavior at high granularity over time, across channels and at the level of individual devices or accounts. Recent reviews underline that the combination of such data with machine learning techniques has shifted marketing analytics from periodic reporting to continuous, event-driven decision-making [6].

AI and machine learning methods provide the tools to extract patterns and predictions from these data streams. Supervised learning models support tasks such as response prediction, churn scoring, fraud detection and demand forecasting [26,27]. Unsupervised methods are used for behavioral segmentation, anomaly detection and journey clustering [28]. Recommender systems combine collaborative and content-based techniques to personalize product and content rankings. Sequence models, such as recurrent networks or attention-based architectures, capture dynamics in click paths, content consumption and payment histories [29]. These capabilities underpin the operational applications explored in Section 4: they determine which offers are shown (Section 4.1), how prices are adjusted over time and across segments (Section 4.2), which customers are prioritized in CRM actions (Section 4.3), how products and services are iteratively improved (Section 4.4) and how suspicious activity is flagged for fraud control (Section 4.5).

At the same time, new data practices are emerging around synthetic data and dark data. Synthetic data techniques use generative models or simulation frameworks to create artificial records that mimic the statistical properties of real consumer datasets, with the aim of enabling experimentation, sharing and model development under stricter privacy constraints [30]. Dark data refers to information that organizations already store but do not actively analyze, such as detailed service logs, free-text complaints, call recordings or sensor streams [31]. Together, these developments extend the scope of marketing analytics beyond traditional data warehouses. They also raise questions about how far firms should go in extracting behavioral signals from every available source, and how much synthetic reconstruction of consumer patterns is acceptable when real data access is constrained. These questions motivate the dedicated discussion of synthetic data (Section 5) and dark data (Section 6) later in this paper.

2.3. Positioning Relative to Prior Reviews

A large body of review work already examines the impact of digital marketing and big data on consumer behavior. Recent articles synthesize how social media, search and omnichannel strategies affect awareness, attitudes, purchase and post-purchase behavior, often emphasizing personalization, engagement metrics and customer journey management [32,33]. Other reviews focus on AI in marketing more broadly, discussing recommender systems, chatbots, programmatic advertising and predictive analytics, sometimes

with an emphasis on financial services or e-commerce [34]. These contributions provide important overviews of technical methods and application domains, but they tend to treat data availability as given and focus less on how emerging data practices reshape the foundations of consumer analysis.

Work on synthetic data and synthetic respondents has started to appear in marketing and market research outlets, mostly in the context of survey research, experimentation and privacy-preserving analytics [35]. Studies and position papers argue that synthetic data and "synthetic customers" can accelerate insight generation, expand coverage of rare segments and reduce dependence on traditional panels, while simultaneously raising concerns about validity, bias and evaluation. Parallel discussions on dark data highlight the strategic value of unexploited logs and unstructured records for customer insight, but these are often framed from a general business analytics perspective rather than from a consumer-behavior standpoint.

This review builds on those streams but takes a different focus. First, it centers explicitly on consumer behavior and decision-making, using established behavioral constructs, such as information overload, privacy calculus, trust and technology acceptance, to interpret how data-driven marketing applications influence actual choices. Second, it brings synthetic data and dark data into the same analytical frame as more conventional big data applications. Sections 4.1–4.5 examine how current big data and AI techniques are deployed in personalization, pricing, CRM, product innovation and fraud detection. Section 5 then discusses how synthetic data reshapes model development, experimentation and collaboration, while Section 6 considers the opportunities and risks of turning dark data into behavioral insight. This positioning distinguishes this article from prior reviews that either survey digital marketing tools without addressing these emerging data practices, or discuss synthetic and dark data primarily from a technical or infrastructural angle rather than from the perspective of consumer behavior.

3. Methods

In methodological terms, this article adopts a narrative, concept-driven review rather than a systematic review protocol. The main reason is the character of the topic: research on big data analytics and AI-driven marketing is dispersed across technical, behavioral, and sector-specific outlets, while the evidence base changes quickly (e.g., new model classes, data practices, and governance requirements). Under these conditions, a systematic review would prioritize exhaustive retrieval and strict inclusion rules, but it would not necessarily yield a clearer integration of mechanisms that link big data/AI practices to consumer cognition, trust, perceived intrusiveness, and behavioral adaptation. The goal here is therefore synthesis and conceptual organization: to map key application domains and big data practices, consolidate consistent findings, and highlight where claims remain under-tested or context-dependent.

Given the breadth of the topic, we balanced breadth and depth by using a structured set of application areas (Section 4) as an organizing lens, and then integrating synthetic data (Section 5) and dark data (Section 6) as cross-cutting methodological and governance themes. We prioritized domains where AI-enabled decisioning is mature and data-rich (e.g., digital commerce/platforms, financial services, and subscription-based services), because these settings provide clearer evidence on both marketing impact and consumer-facing risks. This narrative approach necessarily introduces selection bias and may under-represent low-data, offline, or emerging contexts; we mitigate this by triangulating recurring claims across multiple research streams and by treating contested points as open problems rather than settled conclusions.

Sources were chosen for topical relevance and conceptual contribution to consumer-behavior implications of AI-enabled marketing decisions, rather than exhaustive coverage of every industry or method.

The relevant literature was identified (mainly 2015–2025) through iterative searches in major academic databases, including Scopus, Web of Science, IEEE Xplore, ACM Digital Library and Google Scholar. Search strings combined terms related to big data analytics, AI and machine learning, consumer behavior, personalization, pricing, CRM, fraud detection, synthetic data and dark data. We prioritized peer-reviewed journal articles, full conference papers and influential books that offer clear methodological descriptions. Additional references were located through backward and forward citation tracking. Studies were selected for inclusion based on topical relevance and conceptual contribution; no formal quality scoring, effect-size extraction or quantitative synthesis was undertaken.

4. Applications of Big Data in Consumer Behavior Analysis

The integration of big data analytics into consumer behavior analysis has substantially changed how businesses understand, anticipate and shape customer actions. Businesses use extensive amounts of structured and unstructured data to generate valuable insights which enable them to create personalized marketing campaigns, build predictive models, segment customers, and analyze consumer sentiment. The applications enable businesses to enhance their targeting abilities while making adjustments to their communication and pricing methods, and tracking consumer reactions throughout time. Thus, organizations achieve better customer satisfaction, along with simpler decision-making and longer business relationships through their improved customer engagement and conversion rate enhancement [36,37].

4.1. Personalized Marketing and Recommender Systems

Businesses are trying to develop customized marketing strategies through particular promotional approaches, various recommendation platforms and email marketing campaigns that utilize consumer information and present-day market conditions. Companies use historical consumer actions, together with forecasted consumer tastes and signal information, to generate personalized marketing materials which they then distribute through particular channels, at exact times. This tactic generates higher value from customer interactions while simultaneously boosting conversion rates and strengthening brand loyalty. These personalization mechanisms provide consumers with suitable options that match their needs, which simplifies their search and decision-making process [20,38,39].

Big data technologies enable businesses to create personalized user profiles through the integration of data from various sources which include purchase records, browsing activities, search data and application usage, social media activities and fundamental demographic information. The processed data produces evaluation results and market segments that direct specific business choices about product placement, promotion display, email content and mobile application notifications. These e-commerce platforms show customers additional items based on their purchase history, while streaming services recommend new content through user viewing patterns [40,41].

Recommender systems function as the core element that makes successful personalized marketing possible through their ability to support these methods. Personalization operates through recommender engines because they use user data to determine product, content and advertisement rankings for individual users. When implemented well, recommender mechanisms are associated with better click-through rates, higher platform usage and customer purchases of various products. However, they generate user fatigue and perceived

lack of relevance when they do not understand user preferences and fail to recognize recent behavioral shifts, which is what leads to customer churn [42,43].

Recommender systems work based on a range of machine learning algorithms and predictive modeling techniques that run in the background. The most common approaches include collaborative filtering, content-based filtering and various hybrid models. The more recent systems also integrate deep learning components that capture complex patterns in user behavior and allow recommendations to adapt to constantly changing preferences [44,45].

4.1.1. Collaborative Filtering

Collaborative filtering is one of the most widely used recommendation techniques. It is based on the principle that users who behaved similarly in the past, are likely to show similar preferences and patterns in the future. Collaborative filtering relies on user–item interaction data, such as purchases, clicks, ratings or viewing events, rather than focusing on product attributes [46].

In user-based collaborative filtering, the system recommends items to a target user u by first identifying other users with similar behavior. The similarity between two users u and v can be computed using cosine similarity:

$$Sim(u,v) = \frac{\Sigma_{i \in I} r_{u,i} \cdot r_{v,i}}{\sqrt{\Sigma_{i \epsilon I} r_{u,i}^2} \cdot \sqrt{\Sigma_{i \in I} r_{v,i}^2}} \quad (1)$$

Here, $r_{u,i}$ is the rating (or implicit feedback score) given by user u to item i, and I is the set of items that both users have interacted with. $Sim(u,v)$ summarizes how closely aligned their interaction patterns are. Equation (1) is shown as a representative example; in operational systems, similarity and neighborhood formation are often adapted to the platform's feedback type and sparsity. By prioritizing items endorsed by behaviorally similar users, user-based filtering can reduce search effort, but it can also narrow exposure if similarity is defined too rigidly.

After computing similarities, recommendations are generated by aggregating the preferences of the most similar neighbors for items the target user has not yet consumed. In practice, systems restrict aggregation to a limited set of nearest neighbors to reduce noise, improve scalability, and prevent weakly related users from dominating the recommendation.

Item-Based Collaborative Filtering

Item-based collaborative filtering shifts the focus from users to items. Instead of identifying similar users, it identifies similar items based on the ratings they receive and recommends products that are often co-rated or co-purchased. Item similarity is estimated from overlapping audiences and co-consumption patterns, using standard similarity measures applied to item interaction profiles [47]. This approach is particularly common in e-commerce settings. When many users who purchased item A also purchased item B, and their evaluations of both items are consistently high, the system learns that A and B are related. When a new consumer buys A or repeatedly interacts with it, item-based collaborative filtering suggests B and other similar products. This directly informs cross-selling and upselling decisions and affects how consumers explore product assortments [48].

4.1.2. Content-Based Filtering

Content-based filtering generates recommendations by analyzing item attributes rather than patterns of co-consumption. Each item is described by a set of features, such as product category, brand, technical characteristics, textual description, genre or cast. The system builds a profile of each user by summarizing the attributes of items they have previously interacted with and recommends new items that share similar attributes [49].

To compare items, content-based filtering often uses representations derived from TF–IDF (Term Frequency–Inverse Document Frequency) for textual data and similarity measures such as cosine similarity [50]. Operationally, a content-based recommender scores a candidate item by comparing its feature representation (e.g., keywords, categories, embeddings) to the feature profile derived from the user's previously consumed items. Items that are more similar to the user's profile receive higher ranks, and the scoring can be weighted to reflect stronger signals from items the user repeatedly views, purchases, or rates highly. For example, in a movie recommender system, if a user repeatedly watches science-fiction films with particular actors or directors, the profile will emphasize these attributes, and the system will surface new titles with similar characteristics.

From a consumer-behavior perspective, this attribute-driven matching can increase perceived relevance and reduce search effort, but it may also feel repetitive or overly narrow if the system over-weights a small set of past preferences.

From a decision-making perspective, content-based methods are attractive when detailed item descriptions are available, but user–item interaction data are sparse or new items are frequently added. They help decide which recently released products to highlight for each user and support long-tail discovery without waiting for extensive historical feedback [51].

4.1.3. Hybrid Methods

Hybrid recommender systems combine collaborative and content-based information to mitigate the limitations of each individual approach. They can, for example, use item attributes to handle cold-start situations, while still exploiting user–user and item–item correlations when enough interaction data are present. Another strategy is to compute predictions separately with collaborative and content-based models and then merge them [52].

Conceptually, hybrid methods combine the outputs of collaborative filtering and content-based filtering into a single ranking score. The contribution of each component can be fixed (a single global weighting) or adjusted dynamically depending on data availability, user segment, or item type, so that content signals dominate in cold-start settings while collaborative signals dominate when interaction histories are rich. This blending strategy improves stability across sparse and noisy environments and reduces the risk that recommendations collapse when one signal source is weak [53].

In practice, a music streaming service, for example, may first use content-based filtering to propose songs that share acoustic or stylistic features with tracks the user already likes, and then refine this list using collaborative signals from similar listeners. This improves robustness when some artists or tracks are new and interaction data are still limited. It also stabilizes recommendations when user feedback is noisy or inconsistent. From the consumer perspective, hybridization can increase perceived relevance and variety (less repetition), but poorly calibrated blending can still produce inconsistent recommendations across sessions, which may reduce trust in the system's reliability.

4.1.4. Advanced Predictive Modeling and Deep Learning

Modern recommender systems increasingly incorporate advanced predictive models and deep learning architectures. Traditional collaborative and content-based techniques can struggle with sparse data, rapidly changing preferences and complex interaction effects. Deep learning allows systems to learn non-linear relationships and to integrate heterogeneous data sources, such as text, images and sequences of actions, in a unified framework [54].

Neural Collaborative Filtering (NCF)

Neural Collaborative Filtering (NCF) represents one of the most famous deep learning solution for recommender systems because it uses neural networks to replace conventional similarity functions. The model uses non-linear transformation layers to find intricate user–item connections which go beyond what matrix factorization methods can achieve. It also uses deep learning to discover complex user–item relationships by transforming user and item data into high-dimensional latent vectors [55]. The NCF model produces rating predictions by using the following mathematical equation:

$$\hat{r}_{u,i} = f(W_2 \cdot \sigma(W_1 \cdot [v_u, v_i] + b_1) + b_2) \quad (2)$$

In this formulation, the user and item are represented as latent vectors that are combined and transformed through non-linear layers to learn complex interaction patterns. The weight matrices and bias terms are learned during training, allowing the model to capture relationships that are not well approximated by linear similarity measures. This flexibility is particularly useful in e-commerce and streaming settings where preferences shift over time and feedback is largely implicit [56].

Recurrent Neural Networks (RNNs)

RNNs serve as essential tools for recommender systems which require users to make sequential decisions such as music-streaming, playlist suggestions and YouTube video recommendations. RNNs allow systems to learn from user interaction history because they learn to recognize time-based patterns between events. The hidden states of RNNs store information about previous user interactions which enables the system to generate recommendations that adapt to recent user behavior [57].

Operationally, an RNN updates an internal state after each interaction, so recent actions influence the next recommendation more strongly than older ones. This makes RNN-style models suitable for session-based settings where order and recency matter (e.g., playlists, short video feeds), and where the system must adapt to changes in intent within a single browsing session [57].

In practice, several platform implementations use multiple deep learning components which work together as a system. Product image processing can occur through Convolutional Neural Networks (CNNs), for example, while transformer-based models process textual reviews and descriptions. The analysis of clickstream and viewing history data depends on sequence models, which include RNNs and attention-based architectures. These networks use combined model outputs to determine item rankings, select thumbnails and default options, and schedule notifications [58].

From a customer's perspective, these AI-enhanced recommender systems shape how customers explore their options, build habits and perceive their autonomy. High-quality recommendations can help consumers discover relevant products or content they would not have found easily on their own, and reduce the effort required to navigate large collections of products. At the same time, if recommendations become too narrow or overly persuasive, consumers may feel constrained or manipulated. This tension highlights the need to design recommendation policies that take behavioral responses, fairness considerations and privacy concerns into account, not only predictive accuracy [59].

Overall, the marketing value of recommender systems depends not only on predictive accuracy but on how recommendations are experienced in context. When ranking policies balance relevance with diversity, and provide stable, understandable signals, they can reduce search costs, support exploration, and increase satisfaction and repeat engagement. Conversely, when personalization is overly narrow, volatile across sessions, or perceived as surveillance-driven, it can trigger reactance, reduce perceived autonomy, and weaken trust, which ultimately limits conversion quality and long-term loyalty.

4.2. Dynamic Pricing Strategies

The practice of dynamic pricing involves changing product or service prices based on shifting market demand, supply levels, production expenses and market competition. Companies use a multitude of data available to them, such as transaction data, inventory levels, booking patterns, competitor prices, individual browsing activities and loyalty status to set their prices instead of using fixed rates. This approach uses pricing rules and predictive models to adjust prices based on different time intervals that range from e-commerce hourly changes to ride-hailing platforms that update prices every second [60,61].

In online retailing, a platform may increase the price of a popular item during peak demand periods, when inventories are limited the most, while lowering prices for slow-selling products in order to stimulate sales and clear stock. Airlines, hotels and event organizers use dynamic pricing techniques to vary fares and ticket prices according to booking time, remaining capacity, seasonality, expected demand peaks and special events. Ride-hailing services modify their pricing through real-time adjustments that depend on local market supply, demand levels, travel duration and traffic patterns. In each of these settings, price is no longer a static parameter but a decision variable, a dynamic element that is continuously updated based on incoming data streams [62,63].

Mathematical Models in Dynamic Pricing

Dynamic pricing relies on models that link prices to expected demand and revenue under operational constraints. A common starting point is price elasticity, which summarizes how responsive demand is to price changes and can differ substantially across segments, product categories, and contexts. Revenue management extends this logic by selecting price levels and availability rules over time under capacity constraints (e.g., seats, rooms, limited stock), often using historical booking patterns and contextual signals (seasonality, events, day-of-week) to allocate inventory across customers with different willingness to pay.

Recent work increasingly uses machine learning to predict demand under alternative candidate prices by mapping contextual features (such as competitor prices, promotional activity, device type, and user history) to expected sales volumes [64]. Pricing policies can then be chosen to maximize expected revenue or profit, subject to constraints such as minimum margins, inventory limits, and explicit business rules. Reinforcement learning approaches further adapt pricing policies based on observed outcomes, gradually learning which actions perform better in specific market states [65].

However, predictive performance and revenue optimization do not automatically translate into marketing success, because consumer responses are mediated by reference prices, perceived fairness, and trust.

Consumer Behavior and Perceived Fairness

When it comes down to consumer behavior, dynamic pricing affects not only purchase incidence and timing but also trust and brand image. Consumers often develop their own personal price standards through their past shopping experiences and their observations of what others pay. When prices move beyond what customers expect, they will view it as an unfair practice, especially when there is no obvious reason for such change [66]. For example, a sharp price increase for last-minute tickets during a natural disaster, or a surge multiplier in ride-hailing during an emergency, creates strong negative reactions from customers even if the pricing algorithm is actually working as intended.

By contrast to the above, transparent and predictable pricing strategies, such as early-bird discounts, off-peak fares, "weekday vs. weekend" differences or loyalty-based price tiers, tend to be more acceptable by most customers. Price changes become more understandable for consumers when businesses explain their pricing system behind them, rather than using arbitrary exploitation methods. In competitive markets, when consumers are repeatedly exposed to extreme or opaque price variability, they may be encouraged to search

more actively for alternatives, rely on comparison sites and adopt defensive strategies (e.g., waiting for promotions, switching providers), which in turn affects negatively long-term loyalty and lifetime value [67,68].

Practically, this means that businesses need to find a balance between short-term revenue gains, from aggressive dynamic pricing, and potential long-term costs associated with reduced trust from consumers, heightened perceived risk and negative word-of-mouth. Incorporating behavioral insights into pricing models (for instance, by setting bounds on what price "movements" are allowed, or by respecting psychological thresholds) helps align algorithmic decisions with how consumers actually perceive prices, not only with predicted demand curves.

Accordingly, effective dynamic pricing requires not only accurate demand prediction but also policy constraints and communication choices that protect perceived fairness and long-term relationship value.

4.3. Customer Relationship Management (CRM)

Nowadays, the business world has become very competitive. Customer relationship management (CRM) is now a useful tool for enhancing customer loyalty, keeping or increasing retention rates and building long-term relationships. By using big data, artificial intelligence (AI) and predictive analytics, CRM systems provide businesses with data-driven insights that support more precise decisions about which customers to contact, what to offer them, when to reach out and through which channel [69]. Such ability to understand individual behaviors and their preferences, as well as purchasing patterns, allows companies to design interactions that feel more relevant, while at the same time, monitoring the value and risk profile of each relationship. When this is done well, it strengthens brand attachment and increases customer lifetime value (CLV) [70].

CRM systems work as centralized platforms that are made to collect, process and analyze customer data from multiple touchpoints, including website interactions, mobile app usage, social media engagements, purchase histories, customer service contacts and feedback surveys. By pulling and integrating data across these sources, CRM tools construct comprehensive and extensive customer profiles. These profiles help businesses anticipate future behaviors and cater to personalized actions accordingly [71]. For example, an e-commerce platform may track a customer's browsing behavior and identify a strong interest in a specific product category. Based on this, the CRM system can automatically trigger personalized emails, on-site banners with exclusive bundles, or push notifications recommending products that match the customer's preferences [72]. A banking institution, for instance, can segment clients according to spending patterns and financial goals, then propose targeted savings plans or advisory services that correspond to each segment's needs.

Figure 1 illustrates a typical big-data-enabled CRM architecture, showing how raw data from different channels are consolidated into customer profiles and transformed into decision scores that guide marketing, sales and service actions.

One of the most powerful aspects of modern CRM is its ability to integrate predictive analytics and machine learning in order to anticipate customer needs and likely behaviors [70]. By analyzing historical data, transactional patterns and customer feedback, CRM platforms identify trends and infer which products or services a customer is likely to seek next, which customers have high potential value and which show early signs of disengagement [73]. This predictive capability allows firms to move from reactive responses to proactive engagement: instead of waiting for customers to complain, cancel or reduce activity, the system highlights who should receive preventive offers, service calls or educational content [74].

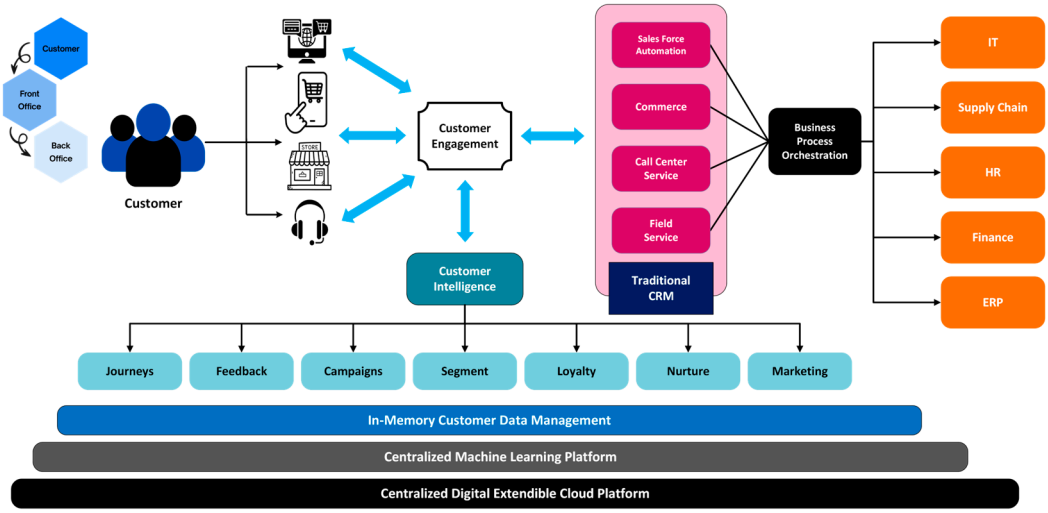

Figure 1. Customer-centric business architecture linking CRM to customer experience.

Predictive CRM also reshapes retention work because it keeps updating its estimates whenever fresh data come in. As more interactions pile up, the underlying models are adjusted so that their forecasts stay closer to what customers currently want and to what is actually happening in the market [73]. This kind of ongoing recalibration matters a lot in fast-moving settings such as online retail, banking and insurance, telecom operators, or streaming and gaming services, where tastes and habits shift quite quickly. On top of that, decision-support components inside modern CRM systems use these outputs to fine-tune day-to-day management: they suggest when it is a good moment to reach out, recommend different channels for different people, rank which customers or tickets should go first to human staff, and highlight which at-risk customers might respond better to discounts, loyalty rewards, or more personalized service [75].

Beyond operational efficiency, these CRM capabilities also have direct implications for how customers perceive the firm and how they respond over time. These mechanisms can significantly shape how people behave. When CRM-based messages are reasonably accurate, not too pushy, and arrive at moments that make sense, customers are more likely to feel that the company understands them and treats them with some respect. That feeling tends to build trust and lowers the effort they think they need to spend to get things done. The opposite also happens: if CRM campaigns keep sending off-target offers, repeat the same promotion again and again, use intrusive tracking, or treat customers differently without a clear reason across channels, many people react with annoyance, feel their privacy is being ignored, and may decide to leave sooner. Because of this, CRM strategy should not rely only on predictive precision or short-term response metrics. It needs to account for contact fatigue, privacy and data-use expectations, and basic fairness concerns, otherwise the system risks optimizing short-run metrics at the expense of long-term relationships [76–78].

4.4. Product Development and Innovation

Bringing consumer data into product development and innovation has reshaped the way firms design, test and adjust their products and services. Instead of depending mainly on small-scale surveys or a few focus group sessions, companies can now track behavior in a much more systematic way, using clickstream paths on websites, in-app interaction logs, purchase and subscription records, star ratings and written reviews, customer support or call-center transcripts, as well as public conversations on social platforms [79,80]. Taken together, these traces offer a richer view of how offerings are used in everyday life: which functions attract repeated use, which options remain almost invisible, the points in the

journey where users get stuck or abandon a process, and how attitudes and preferences shift in the weeks and months after launch. Product-related choices therefore rely less on guesswork or internal opinion and more on patterns that are directly observable in consumer behavior [81].

In digital environments, product teams monitor user journeys in fine detail. They can see where users drop out of a sign-up process, which interface elements are rarely touched, how often specific features are activated and how long sessions last under different configurations. This information feeds into decisions about interface redesigns, onboarding flows, default settings and help content [82]. For example, if a mobile app records that a new feature is opened by many users but quickly abandoned, this pattern may indicate misunderstandings about its purpose, a hidden usability problem or a mismatch with expectations set by marketing messages. In e-commerce, similar analyses of navigation paths and basket contents reveal which combinations of items are frequently bought together, which filters are used and how changes to search or recommendation layouts influence exploration and cart completion [83].

Figure 2 illustrates a data-driven product development cycle, where consumer interactions generate behavioral signals that are aggregated, analyzed and translated into design hypotheses, tests and any possible revisions necessary.

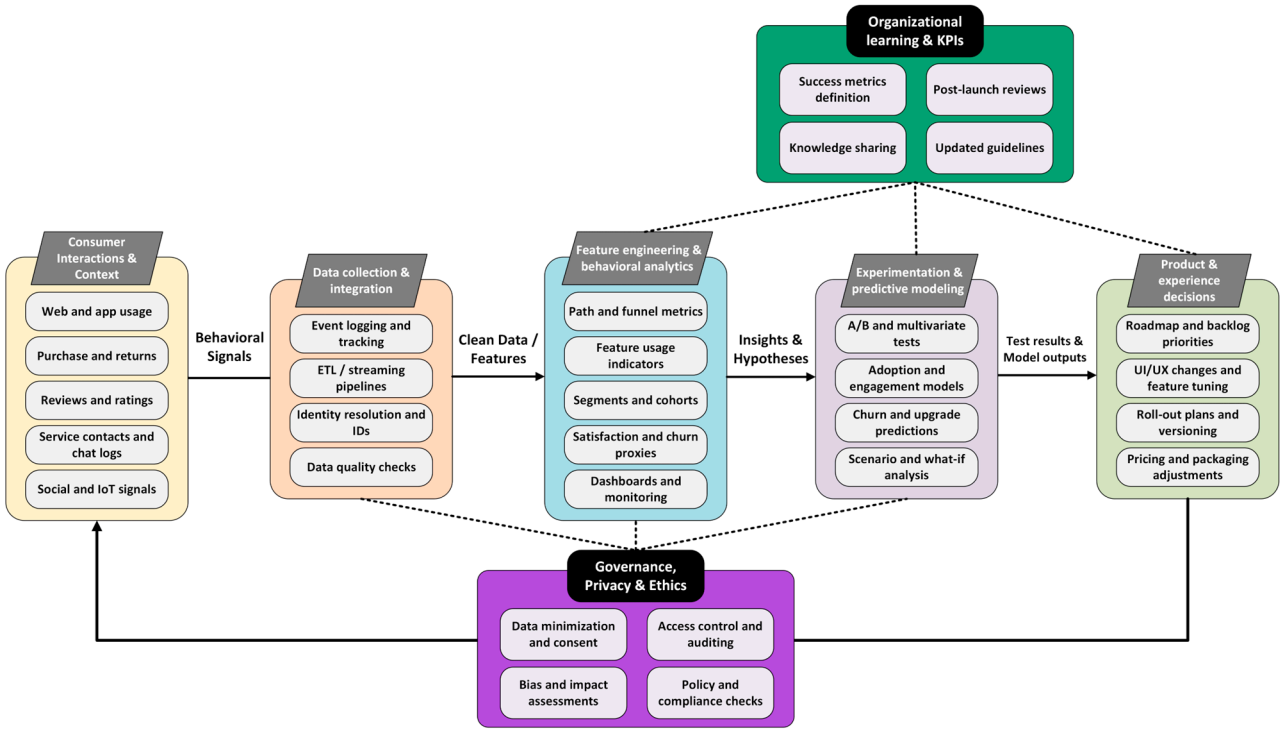

Figure 2. Data-driven product development cycle.

A vital "tool" in this process is experimentation. Many firms run controlled A/B or multivariate tests in which subsets of users are exposed to different versions of a page, feature or price presentation. Key outcomes such as click-through rates, completion of a task, time to purchase and subsequent engagement are compared across variants. When a new design significantly improves these metrics without harming other important indicators, it can be rolled out more widely. When results are ambiguous, teams may refine hypotheses and run additional experiments. This iterative approach allows companies to learn directly from real usage rather than relying only on stated preferences [84,85]. From a consumer standpoint, the same changes can be experienced either as helpful refinement or

as unpredictable 'interface churn', depending on how frequently they occur and whether the rationale is communicated.

Predictive modeling and simulation sit next to experimentation rather than replacing it. When firms bring together past sales records, individual-level attributes and contextual signals like season, device, location or recent campaign exposure, they can estimate what is likely to happen if a specific feature is added, tweaked or removed [86]. On top of that, scenario analyses let teams play out alternative design choices and see how these might influence take-up, satisfaction scores or revenue under different competitive or macro conditions [87]. In subscription models, churn prediction is especially important: by linking cancellations to concrete elements of the experience (for example, onboarding flow, content mix, pricing frictions or technical issues), managers get a clearer view of where product investments are most likely to keep people from leaving [26,88].

Beyond improving design choices internally, these analytics and experimentation practices shape how consumers experience stability, transparency, and control during product evolution.

Looking at it from a consumer behavior angle, data-informed product work has several consequences. When updates clearly respond to real problems that users face—cleaner navigation, functions that match everyday tasks better, more stable performance—people tend to see the product as more useful and less effortful to use. That supports initial adoption, makes repeated use more likely and often leads to more genuine recommendations to others [89]. However, very fast or continuous change can also backfire if the logic behind it is not explained. Interfaces that move around too often or features that appear and disappear without warning can create uncertainty and fatigue. A further concern is that teams may chase easy-to-measure outcomes such as clicks, scroll depth or minutes spent, while ignoring harder but more meaningful aspects of welfare, including mental load, perceived control over one's choices or satisfaction over a longer period. A more balanced product strategy therefore needs to combine behavioral understanding with statistical gains, instead of treating short-term metrics as the only goal [90].

The same logic of big data-driven innovation extends beyond apps and websites. Many physical products now come with embedded sensors or connected services that send usage information back to the producer. With these streams, firms can spot recurring failure modes, see which features are used in unexpected ways and design updated versions that fit real-world routines more closely [91]. At the same time, unstructured feedback from online reviews, fora and social media can push changes in packaging, size, flavor profiles, bundle composition or after-sales services. Across both digital and physical settings, the shared pattern is that consumer behavior is treated as a continuous feedback loop for product development, not as a one-off input collected before launch and then forgotten [92].

Important to mention, treating consumer behavior as a continuous feedback loop also raises questions about perceived legitimacy of data capture. When product improvement is coupled with clear communication, predictable update practices, and meaningful controls (opt-outs, preference settings, data minimization), consumers are more likely to interpret innovation as responsive rather than extractive. Without these safeguards, the same data-driven cycle can be perceived as surveillance or manipulation, which weakens trust and can undermine adoption and long-term loyalty.

4.5. Fraud Detection and Prevention

Fraud detection and prevention are now core parts of consumer analytics, especially in areas like online retail, banking and credit, insurance services, and digital or mobile payments. As fraud tactics keep changing and the number of possible entry points grows, firms increasingly turn to anomaly detection techniques and a range of machine learning

models to flag unusual behavior almost as it happens. Such monitoring models scan past transactions, habitual spending or usage patterns, device fingerprints, login activity and other digital traces to separate normal customer behavior from actions that look suspicious or inconsistent. In this way, they help limit direct financial losses and contribute to a stronger security posture overall. However, their configuration has an immediate impact on customers: if the rules are too strict, genuine payments and account actions may be blocked or delayed; if they are too loose, consumers are left more exposed to unauthorized use and identity abuse [93,94].

Anomaly detection focuses on deviations from normal behavior. In a fraud context, anomalies correspond to unusual transactions, irregular purchasing sequences or atypical access attempts that do not fit the expected profile of a user, a device or a group. Because fraud is typically rare and heterogeneous, rule-based systems alone are insufficient, and statistical or machine learning models are used to detect subtle irregularities. Two major families of such approaches are unsupervised anomaly detection, which works without labeled examples of fraud, and supervised learning, which uses known fraud cases for training [95].

Unsupervised Anomaly Detection

Unsupervised anomaly detection models the distribution of normal transactions and flags data points that are highly unlikely under this distribution. A common strategy is to represent each transaction through a set of behavioral and contextual features and approximate the density of normal data using methods such as Gaussian Mixture Models (GMMs) or Kernel Density Estimation (KDE) [96]. Under density-based approaches such as KDE, the model builds a smooth estimate of what "normal" transactions look like in the observed feature space, and assigns each new transaction a plausibility score. Transactions that receive very low plausibility scores, relative to historical normal behavior, are treated as anomalies and may be routed to additional verification, manual review, or temporary decline. In practice, performance depends strongly on feature design and on calibration choices (e.g., how strict the anomaly threshold is), because these settings determine the trade-off between catching rare fraud patterns and avoiding excessive false alarms [97].

Other unsupervised techniques, such as clustering-based methods or distance-based outlier detection, follow a similar logic: they construct a representation of typical behavior and then identify points that lie far from clusters or have unusually large distances to neighbors. These approaches are particularly useful when labeled fraud data are scarce, when fraud patterns change quickly or when organizations wish to monitor new channels where supervised models are not yet available [98,99].

Supervised learning techniques

Supervised learning methods for fraud detection rely on labeled datasets in which each transaction is annotated as fraudulent or legitimate. The goal is to learn a mapping from transaction features to class labels that generalizes well to unseen data. Common models include logistic regression, decision trees, ensemble methods, support vector machines (SVMs) and various neural network architectures [100].

Decision trees classify transactions by recursively splitting the feature space according to conditions on variables such as transaction amount, merchant category, country, or time of day. At each internal node, the model evaluates a feature and splits transactions into branches based on whether the feature crosses a learned threshold.

The tree grows until a stopping criterion is met, such as a minimum number of samples in each leaf. Each leaf node corresponds to a region of the feature space where the majority of transactions are either fraudulent or legitimate, and new transactions are classified according to the leaf they fall into. Trees are easy to interpret and can reveal intuitive patterns, for example combinations of high amounts, unusual locations and specific merchant types

that tend to indicate fraud [101,102]. In many applications, individual trees are combined into random forests or gradient-boosted ensembles to improve predictive performance. A typical example of a decision tree classifying a transaction as fraudulent or not is depicted below, in Figure 3.

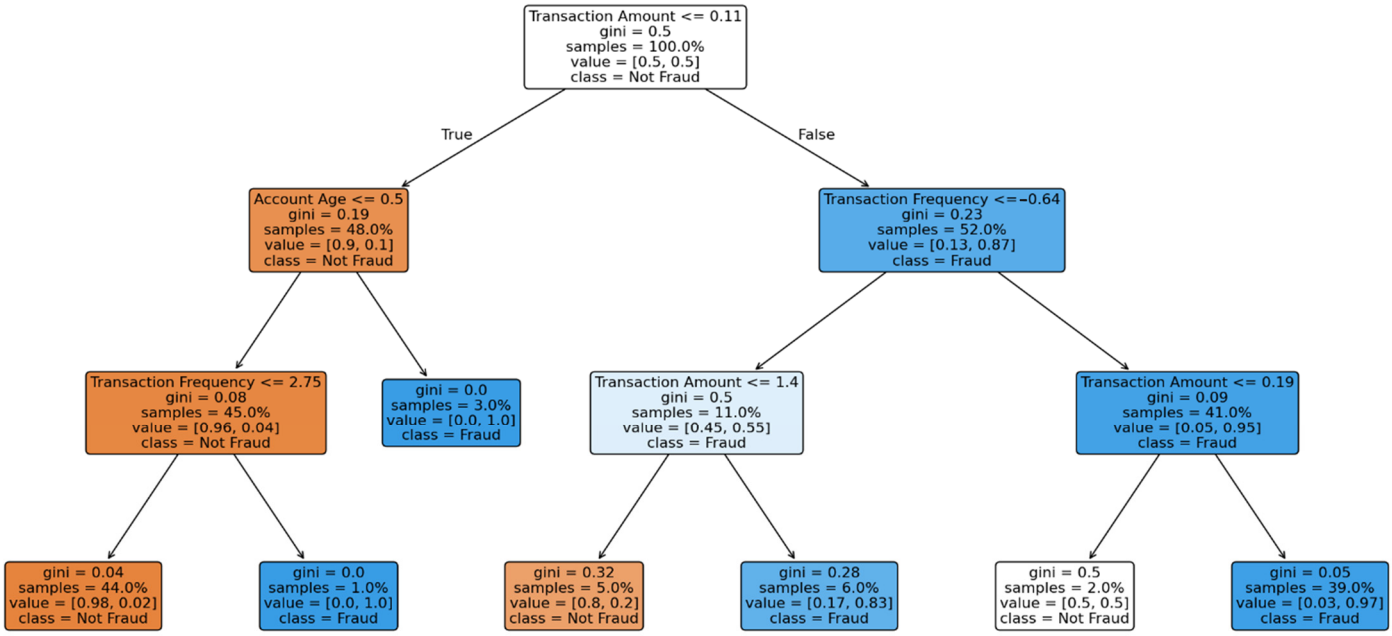

Figure 3. Illustrative decision tree structure for fraud classification using example transaction-related features. Note: The figure is a schematic illustration intended for conceptual explanation, not an empirical model output from this study.

Support vector machines (SVMs) provide another supervised option, especially when fraudulent cases are rare and the boundary between classes is complex. Given labeled data, a SVM seeks a separating hyperplane:

$$w^T X + b = 0 \tag{3}$$

that separates fraudulent from legitimate transactions with maximum margin. Here, X denotes the transaction feature vector, while w and b define the separating boundary. Kernel functions can map inputs into higher-dimensional spaces, enabling non-linear separation when fraud patterns are complex. In practice, SVMs are often combined with techniques for handling class imbalance, such as cost-sensitive training or resampling, to ensure that rare fraud cases influence the model appropriately [103,104]. Figure 4 provides an illustrative visualization of a classifier boundary in a fraud-detection setting.

Lastly, deep learning models, including feedforward neural networks (FNNs), recurrent neural networks (RNNs) and autoencoder-based architectures, are used when very large transaction streams or sequential behaviors must be analyzed [105,106]. A neural network consists of layers of neurons, where each neuron transforms its input using an activation function such as ReLU or sigmoid [107]. For binary fraud detection, these models are trained to produce a probability that a transaction is fraudulent and to reduce prediction error through standard supervised optimization. In practical terms, training adjusts network weights so that combinations of features and sequences that repeatedly co-occur with confirmed fraud are assigned higher risk scores, while normal behavior is assigned low risk. Sequence-aware architectures (e.g., RNN-style models) can additionally exploit temporal dependencies, capturing patterns such as rapid sequences of small test

transactions followed by a large purchase, which often precede card compromise [108]. Figure 5 illustrates a simple example of an FNN trying to mark a possible account takeover.

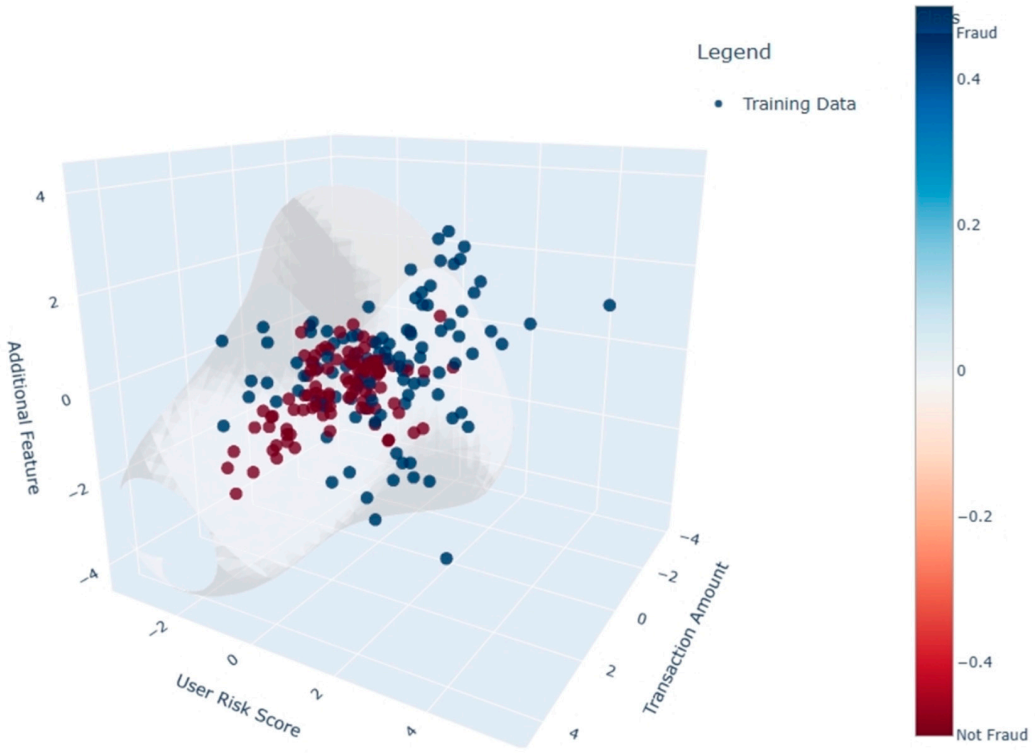

Figure 4. Illustrative 3D visualization of a fraud-detection classifier decision boundary in feature space (conceptual example; not an empirical model output).

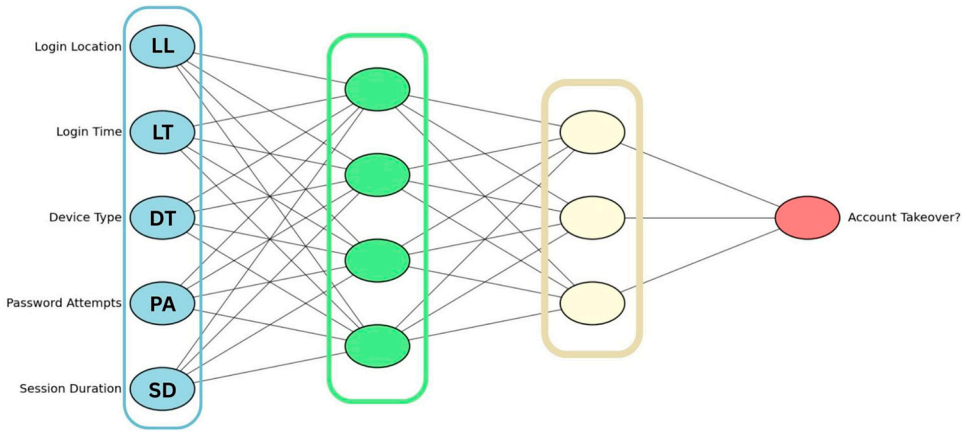

Figure 5. Illustrative neural-network architecture for fraud detection using example behavioral and device-related features. Note: This schematic is included for conceptual explanation and does not represent a specific trained model or empirical results.

Additional Machine Learning Techniques

Several additional techniques are common in fraud analytics; they are listed here briefly for completeness, as their core logic overlaps with methods already discussed above:

- Isolation Forests isolate anomalies by recursively partitioning the feature space at random; fraudulent transactions, being rare and different, tend to be isolated in fewer steps [109].

- Autoencoders learn to reconstruct normal transaction patterns; large reconstruction errors indicate that an input deviates from what the network has seen during training [110].
- Hidden Markov Models (HMMs) model sequential spending behavior and can identify abrupt shifts in transaction sequences that may correspond to account takeover [111].
- Graph-based methods represent relationships among cards, accounts, merchants and devices as networks, highlighting suspicious subgraphs that correspond to coordinated fraud rings [112].

These methods are often used in combination, for instance by running unsupervised detectors first to filter a large stream and then applying supervised models to high-risk candidates.

Applications and behavioral implications

Fraud detection systems are deployed across industries to protect both organizations and consumers. In banking and payments, transaction monitoring engines score each operation in real time and trigger step-up authentication, temporary blocks or manual review for high-risk cases [113]. In e-commerce, anomaly detection models evaluate orders using features such as mismatch between billing and shipping addresses, unusual purchase volumes, atypical device fingerprints and prior chargeback history [114]. Digital platforms and online marketplaces also apply fraud analytics to detect fake account creation, bot activities and abusive promotion usage [115]. Behavioral biometrics, such as typing rhythm, mouse trajectories and touchscreen gestures, add another layer by constructing profiles of how legitimate users typically interact with systems and flagging deviations that suggest hijacked accounts [116].

These mechanisms have direct effects on consumer behavior. Effective fraud detection increases perceived security, which encourages the adoption and continued use of digital payment channels and online services. However, false positives that unnecessarily block cards, decline legitimate orders or force repeated identity checks can cause frustration, embarrassment at the point of sale and eventual switching to competitors [117,118]. The way in which alerts and interventions are communicated—how transparent the rationale appears, how quickly issues are resolved and how much control customers feel they retain—shapes trust as much as the underlying model performance [119,120]. Designing fraud detection and prevention systems therefore involves a trade-off: minimizing losses and deterring attackers, while keeping friction at a level that consumers perceive as reasonable and protective rather than punitive. From a marketing perspective, the objective is therefore not only loss reduction but also maintaining perceived fairness and proportionality of interventions, so that security controls increase confidence without damaging relationship quality.

5. Synthetic Data and Its Role in Consumer Behavior Analysis

5.1. What Is Synthetic Data

Synthetic data are artificially generated information that imitates the statistical and behavioral properties of real datasets without reproducing individual records. A key distinction is that synthetic data can match statistical patterns (distributions, correlations) while still failing to preserve behavioral mechanisms (e.g., reference dependence, habit formation, social influence) that drive real consumer decisions. In consumer behavior analysis, it is used to approximate purchase patterns, browsing trajectories, response histories and other signals that would otherwise be difficult, expensive or risky to share. By constructing realistic but fictitious consumers, firms can train models, test scenarios and stress-test decision rules while limiting direct exposure of sensitive data. Synthetic datasets are particularly useful in domains where privacy regulation is strict, where data access is fragmented across organizations or where the events of interest (e.g., rare types of fraud or very specific customer journeys) are scarce in observed logs [121].

A useful distinction for consumer-behavior work is between technical validity and behavioral validity. Technical validity refers to statistical similarity—whether synthetic data reproduce distributions, correlations, and aggregate patterns observed in real datasets. Behavioral validity is stricter: it asks whether the data preserve decision realism and psychological plausibility, such as stable segment differences, reference-price effects, habit formation, and realistic switching between states (e.g., from browsing, to cart, and then to purchase) [122].

Several families of methods are used to generate synthetic consumer data. A prominent class relies on generative models such as Generative Adversarial Networks (GANs) and variational autoencoders (VAEs). In a GAN, a generator network G produces candidate synthetic samples from random noise, while a discriminator network D tries to distinguish real from synthetic data. Training proceeds as a minimax game in which G improves at fooling D, and D improves at detection. When training converges, the distribution of G(z) for random inputs z approximates the distribution of real observations, such as transaction vectors or clickstream sequences [123]. Figure 6 illustrates the key components and information flow in a typical GAN set-up.

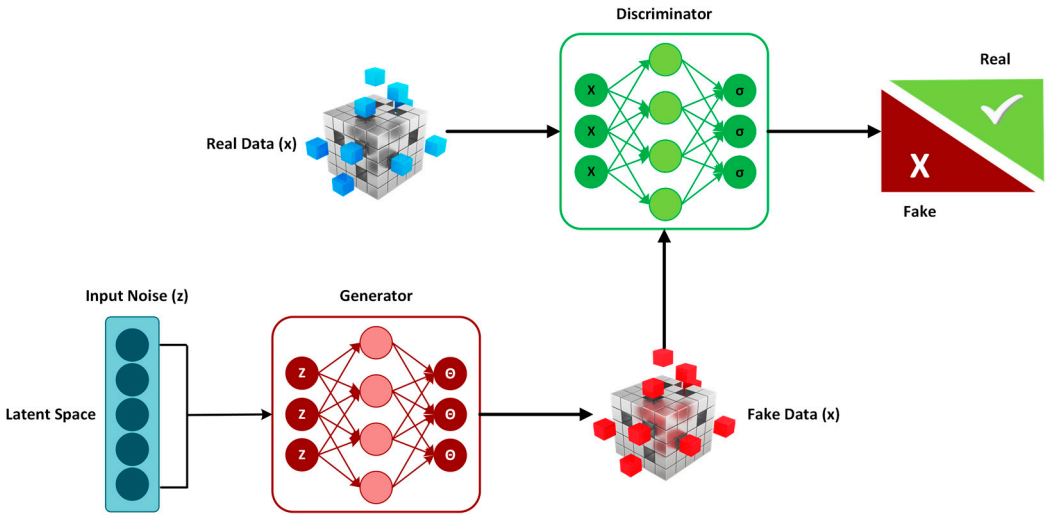

Figure 6. Conceptual diagram of a Generative Adversarial Network (GAN), where a generator transforms latent noise into synthetic samples and a discriminator learns to distinguish real from fake data, with both networks trained together in an adversarial loop.

A second important family is based on agent-based or rule-driven simulations. Here, synthetic consumers are modeled as agents with preferences, budgets, decision rules and social influences. These agents interact with each other and with a virtual marketplace where prices, recommendations and product assortments evolve over time. By running simulations under different conditions, analysts can generate large synthetic panels of customer journeys, responses to promotions or diffusion of product adoption. This approach is attractive when theory or prior research provides a reasonably clear idea of decision mechanisms, but detailed behavioral logs are lacking [124].

Simpler techniques such as bootstrapping, probabilistic graphical models or copula-based methods are also used to create synthetic datasets. They preserve marginal distributions and selected dependences between variables (e.g., between income, age and spending in certain categories) without explicitly modeling the full behavioral process. In practice, organizations often combine several approaches: for example, using empirical distributions for stable attributes such as demographics, GANs for high-dimensional transaction histories and simulation models for future scenario exploration [123,125].

5.2. Advantages of Synthetic Data in Consumer Analysis

Synthetic data offers several advantages for consumer behavior analysis and, by extension, for marketing decision-making. First, it can alleviate privacy and confidentiality concerns. Because synthetic records do not belong to real individuals, firms can share them more easily with external partners, vendors or research teams, while still conveying the broad structure of their customer base. This facilitates collaboration on model development, benchmarking of algorithms and independent auditing of decision rules [126]. Second, synthetic datasets help address class imbalance and rare-event problems. Many phenomena of interest, such as high-value churn, specific types of fraud, or responses to niche campaigns, occur infrequently in operational logs. Generating additional synthetic examples in these regions of the feature space allows models to learn decision boundaries more effectively, improving sensitivity to important but rare behaviors without having to wait for large volumes of real cases [127].

One more benefit is the use of synthetic data for scenario analysis and stress testing. Since these datasets can be produced under different assumptions about the wider economy, competitor moves or regulatory limits, firms can examine "what-if" cases that do not yet show up in their historical records [128]. For example, they can approximate how a revised pricing structure might change demand in different customer groups, or how tighter lending rules could reshape the risk profile and composition of a credit portfolio. This kind of exploration makes strategic planning less dependent on past regularities, which is particularly valuable when existing trends are unstable or unlikely to hold in the future.

Synthetic data can also reduce dependence on limited field experiments. When running large-scale A/B tests is costly or operationally constrained, synthetic logs can provide a first approximate view of potential outcomes. They are not a substitute for real-world experimentation but can help narrow down the set of plausible strategies, refine hypotheses and prioritize which interventions deserve live testing. In addition, synthetic datasets offer educational and prototyping benefits: teams can build and evaluate pipelines, dashboards and decision engines without accessing production systems [129,130].

In consumer behavior research more broadly, synthetic data contributes to reproducibility and openness. Publicly shared synthetic versions of proprietary datasets allow external researchers to test methods, compare algorithms and critique assumptions, even when access to the original data is impossible. Although such datasets inevitably abstract away from some details, they still provide a common reference point for methodological work in recommender systems, customer segmentation or demand modeling [131]. Table 1 summarizes key advantages of synthetic data for consumer behavior analysis, linking each benefit to a concrete marketing use case.

Table 1. Selected advantages of synthetic data for consumer behavior analysis.

Advantage	Short Description	Example Use Case
Privacy and data sharing	Shares structure of customer base without real identities	Bank sharing synthetic cards data with a vendor
Handling rare events	Adds extra samples for infrequent but important behaviors	Extra churn or fraud cases to train classifiers

Table 1. *Cont.*

Advantage	Short Description	Example Use Case
Scenario analysis and stress tests	Simulates "what-if" market or policy conditions	New pricing rules or tighter credit criteria
Cheaper experimentation	Supports early testing when large A/B tests are not feasible	Comparing targeting rules before live rollout
Prototyping and education	Lets teams build pipelines without touching production data	Training analysts on a synthetic clickstream set
Reproducible research	Enables public datasets that mimic proprietary ones	Benchmarking recommender algorithms across studies

5.3. Limitations and Challenges

Despite these benefits, synthetic data come with important limitations that must be acknowledged when using it to support marketing and consumer behavior decisions. A first challenge concerns accuracy and realism. Generating synthetic data that faithfully reflects the complex, multi-layered nature of consumer behavior is difficult. Generative models such as GANs and VAEs can approximate high-dimensional structures, but they may still miss subtle patterns, temporal dependencies or cross-channel interactions that matter for downstream decisions. This matters because many marketing outcomes depend on behavioral dynamics (such as learning, fatigue, trust erosion, or promotion sensitivity) that may not be captured by distributional similarity alone. If these aspects are distorted, models trained solely on synthetic data may perform well in offline tests yet fail when applied to real customers [128,132].

A second concern has to do with how representative the data really are, and with embedded bias. Any synthetic dataset is shaped by the real data and modelling assumptions used to generate it. If the original records largely ignore some demographic groups, focus too heavily on certain regions, or reflect historical discrimination in lending or pricing, the artificial copies will usually carry the same distortions forward. In some situations, generative models may even make this worse, because they tend to emphasize the strongest and most frequent patterns in the training set. So, working with synthetic data does not automatically solve fairness or inclusion issues. On the contrary, it can quietly preserve them if nobody checks [133,134].

Validation is another difficult area. With real data, model quality can be judged by comparing predictions to actual outcomes. With synthetic data, there is no such direct benchmark. Analysts, instead, use indirect checks: they look at how distributions line up, whether correlations are similar, whether clustering structure looks comparable, or whether models trained on real and synthetic samples behave in roughly the same way. These diagnostics help, but they cannot fully guarantee that important behavioral mechanisms have been preserved. In particular, they may miss whether the synthetic data reproduces stable segment differences or realistic switching between states (e.g., browsing-to-cart-to-purchase). Many organizations still struggle to build validation routines that are both statistically serious and feasible in day-to-day practice [135].

The messy nature of human behavior also imposes limits. Consumer choices are shaped by habits, peer influence, emotions, simple rules of thumb, institutional rules and context-specific cues that are difficult to translate into formal parameters. Even advanced simulation frameworks can only approximate these forces. In that sense, synthetic datasets

are more useful for mapping general patterns, running "broad brush" sensitivity checks and exploring scenarios than for predicting detailed outcomes at the individual level [124,136].

Several behavioral risks follow when synthetic data are treated as a substitute for real consumer traces. First, biases present in the original data (unequal coverage, historical targeting practices, under-representation of some groups) can be reproduced and, in some settings, amplified by the generation process, leading models to "learn" exclusionary patterns with higher apparent confidence. Second, rare but behaviorally meaningful edge cases may be smoothed away; events that often matter for marketing decisions under stress, such as abrupt preference shifts, churn after a negative service episode, or atypical but valuable trajectories. Finally, models trained predominantly on synthetic populations can appear stable in internal evaluation while being overconfident when confronted with real-world uncertainty, changing contexts, or strategic consumer adaptation. For this reason, synthetic data are best treated as a complement for prototyping and scenario analysis, alongside explicit checks that decision-relevant behavioral mechanisms remain plausible [133,137].

Ethical and legal issues also remain on the table, even when the records are synthetic. Certain generation techniques can leak information about real people, especially if models are overfitted or if attackers can combine synthetic outputs with other available datasets. Privacy experts and regulators have pointed out that such releases may create a false sense of security. Organizations therefore need explicit governance rules for how synthetic datasets are created, evaluated and shared. This includes clear documentation of the generation process, discussion of residual privacy risks and specification of acceptable use cases. In marketing, these safeguards are particularly important if synthetic data are to support responsible experimentation and product design, rather than undermining consumer protection [138,139].

6. Dark Data and Unused Consumer Insights

6.1. Definition and Types of Dark Data

In the context of consumer behavior, dark data cover any customer-related information that is retained but rarely examined. Typical examples are long-form customer emails, chat transcripts, call-center recordings, open-ended survey responses, free-form complaint texts, refund reasons, product return notes, CCTV footage, in-store sensor logs, web server logs and old campaign files. The defining feature is not the format but the fact that this kind of data sits in "shadows": they exist in storage systems, yet are excluded from regular analytics workflows [140]. From a consumer perspective, the "dark" aspect is often experiential rather than technical: the data exist, but people typically do not expect them to be re-used for profiling, targeting, or automated decision-making.

Several broad categories can be distinguished. Textual dark data includes customer service logs, social media comments collected but never mined, or unstructured feedback fields. Behavioral dark data arises from detailed event logs that are aggregated into simple metrics and then discarded in their raw form, such as fine-grained navigation paths or page-level scroll events. Sensor and image data come from beacons, cameras and IoT devices that monitor store traffic, shelf interactions or product usage but are only used for basic monitoring, if at all. Finally, historical dark data consists of legacy CRM or campaign databases migrated to new systems and left untouched because of format incompatibilities or missing documentation [13,141].

A useful way to clarify governance implications is to distinguish dark data by how consumers typically interpret its use. *Operational* dark data refers to traces generated as a by-product of delivering or maintaining a service (e.g., technical logs, error reports) and is often perceived as legitimate when used for reliability, fraud prevention, or service

improvement [142]. *Experiential* dark data captures narrative or interactional content that reflects lived experience (e.g., complaint text, chats, call transcripts) and can feel more intrusive when repurposed for targeting rather than resolving service problems [143]. *Sensitive or inferred* dark data includes information that is highly personal or constructed through inference (e.g., location-adjacent traces, psychographic proxies, health-adjacent signals), where even formally lawful use may be perceived as surveillance if transparency and control are weak [144].

6.2. Sources, Uses and Marketing Value

When processed responsibly, dark data can enrich the understanding of consumer behavior and support more nuanced marketing actions [145]. For instance, mining call-center transcripts and complaint emails with natural language processing can reveal recurrent friction points in onboarding, billing or product configuration that do not appear in structured fields [146]. Web server logs and detailed click-level traces help reconstruct actual paths through digital properties, showing which content combinations precede purchase, abandonment or support contact [83]. In physical retail, anonymized in-store movement patterns can clarify how shoppers navigate aisles and where they hesitate or backtrack [147].

These insights inform a range of decisions: redesigning service processes to eliminate frequent causes of complaints; adjusting information architecture and recommendation placement on websites; fine-tuning shelf layouts, signage and in-store promotions; prioritizing which product features need simplification. Dark data can also support early warning systems by highlighting weak signals of emerging issues, such as rising sentiment about delivery delays or a new cluster of complaints about a recent software update [148].

Because many of these sources are collected in service or operational contexts, their marketing value must be weighed against perceived intrusiveness and the risk that consumers interpret downstream use as surveillance.

However, not all dark data are equally valuable. Some sources are noisy, redundant or too costly to clean and integrate. A key task for organizations is therefore to identify a small number of dark-data streams that are both rich in behavioral signal and feasible to process with available resources [149]. Table 2 summarizes typical sources and their potential use in consumer behavior analytics, using concise descriptions to keep the focus on practical value.

Table 2. Examples of dark data sources in consumer behavior analytics.

Dark Data Source	Typical Examples	Potential Marketing Value	Key Risks/Challenges
Customer service text	Emails, chat logs, complaint forms	Detect pain points, reasons for churn, UX issues	Sensitive content, re-identification
Call-center recordings	Audio from support and sales calls	Voice tone analysis, script testing, objection data	Heavy anonymization, storage costs
Web and app technical logs	Server logs, error logs, scroll events	Journey reconstruction, friction detection	Volume, noisy events, short retention
Social media and community posts	Comments, reviews, direct messages	Sentiment, themes, peer influence signals	Blurred consent, platform policies

Table 2. *Cont.*

Dark Data Source	Typical Examples	Potential Marketing Value	Key Risks/Challenges
In-store sensor and video data	Footfall sensors, camera feeds, beacons	Path analysis, zone heatmaps, display effectiveness	Strong privacy constraints, regulation
Legacy CRM and campaign files	Old databases, archived lists	Longitudinal views, cohort histories	Poor documentation, format issues

6.3. Challenges, Risks and Ethical Considerations

Perceived surveillance tends to arise when repurposing is unexpected, when inference feels disproportionate to the service being provided, or when consumers cannot reasonably anticipate that a trace will be used for profiling. Consumers are more likely to accept "helpful" inference when it is tightly connected to a clear benefit (e.g., resolving an issue, preventing fraud), bounded in scope, and accompanied by meaningful choices (opt-outs, preference controls). Transparency is of the utmost importance in dark data settings because the ethical issue is often not the data type itself, but the gap between what is collected and what consumers believe is collected [150].

Using dark data in consumer behavior analysis is technically demanding. Many sources are unstructured, high volume and noisy. They require significant effort in data cleaning, annotation and integration before any modeling can begin [151]. For example, turning thousands of call-center recordings into a usable dataset involves transcription, speaker identification, segmentation and domain-specific language handling. Similar preprocessing is needed for logs from multiple systems that use different identifiers or time standards. Without careful preparation, models may learn artefacts of data collection rather than meaningful behavioral patterns [152].

The most difficult issues around dark data are legal and ethical rather than technical. A lot of the information stored in back-end systems was never collected with the expectation that it would later feed into marketing models. When firms start joining together service transcripts, usernames or social media handles, device IDs, approximate locations and similar traces, the resulting profile can feel very close to surveillance from a consumer point of view, even if the company can point to a generic consent clause in the terms and conditions [153,154]. Regulatory frameworks such as the GDPR and related privacy laws in other jurisdictions put hard constraints on data minimization, purpose limitation and how long data can be kept [155]. These rules are directly relevant for long-term stockpiles of dark data that sit outside ordinary reporting processes.

There is also a clear danger of reinforcing existing biases. Contacts with customer service, formal complaints and written feedback are not distributed evenly across the population. Some groups are more likely to call or chat, others tend to send detailed emails or post reviews, while many customers rarely complain at all. If analysts simply treat these dark-data sources as if they were a balanced mirror of the entire customer base, models can end up overweighting the preferences and frustrations of more vocal, digitally active or better-connected segments [156,157]. Social media data suffer from similar skew: they over-represent certain age groups, lifestyles and communication styles, and under-represent others who are less active online. As a result, decisions based on these signals alone may unintentionally prioritize the needs of those who "speak up" the most [158].

Given these concerns, organizations need explicit governance for dark data, not just ad-hoc technical fixes. A basic starting point is to map which dark data sources exist across systems, assess their necessity for the purposes at hand, and judge whether using them

is proportionate to the potential benefits [159]. In some cases, the appropriate decision may simply be deletion or strict archiving rather than analysis. For data that will be used, businesses should define clear internal rules on access, acceptable uses, and conditions for sharing with partners. Privacy-preserving measures—such as aggregation, strong and regularly reviewed anonymization, tight role-based access controls, and conservative retention limits—can substantially reduce risk but cannot fully replace human judgment about what is appropriate [160,161]. When approached carefully, with these safeguards in place, dark data can complement standard customer analytics and support more responsive services and products. When exploited without clear boundaries, it can undermine trust, trigger regulatory scrutiny and damage the organization's reputation.

7. Discussion and Future Directions

7.1. Synthesis of Main Insights

This review suggests that big data and AI tools now influence consumer behavior analysis across almost every stage of the marketing process. At the operational level, systems for personalized recommendations, dynamic and contextual pricing, CRM-driven targeting, data-informed product development, and automated fraud screening all turn fine-grained behavioral traces into concrete actions: which offer to display to a given person, what price or discount to put forward, which customers deserve proactive outreach, which feature should be adjusted or removed, and which payment attempt should be challenged or declined. Such decision-support systems do more than increase predictive accuracy. They reshape the decision environment that consumers encounter by changing which options are visible or salient, how information is framed on screens and in messages, and how much uncertainty or perceived risk is involved in interacting with a firm. This is why the performance of these automated decision tools needs to be evaluated not only in predictive terms but also in terms of how they affect perceived autonomy, perceived fairness, and willingness to stay in the relationship.

Synthetic data and dark data extend this picture. Synthetic datasets allow firms and researchers to build and test models in situations where real data are scarce, sensitive or fragmented, while still approximating key behavioral patterns. Dark data, in turn, highlights the large amount of customer-related information that remains unused in many organizations, from service transcripts and technical logs to sensor streams and legacy CRM files. Together, these developments broaden the data foundation of consumer analytics beyond traditional transaction and campaign databases. They also expose new tensions between analytical ambition, privacy expectations and the practical limits of data quality.

Viewed together, the material in Sections 4–6 suggests a shift from episodic, survey-based views of consumers toward continuous, trace-based representations, where behavior is inferred from streams of digital interactions. This shift creates opportunities for more timely, tailored and context-aware marketing actions, but it also amplifies concerns about intrusiveness, surveillance and fairness. Future work, both academic and managerial, needs to address these opportunities and risks jointly rather than treating them as separate topics.

7.2. Implications for Digital Marketing Practice

For practitioners, one implication is that data-driven marketing should be framed as a design problem, not only as a modeling problem. Personalization, pricing, CRM actions, product changes and fraud controls all influence how consumers perceive the brand and how they experience the decision process itself. Firms need governance mechanisms that connect data science teams with marketing, legal, UX and service design, so that deployment choices (what to optimize, what to constrain, what to explain to users) are evaluated together with behavioral evidence, regulatory constraints and long-term relationship goals.

A second implication concerns the selective use of synthetic and dark data. Synthetic data can be valuable for prototyping, stress testing and collaboration, but it should not become a black box that hides the assumptions embedded in the generation process. Similarly, dark data should be prioritized according to signal value and feasibility, rather than harvested indiscriminately. Starting from a small number of well-justified sources—such as complaint text or call-center logs—and linking them to clear improvement projects (e.g., reducing a specific friction point) is likely to be more effective than broad, unfocused mining initiatives. Across all these uses, transparency toward consumers about data practices and safeguards will be critical for sustaining trust.

7.3. Research Implications and Limitations

Methodologically, this article has clear boundaries that follow from its purpose. It is a narrative, concept-driven review intended to integrate evidence across marketing, information systems, and computer science, rather than to provide an exhaustive census of studies. For that reason, it does not follow PRISMA procedures or include a meta-analysis; sources were selected for topical relevance and conceptual contribution, not through a protocol-based search with reproducible inclusion/exclusion rules. To mitigate these constraints, the synthesis prioritizes peer-reviewed sources, triangulates recurring claims across multiple research streams, and treats contested issues as open problems rather than settled conclusions. The scope is intentionally centered on digital, data-rich commercial environments, with limited attention to offline, low-data, or non-commercial contexts (e.g., public services and non-profits). In addition, the examples are illustrative and emphasize well-documented application areas; they are not intended to cover every industry where big data and AI shape consumer decisions.

From a research perspective, several directions follow from this review. The growing reliance on behavioral traces and synthetic data calls for new ways of validating models and theories of consumer behavior. There is a need for studies that combine laboratory or field experiments with log data, so that psychological constructs such as perceived intrusiveness, fairness or trust can be linked more tightly to observable click paths, purchases and service interactions. Work on synthetic consumers and agent-based models could be used to explore under which conditions certain marketing strategies are likely to produce unintended distributional effects across segments, or to test how robust personalization policies are under shifts in context.

The future directions summarized in Table 3 should be read in this light. They highlight broad trajectories (such as deeper integration of AI and machine learning, increased use of edge computing, stronger consumer participation, more explicit ethical frameworks and exploratory work on quantum methods) rather than a complete map of all possible developments. These themes point to areas where both empirical and conceptual research can contribute: by examining how AI-enabled decisions are perceived in practice, by studying the effects of local (edge) analytics on behavior and by clarifying what meaningful consumer control over data use might look like in different contexts.

Finally, the rapid pace of technological and regulatory change means that any review of this type is necessarily provisional. New data sources, modeling techniques and legal constraints will continue to appear. Periodic updates, domain-specific extensions (for instance, in financial services, healthcare or retail) and cross-country comparisons would be useful to refine and challenge the conclusions drawn here, and to keep the discussion aligned with evolving consumer expectations.

Table 3. Selected future directions in consumer behavior analytics.

Future Direction	Short Description
Integration of Artificial Intelligence and Machine Learning	Broader use of advanced models to personalize decisions and automate analytics.
Rise of Edge Computing	Moving parts of analytics to devices and stores for faster, local decisions.
Enhanced Consumer Participation	Giving users more control, insight and benefits from the data they share.
Development of Ethical Frameworks	Establishing clear rules for profiling, targeting and automated interventions.
Quantum Computing and Big Data Analytics	Testing quantum methods for complex optimization and pattern discovery tasks.

8. Conclusions

This article examines how big data and AI technologies are transforming consumer behavior analysis in digital marketing, and how new data practices are starting to alter the foundations on which this analysis rests. By focusing on five core application domains—personalized recommendations, dynamic pricing, customer relationship management, data-driven product development and fraud detection—this review shows that algorithmic systems now play a central role in shaping the information, options and constraints that consumers encounter. These AI-enabled decision tools do not simply predict choices; they actively participate in constructing the choice environment, influencing what is noticed, when action is taken and how fair or trustworthy a firm is perceived to be.

The sections on synthetic data and dark data broaden this picture beyond the usual big-data storyline. Synthetic datasets open up alternative ways to train and evaluate models, to exchange findings with partners, and to run "what-if" experiments when direct access to detailed behavioral records is limited or tightly regulated. Dark data, in turn, highlights that many organizations already hold large amounts of potentially useful consumer information that never reach their regular dashboards or modeling workflows. Both trends can increase analytical flexibility and uncover new patterns, but they also raise additional concerns about how valid the insights really are, how bias might be reinforced, and how privacy and governance are handled in practice. Ultimately, the usefulness of these data strategies depends less on how advanced the underlying techniques appear and more on whether they are tied to clear marketing goals and remain broadly in line with what consumers could reasonably expect to happen with their data.

For practitioners, this review underscores that data-driven marketing should be approached as an ongoing design challenge rather than as a one-time technical upgrade. Personalization policies, pricing algorithms, CRM workflows, product experiments and fraud controls all have measurable effects on key outcomes, but they also carry less visible consequences for perceptions of autonomy, intrusion and fairness. Treating these behavioral aspects as first-order design criteria, and not just as side effects, is essential for building and maintaining trust in increasingly automated environments. For researchers, the analysis points to the need for work that connects log-level behavior and synthetic datasets with established theories of decision-making, trust and technology acceptance, and that develops stronger validation procedures for models trained on artificial or highly processed data.

This review itself has limitations. It is narrative and concept-driven, not a PRISMA-style systematic review or a meta-analysis, and it concentrates on data-rich digital contexts.

The examples and application areas discussed are illustrative rather than exhaustive. As technologies, regulations and social norms continue to evolve, periodic reassessment of the topics covered here will be necessary. Nevertheless, by bringing together applications of big data and AI with emerging synthetic and dark data practices under a consumer-behavior lens, this article aims to provide a structured basis for both critical reflection and further research on how marketing decisions are made in a world of continuous digital traces and increasingly automated interventions.

Author Contributions: Conceptualization, L.T. and A.T.; methodology, L.T. and A.T.; formal analysis, L.T. and A.T.; investigation, A.T. and C.K.; writing—original draft preparation, A.T. and C.K.; writing—review and editing, L.T. and A.T.; visualization, A.T. and C.K.; supervision, L.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lamberton, C.; Stephen, A.T. A thematic exploration of digital, social media, and mobile marketing research's evolution from 2000 to 2015 and an agenda for future research. *J. Mark.* **2016**, *80*, 146–172. [CrossRef]
2. Schweidel, D.A.; Bart, Y.; Inman, J.J.; Stephen, A.T.; Libai, B.; Andrews, M.; Babić Rosario, A.; Chae, I.; Chen, Z.; Kupor, D.; et al. How consumer digital signals are reshaping the customer journey. *J. Acad. Mark. Sci.* **2022**, *50*, 1257–1276. [CrossRef]
3. Akter, S.; Fosso Wamba, S. Big data analytics in e-commerce: A systematic review and agenda for future research. *Electron. Mark.* **2016**, *26*, 173–194. [CrossRef]
4. Saura, J.R. Using data sciences in digital marketing: Framework, methods, and performance metrics. *J. Innov. Knowl.* **2021**, *6*, 92–102. [CrossRef]
5. Basu, R.; Basu, A.; Batra, R. Marketing analytics: The bridge between customer data and strategic decisions. *Psychol. Mark.* **2023**, *40*, 1796–1814. [CrossRef]
6. Rosário, A.T.; Dias, J.C. How has data-driven marketing evolved: Challenges and opportunities with emerging technologies. *Int. J. Inf. Manag. Data Insights* **2023**, *3*, 100203. [CrossRef]
7. Saura, J.R.; Škare, V.; Dosen, D.O. Is AI-based digital marketing ethical? Assessing a new data privacy paradox. *J. Innov. Knowl.* **2024**, *9*, 100597. [CrossRef]
8. Naz, H.; Kashif, M. Artificial intelligence and predictive marketing: An ethical framework from managers' perspective. *Span. J. Mark.-ESIC* **2025**, *29*, 22–45. [CrossRef]
9. Kaponis, A.; Maragoudakis, M.; Sofianos, K.C. Enhancing user experiences in digital marketing through machine learning: Cases, trends, and challenges. *Computers* **2025**, *14*, 211. [CrossRef]
10. Riandhi, A.N.; Arviansyah, M.R.; Sondari, M.C. AI and consumer behavior: Trends, technologies, and future directions from a scopus-based systematic review. *Cogent Bus. Manag.* **2025**, *12*, 2544984. [CrossRef]
11. Lomotey, R.K.; Kumi, S.; Ray, M.; Deters, R. Synthetic data digital twins and data trusts control for privacy in health data sharing. In Proceedings of the 2024 ACM Workshop on Secure and Trustworthy Cyber-Physical Systems Porto, Porto, Portugal, 21 June 2024; pp. 1–10. [CrossRef]
12. Eigenschink, P.; Reutterer, T.; Vamosi, S.; Vamosi, R.; Sun, C.; Kalcher, K. Deep generative models for synthetic data: A survey. *IEEE Access* **2023**, *11*, 47304–47320. [CrossRef]
13. Olaitan, O.F.; Adebanjo, T.A.; Obozokhai, L.I.; Iwerumoh, A.N.; Balogun, I.O.; Ojo, D.A. Dark data in business intelligence: A systematic review of challenges, opportunities, and value creation potential. *J. Econ. Bus. Commer.* **2025**, *2*, 135–142. [CrossRef]
14. Lemon, K.N.; Verhoef, P.C. Understanding customer experience throughout the customer journey. *J. Mark.* **2016**, *80*, 69–96. [CrossRef]
15. Haridasan, A.C.; Fernando, A.G.; Saju, B. A systematic review of consumer information search in online and offline environments. *RAUSP Manag. J.* **2021**, *56*, 234–253. [CrossRef]

16. Sachdeva, R. The Coronavirus shopping anxiety scale: Initial validation and development. *Eur. J. Manag. Bus. Econ.* **2022**, *31*, 409–424. [CrossRef]
17. Yang, H.-P.; Fan, W.-S.; Tsai, M.-C. Applying Stimulus–Organism–Response Theory to Explore the Effects of Augmented Reality on Consumer Purchase Intention for Teenage Fashion Hair Dyes. *Sustainability* **2024**, *16*, 2537. [CrossRef]
18. Kelly, S.; Kaye, S.A.; Oviedo-Trespalacios, O. What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telemat. Inform.* **2023**, *77*, 101925. [CrossRef]
19. Kim, Y.; Blazquez, V.; Oh, T. Determinants of generative AI system adoption and usage behavior in Korean companies: Applying the UTAUT model. *Behav. Sci.* **2024**, *14*, 1035. [CrossRef] [PubMed]
20. Yin, J.; Qiu, X.; Wang, Y. The Impact of AI-Personalized Recommendations on Clicking Intentions: Evidence from Chinese E-Commerce. *J. Theor. Appl. Electron. Commer. Res.* **2025**, *20*, 21. [CrossRef]
21. Kim, Y.; Kim, S.H.; Peterson, R.A.; Choi, J. Privacy concern and its consequences: A meta-analysis. *Technol. Forecast. Soc. Change* **2023**, *196*, 122789. [CrossRef]
22. Kezer, M.; Dienlin, T.; Baruh, L. Getting the privacy calculus right: Analyzing the relations between privacy concerns, expected benefits, and self-disclosure using response surface analysis. *Cyberpsychol. J. Psychosoc. Res. Cyberspace* **2022**, *16*, 1. [CrossRef]
23. Li, Y.; Deng, X.; Hu, X.; Liu, J. The Effects of E-Commerce Recommendation System Transparency on Consumer Trust: Exploring Parallel Multiple Mediators and a Moderator. *J. Theor. Appl. Electron. Commer. Res.* **2024**, *19*, 2630–2649. [CrossRef]
24. Teodorescu, D.; Aivaz, K.-A.; Vancea, D.P.C.; Condrea, E.; Dragan, C.; Olteanu, A.C. Consumer Trust in AI Algorithms Used in E-Commerce: A Case Study of College Students at a Romanian Public University. *Sustainability* **2023**, *15*, 11925. [CrossRef]
25. Aquilino, L.; Di Dio, C.; Manzi, F.; Massaro, D.; Bisconti, P.; Marchetti, A. Decoding Trust in Artificial Intelligence: A Systematic Review of Quantitative Measures and Related Variables. *Informatics* **2025**, *12*, 70. [CrossRef]
26. Imani, M.; Joudaki, M.; Beikmohammadi, A.; Arabnia, H.R. Customer Churn Prediction: A Systematic Review of Recent Advances, Trends, and Challenges in Machine Learning and Deep Learning. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 105. [CrossRef]
27. Compagnino, A.A.; Maruccia, Y.; Cavuoti, S.; Riccio, G.; Tutone, A.; Crupi, R.; Pagliaro, A. An Introduction to Machine Learning Methods for Fraud Detection. *Appl. Sci.* **2025**, *15*, 11787. [CrossRef]
28. De Mauro, A.; Sestino, A.; Bacconi, A. Machine learning and artificial intelligence use in marketing: A general taxonomy. *Ital. J. Mark.* **2022**, *2022*, 439–457. [CrossRef]
29. Hasan, E.; Rahman, M.; Ding, C.; Huang, J.X.; Raza, S. Based recommender systems: A survey of approaches, challenges and future perspectives. *ACM Comput. Surv.* **2025**, *58*, 1–41. [CrossRef]
30. Figueira, A.; Vaz, B. Survey on Synthetic Data Generation, Evaluation Methods and GANs. *Mathematics* **2022**, *10*, 2733. [CrossRef]
31. Freitas, N.; Rocha, A.D.; Barata, J. Data management in industry: Concepts, systematic review and future directions. *J. Intell. Manuf.* **2025**, 1–29. [CrossRef]
32. Chadwick, L.B.; Krishnamoorthy, A.; Yadav, S.; Dixon, H.E.T. Scrolling Into Choice: The Psychology and Practice of Social Media Consumerism. *Adv. Consum. Res.* **2025**, *2*, 190–211. [CrossRef]
33. Kabir, M.H.; Sultana, S.; Hossain, M.M.; Islam, S.M.A. The Impact of Digital Marketing Strategies on Consumer Behavior: A Comprehensive Review. *Bus. Soc. Sci.* **2025**, *3*, 1–8. [CrossRef]
34. Hariguna, T.; Ruangkanjanases, A. Assessing the impact of artificial intelligence on customer performance: A quantitative study using partial least squares methodology. *Data Sci. Manag.* **2024**, *7*, 155–163. [CrossRef]
35. Chakradhar, B.S.; Vijey, M.; Tunk, S. Synthetic Data Generation for Marketing Insights. In *Predictive Analytics and Generative AI for Data-Driven Marketing Strategies*; Chapman and Hall/CRC: Abingdon, UK, 2024; pp. 195–215. [CrossRef]
36. Jae, Y.I.; Hwa, P.I. Personalized Digital Marketing Strategies: A Data-Driven Approach Using Marketing Analytics. *J. Manag. Inform.* **2025**, *4*, 668–686. [CrossRef]
37. Theodorakopoulos, L.; Theodoropoulou, A.; Klavdianos, C. Interactive Viral Marketing Through Big Data Analytics, Influencer Networks, AI Integration, and Ethical Dimensions. *J. Theor. Appl. Electron. Commer. Res.* **2025**, *20*, 115. [CrossRef]
38. Chandra, S.; Verma, S.; Lim, W.M.; Kumar, S.; Donthu, N. Personalization in personalized marketing: Trends and ways forward. *Psychol. Mark.* **2022**, *39*, 1529–1562. [CrossRef]
39. Tyrväinen, O.; Karjaluoto, H.; Saarijärvi, H. Personalization and hedonic motivation in creating customer experiences and loyalty in omnichannel retail. *J. Retail. Consum. Serv.* **2020**, *57*, 102233. [CrossRef]
40. Maraj, D.; Vuković, M.; Hotovec, P. A Survey on User Profiling, Data Collection, and Privacy Issues of Internet Services. *Telecom* **2024**, *5*, 961–976. [CrossRef]
41. Trusov, M.; Ma, L.; Jamal, Z. Crumbs of the cookie: User profiling in customer-base analysis and behavioral targeting. *Mark. Sci.* **2016**, *35*, 405–426. [CrossRef]
42. Sagtani, H.; Jhawar, M.G.; Gupta, A.; Mehrotra, R. Quantifying and leveraging user fatigue for interventions in recommender systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), Taipei, China, 23–27 July 2023*; ACM: New York, NY, USA, 2023. [CrossRef]

43. Wu, Y.; Yusof, Y. Emerging trends in real-time recommendation systems: A deep dive into multi-behavior streaming processing and recommendation for e-commerce platforms. *J. Internet Serv. Inf. Secur.* **2024**, *14*, 45–66. [CrossRef]
44. Zhou, H.; Xiong, F.; Chen, H. A Comprehensive Survey of Recommender Systems Based on Deep Learning. *Appl. Sci.* **2023**, *13*, 11378. [CrossRef]
45. Sami, A.; El Adrousy, W.; Sarhan, S.; Elmougy, S. A deep learning based hybrid recommendation model for internet users. *Sci. Rep.* **2024**, *14*, 29390. [CrossRef]
46. Srifi, M.; Oussous, A.; Ait Lahcen, A.; Mouline, S. Recommender Systems Based on Collaborative Filtering Using Review Texts—A Survey. *Information* **2020**, *11*, 317. [CrossRef]
47. Zhao, W.; Tian, H.; Wu, Y.; Cui, Z.; Feng, T. A new item-based collaborative filtering algorithm to improve the accuracy of prediction in sparse data. *Int. J. Comput. Intell. Syst.* **2022**, *15*, 15. [CrossRef]
48. Abdalla, H.I.; Amer, A.A.; Amer, Y.A.; Nguyen, L.; Al-Maqaleh, B. Boosting the item-based collaborative filtering model with novel similarity measures. *Int. J. Comput. Intell. Syst.* **2023**, *16*, 123. [CrossRef]
49. Parthasarathy, G.; Sathiya Devi, S. Hybrid recommendation system based on collaborative and content-based filtering. *Cybern. Syst.* **2023**, *54*, 432–453. [CrossRef]
50. Lumintu, I. Content-Based Recommendation Engine Using Term Frequency-Inverse Document Frequency Vectorization and Cosine Similarity: A Case Study. In *Proceedings of the 2023 IEEE 9th Information Technology International Seminar (ITIS), Surabaya, Indonesia, 18–20 October 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6. [CrossRef]
51. Qin, J. A survey of long-tail item recommendation methods. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 7536316. [CrossRef]
52. Çano, E.; Morisio, M. Hybrid recommender systems: A systematic literature review. *Intell. Data Anal.* **2017**, *21*, 1487–1524. [CrossRef]
53. Widayanti, R.; Chakim, M.H.R.; Lukita, C.; Rahardja, U.; Lutfiani, N. Improving recommender systems using hybrid techniques of collaborative filtering and content-based filtering. *J. Appl. Data Sci.* **2023**, *4*, 289–302. [CrossRef]
54. Gheewala, S.; Xu, S.; Yeom, S. In-depth survey: Deep learning in recommender systems—Exploring prediction and ranking models, datasets, feature analysis, and emerging trends. *Neural Comput. Appl.* **2025**, *37*, 10875–10947. [CrossRef]
55. He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; Chua, T.-S. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17), Perth, Australia, 3–7 April 2017*; ACM: New York, NY, USA, 2017; pp. 173–182. [CrossRef]
56. Ayemowa, M.O.; Ibrahim, R.; Bena, Y.A. A systematic review of the literature on deep learning approaches for cross-domain recommender systems. *Data Anal.* **2024**, *4*, 100518. [CrossRef]
57. Donkers, T.; Loepp, B.; Ziegler, J. Sequential user-based recurrent neural network recommendations. In Proceedings of the RecSys '17 11th ACM Conference on Recommender Systems, Como, Italy, 27–31 August 2017; pp. 152–160. [CrossRef]
58. Kim, N.; Lim, H.; Li, Q.; Li, X.; Kim, S.; Kim, J. Enhancing Review-Based Recommendations Through Local and Global Feature Fusion. *Electronics* **2025**, *14*, 2540. [CrossRef]
59. Bonicalzi, S.; De Caro, M.; Giovanola, B. Artificial intelligence and autonomy: On the ethical dimension of recommender systems. *Topoi* **2023**, *42*, 819–832. [CrossRef]
60. Nowak, M.; Pawłowska-Nowak, M. Dynamic Pricing Method in the E-Commerce Industry Using Machine Learning. *Appl. Sci.* **2024**, *14*, 11668. [CrossRef]
61. Yuan, E.; Van Hentenryck, P. Real-time pricing optimization for ride-hailing quality of service. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21), Virtual, 19–26 August 2021*; Zhou, Z.-H., Ed.; International Joint Conferences on Artificial Intelligence Organization: Montreal, QC, Canada, 2021; pp. 3742–3748. [CrossRef]
62. Battifarano, M.; Qian, Z.S. Predicting real-time surge pricing of ride-sourcing companies. *Transp. Res. Part C Emerg. Technol.* **2019**, *107*, 444–462. [CrossRef]
63. Yan, C.; Zhu, H.; Korolko, N.; Woodard, D. Dynamic pricing and matching in ride-hailing platforms. *Nav. Res. Logist.* **2020**, *67*, 705–724. [CrossRef]
64. Ban, G.Y.; Keskin, N.B. Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Manag. Sci.* **2021**, *67*, 5549–5568. [CrossRef]
65. Zhou, Q.; Yang, Y.; Fu, S. Deep reinforcement learning approach for solving joint pricing and inventory problem with reference price effects. *Expert Syst. Appl.* **2022**, *195*, 116564. [CrossRef]
66. Neubert, M. A systematic literature review of dynamic pricing strategies. *Int. Bus. Res.* **2022**, *15*, 1–17. [CrossRef]
67. Theodorakopoulos, L.; Theodoropoulou, A. Leveraging big data analytics for understanding consumer behavior in digital marketing: A systematic review. *Hum. Behav. Emerg. Technol.* **2024**, *2024*, 3641502. [CrossRef]
68. Vomberg, A.; Homburg, C.; Sarantopoulos, P. Algorithmic pricing: Effects on consumer trust and price search. *Int. J. Res. Mark.* **2024**, *42*, 1166–1186. [CrossRef]
69. Gaidhani, Y.; Venkata Naga Ramesh, J.; Singh, S.; Dagar, R.; Rao, T.S.M.; Godla, S.R.; El-Ebiary, Y.A.B. AI-driven predictive analytics for CRM to enhance retention, personalization and decision-making. *Int. J. Adv. Comput. Sci. Appl.* **2025**, *16*, 553–563. [CrossRef]

70. Wong, A.K.S.; Viloria Garcia, A.; Lim, Y.-W. A data-driven approach to customer lifetime value prediction using probability and machine learning models. *Decis. Anal. J.* **2025**, *16*, 100601. [CrossRef]
71. Del Vecchio, P.; Mele, G.; Siachou, E.; Schito, G. A structured literature review on Big Data for customer relationship management (CRM): Toward a future agenda in international marketing. *Int. Mark. Rev.* **2022**, *39*, 1069–1092. [CrossRef]
72. Lin, J.-Y.; Chen, C.-C. Driving Innovation Through Customer Relationship Management—A Data-Driven Approach. *Sustainability* **2025**, *17*, 3663. [CrossRef]
73. Ibitoye, A.O.; Kolade, O.; Onifade, O.F. Customer retention model using machine learning for improved user-centric quality of experience through personalised quality of service. *J. Bus. Anal.* **2025**, 1–19. [CrossRef]
74. Hasan, M.S.; Siam, M.A.; Ahad, M.A.; Hossain, M.N.; Ridoy, M.H.; Rabbi, M.N.S.; Hossain, A.; Jakir, T. Predictive Analytics for Customer Retention: Machine Learning Models to Analyze and Mitigate Churn in E-Commerce Platforms. *J. Bus. Manag. Stud.* **2024**, *6*, 304–320. [CrossRef]
75. Alnofeli, K.K.; Akter, S.; Yanamandram, V.; Hani, U. AI-powered CRM capability model: Advancing marketing ambidexterity, profitability and competitive performance. *Int. J. Inf. Manag.* **2026**, *86*, 102981. [CrossRef]
76. Shi, X.; Zhang, Y.; Yu, M.; Zhang, L. Revolutionizing market surveillance: Customer relationship management with machine learning. *PeerJ Comput. Sci.* **2024**, *10*, e2583. [CrossRef]
77. Sharma, N. Predictive customer intelligence: A synthetic data-driven evaluation of machine learning and NLP integration for CRM churn prediction and lifetime value forecasting. *Int. J. Comput. Appl.* **2025**, *187*, 64–70. [CrossRef]
78. Hardcastle, K.; Vorster, L.; Brown, D.M. Understanding customer responses to AI-driven personalized journeys: Impacts on the customer experience. *J. Advert.* **2025**, *54*, 176–195. [CrossRef]
79. Monti, T.; Montagna, F.; Cascini, G.; Cantamessa, M. Data-driven innovation: Challenges and insights of engineering design. *Des. Sci.* **2025**, *11*, e26. [CrossRef]
80. Nasrabadi, M.A.; Beauregard, Y.; Ekhlassi, A. The implication of user-generated content in new product development process: A systematic literature review and future research agenda. *Technol. Forecast. Soc. Change* **2024**, *206*, 123551. [CrossRef]
81. Park, S.; Kim, H. Data-driven analysis of usage-feature interactions for new product design. *Expert Syst. Appl.* **2026**, *296*, 128932. [CrossRef]
82. Giannakoulopoulos, A.; Pergantis, M.; Lamprogeorgos, A. User Experience, Functionality and Aesthetics Evaluation in an Academic Multi-Site Web Ecosystem. *Future Internet* **2024**, *16*, 92. [CrossRef]
83. Sakalauskas, V.; Kriksciuniene, D. Personalized Advertising in E-Commerce: Using Clickstream Data to Target High-Value Customers. *Algorithms* **2024**, *17*, 27. [CrossRef]
84. Quin, F.; Weyns, D.; Galster, M.; Silva, C.C. A/B testing: A systematic literature review. *J. Syst. Softw.* **2024**, *211*, 112011. [CrossRef]
85. Mandić, M.; Gregurec, I.; Vujović, U. Measuring the Effectiveness of Online Sales by Conducting A/B Testing. *Market-Tržište* **2023**, *35*, 223–249. [CrossRef]
86. Alshaketheep, K.; Al-Ahmed, H.; Mansour, A. Beyond purchase patterns: Harnessing predictive analytics to anticipate unarticulated consumer needs. *Acta Psychol.* **2025**, *242*, 105089. [CrossRef]
87. Conde, R. Necessary condition analysis for sales funnel optimization. *J. Mark. Anal.* **2025**, 1–13. [CrossRef]
88. Becerril-Castrillejo, I.; Nieto-García, M.; Muñoz-Gallego, P.A. Do satisfaction and satiation both drive immediate and delayed subscription cancellation? Implications for subscription video-on-demand services. *J. Retail. Consum. Serv.* **2026**, *89*, 104624. [CrossRef]
89. Pires, P.B.; Perestrelo, B.M.; Santos, J.D. Measuring Customer Experience in E-Retail. *Adm. Sci.* **2025**, *15*, 434. [CrossRef]
90. Docherty, N.; Biega, A.J. (Re) Politicizing digital well-being: Beyond user engagements. In Proceedings of the CHI '22: 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 30 April–6 May 2022; pp. 1–13. [CrossRef]
91. Deng, Q.; Thoben, K.-D. A Systematic Procedure for Utilization of Product Usage Information in Product Development. *Information* **2022**, *13*, 267. [CrossRef]
92. Quan, H.; Li, S.; Zeng, C.; Wei, H.; Hu, J. Big Data and AI-Driven Product Design: A Survey. *Appl. Sci.* **2023**, *13*, 9433. [CrossRef]
93. Vanini, P.; Rossi, S.; Zvizdic, E.; Domenig, T. Online payment fraud: From anomaly detection to risk management. *Financ. Innov.* **2023**, *9*, 66. [CrossRef]
94. Hernandez Aros, L.; Bustamante Molano, L.X.; Gutierrez-Portela, F.; Moreno Hernandez, J.J.; Rodríguez Barrero, M.S. Financial fraud detection through the application of machine learning techniques: A literature review. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 1130. [CrossRef]
95. Hilal, W.; Gadsden, S.A.; Yawney, J. Financial fraud: A review of anomaly detection techniques and recent advances. *Expert Syst. Appl.* **2022**, *193*, 116429. [CrossRef]
96. Goldstein, M.; Uchida, S. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS ONE* **2016**, *11*, e0152173. [CrossRef] [PubMed]

97. Wahid, A.; Rao, A.C.S. An outlier detection algorithm based on KNN-kernel density estimation. In *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020*; IEEE: Piscataway, NJ, USA, 2020; pp. 1–8. [CrossRef]
98. Sánchez Vinces, B.V.; Schubert, E.; Zimek, A.; Cordeiro, R.L. A comparative evaluation of clustering-based outlier detection. *Data Min. Knowl. Discov.* **2025**, *39*, 13. [CrossRef]
99. Kim, Y.; Vasarhelyi, M. Anomaly detection with the density based spatial clustering of applications with noise (DBSCAN) to detect potentially fraudulent wire transfers. *Int. J. Digit. Account. Res.* **2024**, *24*, 57–91. [CrossRef]
100. Alarfaj, F.K.; Malik, I.; Khan, H.U.; Almusallam, N.; Ramzan, M.; Ahmed, M. Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. *IEEE Access* **2022**, *10*, 39700–39715. [CrossRef]
101. Adewumi, A.O.; Akinyelu, A.A. A survey of machine-learning and nature-inspired based credit card fraud detection techniques. *Int. J. Syst. Assur. Eng. Manag.* **2017**, *8*, 937–953. [CrossRef]
102. Dornadula, V.N.; Geetha, S. Credit card fraud detection using machine learning algorithms. *Procedia Comput. Sci.* **2019**, *165*, 631–641. [CrossRef]
103. Carcillo, F.; Dal Pozzolo, A.; Le Borgne, Y.A.; Caelen, O.; Mazzer, Y.; Bontempi, G. Scarff: A scalable framework for streaming credit card fraud detection with spark. *Inf. Fusion* **2018**, *41*, 182–194. [CrossRef]
104. Odeyale, K.M.; Moruff, O.A.; Taofeekat, S.I.T.; Kayode, S.M. A support vector machine credit card fraud detection model based on high imbalance dataset. *J. Comput. Soc.* **2024**, *5*, 85–94. [CrossRef]
105. Fanai, H.; Abbasimehr, H. A novel combined approach based on deep autoencoder and deep classifiers for credit card fraud detection. *Expert Syst. Appl.* **2023**, *217*, 119562. [CrossRef]
106. Forough, J.; Momtazi, S. Ensemble of deep sequential models for credit card fraud detection. *Appl. Soft Comput.* **2021**, *99*, 106883. [CrossRef]
107. Apicella, A.; Donnarumma, F.; Isgrò, F.; Prevete, R. A survey on modern trainable activation functions. *Neural Netw.* **2021**, *138*, 14–32. [CrossRef]
108. Banu, S.R.; Gongada, T.N.; Santosh, K.; Chowdhary, H.; Sabareesh, R.; Muthuperumal, S. Financial fraud detection using hybrid convolutional and recurrent neural networks: An analysis of unstructured data in banking. In *Proceedings of the 2024 10th International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, India, 25–27 April 2024*; IEEE: Piscataway, NJ, USA, 2024; pp. 1027–1031. [CrossRef]
109. Cao, Y.; Xiang, H.; Zhang, H.; Zhu, Y.; Ting, K.M. Anomaly detection based on isolation mechanisms: A survey. *Mach. Intell. Res.* **2025**, *22*, 849–865. [CrossRef]
110. Misra, S.; Thakur, S.; Ghosh, M.; Saha, S.K. An autoencoder based model for detecting fraudulent credit card transaction. *Procedia Comput. Sci.* **2020**, *167*, 254–262. [CrossRef]
111. Ogundile, O.; Babalola, O.; Ogunbanwo, A.; Ogundile, O.; Balyan, V. Credit Card Fraud: Analysis of Feature Extraction Techniques for Ensemble Hidden Markov Model Prediction Approach. *Appl. Sci.* **2024**, *14*, 7389. [CrossRef]
112. Cheng, D.; Zou, Y.; Xiang, S.; Jiang, C. Graph neural networks for financial fraud detection: A review. *Front. Comput. Sci.* **2025**, *19*, 199609. [CrossRef]
113. Immadisetty, A. Real-Time Fraud Detection Using Streaming Data in Financial Transactions. *J. Recent Trends Comput. Sci. Eng. (JRTCSE)* **2025**, *13*, 66–76. [CrossRef]
114. Reddy, S.R.B.; Kanagala, P.; Ravichandran, P.; Pulimamidi, R.; Sivarambabu, P.V.; Polireddi, N.S.A. Effective fraud detection in e-commerce: Leveraging machine learning and big data analytics. *Meas. Sens.* **2024**, *33*, 101138. [CrossRef]
115. Latah, M. Detection of malicious social bots: A survey and a refined taxonomy. *Expert Syst. Appl.* **2020**, *151*, 113383. [CrossRef]
116. Verma, A.; Moghaddam, V.; Anwar, A. Data-Driven Behavioural Biometrics for Continuous and Adaptive User Verification Using Smartphone and Smartwatch. *Sustainability* **2022**, *14*, 7362. [CrossRef]
117. Hossain, M.A. Security perception in the adoption of mobile payment and the moderating effect of gender. *PSU Res. Rev.* **2019**, *3*, 179–190. [CrossRef]
118. Vorobyev, I.; Krivitskaya, A. Reducing false positives in bank anti-fraud systems based on rule induction in distributed tree-based models. *Comput. Secur.* **2022**, *120*, 102786. [CrossRef]
119. Raza, A.; Tsiotsou, R.; Sarfraz, M.; Ishaq, M.I. Trust recovery tactics in financial services: The moderating role of service failure severity. *Int. J. Bank Mark.* **2023**, *41*, 1611–1639. [CrossRef]
120. Zhang, J.; Luximon, Y.; Song, Y. The Role of Consumers' Perceived Security, Perceived Control, Interface Design Features, and Conscientiousness in Continuous Use of Mobile Payment Services. *Sustainability* **2019**, *11*, 6843. [CrossRef]
121. Bickley, S.J.; Chan, H.F.; Dao, B.; Torgler, B.; Tran, S.; Zimbatu, A. Comparing human and synthetic data in service research: Using augmented language models to study service failures and recoveries. *J. Serv. Mark.* **2025**, *39*, 36–52. [CrossRef]
122. Mochon, D.; Schwartz, J. The importance of construct validity in consumer research. *J. Consum. Psychol.* **2020**, *30*, 208–214. [CrossRef]

123. Goyal, M.; Mahmoud, Q.H. A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. *Electronics* **2024**, *13*, 3509. [CrossRef]
124. Rand, W.; Stummer, C. Agent-based modeling of new product market diffusion: An overview of strengths and criticisms. *Ann. Oper. Res.* **2021**, *305*, 425–447. [CrossRef]
125. Jeong, B.; Lee, W.; Kim, D.-S.; Shin, H. Copula-based approach to synthetic population generation. *PLoS ONE* **2016**, *11*, e0159496. [CrossRef] [PubMed]
126. Wang, P.; Loignon, A.C.; Shrestha, S.; Banks, G.C.; Oswald, F.L. Advancing organizational science through synthetic data: A path to enhanced data sharing and collaboration. *J. Bus. Psychol.* **2025**, *40*, 771–797. [CrossRef]
127. Suguna, R.; Suriya Prakash, J.; Aditya Pai, H.; Mahesh, T.R.; Vinoth Kumar, V.; Yimer, T.E. Mitigating class imbalance in churn prediction with ensemble methods and SMOTE. *Sci. Rep.* **2025**, *15*, 16256. [CrossRef]
128. Assefa, S.; Dervovic, D.; Mahfouz, M.; Tillman, R.; Reddy, P.; Veloso, M. Generating synthetic data in finance: Opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance (AIFin '20), New York, NY, USA, 15–16 October 2020*; ACM: New York, NY, USA, 2020. [CrossRef]
129. Ekstrand, M.D.; Chaney, A.; Castells, P.; Burke, R.; Rohde, D.; Slokom, M. Simurec: Workshop on synthetic data and simulation methods for recommender systems research. In Proceedings of the 15th ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September–1 October 2021; pp. 803–805. [CrossRef]
130. Slokom, M. Comparing recommender systems using synthetic data. In Proceedings of the 12th ACM Conference on Recommender Systems, Vancouver, BC, Canada, 27 September 2018; pp. 548–552. [CrossRef]
131. Rezaeinia, S.M.; Rahmani, R. Recommender system based on customer segmentation (RSCS). *Kybernetes* **2016**, *45*, 946–961. [CrossRef]
132. Patki, N.; Wedge, R.; Veeramachaneni, K. The synthetic data vault. In *Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada, 17–19 October 2016*; IEEE: Piscataway, NJ, USA, 2016; pp. 399–410. [CrossRef]
133. Wyllie, S.; Shumailov, I.; Papernot, N. Fairness feedback loops: Training on synthetic data amplifies bias. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), Rio de Janeiro, Brazil, 3–6 June 2024; pp. 2113–2147. [CrossRef]
134. Barbierato, E.; Vedova, M.L.D.; Tessera, D.; Toti, D.; Vanoli, N. A Methodology for Controlling Bias and Fairness in Synthetic Data Generation. *Appl. Sci.* **2022**, *12*, 4619. [CrossRef]
135. Ji, E.; Ohn, J.H.; Jo, H.; Park, M.J.; Kim, H.J.; Shin, C.M.; Ahn, S. Evaluating the utility of data integration with synthetic data and statistical matching. *Sci. Rep.* **2025**, *15*, 19627. [CrossRef] [PubMed]
136. Willman-Iivarinen, H. The future of consumer decision making. *Eur. J. Futures Res.* **2017**, *5*, 14. [CrossRef]
137. Zherdeva, L.; Zherdev, D.; Nikonorov, A. Prediction of human behavior with synthetic data. In *Proceedings of the 2021 International Conference on Information Technology and Nanotechnology (ITNT), Samara, Russia, 20–24 September 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6. [CrossRef]
138. Hyrup, T.; Lautrup, A.D.; Zimek, A.; Schneider-Kamp, P. Sharing is CAIRing: Characterizing principles and assessing properties of universal privacy evaluation for synthetic tabular data. *Mach. Learn. Appl.* **2024**, *18*, 100608. [CrossRef]
139. Beduschi, A. Synthetic data protection: Towards a paradigm change in data regulation? *Big Data Soc.* **2024**, *11*, 20539517241231277. [CrossRef]
140. Bhatia, S.; Alojail, M. A Novel Approach for Deciphering Big Data Value Using Dark Data. *Intell. Autom. Soft Comput.* **2022**, *33*, 1261–1271. [CrossRef]
141. Marumolwa, L.; Marnewick, C. Unveiling Dark Data in Organisations: Sources, Challenges, and Mitigation Strategies. *Int. J. Serv. Sci. Manag. Eng. Technol. (IJSSMET)* **2025**, *16*, 1–32. [CrossRef]
142. Corallo, A.; Crespino, A.M.; Del Vecchio, V.; Lazoi, M.; Marra, M. Understanding and defining dark data for the manufacturing industry. *IEEE Trans. Eng. Manag.* **2021**, *70*, 700–712. [CrossRef]
143. Gimpel, G. Bringing dark data into the light: Illuminating existing IoT data lost within your organization. *Bus. Horiz.* **2020**, *63*, 519–530. [CrossRef]
144. Shin, S.I.; Kwon, M.M. Dark data: Why What You Don't Know Matters: Dark Data: Why What You Don't Know Matters, by David J. Hand, New Jersey, US, Princeton University Press, 2020, 330 pp. *J. Inf. Technol. Case Appl. Res.* **2023**, *25*, 112–118. [CrossRef]
145. Schembera, B.; Durán, J.M. Dark data as the new challenge for big data science and the introduction of the scientific data officer. *Philos. Technol.* **2020**, *33*, 93–115. [CrossRef]
146. Jurn, S.; Kim, W. Improving Text Classification of Imbalanced Call Center Conversations Through Data Cleansing, Augmentation, and NER Metadata. *Electronics* **2025**, *14*, 2259. [CrossRef]

147. Alfian, G.; Octava, M.Q.H.; Hilmy, F.M.; Nurhaliza, R.A.; Saputra, Y.M.; Putri, D.G.P.; Syahrian, F.; Fitriyani, N.L.; Atmaji, F.T.D.; Farooq, U.; et al. Customer Shopping Behavior Analysis Using RFID and Machine Learning Models. *Information* **2023**, *14*, 551. [CrossRef]
148. Cai, H.; Dong, T.; Zhou, P.; Li, D.; Li, H. Leveraging Text Mining Techniques for Civil Aviation Service Improvement: Research on Key Topics and Association Rules of Passenger Complaints. *Systems* **2025**, *13*, 325. [CrossRef]
149. Gimpel, G. Dark data: The invisible resource that can drive performance now. *J. Bus. Strategy* **2021**, *42*, 223–232. [CrossRef]
150. Martin, K.D.; Borah, A.; Palmatier, R.W. Data privacy: Effects on customer and firm performance. *J. Mark.* **2017**, *81*, 36–58. [CrossRef]
151. Singh, J.; Gebauer, H. Clean Customer Master Data for Customer Analytics: A Neglected Element of Data Monetization. *Digital* **2024**, *4*, 1020–1039. [CrossRef]
152. Dakic, D.; Stefanovic, D.; Vuckovic, T.; Zizakov, M.; Stevanov, B. Event Log Data Quality Issues and Solutions. *Mathematics* **2023**, *11*, 2858. [CrossRef]
153. Reviglio, U. The untamed and discreet role of data brokers in surveillance capitalism: A transnational and interdisciplinary overview. *Internet Policy Rev.* **2022**, *11*, 1–27. [CrossRef]
154. Samuelsson, L.; Cocq, C.; Gelfgren, S.; Enbom, J. *Everyday Life in the Culture of Surveillance*; Nordicom: Gothenburg, Sweden, 2023. [CrossRef]
155. Yang, Q.; Lepore, C.; Eynard, J.; Laborde, R. From theory to practice: Data minimisation and technical review of verifiable credentials under the GDPR. *Comput. Law Secur. Rev.* **2025**, *57*, 106138. [CrossRef]
156. Kerkhof, A.; Münster, J. Detecting coverage bias in user-generated content. *J. Media Econ.* **2019**, *32*, 99–130. [CrossRef]
157. Zhu, Q.; Lo, L.Y.H.; Xia, M.; Chen, Z.; Ma, X. Bias-aware design for informed decisions: Raising awareness of self-selection bias in user ratings and reviews. In Proceedings of the ACM on Human-Computer Interaction, 6(CSCW2), Taipei, China, 8–22 November 2022; pp. 1–31. [CrossRef]
158. Olteanu, A.; Castillo, C.; Diaz, F.; Kıcıman, E. Social data: Biases, methodological pitfalls, and ethical boundaries. *Front. Big Data* **2019**, *2*, 13. [CrossRef]
159. Finck, M.; Biega, A.J. Reviving purpose limitation and data minimisation in data-driven systems. *Technol. Regul.* **2021**, *2021*, 44–61. [CrossRef]
160. Chereja, I.; Erdei, R.; Delinschi, D.; Pasca, E.; Avram, A.; Matei, O. Privacy-Conducive Data Ecosystem Architecture: By-Design Vulnerability Assessment Using Privacy Risk Expansion Factor and Privacy Exposure Index. *Sensors* **2025**, *25*, 3554. [CrossRef] [PubMed]
161. Rupp, V.; von Grafenstein, M. Clarifying "personal data" and the role of anonymisation in data protection law: Including and excluding data from the scope of the GDPR (more clearly) through refining the concept of data protection. *Comput. Law Secur. Rev.* **2024**, *52*, 105932. [CrossRef]