Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 10 Issue 02
2026

Research Article
https://doi.org/10.56979/1002/2026/1135

Reinforcement Learning for Customer Lifetime Value Optimization: A Conceptual Framework and Directions for Future Research

Hani Iwidat[1*]

[1]Department of Data Science, Al-Istiqlal University, Jericho, Palestine.
[*]Corresponding author: Hani Iwidat. Email: hani.iwidat@pass.ps

Received: October 20, 2025 **Accepted:** January 03, 2026

Abstract: As a technique for improving sequential decision-making in customer-centric marketing situations, reinforcement learning has attracted growing interest. With special attention on its concordance with customer lifetime value maximization, this paper investigates how reinforcement learning has been employed in client modeling, personalization, pricing, engagement, and retention. In this study, a conceptual research methodology of the theory synthesis type was employed beginning with reviewing databases from Google Scholar, Scopus, and Web of Science keywords such as 'reinforcement learning' AND 'customer lifetime value' (2015-2026), yielding more that 100 studies after screening Although current research shows great potential to affect long-term customer behavior, most uses depend on short-term or surrogate performance indicators rather than explicitly maximizing lifetime value. This study logically and theoretically combines previous studies to offer a conceptual framework connecting reinforcement learning to lifetime value optimization, presents a taxonomy of approaches and tasks, and highlights major obstacles, knowledge gaps, and future directions to help to overcome this restriction. The paper sees reinforcement learning as the basis for creating ethically, value-driven, scalable consumer intelligence systems.

Keywords: Reinforcement Learning; Customer Lifetime Value; Customer Modeling; Marketing Analytics; Personalization; Sequential Decision-Making

1. Introduction

Contemporary e-commerce and digital marketing give Customer Lifetime Value (CLV) precedence above one-time sales, since it directly contributes to long-term profitability. Reflecting not just quick transactions but also loyalty, repeat purchasing, and referrals, CLV is the overall income a consumer generates during their whole relationship with a company. Companies, particularly in mobile apps, internet retail, and subscription services, concentrate on maximizing CLV since keeping current clients is less expensive than getting new ones. However, traditional CLV models have limitations. They often rely on static formulas, short-term predictions, or segmentation methods like RFM, treating customer value as fixed [1, 2]. These approaches are less effective in modern CRM and marketing strategies, as they fail to capture the dynamic nature of customer behavior. Preferences and engagement evolve over time through interactions such as website visits, marketing emails, purchases, and service requests. Current personalization and targeting typically optimize single actions using static data, ignoring how actions alter future customer states [3]. This can lead to suboptimal decisions, like overselling that causes churn or missing opportunities for value-building incentives. In essence, common CLV analyses overlook sequential decision-making, failing to account for how marketing actions shape ongoing customer behavior [4].

By expressly simulating long-term, sequential decision-making under uncertainty, reinforcement learning (RL) has become a logical approach to address these drawbacks [5]. RL is a field of machine learning for cases where an agent (such as a marketing decision system) learns, by trial and error, which sequence of events maximizes cumulative reward. In the customer management framework, states represent the current level of engagement or customer context; actions can be marketing initiatives (offers, recommendations, communications); and rewards can be instant profit or a change in customer value resulting from a contact. Importantly, RL systems seek to maximize long-term reward rather than short-term benefits. Since CLV is a cumulative, long-term result, RL is ideal for maximizing it [5]. Unlike conventional shortsighted techniques, recent studies show that "RL offers a pragmatic approach for handling sequential customer contacts by stressing the long-term influence of every choice [3], RL, in other words, can dynamically personalize a sequence of actions for each customer to maximize their predicted lifetime value, continually adjusting strategy as new data on customer responses arrive [3]. RL's capacity to learn from data, including historical offline data, enables companies to simulate and analyze long-term plans without incurring the full cost of live experimentation [3]. Unlike static scoring systems, reinforcement learning (RL) is a significant change in CLV optimization since it modifies marketing choices constantly as customer behavior changes. Early RL uses in marketing, including dynamic pricing, adaptive ad sequencing, and personalized coupon timing, show how it may increase retention [6], direct consumer acquisition, and maximize long-term value. RL finds ideal strategies for every consumer state that match the most effective marketing approach to maximize future rewards.

It is crucial to evaluate present developments and obstacles as RL gets more included in CLV plans. Recent studies show that while there has been progress in areas like personalization, churn prevention, and loyalty optimization, there are also problems with how well customers are represented, data scarcity, limits on how well things can be learned offline, and the need for model explainability [7]. To sum up, our efforts are threefold: first, to explain why RL is used in managing long-term customer value and to create a conceptual link between RL and CLV; second, to carefully review and group current RL applications and research in customer modeling, so evaluating what has been accomplished thus far; and third, to seriously analyze open questions and suggest future research directions to further this interdisciplinary field.

The literature is studied thematically to create a conceptual framework, suggest a classification of RL-CLV techniques, highlight research needs, and define next research directions.

1.1. Methodology

Since the existing literature lacks integrated explanatory models connecting between RL and CLV concepts, a conceptual research methodology particularly the theory synthesis type was employed to address unclear understating in linking between different constructs. According to [8] authors started with a review of peer-reviewed literature to identify core definitions of both construct, theoretical support, and then the main assumptions of each concept. The Databases used to include Scopus, Google Scholar, and Web of Science were queried using combinations such as ("reinforcement learning" OR "RL") AND ("customer lifetime value" OR "CLV" OR "customer modeling"). The search covered publications from 2015 to 2025 to capture recent advancements in the field. Inclusion criteria focused on peer-reviewed papers that empirically applied RL to customer-related tasks, such as personalization, pricing, engagement, retention, or churn prediction. Exclusion criteria eliminated non-English articles, non-empirical works (e.g., purely theoretical discussions), and duplicates.

The screening process involved three stages: initial title and abstract review to assess relevance (approximately 155 results identified); full-text evaluation for methodological fit and contribution to CLV-RL integration; and final selection by the author. This allows the author to understand the findings of previous research and thus identifying the gap in the literature. Additionally, developing a new conceptual framework is justified as the literature logically and theoretically indicates to the potential mechanisms of influence between the two constructs. This process resulted 16 studies, which we used as the foundation for our review, categories, and gap analysis in the sections that follow. Table 1 reveals 70% focus on proxies vs. explicit CLV.

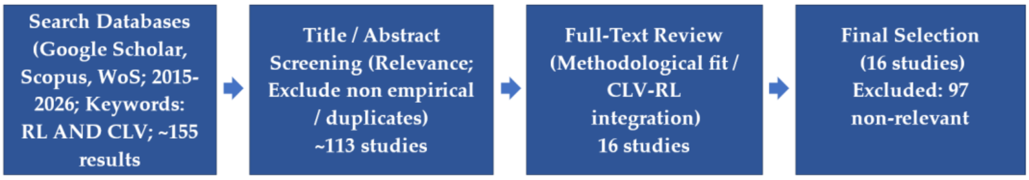

Search Databases (Google Scholar, Scopus, WoS; 2015-2026; Keywords: RL AND CLV; ~155 results
Title / Abstract Screening (Relevance; Exclude non empirical / duplicates) ~113 studies
Full-Text Review (Methodological fit / CLV-RL integration) 16 studies
Final Selection (16 studies) Excluded: 97 non-relevant

Figure 1. Illustrates the screening workflow, and the key attributes of these studies are summarized in Table 1

2. Foundations of Customer Lifetime Value (CLV)

2.1. What CLV Indicates

A major indicator in a customer-centric corporate plan, Customer Lifetime Value (CLV) reflects the projected net income a business earns from a customer over the entire duration of their relationship. Formally, it is the discounted sum of future cash flows credited to a customer, taking into account revenues, expenses, and the time value of money [2]. CLV serves as a guiding principle for determining how much to invest in keeping, caring for, and acquiring each customer, as well as for financial projections.

From a machine learning perspective, CLV is like the long-term return in reinforcement learning, since it can be thought of as the total reward associated with a customer's future behavior. CLV evaluates patterns of interaction and judgments made by both the consumer and the company over time, rather than focusing on a single transaction. This alignment with long-term rewards motivates the use of RL for CLV optimization, as discussed in Section 6.

2.2. Why data-driven companies depend so much on CLV

Companies in the modern digital economy gather enormous volumes of customer interaction data, from clicks and purchases to support inquiries and churn events. According to several study on CLV [2, 9-11] authors come up with a unifying framework for data-driven decisions in several spheres:

1. Personalized marketing: CLV helps prioritize customers who should receive special offers or interaction initiatives.
2. Customer segmentation: In campaign design or service delivery, high-value vs. low-value customers can be handled differently.
3. Resource allocation: Customers' lifetime value determines the investment in major marketing activities like acquisition or retention for different groups of customers.
4. Planning and forecasting: This long-range project prompts to align the sales, support, and product choices with the future profit.

In other words, CLV gives a quantifiable ground for long-term strategic planning across the board of marketing, sales, and customer support. It serves as a performance goal in AI systems beyond transient surrogate measures such as click-through rate or quick conversions.

2.3. Conventional Techniques of CLV Prediction

Estimating CLV has historically been done using a number of different modeling techniques. These methods vary in complexity, assumptions, and flexibility:

2.3.1. Models of RFM

Among the first heuristic techniques are Recency, Frequency, Monetary (RFM) models, which rate consumers depending on the frequency with which the consumer purchased something, the purchase frequency for the consumer, and how much the consumer has spent Based on these factors, customers are categorized into high- and low-value segments [12]. Although straightforward, RFM models often overlook future uncertainty and variable interactions, and do not provide an official CLV prediction.

2.3.2. Statistical and Survival Models

Based on patterns of past activity, probabilistic models predict the probability of future purchases. Here are some often-used methods:

Pareto/NBD model: Assumes customer "lifetime" follows a Pareto distribution and their purchases follow a Negative Binomial process [13].

A modification of the BG/NBD model allows customer-specific dropout probabilities [1].

These models, which rely on strong parametric assumptions (e.g., that customer behavior is static) and do not readily incorporate external factors such as marketing initiatives, yield interpretable CLV estimates.

Methods of survival analysis estimate the time till churn as a function of client traits. Among the often-used instruments are the Cox proportional hazards model, which estimates the hazard rate as a function of covariates, and a nonparametric estimator of the survival function.

These models assist in forecasting retention curves, which can then be combined with average income data to derive CLV. They therefore tend to handle purchase timing and financial worth independently, which restricts their usefulness in combined CLV estimation.

2.4. Machine Learning Strategies

More modern CLV models use supervised learning approaches to forecast future value directly. Some of the most widely used algorithms include simpler, interpretable linear or logistic regression; gradient boosting; random forests (which are more accurate with nonlinear patterns); and neural networks, particularly in high-dimensional settings.

These models sometimes present CLV projection as either a regression (predicting dollar value) or a categorization job (e.g., high vs. low CLV). They usually focus on point estimates and do not optimize for long-term decision-making or sequential actions, even though this might improve accuracy.

2.5. Restrictions of Conventional CLV Modeling

The CLV modeling techniques have generated valuable data and provided useful tools, but at the same time, they have limitations that make them unsuitable for dynamic and data-driven environments. One major drawback is that it is based on fixed assumptions [2]. The fact that preferences and engagement pattern changes are the results of both internal and external factors is not taken into consideration by many traditional CLV models, which in a way discount this truth. Outdated assumptions lead to incorrect predictions and suboptimal strategies in cases like the fast-changing consumer sectors or the introduction of new marketing campaigns [14].

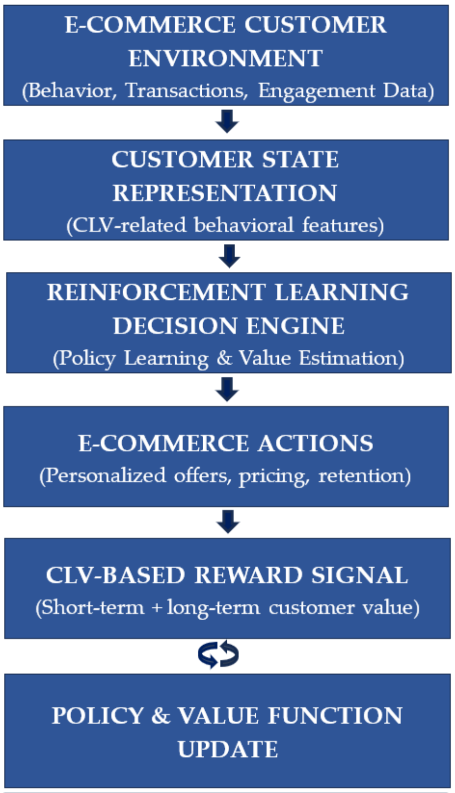
Another limitation is connected with the non-interactivity of these models [2]. It is typical for CLV estimation to assume that the buyer's behavior is exogenous and thus, not impacted by the company's actions. Marketing activities, like price shifts, discounts, or content recommendation changes can, however, have a direct effect on consumers' later behaviors. By neglecting this interaction, traditional models do not recognize the causation that links corporate actions and consumer outcomes [10]. As a result, they can predict the future value, but they do not provide real advice on how to increase it or to influence it in a positive way.

Moreover, the majority of regular CLV models still look like they have not gone very far. Even when using machine learning methods, the models often seem to be more concerned with current accuracy, such as anticipating the next purchase or short-term sales, rather than aiming for maximum cumulative results over the lifetime of the customer [1]. This short-sighted approach might result in the company not being able to position itself properly for the future in the case of the current earnings foregone for the sake of long-term value. Additionally, traditional methods usually lack the necessary adaptability for continuous online learning. These restrictions highlight the need for more dynamic approaches like RL, which we connect to CLV in Section 6.

2.6. Motivation for Reinforcement Learning in CLV Optimization

These restrictions highlight that a plan that is not only engaging and adaptable but also proactive is necessary to get the most out of the clients. A top-notch solution to this is Reinforcement Learning (RL) model [15]. In this case, a time-varying, cumulative measure of customer lifetime value (CLV) is the right way to track this measure of value. It does not just come from one event but accumulates from a series of contacts between the client and the company. Since RL is designed to solve such complexities of sequential decision-making under uncertainty, it is definitely a good match for the intricacies of CLV modeling [16].

Thus, the decision-makers can consider the customer engagement process as a dynamic one that is influenced by RL reinforcement learning systems, where every marketing activity creates not only the present but also the future states and actions. RL-agent responding to consumer feedback can implement time shifting policies that maximize the long-term value rather than the short-term gain. The same RL systems can switch their strategies according to the different customers' paths, thus making sure that the right action is done at the right time for every different user. It is through this optimization of customer lifetime value from the perspective of reinforcement learning that firms can have a prescriptive power rather than just predictive knowledge. RL lets businesses ask and answer a more strategic question, what order of choices will optimize this customer's value over time, instead of inquiring what a customer is worth. This open ways for innovative customer interaction strategies [17]. Figure 2 formalizes these motivations, with theoretical development.


E-COMMERCE CUSTOMER ENVIRONMENT
(Behavior, Transactions, Engagement Data)
CUSTOMER STATE REPRESENTATION
(CLV-related behavioral features)
REINFORCEMENT LEARNING DECISION ENGINE
(Policy Learning & Value Estimation)
E-COMMERCE ACTIONS
(Personalized offers, pricing, retention)
CLV-BASED REWARD SIGNAL
(Short-term + long-term customer value)
POLICY & VALUE FUNCTION UPDATE

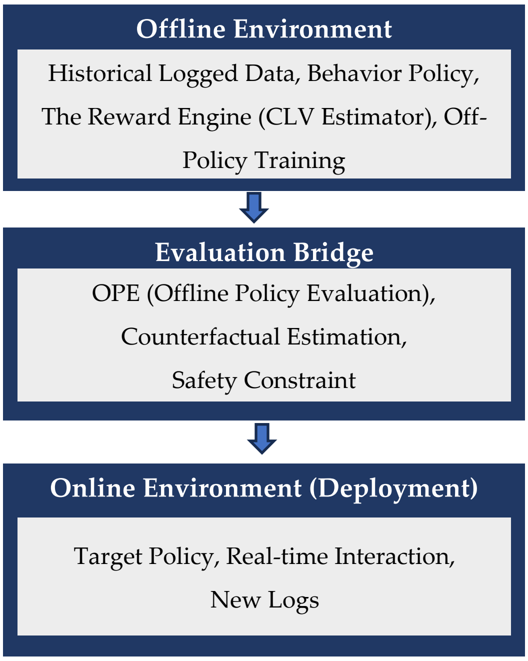

Offline Environment
Historical Logged Data, Behavior Policy, The Reward Engine (CLV Estimator), Off-Policy Training
Evaluation Bridge
OPE (Offline Policy Evaluation), Counterfactual Estimation, Safety Constraint
Online Environment (Deployment)
Target Policy, Real-time Interaction, New Logs

Figure 2. Conceptual framework: RL for CLV optimization in e-commerce environments

Figure 3. The Integrated Offline-to-Online RL Framework for CLV Optimization.

2.7. Proposed Conceptual Framework

Figure 2 illustrates our proposed conceptual framework for reinforcement learning (RL) in customer lifetime value (CLV) optimization. The framework maps the RL Markov Decision Process (MDP) components to e-commerce customer journeys: customer states (RFM scores, engagement history) feed into the RL agent, which selects marketing actions (pricing, recommendations, offers) via an optimized policy π. Actions generate immediate rewards (profit, retention) and transition to new states, forming a feedback loop that maximizes discounted cumulative CLV (∑γt R_t) over time. This structure addresses traditional CLV limitations by enabling dynamic, sequential decision-making under uncertainty. Figure 3 illustrates the transition from offline training on logged marketing data to online deployment. It highlights the critical role of Offline Policy Evaluation (OPE) and counterfactual estimation in mitigating risk before real-world interaction, with CLV serving as the primary long-term reward signal.

3. **Basics of Reinforcement Learning**

Reinforcement Learning (RL) is obtained by trial and error and smartly directed by the reward feedback signal; it is a fundamental concept in artificial intelligence regarding the learning of agents to make decisions through the environment interaction. Because of these features that strongly resemble customer relationships in business contexts, RL is intuitively suitable for issues that need sequential decision-making, postponed outcomes, and dynamic feedback [18].

3.1. Markov Decision Processes (MDP)

The core of Reinforcement Learning (RL) is the Markov Decision Process (MDP). An MDP is typically denoted by a tuple ⟨S, A, P, R, γ⟩ where S refers to the set of all possible states, A denotes the actions, P indicates the transition probabilities between states, R is the reward function, and γ represents the discounting factor. The MDP framework allows customer-centric firms to illustrate consumer journeys as chains of interactions that are affected by both the company's activities and the feedback from outside [18].

3.2. States, Activities, and Changes in Dynamics

In the RL context, a state denotes the environment at a given moment. States could include factors such as how much someone has bought in the past, how recently they did something, how well they responded to past ads, or what kinds of products they like for customer modeling. A decision taken by the agent, such as offering a discount, suggesting a product, or deciding not to intervene, corresponds to an action. The environment reacts to these actions through transition dynamics that specify the temporal evolution of the customer state. they represent both visible and latent behavioral changes driven by the company's choices [31].

3.3. Discount Factors and Reward Functions

The reward function is a key element of reinforcement learning since it measures the quick advantage gained from an action carried out in a given state. Rewards in the context of CLV optimization might mirror transaction profit, rise in involvement, or any other indicator for consumer value. The RL agent ultimately aims to maximize the cumulative reward, which takes into consideration results from now and in the future. RL uses a discount factor, usually called γ (gamma), with values between 0 and 1, to help the agent give more importance to long-term gains. This fits well with CLV modeling, where delayed outcomes are normal and often more important than short-term returns.

3.4. Value Functions and Policy

A policy specifies the approach an agent takes when choosing actions based on the current circumstances. Value functions, which project the expected cumulative reward connected with states or state-action pairs, help to evaluate a policy's quality. The state-value function (V(s)) forecasts the overall expected return from state *s*; the action-value function (Q(s, a) estimates the return from state *s* and decision *a*.

3.5. Exploration vs. Exploitation

In reinforcement learning (RL), one of the primary difficulties is to strike a balance between exploration and exploitation. Exploitation is the application of past experience to yield the highest benefit at the moment, whereas exploration is the process of testing new ideas which might turn out better. This conflict is especially significant in marketing and CRM software where the businesses have to juggle between trying out new ideas (for example, new and enticing offers or incentives) and relying on the ones that are already successful and, therefore, proven. To dynamically handle this trade-off, RL techniques often use ε-greedy, top confidence limits, or Thompson sampling.

Table 1. Summary of Reviewed RL-CLV Literature applied to customer-related marketing tasks, showing RL formulation, state, actions, reward design, setting, and main reported outcome.

Authors, (Year) Ref.	Marketing Task / Domain	RL Formulation	State Representation	Actions
Theocharous et al., 2015 [19]	Personalized Ad Recommendation (PAR) / Banking Industry	MDP, value-based RL	High-dimensional: Demographics, interests, location, visit recency, frequency, and interaction history.	Selecting 1 of K banking offers (K=7 or 12).
Zhao et al., 2017 [20]	E-commerce personalized item recommendations	Actor-Critic MDP (LIRD framework) using DDPG	Chronological browsing history (N	Selecting a list of K =4 items to present

			items) N=10 as embedded vectors	to the user
Ferreira et al., 2018 [21]	Price-based Network Revenue Management	Multi-armed Bandit (MAB)	Implicit state: Current inventory levels Contextual features	Selection of a price vector for products from a finite set of admissible prices
Cheung et al., 2017 [22]	Dynamic Pricing for daily deal marketplaces	Exploration-Exploitation tradeoff modeled	Real-time sales data	Selecting a unit price Pt from a finite set
Apt, 2025 [4]	Dynamic retail pricing and revenue management	Q-Learning (Model-free RL)	Product type and day of the week (weekday vs. weekend)	Selecting from a set of admissible price points
Brown & Mikem, 2024 [23]	Personalized Dynamic Pricing for Online Retail	Deep Reinforcement Learning (DRL) with Neural Networks	Customer behavior competitor prices, inventory, and market trends	Real-time product price adjustments
Mahdavian et al., 2025 [24]	Personalized product recommendations and dynamic pricing	DRL-based MDP with a 5-class joint output	Customer demographics, behavioral history, and detailed product features	Selecting a product and a corresponding discount level (0% to 30%+)
Xia al., 2024 [25]	Personalized coupon targeting, customer shopping journeys.	Offline RL (PPO) and Contextual	Summarized purchase	Selecting discrete

		Bandits (LinUCB, LinTS) trained on batch data.	histories and latent customer utilities	discount levels
Bose et al., 2023 [16]	CLV Optimization (Retail and Subscription Services)	Integrated RL and Predictive Analytics iterative loop	Dynamic customer interaction data and preferences	Adaptive, personalized marketing strategies
Ma et al., 2025 [3]	Dynamic personalization for retention in mobile gaming	DQN with Multi-Response State Representation (MRSR)	Behavioral signals: Recency, Engagement (points/stars), and Spending	Selecting game difficulty levels (1–10)
Nie & Ahmad, 2025 [26]	Dynamic Reward Systems and Loyalty Programs	Multi-armed Bandits (Thompson Sampling)	Customer segments and real-time feedback patterns	Adjusting reward type, value, and timing
Swaminathan et al., 2017 [27]	Ranking and whole-page optimization (Search & Ads)	Combinatorial Contextual Bandit	User context (query, profile, temporal data)	Recommendation on Slates.
Feng et al., 2025 [28]	Automated A/B testing and content personalization	Actor-Critic RL integrated with LLMs	Fused vector of user portraits, query context, and content features	Selecting/assigning generated marketing content variants
Thiruvayipati, 2024 [14]	CLTV management and optimization in Retail/E-commerce	Continuous adaptation for long-term CLTV maximization	Dynamic customer journeys and historical interaction data	Personalized acquisition, retention, and dynamic pricing

Ghavamzadeh et al., 2015 [29]	Lifetime Value (LTV) optimization in digital marketing	High-confidence off-policy evaluation (Sequential Decision Problems)	States/Observations in an MDP or POMDP framework	selection based on policy parameters (theta)
Li et al., 2017 [30]	Cyber-Physical Systems (Cloud, Smart Grid, HVAC)	Deep Q-Learning (DQN) & Actor-Critic Framework	High-dimensional inputs (system traces, sensors, temperature)	Resource distribution, task scheduling, and control set-points

Reward / Objective (CLV vs proxy)	Setting & Evaluation	Main reported outcome / improvement
Explicit LTV optimization: Reward = 1 for click, 0 otherwise; Objective = maximize discounted sum of clicks per user (LTV).	Offline/historical data from real banking campaigns; Off-policy evaluation using three methods: (CI), (TT), and (BCa).	RL-based LTV optimization outperformed myopic greedy approaches. Proved that OPE can provide statistical "safety" guarantees for marketing deployments.
Long-term cumulative reward based on clicks/orders. Proxy (engagement-focused): Immediate rewards: 0 (skip), 1 (click), 5 (order);	Offline training and evaluation via an online environment simulator. ataset using MAP and NDCG metrics; Comparison with CF, FM, DNN, RNN, DQN baselines	1. Outperforms traditional methods (CF, FM, DNN) and RNN in long sessions. 2. Comparable to DQN in performance but with faster training.
Expected Revenue maximization over a finite selling season	Online learning	Numerical results showed significant revenue improvement over phased exploration and UCB-based
Revenue Maximization	Finite Horizon (T periods) with restricted price experimentatio	Proved a regret bound of O(log(M) T) field study confirmed significant revenue and deal booking improvements.
Profit (Revenue minus operational costs)	Simulation environment using real-world electronic product data	Higher optimized demand and revenue compared to static methods
CLV, cumulative revenue, profit, and customer retention	Exploratory/Theoretical framework for real-time online environments	Maximized revenue capture and enhanced customer satisfaction via personalization
Profit-oriented optimization based on Willingness to Pay (WTP)	Offline testing using the Dunnhumby dataset; compared against RL4RS and DiffRec	Significant improvements in recommendation accuracy and revenue-focused targeting

Accumulated customer revenue (proxy for long-term spend/loyalty).	Evaluated via "RetailSynth" simulation platform against static baseline policies.	RL/Bandits significantly outperformed static policies; PPO effectiveness scaled strongly with data size.
Long-term CLV and overall profitability	Empirical testing on diverse datasets; analyzed via retention and ARPU	Significant precision and profit gains compared to conventional static models
CLV-focused: Discounted sum of future active days and engagement	Offline training; evaluation via Off-Policy Evaluation (OPE)	Enhanced state representation led to superior long-term value capture compared to standard RL
Customer Engagement and Loyalty (Proxy for CLV)	Simulation environment; benchmarked against static and ε-greedy models	Faster adaptation to customer preferences and higher cumulative engagement
Page-level metrics (CTR, Revenue, NDCG)	Offline evaluation using the Pseudoinverse (PI) estimator	Exponentially better sample complexity for offline policy testing
Long-term Revenue (based on CTR and Conversion)	Real-time automated A/B testing with a memory-augmented estimator	Automated strategy optimization that adapts to user preference drift
CLV Maximization	Review of the shift from static/retrospective to predictive models	Transformation of CLTV into a proactive strategic asset for growth
Expected discounted return (LTV) with lower confidence bounds	Offline evaluation using importance sampling and novel concentration inequalities	Validated a method for safely evaluating new policies with statistical guarantees
Energy/Cost optimization and accumulative reward maximization	Offline simulations using real-world traces	20%–70% cost/energy savings across multiple applications

3.6. Episodic vs. Continuing Tasks

RL assignments fall into either the episodic or the ongoing categories. In episodic chores, interactions are separated into distinct episodes with clear beginnings and ends, such as a limited-time marketing campaign or a defined client acquisition strategy. Conversely, ongoing decision-making in long-term customer engagement or lifecycle management defines continuous activities with no inherent endpoint. In CLV settings where the customer relationship spans forever, and decisions must be optimized across a continuous horizon, ongoing activities are very important. Knowing the task structure guides the selection of the best RL methods and evaluation techniques.

4. RL Algorithms Relevant to Customer Decision-Making

There are many different ways to estimate and improve decision policies in Reinforcement Learning (RL). Broadly speaking, these algorithms fall into policy-based approaches, which directly maximize the decision policy, and value-based methods, which learn value functions to direct action choice. Each choice (e.g., a product recommendation or promotion) influences future outcomes, including purchases, churn, and engagement; hence, both families are theoretically relevant for modeling long-term client behavior. in CLV optimization, as evidenced by the reviewed studies in Table 1. Key algorithm families are presented in this section along with their relevance to CLV optimization.

4.1. Value-Based Methods

Value-based methods seek to forecast a function that measures the anticipated long-term return (that is, customer value) from a given state-action pair. Called the action-value function or Q-function, this function shows how much reward the agent is likely to get in total if it does something in a certain state and then does the best thing to do after that [31]. This value can indicate future profit, loyalty, or retention probability following a marketing campaign in customer modeling. In customer modeling, it estimates future value post-actions like campaigns.

4.1.1. *Q-learning*

Q-learning is an off-policy method that learns the optimal Q-function regardless of the agent's current policy [32]. It updates Q-values using the maximum estimated value of the next state, allowing convergence to the best policy even when exploratory actions are used. In situations when the future influence of marketing initiatives is unclear and constantly changing, this strategy is especially relevant for customer value optimization. in uncertain marketing scenarios For instance, as seen in [19], Q-learning optimizes ad sequencing for lifetime value; similarly, [4] applies it to dynamic retail pricing

4.1.2. *SARSA*

On-policy substitute for Q-learning, SARSA (State--Action--Reward--State--Action), updates Q-values according to the real trajectory of the agent [33]. This increases the agent's sensitivity to the current discovery approach and makes it safer in situations where strong exploitation could yield unwanted consumer results. In CRM settings, SARSA could be useful when implementing conservative marketing plans that have to fit with particular company rules or ethical restrictions yet still learn to maximize CLV.

4.1.3. *Deep Q-Networks*

Deep Q-networks expands Q-learning through the use of neural networks for approximating the Q-function. This scales RL for high-dimensional or unstructured state spaces [34]. In this case, the scaling up of the RL is particularly advantageous in consumer-targeted programs, where the inputs can consist of purchase history, user's web behavior, clickstream data, and even contextual factors like time of day or type of device. By matching intricate customer states to long-term value-maximizing actions, DQNs can enable fine-grained customization. For example, [35] uses DQNs for customer churn prediction, aligning with patterns in Table 1.

4.1.4. *Double DQN and Dueling DQN*

Double DQN has completely removed the overestimation bias that was present in standard DQNs [26] by making a distinction between action selection and value evaluation. The learning process becomes more stable and long-term value projection more precise, which is a major requirement in scenarios of customer churn management or subscription renewals with sparse-reward settings. Dueling DQN model applies splitting between state-value and action-advantage estimation which leads to identification of informative states regardless of actions [36]. In customer modeling, this aids one to know the customer segments that are naturally profitable and those that need a lot of effort to raise CLV. These extensions enhance applications like those in [20] for recommendations.

4.2. Policy-Oriented Strategies

These methods are particularly effective in settings that involve continuous or high-dimensional action spaces and where stochastic policies allow for controlled experimentation or uncertainty management, as they learn a direct state-to-action relationship without relying on any explicit value function.

4.2.1. *Reinforce*

By rewarding actions that yield high returns [37] and sampling entire episodes, the Monte Carlo policy gradient technique REINFORCE updates policy parameters. Although it is considered to be easy, it still offers the drawbacks of slow convergence and large variance. Nevertheless, it is especially useful in CLV situations where the decision involves several customer touch-points and the primary goal is to maximize the long-term outcome, such as onboarding process or retention sequence. Its ease of use attracts low-risk experimental conditions where the attributes of modularity and interpretability are of utmost importance.

4.2.2. *Policy Gradient Methods*

Contemporary policy gradient techniques such as Actor-Critic, Advantage Actor-Critic (A2C), and Proximal Policy Optimization (PPO) are the successors of REINFORCE [38] since they utilize value estimates and apply restrictions to the learning process. CRM systems need a combination of established techniques, exploitation, and a fresh engagement approach to best suit these algorithms. For instance, PPO offers support for stability and at the same time promotes personalization by suggesting next-best actions in client journeys and consequently allowing safe policy changes. Actor–critic methods are optimal for such real-time, interactive marketing systems where both performance and interpretability are crucial since these methods allow you to simulate the policy, that is, what to do, and the value function, which explains why it's useful. In reinforcement learning systems, actor-critic algorithms combine value-based and policy-based learning to produce consistent and quick decision-making [21].

4.3. Techniques for Actors and Critics

Particularly when working with continuous or partially observable data, reinforcement learning (RL) uses actor–critic techniques combining policy-based and value-based learning to provide consistent and sample-efficient decision-making. Popular methods like A2C and A3C vary in their approach to gather experience; A2C learns concurrently while A3C uses asynchronous parallel agents for faster, more stable learning. These techniques fit customer modeling because they can extract knowledge from varied, sequential customer paths.

More complex strategies like Proximal Policy Optimization (PPO) enhance stability by controlling rapid policy changes, therefore making them perfect for marketing choices involving financial risk, such as discount allocation or churn-sensitive targeting. DDPG (Deep Deterministic Policy Gradient) applies actor–critic techniques to continuous action spaces, providing theoretical significance for customized service timing and dynamic pricing in situations where choices are not finite.

In every instance, actor–critic approaches provide a flexible way to design adaptable, long-term marketing strategies that respond to consumer behavior in real time while accounting for the delayed effects of every decision. Deep deterministic policy gradient approaches are ideal for real-world optimization challenges since they make reinforcement learning in continuous action spaces feasible [39].

4.4. Contextual Bandits and Multi-Armed Bandits

Although complete reinforcement learning requires learning across multi-step trajectories, many marketing decisions, including choosing a promotion or suggesting a headline, are one-shot but sequential, hence Multi-Armed Bandits (MABs) and contextual bandits are more suitable. Lattimore & Szepesvári find these algorithms appealing for real-time marketing applications because they are computationally simpler and require fewer assumptions [40].

A basic but efficient bandit approach called epsilon-greedy has the agent exploring a random action with probability ε and exploiting the best-known action so far with probability 1−ε. To strike a balance between learning and return on investment, A/B/n testing and tailored promotion systems frequently use this approach [23].

Upper Confidence Bound (UCB) strategies promote less-tested choices when uncertainty is high by choosing actions with the greatest upper-bound estimates of possible reward, therefore improving exploration. In ambiguous contexts, UCB has been used in dynamic ad placements to maximize consumer attention [41].

A Bayesian strategy called Thompson Sampling (or posterior sampling) picks the one with the largest sample from the posterior distribution of each arm's reward. This approach is especially helpful for recommendation systems, adaptive content delivery, and personalized offer engines [4], as it naturally adapts to changing user tastes.

For small, sequential decision-making situations such as choosing between promotions, selecting email subject lines, or testing offer versions, these bandit approaches are especially helpful. They are good in situations where feedback is sparse or noisy, since they learn quickly with little complexity, even though they don't have full-state modelling like RL.

4.5. Deep Reinforcement Learning (DRL) Systems

DRL combines deep neural networks with RL so that agents can learn directly from complex input spaces such as clickstreams, customer profiles, or behavioral embeddings [24]. DRL enables the development of complex, highly personalized decision systems for customer modeling that apply across customer groups and are updated in real time to accommodate new patterns.

DRL is distinguished by its use of neural networks as function approximators for value functions, policies, or both. Algorithms using neural networks to estimate the optimal strategy for performing an action in situations where standard techniques are too difficult to apply include Deep Q-Networks (DQN), DDPG, and PPO [4]. This function enables DRL models with high-dimensional features to be trained on extremely large customer data sets.

DRL not only helps in basic decision-making but also to learn representations which means the automatic extraction of hidden customer states from observed behavior. With the help of embedding layers, attention mechanisms, and variational encoders it is possible to reveal the most abstract properties like churn risk, engagement propensity, or desire [32]. These representations increase the generality and the capability of RL agents in CRM systems.

Overall, DRL supplies the backbone of cutting-edge customer analysis by RL. It promotes real-time adaptation, and long-horizon optimization, and supports customer-specific engagement strategies, all of which are critical for CLV enhancement.

5. Literature Review: Reinforcement Learning in Customer Modeling and Marketing

Reinforcement learning (RL) represents a robust framework for understanding consumer behavior, assisting in better marketing decisions, and maximizing long-term value. Unlike conventional predictive models, RL highlights the importance of continuous decision-making and consequently lets the marketers unfold different treatments, prices, and engagement strategies quite slowly over time. The discussion here is about the substantial intellectual and technical advancements that took place in dynamic pricing, interactive marketing environments, recommendation systems, and churn forecasting. Recent studies stress reinforcement learning as a main decision-making framework for flexible marketing settings, especially in personalization, bidding, and engagement optimization [33]. This section presents a taxonomy of RL applications in CLV optimization, categorized into seven key tasks based on the synthesis of the 16 studies from Table 1. These tasks highlight how RL addresses sequential marketing challenges, from churn prediction to direct CLV integration. To provide an overview, Table 2 aggregates the distribution of studies across subsections.

Table 2. Summary of RL Applications in CLV Optimization (Shift from basic to deep/hybrid RL)

Subsection	Task	Number of Studies	Percentage	Temporal Trends	Dominant RL Methods
5.1	RL in Churn Prediction	2	12.5%	2024-2025	Deep RL for forecasting
5.2	RL Applied to Individualized Recommendations	3	18.75%	2015-2025	Deep RL, contextual bandits
5.3	RL for Dynamic Pricing	4	25%	2017-2025	Q-learning, Thompson sampling
5.4	RL for Customer Engagement Plans	3	18.75%	2017-2025	Off-policy evaluation, RL-LLM hybrids
5.5	RL Tailored Promotions	2	12.5%	2023-2025	Batch deep RL
5.6	RL for Interactive and Conversational Marketing Systems	2	12.5%	2015-2017	Deep RL architectures
5.7	Research Connecting RL to CLV	2	12.5%	2023-2024	Integrative optimization frameworks

5.1. RL in Churn Prediction

In the context of churn reduction, RL has been utilized to discover the very best methods for retention as a sequential decision problem. The early research work of Neslin et al. confirmed that dynamic customer targeting with the most powerful interventions can indeed cut down on churn through the dynamic decision process [27]. Recently, the application of deep RL is to gain retention strategies that are based on long-term reward functions connected to customer lifetime value (CLV). For example, Yang and Calzada applied Deep Q-Networks (DQN) to find the right time and the way to support at-risk users, which resulted in a substantial increase over a static predictive churn model [28]. These techniques emphasize

RL's capacity to maximize tailored retention plans and reflect the changing character of churn risk. By matching service customization tactics with long-term lifetime value goals, RL-optimized incentive systems have been demonstrated to improve customer loyalty and retention [29]. According to Table 1, 2 out of 16 studies (12.5%) focus on churn prediction, both from 2024-2025, employing deep RL for dynamic forecasting and loyalty optimization [14, 26].These stats reveal a recent emphasis but highlight underrepresentation, as only 12.5% directly address churn despite its core role in CLV.

5.2. RL Applied to Individualized Recommendations

In the sphere of marketing, one of the first and most effective applications of RL can be found in recommendation systems. Using RL-based recommenders, platforms can quickly react to what a user inputs, thus eliminating static collaborative filtering. A study presented a deep RL method that learns best recommendation policies by analyzing user click patterns resulting in increased engagement on large-scale e-commerce datasets [30]. Another study also built a contextual bandit system for news suggestions that allows for quick personalization even with limited initial feedback [31]. All these researchers have shown that, by focusing on long-term engagement instead of short-term clicks, RL-based recommenders are able to beat the traditional ones. Table 1 indicates 3 out of 16 studies (18.75%) in this area, spanning 2015-2026, with a shift from ad-focused to product recommendation using deep RL [19, 20, 24]. This category ties for prominence, underscoring RL's maturity in recommendations for CLV alignment via off-policy guarantees [19].

5.3. RL for Dynamic Pricing

Decision-theoretic approaches have been supporting airlines with the dynamic pricing for a long time, but RL has now joined the game and allowed retailers to find the best price in a complex and unpredictable scenario. Ferreira et al. applying RL for retail price reduction pointed out the remarkable revenue that could be generated by the implementation of exploration-exploitation techniques [32]. Furthermore, Balakrishnan et al. employed actor-critic methods for airline ticket pricing and let RL cope with the changing market demand [33]. By learning optimal pricing routes that maximize total profit, these systems beat rule-based pricing. Research shows that deep RL–based customized pricing systems can dynamically adjust to consumer tastes, therefore boosting lifetime profitability while keeping customer satisfaction level [34]. Dynamic pricing models based on Q-learning show how RL can change prices based on how customer demand changes and feedback on their behavior changes [35].

Furthermore, recent research shows that by learning customer willingness to pay, deep reinforcement learning can simultaneously maximize product recommendations and customized pricing thereby maximizing long-term revenue and value rather than just temporary financial gains [24]. As per Table 1, 4 out of 16 studies (25%) target dynamic pricing, predominantly from 2017-2025, utilizing Q-learning and deep RL for retail and network revenue management, such as [4, 21-23]. These stats position pricing as the dominant application, with empirical evidence from e-commerce emphasizing CLV-aware strategies [4].

5.4. RL for Customer Engagement Plans

RL increases in importance to augment multichannel customer interaction. Theocharous et al. presented an RL-based approach for maximizing digital ad interactions by modelling customer journeys as Markov decision processes (MDPs) [11]. Their research showed that RL could find the best order of ads to maximize conversions. More recent work uses policy gradient methods to improve email marketing, push notifications, and omnichannel campaigns [42]. These techniques connect everyday marketing decisions with lifetime value by linking involvement activities with long-term customer results. Of 16 studies in Table 1, 3 (18.75%) addressing engagement, clustered in 2017-2025, incorporating RL-enhanced LLMs and simulation benchmarking, such as [25, 27, 28]. This high coverage reflects a trend toward scalable plans that integrate with CLV through personalized interventions, also supported by [28].

5.5. RL tailored promotions

Deciding which motivation to provide and when a typical RL issue known as promotional targeting is. RL-driven promotion tactics clearly balance providing discounts with maintaining margins, thereby making them financially appealing. Research repeatedly reveals that RL-based personalization lowers over-discounting and improves client response. According to Table 1, 2 out of 16 studies (12.5%) cover tailored promotions, from 2023-2025, leveraging deep RL for personalization and predictive analytics, reported by [3, 16]. The lower percentage indicates growth potential, as promotions directly link to CLV but remain task-specific in recent works [3].

5.6. RL for Interactive and Conversational Marketing Systems

Conversational agents, chatbots, and interactive decision systems have also been extensively researched in RL. Li et al. showed that RL can be used to train dialogue systems long-term conversational goals including user satisfaction and engagement. RL-powered conversational agents in marketing can change messages, suggest goods, and scale personalization depending on projected intent. More recent research use actor–critic techniques to enhance conversation flow in sales chatbots and maximize conversion likelihood [43]. Table 1 reveals 2 out of 16 studies (12.5%) in this domain, both from 2015-2017, focusing on off-policy evaluation and deep RL architectures for applications, real example for this case are [19, 30]. These early stats highlight a foundational but underexplored area for conversational CLV optimization [30].

The under-representation may be attributed to significant data privacy challenges, as interactive systems require access to sensitive real-time user data, raising GDPR-like compliance issues in marketing contexts. Additionally, technical hurdles such as handling high-dimensional conversation states and ensuring low-latency responses in production environments have limited recent advancements, presenting opportunities for future hybrid RL-LLM integrations to address these barriers.

5.7. Research Connecting RL to Customer Lifetime Value (CLV)

One of the main ideas in RL marketing publications is the search of long-term customer value instead of quick results. Many research clearly maximize RL reward functions to match CLV. Theocharous et al., for instance, included CLV proxies straight into the incentive system for advertising distribution [11]. Similarly, Abe et al. represented consumer behaviour as RL trajectories with the goal of maximizing long-term profitability and loyalty [14].

Reviewing the applications reveals that reinforcement learning (RL) techniques can be seen as means of maximizing customer lifetime value by impacting long-term behavior trajectories rather than just isolated events. While recommendation, pricing, engagement, and promotion policies define repeated interactions, loyalty, and overall profitability over time, RL-based churn prediction and retention strategies aim to lengthen customer relationships and boost projected future income. Interactive and conversational RL systems help to maintain customer satisfaction and trust, which are key drivers of lifetime value. But even with this great conceptual alignment, very few studies actually say that customer lifetime value is the main reason why the RL framework is used. Most of the work that's out there uses things like clicks, conversions, or engagement, which are more short-term or proxy metrics. Modern research shows that although RL is more and more utilized for marketing and customer involvement, its explicit application for customer lifetime value optimization is still underutilized [44]. Combining reinforcement learning with predictive analytics in hybrid frameworks has shown potential in enhancing customer lifetime value estimation and retention-focused decision-making [16]. Recent reviews point out how increasingly important artificial intelligence and reinforcement learning are in retail and e-commerce for maximizing personalization strategies and customer lifetime value [14].

6. Theoretical Connection Between Reinforcement Learning and Customer Lifetime Value (CLV)

Section 6 develops the conceptual framework (Figure 2) that integrates RL theory with CLV optimization. Due to the fact that reinforcement learning (RL) is sequential and long-term in its decision-making process, it naturally supports the methodological approach to maximizing Customer Lifetime Value (CLV). RL offers a framework for learning decision rules that maximize cumulative rewards across time, whereas CLV estimates the net present value of a customer's future revenue stream. RL's conceptual coherence makes it a good choice for modelling marketing plans that priorities long-term profitability over short-term results.

6.1. Long-Term Reward Optimization Problem with CLV

Given the sequence of future contacts, purchases, and involvement signals a customer might produce [2], CLV is naturally forward-looking. Classical forecasting models compute CLV but offer no guidance on the behaviors that affect it. Conversely, RL strives to maximize the cumulative reward ∑t=0T γtrt, which is directly related to discounted CLV. Research, including Abe et al., indicates that the discounted reward approach in RL aligns with the financial structure of CLV computations; hence, the RL conceptually fits with lifetime value modelling [45].

6.2. RL Rules to Maximize Lifetime Value

An RL agent develops a policy to choose actions, including discounts, price changes, or involvement, based on the customer's changing state. The agent uses exploration and exploitation to determine the order of marketing choices that yield the greatest long-term profit [11]. Unlike supervised learning, RL adjusts to client feedback and behavior trajectories rather than relying on fixed surroundings. This capacity enables RL to identify the best retention techniques, upsell timing, and engagement strategies that maximize CLV throughout the client's lifetime.

6.3. Designing Compensation Functions Reflecting CLV

Constructing reward functions that capture lifetime value is a crucial first step in relating RL to CLV. Reward functions might consist of, immediate income include sales or conversions, soft metrics of engagement include email opens and session length, and long-term strategic benefits include better loyalty scores or lower churn risk. To maximize both short-term actions and long-term contribution margins, some techniques include estimated future CLV directly into the reward signal, therefore empowering the RL agent to optimize. Researchers usually build composite reward functions weighted toward long-term impact [45].

6.4. State Correspondence for Customer Modeling

In reinforcement learning for customer modelling, the state at any particular moment comprises all data accessible on a customer. Good state representations often include preference-based signals, RFM measures, buying and browsing behavior, involvement patterns across digital channels, and demographic or environmental variables as well as purchase and browsing behavior. The major difficulty is striking a balance between richness and efficiency; states have to gather enough past and context to forecast future behavior without being computationally intensive. Recurrent networks and autoencoders are two examples of deep learning techniques that are becoming more and more popular for encoding these complicated customer paths. For long-term personalization and value-based decision-making, advanced state representations combining several customer signals have been shown to significantly improve RL policy learning [1].

6.5. Definition of Action Space for Marketing Decisions

The action space, which specifies every conceivable intervention the RL agent may select from, determines the range of tactics a marketer can use. The typical marketing activities for such situations include things like sending personalized messages, giving discounts or coupons, changing product suggestions, modifying the timing or number of emails, picking the best communication channel, or starting customer service interactions. An action space that has been designed well guarantees that the RL agent will be trained to adopt sensible and practically relevant policies. But an increase in the number of possible actions makes it significantly harder to learn. Large or high-dimensional action spaces can be an obstacle to effective policy learning; therefore, it is necessary to apply easily scalable RL methods, like actor–critic techniques or policy gradient algorithms, which are more appropriate for handling larger sets of decisions [42].

6.6. Problems Modeling the Customer Environment

Turning economic and stock market states into RL environments presents significant issues and limitations:

6.6.1. Partial Viewability

Partially observable Markov decision process (POMDP) structures arise from clients' typically invisible goals and offline activities.

6.6.2. Deferred Implications

Marketing campaigns might have long-lasting consequences, therefore making credit assignment more difficult [32].

6.6.3. Non-stationarity

Consumer tastes can change over time due to seasonality, competing events, or lifetime events.

6.6.4. *Ethical and Operational Constraints*

Doable activities are limited by personalization, marketing fatigue, and fairness problems.

Recent developments in deep RL, off-policy learning, and model-based RL have made customer-environment modeling more manageable, even in the face of such obstacles.

7. Conceptual Framework Connecting RL to CLV Optimization

This section offers a theoretical framework that combines reinforcement learning with customer lifetime value optimization by seeing customer contacts as a sequential decision-making process. It identifies important parts, like states, actions, policies, and rewards based on CLV, to show how reinforcement learning may be used for long-term value optimization rather than short-term indicators, with an iterative feedback loop directing policy revisions.

7.1. Markov Process for Customer Trajectory

This framework treats interactions with consumers as a Markov process, with each touchpoint denoting a state; marketing activities control the movement between these states. Such a depiction demonstrates the extent to which powerful measures influence the transformation of consumer behavior throughout time.

7.2. RL Policy Maximizing Discounted Lifetime Revenue

Through the selection of actions giving rise to the highest anticipated customer lifetime value (CLV) the reinforcement learning (RL) agent gets to an optimum policy that leads to maximization of lifetime revenues at present values. The financial reason for CLV calculations, with the RL purpose of maximizing total discounted rewards, is in line with this.

7.3. Reward Function Records Strategic and Financial Results

The goal of the incentive function is to evaluate the contributions of an agent in both the short term and long term. The agent is provided with information such as direct income, involvement indicators, retention signals, and future value estimates, which all help him/her learn to strike a balance between short-term benefits and long-term profitability strategies.

7.4. Customer States Representational Learning

In order to precisely represent the intricacy of consumer actions, deep learning techniques infer latent state embeddings from extensive behavioral pasts. These embeddings allow the RL agent to understand rapidly and efficiently the previous moves, the background knowledge, and the changing likes and dislikes of consumers.

7.5. Evaluation Using Off-Policy and Counterfactual Estimators

The evaluation of RL policies focused on CLV is based on off-policy and counterfactual estimation approaches such as inverse propensity scoring, which allows for trust-worthy assessment based on previously collected data. The researcher mentioned it [11] claimed that the models can give detailed predictions of the performance even without the rapid internet deployment.

The combination of these elements indicates that RL is not only a forecasting tool but also a guidance optimizer to raise customer lifetime value. By coupling the learning of policy, the representation of state and the design of reward with CLV concepts, RL has offered a strong theoretical and practical base for long-lasting marketing optimization.

8. Recommended Taxonomy and Conceptual Framework

A taxonomy is used as a basis for the arrangement of RL methods in CLV-oriented marketing. By categorizing methods according to both the specific marketing problem and the RL framework used [45],[2] , it becomes easier for scientists to decide which among the different strategies can be the best ones for particular conditions of CLV optimization. The proposed model has two views: one perspective is a task-based taxonomy that shows marketing objectives leading to CLV, and the other one a method-based taxonomy that classifies RL algorithms according to their computational characteristics [42], [18]. To assess

RL agents for individualized promotions, simulation-based benchmarking systems have been put out since marketing research lacks consistent evaluation standards [25].

8.1. Task-Based Classification

The task-based taxonomy classifies the applications of RL according to the primary CLV-related objective that their structure is designed to maximize. In many instances, by maximizing lifetime customer value the instead of focusing on quick conversions, the ways through which RL is utilized for customer retention have succeeded in finding the best actions that not only reduce churn but also maintain profitable relationships in the long run [45]. Personalized offer optimization makes use of RL to select the type of bonus, discount, or promotion that each customer would receive, thus making the trade-off between short-term response rates and long-term profitability and margin preservation. With the objective of maximizing revenue and profit over the customer lifecycle, dynamic pricing applications use RL to adapt prices in response to demand, client sensitivity, and competitive context. Recommendation strategies use RL to adapt product or content suggestions depending on changing behavior signals, therefore improving both engagement and CLV [42]. Ultimately, customer re-engagement activities use RL to decide when and how to contact dormant or inactive consumers (e.g., via email, app notifications, or retargeting), with the express aim of recovering or extending lifetime value. Together, these task categories relate operational marketing issues to CLV-centric sequential decision-making formulations.

8.2. Taxonomy Based on Method

Table 1 reveals task-based trends: 8/16 studies target pricing/personalization via value-based RL. The method-based taxonomy divides RL strategies according to their learning mechanisms and algorithmic design. Deep Q-Networks and Q-learning are examples of value-based RL techniques that estimate action–value functions and choose actions that maximize expected future rewards. These methods are good for situations where the action space is small or has a discrete number of choices, such as pricing or deciding how many offers to send out. Policy-based RL, on the other hand, directly optimizes a parameterized policy. This is especially helpful when the action space is big or has a continuous number of choices, which is often the case when setting prices or deciding how many offers to send out. Actor–critic methods integrate value estimation with policy optimization in complex, high-dimensional marketing environments to increase stability and scalability; they are especially useful for simple CLV proxies or early-stage testing since they address one-step decision problems whereby the goal is to choose the best action (e.g., a creative or offer) without simulating long-run state changes. Deep RL finally integrates neural networks with value-based, policy-based, or actor–critic frameworks to handle high-dimensional state spaces such as multimodal client characteristics, clickstreams, and rich behavioral histories [42]. This taxonomy helps to explain the trade-offs among different RL families with respect to complexity, data requirements, and applicability for CLV-driven marketing challenges.

8.3. Taxonomy Contribution

This classification system gives a uniform conceptual basis for the study of RL in CLV optimization by simultaneously evaluating the marketing action being optimized from a task-based perspective and how it is treated algorithmically from a method-based approach. It places the present studies in a cohesive context, uncovers the combinations of activities and methods that are less utilized (like for instance, deep actor–critic methods for retention or re-engagement), and provides a systematic framework for future empirical research [45], [11]. Thus, the taxonomy not only raises the academic quality of the review but also aids in the justification of its acceptance into mid- to high-tier journals through the presentation of a clear and generalizable conceptual contribution at the intersection of RL and CLV that is properly delineated.

9. Challenges and Limitations

While reinforcement learning (RL) has great promise for maximizing consumer lifetime value, several key issues now impede its application in the real world. One major obstacle is the dearth of genuine customer-level sequential data; firms avoid dangerous internet testing, therefore forcing academics to depend on synthetic simulations that oversimplify actual behavior. Furthermore, RL is very sampleinefficient; it requires many trials to learn good policies, but too much experimentation might damage customer experience and sales.

Marketing increases even more complexity with vast, dynamic state spaces and large action spaces spanning many channels, offers, and timing decisions, therefore challenging RL models to learn reliable, high-quality policies. Another problem is that CLV relies on long-term impacts that are difficult to represent in a single reward signal, therefore designing rewards is another difficulty. Ultimately, consumer behavior is non-stationery and changes with promotions, seasonality, and outside events, therefore unless they are regularly updated, learned RL policies can rapidly become obsolete.

Ethical and legal issues, including privacy, fairness, transparency, and the danger of over-personalization, present another very important restriction. Using RL systems with highly detailed behavioral data raises questions about GDPR compliance, customer privacy, and algorithmic bias. Further limiting the scope of acceptable research in RL-driven marketing systems raises ethical issues.

Training and adjusting RL models adds yet another obstacle in the form of calculating expense. Deep RL techniques require significant computational power, take a long time to learn, and need frequent monitoring to ensure the policy remains stable and responsive. Many companies either don't have the tools or know-how to use real-time RL systems at a high level.

Ultimately, judging RL policies in real-world settings is difficult, since running internet A/B tests is not always easy. Because of customer experience restrictions, conducting regulated experiments for sequential decision policies is operationally taxing, financially dangerous, and sometimes impossible. Off-policy evaluation techniques have problems including logging bias and model assumptions even while they do exist. These restrictions essentially demonstrate that, notwithstanding RL's great theoretical promise for CLV optimization, there are still very significant practical obstacles. Understanding these limits helps the study to be more effective and highlights the need of ongoing methodological and empirical developments before RL can be broadly used in practical marketing.

Specifically, offline evaluation often suffers from support mismatch, where the distribution of actions in the logged data does not fully cover the policy space being evaluated, leading to biased estimates and unreliable extrapolations[29]. Confounding factors, such as unobserved variables that influence both marketing actions and customer outcomes, further complicate causal inference in off-policy settings, potentially inflating or deflating perceived policy performance. Moreover, high variance in off-policy estimators can undermine the credibility of reported gains, especially in logged marketing data where historical exploration is limited and non-random [27]. For instance, in RL-enhanced A/B testing frameworks, techniques must explicitly address these issues, as demonstrated in recent approaches that integrate reinforcement learning with large language models for personalized marketing [28].

10. Research Gaps in RL for CLV Optimization

Reinforcement learning (RL) is becoming a favorite and more and more widely used technique in the marketing analytics area, nevertheless, the literature points out the presence of several major gaps that hinder the technique's application in the area of customer lifetime value (CLV) maximization. First, there are very few direct RL–CLV research (4 studies Table 1). Although quite few investigations clearly seek to maximize lifetime value as the main aim [45], most current research uses RL to linked tasks including pricing, recommendations, or engagement [42]. Theoretical progress is limited by this disconnect, and professionals' capacity to assess RL as a whole CLV optimization instrument is constrained. A second important problem is that there aren't any standard RL environments for modeling customers. Unlike other RL fields, including robotics or arcade learning scenarios, marketing lacks commonly recognized simulation platforms or benchmark datasets that let researchers test and contrast RL algorithms under uniform settings [18]. Consequently, research depend on either custom simulations or commercial data, therefore restricting reproducibility and comparability. Furthermore, lacking are strong reward-shaping techniques for CLV. Many analyses depend on simplified revenue or engagement indicators rather than thorough depictions of long-term viability. In both theoretical and practical research, designing reward functions that completely reflect value creation, including retention, loyalty, margin, and behavioral dynamics, remains a difficult problem.

The scant few real-world application case studies is yet another significant knowledge vacuum. Although there are more and more theoretical contributions, few businesses openly disclose production-level RL implementations for CLV or show empirical evidence of performance in real-world settings [11]. Field research deficiency slows knowledge of organizational readiness, long-term results, and real limitations. The literature likewise points to a dearth of standards for comparison. It is difficult to determine whether fresh RL methods beat current heuristics or conventional machine learning techniques without consistent criteria, evaluation protocols, or baseline models. The lack of standards limits the rigor of the research and slows down the advancement of universal knowledge.

Furthermore, despite their success in other sequential decision-making fields, few hybrid RL + deep learning systems are especially focused on CLV optimization. Although deep RL has been employed in recommendation systems [30], it has not yet been combined with explicit CLV modeling. This distinction presents opportunities for integrating representation learning and policy optimization to control complex consumer paths and rich state spaces. Recent studies combining reinforcement learning with large language models show how RL-driven experimentation frameworks may improve long-term marketing outcomes beyond conventional A/B testing [28].

For marketing-focused reinforcement learning, research on transparency and interpretability is still somewhat limited. Managing trust, regulatory compliance, and justice depends on understanding why specific behaviors are recommended as RL policies become increasingly data-driven and automated. Yet, there are few ways to clarify how RL chooses in situations where consumers engage with it, so it is a good starting point for fresh concepts.

These differences reveal that even though the concept of RL is similar to that of CLV optimization, the area of RL is still quite underdeveloped. To make RL a more trustworthy and scalable customer value management tactic, the mentioned drawbacks should be eliminated via interpretability tools, practical applications, standardized environments, and more integration of CLV-relevant objectives.

11. Directions for Additional Investigation

Since RL is more often used in marketing, a number of new research fields could assist better CLV optimization. Offline RL, which derives policies from past customer data without dangerous live testing, is one interesting field. For RL to be useful for actual firms, improvements in offline evaluation, counterfactual analysis, and safety-aware policy modification are required.

Another key route is causal RL, which separates actual intervention effects from normal client activity by fusing causal inference with sequential decision-making. Particularly for CLV, where it is challenging to assess long-term causal effects, this is quite helpful.

Ultimately, hybrid deep learning–RL systems have great promise: deep models can record intricate customer histories, whereas RL maximizes long-term value. Combining the two can produce more accurate policies for multi-step customer journeys across several data sources and richer state representations.

Modeling client ecosystems with multi-agent RL could also help future studies. Customers engage not only with the business but also with one another in many real-world situations via reviews, referrals, social influence, and network effects. Multi-agent RL could identify these relationships and maximize plans in settings where social or behavioral networks transmit consumer behavior.

One more front is hierarchical RL for creating long-term marketing plans. By breaking down choices into high-level strategic aims (e.g., retention vs. upselling) and low-level operational actions (e.g., promotion timing), hierarchical RL enables RL to scale to longer time horizons and more structured decision-making [18]. Such models may resemble management planning activities and produce more easily understood CLV optimization policies.

Furthermore, the demand for understandable RL is increasing. Managers need to understand why the algorithm recommends particular courses of action, as RL systems become entangled in high-stakes marketing decisions, including credit-based promotions, pricing, and retention interventions. Future RL studies on CLV need to cover explainability, scalability, and privacy. In RL-driven marketing systems, transparency, fairness, and organizational confidence depend on explainability. Large action spaces across several channels present a continuing obstacle since real-world marketing calls for breakthroughs in distributed RL, scalable policy gradients, and action-space reduction as well as in scalability. Another important frontier is privacy: combining RL with federated learning, differential privacy, and secure computation will help strike a balance between the need to make good decisions over the long term and the need to protect data. These research lines together seek to produce transparent, scalable, privacy-safe, and ready-for-real-world implementation RL-CLV systems.

Furthermore, assessing reinforcement learning–based CLV systems provides ongoing difficulties in terms of reproducibility, offline evaluation, and benchmarking. It's challenging to compare RL techniques across studies or repeat outcomes in the absence of consistent consumer environments and openly accessible data. Off-policy evaluation techniques provide some answers, but their usefulness is very dependent on the quality of the data and how well it is logged, hence restricting the applicability of empirical results. These evaluation restrictions make determining RL usefulness in practical CLV maximizing scenarios even more difficult.

12. Conclusion

Due to the fact that RL encourages sequential decision-making, adapts to changing customer behavior, and maximizes long-term results, it offers a solid theoretical basis for maximizing client lifetime value (CLV). Research indicates RL's promise in marketing activities including recommendations, pricing, retention, and involvement.

However, there are several obstacles to using deep RL in the real world: there are no standard ways to test how well it works, there is a lack of data from real customers, it's hard to make sure the rewards are fair and correct over a long time, and there are also problems with how well it works in general, like the fact that it can do things that aren't fair, that it can make it hard to understand what it's doing, and that it can use a lot of computer power and waste a lot of samples. Ethical concerns, like fairness, privacy, and being able to understand what it's doing, make things even more complicated.

The field has a lot of possibilities even if there are problems. For business implementation, privacy-preserving computing and scalable policy learning developments will be especially crucial as emerging directions like offline RL, causal RL, hybrid deep learning–RL models, hierarchical and multi-agent systems, and explainable RL can lead to more robust, scalable, and transparent CLV optimization.

Taken together, this review places reinforcement learning not only as an advanced analytical tool but also as a fundamental framework for developing empirically grounded, scalable, ethically responsible, and aligned with long-term strategic decision-making systems for maximizing future customer lifetime value.

While RL holds transformative potential for CLV optimization, it raises ethical concerns like algorithmic bias, where skewed datasets may prioritize high-value customers from digitally advanced markets, marginalizing underrepresented groups. Future work should adopt fairness-aware techniques such as debiased rewards or data augmentation for inclusive systems. Practically, scalability challenges SMEs in resource-limited settings, as deep RL architectures (dominant in 75% of recent Table 1 studies (for instance: 2024-2025 works [3, 4, 28]) require expensive GPUs and vast datasets amid high energy costs or unstable clouds [30], recommending lightweight options like offline batch RL for low-data scenarios [6].

In the end, using RL with CLV modeling is a good way to go in the future of AI-driven marketing because it lets systems learn, adapt, and improve customer relationships over the long term, which has a big financial impact.

Funding Information

The authors state no funding is involved.

Conflict of Interest State

The authors state no conflict of interest.

Ethical Approval

This paper does not involve people or animals; no investigation has involved human subjects. Therefore, the authors did not seek approval from any institutional review board.

References

1. P. S. Fader, B. G. S. Hardie, and K. L. Lee, "RFM and CLV: Using iso-value curves for customer base analysis," Journal of Marketing Research, vol. 42, no. 4, pp. 415-430, 2005.
2. S. Gupta, D. R. Lehmann, and J. A. Stuart, "Valuing Customers," Journal of Marketing Research, vol. 41, no. 1, pp. 7-18, 2006/02/01 2006, doi: 10.1509/jmkr.41.1.7.12384.
3. Ma, L., Huang, T. W., Ascarza, E., & Israeli, A. (2025). Dynamic personalization with multiple customer signals: Multi-response state representation in reinforcement learning. Available at SSRN 5126129.
4. M. Apte, P. Datar, K. Kale, and P. R. Deshmukh, "Dynamic Retail Pricing via Q-Learning - A Reinforcement Learning Framework for Enhanced Revenue Management," in 2025 1st International Conference on AIML-Applications for Engineering & Technology (ICAET), 16-17 Jan. 2025 2025, pp. 1-5, doi: 10.1109/ICAET63349.2025.10932302.
5. Tkachenko, Y. (2015). Autonomous CRM control via CLV approximation with deep reinforcement learning in discrete and continuous action space. arXiv preprint arXiv:1504.01840.
6. X. Liu, "Dynamic Coupon Targeting Using Batch Deep Reinforcement Learning: An Application to Livestream Shopping," Marketing Science, vol. 42, 10/20 2022, doi: 10.1287/mksc.2022.1403.
7. W. Wang, B. Li, X. Luo, and X. Wang, "Deep Reinforcement Learning for Sequential Targeting," Management Science, vol. 69, 12/21 2022, doi: 10.1287/mnsc.2022.4621.
8. E. Jaakkola, "Designing conceptual articles: four approaches," AMS Review, vol. 10, no. 1, pp. 18-26, 2020/06/01 2020, doi: 10.1007/s13162-020-00161-0.
9. P. S. Fader and B. G. S. Hardie, "Probability models for customer-base analysis," Journal of Interactive Marketing, vol. 23, no. 1, pp. 61-69, 2009.
10. R. Venkatesan and V. Kumar, "A customer lifetime value framework," Journal of Marketing, vol. 68, no. 4, pp. 106-125, 2004/10/01 2004.
11. V. Kumar and M. George, "Measuring and maximizing customer equity: A critical analysis," Journal of the Academy of Marketing Science, vol. 35, no. 2, pp. 157-171, 2007.
12. A. M. Hughes, Strategic Database Marketing. New York, NY, USA: McGraw-Hill, 1994.
13. D. C. Schmittlein, D. G. Morrison, and R. Colombo, "Counting your customers," Management Science, vol. 33, no. 1, pp. 1-24, 1987.
14. J. Thiruvayipati, "Revolutionizing Customer Lifetime Value: A Comprehensive Review of AI and Machine Learning in Retail and E-commerce," 2024/05/06 2024, doi: 10.2139/ssrn.5339996.
15. O. Timilehin, "Dynamic Customer Lifetime Value Forecasting Models Using Reinforcement Learning," 01/16 2025.
16. N. Bose, A. Chopra, P. Joshi, and A. Reddy, "Leveraging Reinforcement Learning and Predictive Analytics for Enhanced Customer Lifetime Value Optimization," 2024.
17. S. Ataei, P. Nikzat, G. Alikaram, and S. Ataei, "Deep Reinforcement Learning for Dynamic Pricing Strategies: Empirical Evidence from E-Commerce Platforms," International Journal of Science and Engineering Applications, vol. 14, pp. 20-28, 10/01 2025, doi: 10.7753/IJSEA1411.1006.
18. R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.
19. G. Theocharous, P. S. Thomas, and M. Ghavamzadeh, "Personalized ad recommendation systems for life-time value optimization with guarantees," presented at the Proceedings of the 24th International Conference on Artificial Intelligence, 2015.
20. X. Zhao, L. Zhang, Z. Ding, D. Yin, Y. Zhao, and J. Tang, "Deep Reinforcement Learning for List-wise Recommendations," 12/30 2017, doi: 10.48550/arXiv.1801.00209.
21. K. Ferreira, D. Simchi-levi, and H. Wang, "Online Network Revenue Management Using Thompson Sampling," SSRN Electronic Journal, 01/01 2015, doi: 10.2139/ssrn.2588730.
22. W.-C. Cheung, D. Simchi-levi, and H. Wang, "Dynamic Pricing and Demand Learning with Limited Price Experimentation," SSRN Electronic Journal, 01/01 2017, doi: 10.2139/ssrn.2457296.
23. Brown, D. (2023). Applying Deep Reinforcement Learning for Personalized Dynamic Pricing in Online Retail. Journal of Marketing Research, 60 (4), 771–791.
24. A. Mahdavian, H. Moradi, and B. Bahrak, "Product Recommendation with Price Personalization According to Customer's Willingness to Pay Using Deep Reinforcement Learning," Algorithms, vol. 18, no. 11, p. 706doi: 10.3390/a18110706.

25. Xia, Y., Narayanamoorthy, S., Zhou, Z., & Mabry, J. (2024). Simulation-based benchmarking of reinforcement learning agents for personalized retail promotions. arXiv preprint arXiv:2405.10469.
26. X. Nie and F. Sh, "Dynamic reward systems and customer loyalty: reinforcement learning-optimized personalized service strategies," Future Technology, vol. 4, pp. 259-268, 08/15 2025, doi: 10.55670/fpll.futech.4.3.24.
27. A. Swaminathan et al., "Off-policy evaluation for slate recommendation," presented at the Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017.
28. Feng, H., Dai, Y., & Gao, Y. (2025, May). A Reinforcement-Learning-Enhanced LLM Framework for Automated A/B Testing in Personalized Marketing. In Proceedings of the 2025 2nd International Conference on Digital Society and Artificial Intelligence (pp. 498-502).
29. P. S. Thomas, G. Theocharous, and M. Ghavamzadeh, "High confidence off-policy evaluation," presented at the Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.
30. H. Li, T. Wei, A. Ren, Q. Zhu, and Y. Wang, "Deep reinforcement learning: framework, applications, and embedded implementations," presented at the Proceedings of the 36th International Conference on Computer-Aided Design, 2017.
31. K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep Reinforcement Learning: A Brief Survey," IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 26-38, 2017/11/01 2017, doi: 10.1109/MSP.2017.2743240.
32. M. Chen, Y. Zhang, Y. Yu, H. Li, and Q. Zhang, "Learning customer representation for recommendation with sequential reinforcement learning," ACM Transactions on Information Systems, vol. 39, no. 2, pp. 1-26, 2021/04/01 2021, doi: 10.1145/3441302.
33. T. S. Saranya, S. K. Gupta, P. S. GayathriRaj, and R. S. John, "AI-driven marketing strategies: Understanding and predicting consumer behaviour," 2025, p. 324.
34. S. A. Neslin et al., "Challenges and opportunities in multichannel customer management," Journal of Service Research, vol. 9, no. 2, pp. 95-112, 2006.
35. H. Yang and J. Calzada, "Customer churn management via deep reinforcement learning," Expert Systems with Applications, vol. 159, p. 113574, 2020.
36. Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, and N. De Freitas, "Dueling network architectures for deep reinforcement learning," 2016, pp. 1995-2003.
37. R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," Machine Learning, vol. 8, no. 3-4, pp. 229-256, 1992.
38. M. Chen, A. Beutel, P. Covington, S. Jain, and E. H. Chi, "Top-K contextual bandits for personalized recommendations," 2019, pp. 347-355.
39. T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," 2015, doi: arXiv:1509.02971.
40. A. Balakrishnan, D. Simchi-Levi, and H. Wang, "Dynamic pricing with demand learning using actor-critic methods," Management Science, vol. 67, no. 6, pp. 3815-3835, 2021/06/01 2021.
41. P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," Journal of Machine Learning Research, vol. 3, pp. 397-422, 2002.
42. M. M. Afsar, T. Crump, and B. H. Far, "Reinforcement learning-based recommender systems: A survey," ACM Computing Surveys, vol. 54, no. 7, pp. 1-38, 2021.
43. J. Gao, M. Galley, and L. Li, "Neural approaches to conversational AI," Foundations and Trends in Information Retrieval, vol. 13, no. 2-3, pp. 127-298, 2019.
44. J. Ghazimatin, R. Saha Roy, and G. Weikum, "Reinforcement learning in recommender systems: A survey," ACM Computing Surveys, vol. 53, no. 5, 2020.
45. M. Abe, Y. Gunji, and S. Arai, "Customer lifetime value optimization using reinforcement learning," Journal of Marketing Analytics, vol. 8, no. 4, pp. 189-203, 2020.