Preference-Controllable Multi-Objective Deep Reinforcement Learning for Human-Robot Task Allocation in Service Environments

Asmaa Rashed Alahmari*, Wadee Alhalabi
Computer Science Department-Faculty of Computing & Information Technology,
King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—Human–Robot Collaboration (HRC) has gained increasing attention as it expands from industrial environments to service-oriented settings, where dynamic conditions and diverse operational objectives pose significant challenges for task allocation. Unlike controlled industrial environments, service contexts are characterized by frequent changes, uncertainty, and time-varying priorities, rendering static task allocation strategies ineffective. This paper proposes a method to address the problem of determining the optimal balance between human and robotic task allocation in dynamic service-oriented HRC systems. A preference-controllable multi-objective deep reinforcement learning framework is introduced to formulate task allocation as a dynamic, preference-dependent decision-making process. The proposed approach explicitly captures trade-offs among multiple, potentially conflicting objectives and enables adaptive task allocation under changing operational conditions and service priorities. The framework is evaluated through simulation-based experiments and comparative analysis with baseline strategies using multiple evaluation metrics, complemented by additional validation using external datasets. Experimental results demonstrate the effectiveness and adaptability of the proposed approach across varying preference configurations and workload conditions, supporting its applicability in real-world smart service environments.**

***Keywords*—Deep reinforcement learning; human–robot collaboration; preference-controllable reinforcement learning; smart service environments; task allocation**

I. Introduction

Human–Robot Collaboration (HRC) has demonstrated significant success in industrial environments, where humans and robots work collaboratively to enhance productivity, efficiency, and safety [1–3]. Building on this success, HRC has recently expanded beyond manufacturing into service-oriented sectors such as hospitality and tourism [4–12]. An increasing body of research has demonstrated the feasibility and benefits of deploying collaborative robots in service environments, reporting improvements in operational efficiency and service quality [4,6,11,12].

Among the core challenges in HRC systems, task allocation between human workers and robots has emerged as a critical research problem [13,14]. Determining which tasks should be performed by humans, robots, or jointly has been extensively investigated in industrial settings, both theoretically and experimentally [13–16].

Existing studies have proposed a wide range of task allocation approaches [13–21], including rule-based strategies, optimized heuristic methods, and more advanced solutions such as digital twins and machine learning–based models. More recently, deep reinforcement learning (DRL) has attracted growing attention as a promising approach for addressing complex and dynamic task allocation problems [15,22].

These approaches aim to optimize multiple objectives, such as efficiency, cost, performance, safety, and human workload [1,23–25]. While some studies focus on optimizing a single objective, others consider multiple objectives simultaneously [26–28]. Importantly, these objectives may be complementary in some cases, yet conflicting in others, leading to inherent trade-offs commonly described through Pareto optimality [29,30]. Moreover, task allocation strategies may be static or dynamic, depending on how objectives are defined and prioritized [31,32].

While task allocation in industrial environments benefits from relatively controlled and predictable settings, service environments present fundamentally different challenges [10]. Service systems are highly dynamic, characterized by frequent and unpredictable changes, diverse and often conflicting objectives, and time-varying priorities. For instance, speed and efficiency may be critical during peak hours, whereas customer experience and human touch become more important during off-peak periods. Unexpected events such as staff absence, equipment maintenance, or sudden demand fluctuations further complicate decision-making [5,8].

Despite the importance and complexity of task allocation in service-oriented HRC, existing research remains limited and lacks adaptive frameworks capable of dynamically balancing multiple conflicting objectives under changing preferences in real-world service settings [8,10].

To address these challenges, this paper proposes a preference-controllable multi-objective deep reinforcement learning approach for human–robot task allocation in service environments. Inspired by task allocation strategies developed for industrial HRC, the proposed framework enables dynamic adaptation to changing environmental conditions and shifting objective priorities. By explicitly modeling trade-offs among multiple objectives and allowing preference control, the framework aims to achieve a balanced allocation that improves efficiency, cost, and overall system performance in service-oriented contexts.

*Corresponding author.

The main contributions of this paper are as follows:

- Proposing a preference-controllable multi-objective deep reinforcement learning approach for dynamic human–robot task allocation in service environments.
- Modeling task allocation as a preference-dependent decision-making problem that captures trade-offs among conflicting objectives.
- Demonstrating adaptive and controllable allocation behavior under varying operational conditions and preference configurations.
- Validating the proposed approach using both simulation-based experiments and external real-world datasets.

This work contributes to determining the optimal balance between human and robotic task allocation in dynamic service-oriented HRC systems. Rather than assuming a fixed allocation, optimality is defined as a dynamic and preference-dependent trade-off among multiple objectives, enabling adaptive task allocation decisions under changing operational conditions and service priorities.

The remainder of this paper is organized as follows. Section II reviews the background and related work. Section III describes the proposed methodology. Section IV presents and discusses the experimental results. Section V outlines the limitations and directions for future research. Finally, Section VI concludes the paper.

II. Background and Related Work

This section presents the foundational concepts of the research and reviews existing studies on human–robot task allocation, multi-objective optimization, and preference-aware reinforcement learning, with a focus on identifying limitations in current approaches.

A. Task Allocation in Human–Robot Collaboration

Human–Robot Collaboration (HRC) refers to systems in which humans and robots work together within a shared workspace [13]. Research began in the early 1990s, initially focusing on human-compatible robotic hardware [18,33], and later expanded to include control modalities, interaction, and task representation [18], with comprehensive reviews provided in [34].

With Industry 4.0, HRC has rapidly grown in industrial applications, supported by safety standards such as ISO/TS 15066 and advances in machine learning for improved adaptability [18,35]. More recently, HRC has extended to service environments such as hospitality and retail [4–12], which are less structured, more dynamic, and strongly influenced by human behavior and demand, introducing additional challenges in coordination and decision-making [5,8].

Task allocation plays a central role in enabling effective HRC in both industrial and service environments, determining how responsibilities are shared between humans and robots while balancing objectives such as time, cost, safety, and workload [32]. Although widely studied in multi-robot systems [36], HRC introduces additional complexity due to human variability, making optimal allocation dynamic and context-dependent.

Early approaches focused on system-centric objectives such as makespan, cycle time, and cost [37–39], using methods including mixed-integer linear programming (MILP), meta-heuristics, and reinforcement learning [37–41]. More recent work has incorporated human-centered factors such as workload, ergonomics, and adaptability [42–47], with fatigue modeling approaches such as Learning-Forgetting-Fatigue-Recovery LFFR used to capture human performance dynamics [18,48–50].

Despite these advances, existing studies still lack worker personalization and balanced workload distribution, highlighting the importance of multi-objective optimization to balance human and system-level objectives [23,28]. Task allocation strategies can be categorized as static or dynamic [31,32]. Static approaches rely on predefined conditions and are sensitive to disruptions, while dynamic approaches adapt to real-time changes, making them more suitable for service environments such as cafés, where demand is variable and time-sensitive [32].

Among available techniques for enabling dynamic task allocation under uncertainty, reinforcement learning has emerged as a promising approach [32], including integration with digital twins and hybrid learning frameworks [20,51]. However, most existing work remains focused on industrial settings and does not address preference-controllable task allocation in dynamic service environments.

B. Multi-Objective Task Allocation and Pareto-Based Approaches

Multi-objective optimization is increasingly relevant in collaborative systems, as many real-world decision-making problems cannot be adequately represented by a single performance criterion. Prior work emphasizes that many real-world decision problems are characterized by multiple conflicting objectives that must be balanced based on their relative importance, and that most real-life problems are more naturally expressed with multiple objectives [29, 52]. In HRC, traditional approaches focus on system-centric metrics such as cycle time [28], which may negatively impact human factors such as fatigue and well-being.

Multi-objective task allocation addresses this by modeling trade-offs among competing objectives. Rather than a single optimal solution, a set of Pareto-optimal solutions is obtained, representing different trade-offs [28,29]. These solutions reflect varying priorities, where improving one objective may degrade another. Within HRC, Pareto-based approaches have been used to balance system efficiency and human factors, demonstrating that task allocation should be treated as a trade-off problem rather than a single-objective optimization [28].

Multi-objective reinforcement learning (MORL) extends this concept by learning policies that capture different trade-offs among objectives [29]. Approaches include single-policy methods, multi-policy methods, and preference-conditioned methods. While effective, these methods are often computationally expensive and not tailored to domain-specific

applications such as service-oriented HRC. Overall, although multi-objective approaches provide strong theoretical foundations, they are often abstract and lack direct application to dynamic, service-oriented human–robot task allocation problems.

C. Preference-Aware and Preference-Controllable Learning

In multi-objective problems, preferences are typically represented as weights indicating the importance of each objective [54]. Traditional approaches assume fixed weights, allowing reduction to single-objective formulations [52]. However, in practice, preferences are often unknown or change over time [30,52]. Preference-controllable reinforcement learning addresses this limitation by enabling policies to adapt to varying preference inputs [30]. Rather than learning separate policies, a single model is conditioned on preference vectors, allowing different trade-offs to be achieved dynamically [30,54].

In HRC systems, preferences are particularly important as they reflect human priorities and expectations. Prior studies have explored learning preferences through interaction and feedback [53], but these are generally not formulated as explicit task allocation problems and are not designed for service-oriented environments. Overall, existing approaches provide strong theoretical support for modeling dynamic preferences and adaptive behavior in multi-objective systems. However, these methods are not directly applied to task allocation in dynamic service-oriented HRC systems, where preferences may shift rapidly in response to workload, customer demand, or operational conditions.

D. Limitations of Existing Approaches

Despite progress, several limitations remain. First, most task allocation frameworks are designed for industrial environments and are not easily transferable to dynamic service settings [23,28]. Second, many approaches rely on fixed or implicit preference assumptions, limiting adaptability in real-world scenarios [52,54]. Third, there is a lack of unified frameworks that integrate task allocation, multi-objective optimization, and preference control into a single system.

Overall, existing studies address these aspects in isolation, such as industrial task scheduling, Pareto-based optimization, or preference-aware learning, but do not provide an integrated, preference-controllable framework for dynamic task allocation in service-oriented HRC systems, which motivates the proposed approach.

III. Methodology and Experimental Setup

This study proposes a formulation of human–robot task allocation as a preference-controllable multi-objective decision-making problem, where service priorities may vary dynamically across operational contexts. Rather than optimizing a single fixed objective or training multiple separate policies for different trade-offs, the proposed approach follows the principles of Preference-Controllable Reinforcement Learning (PreCo) [30]. In this framework, task allocation decisions are conditioned on a continuous preference vector that encodes the relative importance of operational efficiency, preservation of human-centric interactions, and effective resource utilization. By conditioning the policy on preferences, the agent can represent a continuum of task allocation strategies within a single model, enabling flexible adaptation to changing service priorities. Fig. 1 illustrates the overall preference-controllable learning framework adopted in this study, showing how user-defined trade-offs and estimated objective outcomes guide the agent's interaction with the environment.

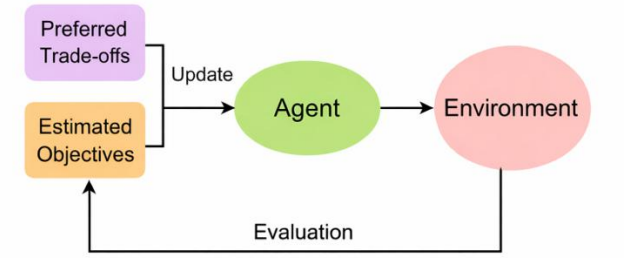

Fig. 1. Overview of the preference-controllable reinforcement learning (PreCo) framework adopted in this study.

A. Problem Formulation and Environmental Design

The task allocation problem is defined within a dynamic service-oriented HRC café environment, as a representative a real-world service scenario, where a human barista and a robotic agent jointly execute café service tasks under uncertain demand and variable execution conditions. At each decision step, the agent observes the current environment state and assigns the incoming order either to the human or to the robot, enabling adaptive allocation strategies as operational conditions evolve.

A custom simulation environment, CafeEnvironment, is developed using the Gymnasium interface [55]. Each episode represents a short operational window of 50 customer orders sampled from a predefined dataset comprising 20 heterogeneous order types (e.g., espresso-based drinks, specialty beverages, desserts, and packaged items). Each order is characterized by estimated human service time, robot service time, an automatability score $[0,1]$, and a human-touch importance score $[0,1]$. For food items, the estimated service times are modeled to be comprehensive, accounting for potential heating and custom preparation rather than mere handover. This task abstraction aligns with PreCo's vector-valued outcome formulation, where actions contribute differently to multiple objectives [30]. The dataset was constructed and validated through interviews with three café-domain experts (Table I).

B. Multi-Objective Learning Formulation

Task allocation performance is evaluated with respect to three potentially conflicting objectives: operational efficiency, preservation of human-centric interactions, and resource utilization, which represent core operational priorities in service-oriented café environments. These objectives are modeled using a vector-valued reward rather than a single scalar signal, explicitly capturing inherent trade-offs. Optimality is therefore defined relative to a preference vector specifying the relative importance of each objective, framing the problem as a preference-dependent multi-objective decision-making task. The policy is modeled as a conditional policy,

$$\pi(a \mid s, p),$$

TABLE I. ORDER TYPES AND TASK CHARACTERISTICS USED IN CAFEENVIRONMENT

Order Type	Human Service Time (s)	Robot Service Time (s)	Automatability	Human-Touch Importance
Espresso (single shot)	30	15	0.95	0.05
Double espresso / Lungo / Ristretto	30	16	0.92	0.08
Americano / Long Black	40	20	0.88	0.12
Cappuccino	60	35	0.80	0.20
Latte / Flat White / Macchiato / Cortado	60	36	0.78	0.22
Mocha / Hot Chocolate / Specialty hot drinks	90	45	0.72	0.30
Iced Coffee / Iced Latte / Iced Tea	90	50	0.70	0.28
Blended / Frappé beverages	90	55	0.60	0.35
Cold Brew / Nitro Cold Brew	30	30	0.40	0.70
Tea (bagged)	30	15	0.85	0.15
Tea (brewed, e.g., matcha latte)	60	30	0.68	0.32
Affogato / Dessert-coffee combinations	90	50	0.50	0.85
Muffin	180	120	0.30	0.55
Biscotti	180	120	0.30	0.45
Bagel	180	120	0.35	0.50
Croissant / Danish / Pastry	180	120	0.35	0.60
Cookie / Brownie / Cake slice	180	120	0.30	0.45
Sandwich	180	120	0.35	0.55
Wrap / Salad bowl	180	120	0.35	0.50
Yogurt parfait / Fruit cup	180	120	0.35	0.50

where s denotes the environment state, a the allocation action, and p the preference vector. Conditioning the policy on preferences follows the preference-conditioned policy formulation adopted in [30] and allows a single model to represent a continuum of allocation strategies without retraining.

C. State, Action, and Preference Modeling

At each decision step t, the action space is discrete with two possible actions:

$$a_t \in \{0,1\},$$

where $a_t = 0$ assigns the incoming order to a human barista and $a_t = 1$ assigns it to the robotic system.

The environment state is represented as a 10-dimensional continuous vector normalized to $[0,1]$, combining operational context, task characteristics, system utilization, and preference information:

$$s_t = [w_t, o_t, h_t, m_t, u_H, u_R, p_{\text{eff}}, p_{\text{touch}}, p_{\text{util}}, c],$$

Here, w_t denotes the remaining workload in the episode expressed as the fraction of unprocessed orders; o_t represents the normalized categorical index of the current order type; h_t and m_t correspond to the human-touch importance and automatability scores of the task, respectively. The variables u_H and u_R denote the current utilization ratios of the human worker and the robotic system. The preference vector $p = (p_{\text{eff}}, p_{\text{touch}}, p_{\text{util}})$ encodes the relative importance assigned to operational efficiency, preservation of human-centric interactions, and resource utilization. Finally, c represents the operational load context, discretized into low, medium, and high demand levels. Including preferences directly in the observation space follows the core principle of preference-conditioned reinforcement learning adopted in PreCo [30], enabling the policy to adapt its behavior to different trade-off configurations without modifying the underlying environment dynamics. At the beginning of each episode, a preference vector is sampled and kept fixed throughout the episode, allowing the policy to learn consistent allocation behavior under a specific preference setting.

D. Reward Design and Multi-Objective Structure

At each decision step t, the environment provides a vector-valued reward that captures multiple aspects of task allocation performance, including service efficiency, preservation of human-centric interactions, and resource utilization balance. Representing rewards in vector form allows each objective to be modeled and analyzed independently before aggregation. To ensure transparency, the individual reward components of operational efficiency (r_t^{eff}), human-touch preservation (r_t^{touch}), and resource utilization (r_t^{util}) are formally defined based on the environment dynamics. The efficiency reward is defined as:

$$r_t^{eff} = \gamma_c \left(1 - \min\left(\frac{T(a_t)}{T_{\max}}, 1\right)\right)$$

where $T(a_t)$ denotes the service time of the selected agent (human or robot), $T_{\max}$ is the maximum service time observed in the system, and γ_c is a context-dependent scaling factor reflecting workload conditions (γ_c= 1.2 for high load, 1.0 for medium load, and 0.9 for low load).

The human-touch reward captures the importance of human interaction (h_t) and is defined conditionally based on the selected agent a_t. When a human is assigned ($a_t = 0$):

$$r_t^{touch} = \min\left(h_t \cdot (1 + 0.5\, p_{touch}) + \delta, 1\right)$$

where $\delta = 0.12\, p_{touch}$ if $p_{touch} > 0.6$, and $\delta = 0$ otherwise. When a robot is assigned ($a_t = 1$):

$$r_t^{touch} = 1 - \min\left(\text{penalty}, 1\right)$$

where the penalty term is defined as: penalty = $0.7\, h_t\, (p_{touch} - 0.6) \times 4.5$, if $p_{touch} > 0.6$, penalty = $0.2\, h_t p_{touch}$, otherwise.

The utilization reward promotes balanced workload distribution between human and robot agents and is defined as:

$$r_t^{util} = 1 - |\, u_H - u_R\, |$$

where u_H and u_R denote the current utilization ratios of the human and robot, respectively.

Finally, let $p = (p_{\text{eff}}, p_{\text{touch}}, p_{\text{util}})$ represent the sampled preference vector specifying the relative importance of each objective. The overall scalar reward used for policy optimization is computed through preference-weighted linear scalarization as follows:

$$R_t = p_{\text{eff}}\, r_t^{\text{eff}} + p_{\text{touch}}\, r_t^{\text{touch}} + \beta_{\text{util}}\, p_{\text{util}}\, r_t^{\text{util}},$$

where β_{util} is a scaling parameter controlling the contribution of the utilization term relative to the other objectives.

Preference-awareness is further incorporated during training through a lightweight shaping mechanism applied outside the environment. Following the Preference-Controllable Reinforcement Learning (PreCo) paradigm [30], an additional preference-alignment signal based on changes in cosine similarity between the cumulative objective vector and the sampled preference vector is used to guide learning. This shaping mechanism is implemented at the wrapper level and does not modify the PPO loss function or the underlying network architecture.

E. Learning Algorithm and Evaluation

The agent is trained using Proximal Policy Optimization (PPO) as implemented in Stable-Baselines3 [56], chosen for its stability and robustness in stochastic environments. Observation and reward normalization are applied using VecNormalize. Key hyperparameters include a learning rate of 3×10^{-4}, discount factor $\gamma = 0.99$, GAE parameter $\lambda = 0.95$, clip range 0.2, and batch size 128. These hyperparameters were selected based on established best practices for PPO and widely adopted configurations in prior reinforcement learning studies, and were empirically validated to ensure stable convergence in the proposed task allocation environment. Training is conducted for 300,000 interaction steps, corresponding to approximately 6,000 episodes.

Following common practice in preference-controllable reinforcement learning, evaluation focuses on two main requirements [30]. The first is the ability of the agent to explore the Pareto front and represent diverse trade-offs among objectives. The second is controllability, defined as the ability of the learned policy to produce a performance trade-off that aligns with the input preference. These requirements are evaluated using Hypervolume (HV) to assess Pareto front coverage and a similarity measure $\Psi(p, v_\pi)$ to quantify preference alignment. In addition to these preference-based metrics, task-level performance is assessed using episodic measures such as average task completion time, makespan, and human–robot allocation ratios, which are commonly used evaluation metrics in human–robot collaborative task allocation studies. All experiments are conducted with fixed random seeds to ensure reproducibility.

IV. Results and Discussion

This section presents and analyzes the experimental results obtained from both internal simulation-based evaluations and external validation experiments, along with their implications for the proposed framework.

A. Internal Evaluation Results

The results in this section are obtained after training the proposed preference-controllable policy, demonstrating stable and consistent learning behavior over 6,000 training episodes. To evaluate the controllability of the learned policy beyond training-time preference sampling, we conduct a controlled evaluation under three representative preference scenarios, designed to reflect realistic operational conditions in a service environment. These scenarios correspond to three dominant service modes commonly encountered in practice.

The first scenario is efficiency-oriented, representing high-demand periods (peak-hour scenarios) where rapid service and throughput are prioritized, and robotic execution is preferred for handling the majority of tasks. The second scenario is human-touch-oriented, emphasizing customer interaction and human touch, where task allocation favors human baristas while still allowing a balanced distribution with robotic assistance when appropriate. The third scenario is balanced, targeting a balanced utilization of human and robotic resources to ensure efficient operation without excessive workload concentration or prolonged idleness.

This evaluation protocol mirrors the controlled preference experiments commonly adopted in preference-controllable reinforcement learning studies, where fixed and interpretable preference vectors are used to examine policy behavior under specific trade-off configurations, as discussed in [30]. To analyze how the learned agent adapts its decisions under each scenario, we report the following evaluation metrics: preference alignment (Ψ), hypervolume (HV), and human–robot task allocation behavior.

In the following subsections, we analyze each of these aspects in detail to demonstrate the controllability, interpretability, and robustness of the proposed framework.

1) Preference alignment (Ψ): Preference alignment (Ψ) is used to evaluate how closely the learned policy follows the intended preference direction under fixed and interpretable preference configurations. It is computed as the cosine similarity between the cumulative achieved objective vector and the corresponding preference vector, providing a directional measure of controllability consistent with preference-controllable reinforcement learning.

As shown in Fig. 2, the balanced preference configuration exhibits the strongest alignment ($\Psi = 0.983$), indicating that the policy effectively tracks the intended trade-off when multiple objectives are jointly emphasized. In contrast, lower alignment values are observed under efficiency-oriented ($\Psi = 0.513$) and human-touch-oriented ($\Psi = 0.675$) preferences. These configurations correspond to more asymmetric preference settings, where emphasis is placed predominantly on a single objective.

This behavior is expected and reflects a fundamental characteristic of preference-controllable reinforcement learning as discussed in [30]. During training, preferences are sampled continuously across the preference simplex, encouraging the policy to remain responsive across a broad range of trade-offs rather than specializing toward extreme or corner configurations. Consequently, alignment under highly asymmetric preferences is naturally lower than under balanced configurations, without indicating a loss of controllability.

Overall, the variation in preference alignment across scenarios demonstrates that the learned policy responds systematically and directionally to changes in the preference input. Rather than converging to a fixed objective or ignoring preference information, the policy adapts its behavior in accordance with the specified preference configuration, satisfying a core requirement of controllable multi-objective decision-making in human–robot collaboration settings.

2) Pareto Quality (Hypervolume, HV): To complement the preference alignment analysis, we evaluate the Pareto solution quality of the learned policy using the hypervolume (HV) metric, which measures the volume of the objective space dominated by the achieved solutions. Hypervolume is a standard indicator in multi-objective optimization, capturing both convergence and coverage of the Pareto front.

Fig.3 reports the HV values obtained under the three controlled preference scenarios. The balanced configuration achieves the highest hypervolume ($HV \approx 0.41$), followed by the efficiency-oriented ($HV \approx 0.31$) and human-touch-oriented ($HV \approx 0.26$) scenarios. This pattern indicates that balanced preferences enable solutions that preserve more favorable trade-offs across objectives, resulting in broader Pareto coverage.

In contrast, preference configurations that strongly emphasize a single objective naturally constrain the solution space, leading to reduced hypervolume. This behavior is consistent with multi-objective optimization theory, where prioritizing one objective limits the achievable trade-offs among the remaining objectives. Importantly, lower HV values under efficiency- or human-touch-oriented preferences do not indicate suboptimal learning, but rather reflect the intentional focus imposed by the specified preference direction. These results align with observations reported in preference-controllable reinforcement learning literature, where balanced or moderately weighted preferences tend to yield higher Pareto quality, while extreme preferences trade overall solution diversity for targeted objective satisfaction.

3) Human–robot task allocation behavior: To examine how preference alignment and Pareto quality translate into operational behavior, we analyze the resulting human–robot task allocation under the controlled preference scenarios. Fig. 4 illustrates the proportion of tasks assigned to the robot and the human agent for efficiency-oriented, human-touch-oriented, and balanced preferences.

Under the efficiency-oriented configuration, the policy allocates a substantially larger share of tasks to the robot (approximately 0.85 robot utilization), reflecting an automation-driven strategy that prioritizes service speed and throughput during peak operational periods. This behavior is consistent with the specified preference direction and confirms that the policy leverages robotic execution when efficiency is emphasized.

In contrast, the human-touch-oriented configuration results in a pronounced shift toward human execution, with robot utilization decreasing to approximately 0.36. Tasks associated with higher interaction importance are preferentially handled by human baristas, demonstrating that the policy internalizes task semantics and adjusts allocation decisions to preserve human-centric service quality when required.

The balanced configuration converges to a moderate allocation (approximately 0.54 robot utilization), representing a stable compromise between automation and human involvement. This allocation pattern aligns with the higher preference alignment and hypervolume observed under balanced preferences, indicating that joint emphasis on multiple objectives yields both robust Pareto performance and interpretable allocation behavior.

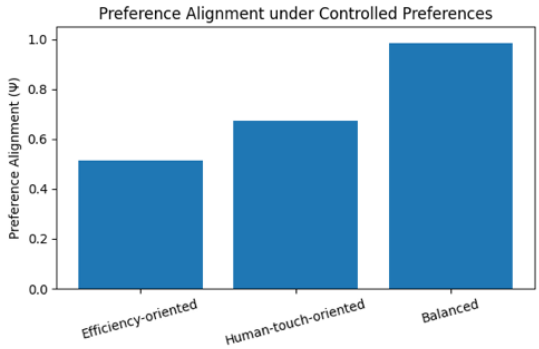

Fig. 2. Preference alignment (Ψ) under controlled preference configurations.

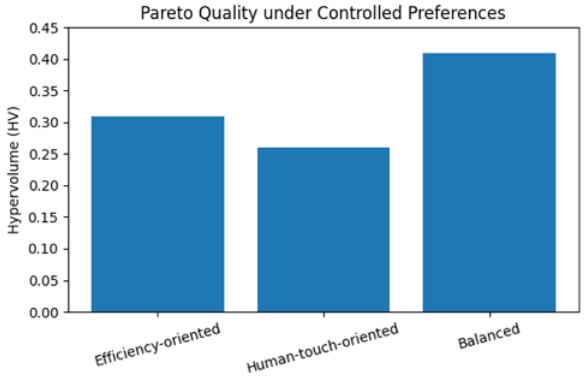

Fig. 3. Pareto quality measured by hypervolume (HV) under controlled preference configurations.

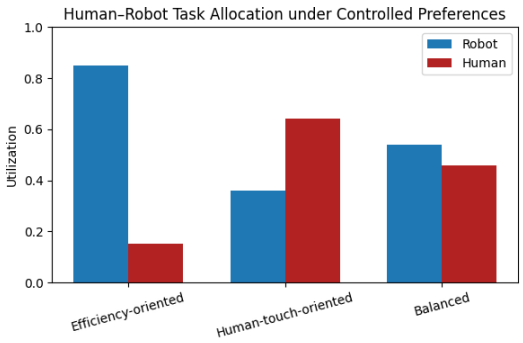

Fig. 4. Human–robot task allocation behavior under controlled preference configurations.

Notably, robot participation remains non-negligible across all scenarios. This outcome reflects the structure of the task environment, in which a subset of orders exhibits high automatability and favorable robot execution times. Rather than suppressing automation under human-centric preferences, the proposed framework permits robot involvement whenever it remains consistent with task characteristics and the specified preference constraints. Consequently, observed changes in utilization reflect relative behavioral adaptation rather than absolute enforcement of automation or manual operation.

Overall, these results demonstrate that the learned policy does not converge to a fixed automation ratio nor rely on static heuristics. Instead, it dynamically adjusts human–robot roles in a directionally consistent and interpretable manner in response to changing preferences, satisfying a central requirement of controllable human–robot collaboration in service-oriented environments.

The observed evaluation patterns are consistent with prior findings in preference-controllable reinforcement learning, where both preference alignment (Ψ) and Pareto quality (HV) tend to decrease under extreme or single-objective-dominant preference configurations, while more balanced preferences yield higher alignment and more favorable multi-objective trade-offs. In our experiments, this behavior manifests clearly as increased robot utilization under efficiency-oriented preferences, reduced robot involvement under human-touch-oriented preferences, and higher Ψ and HV under balanced configurations, demonstrating that the proposed framework exhibits controllable, preference-aware behavior in a service-oriented human–robot collaboration environment with interpretable task allocation decisions.

4) Comparison with baseline strategies: To assess the effectiveness of the proposed framework, we conduct a comparative evaluation against several baseline strategies. The comparison focuses on efficiency-oriented performance, measured by mean makespan and average task execution time, as well as the resulting human–robot allocation ratios.

Specifically, the proposed PreCo+PPO approach is evaluated under the three representative preference configurations defined in the previous section, which are efficiency-oriented, human-touch-oriented, and balanced, to reflect different operational priorities in service environments. These configurations are compared against four baseline strategies commonly used in practice: a rule-based two-threshold heuristic, full human execution, full robot execution, and a static 50/50 allocation. The rule-based two-threshold strategy assigns tasks using fixed decision rules based on task characteristics, favoring robot execution for highly automatable tasks and human execution for tasks requiring higher human involvement, without adapting to changing workloads or preferences. The full human and full robot strategies assign all tasks exclusively to a single agent, representing fully manual and fully automated operations, respectively. The static 50/50 strategy maintains an equal task split between human and robot throughout the episode, without considering task characteristics or operational context.

This comparison aims to examine whether the proposed approach can achieve competitive efficiency relative to automation-heavy strategies, while simultaneously maintaining meaningful human involvement and adaptive task allocation behavior that other baseline methods cannot provide. The results in Fig. 5 and Table II highlight the advantages of the proposed preference-controllable approach over baseline strategies. Under the efficiency-oriented configuration, the PreCo+PPO policy achieves a substantially lower mean makespan compared to rule-based and static allocation strategies, approaching the performance of the full-robot baseline while avoiding the complete elimination of human involvement. This demonstrates that high operational efficiency can be achieved without resorting to extreme automation, which has been associated with known limitations in prior studies.

When prioritizing human touch, the framework intentionally increases reliance on human execution to support customer interaction and service quality. As expected, this shift leads to a higher makespan due to reduced automation. However, compared to the full-human baseline, which results in the longest makespan and average task time, the proposed approach still benefits from selective robotic assistance to reduce overall service time while maintaining a strong level of human participation. This indicates that the framework does not simply replace automation with manual execution, but instead balances human involvement with robotic support to mitigate known challenges of fully manual operation, such as prolonged processing times, workload variability, and sensitivity to human-related disruptions reported in prior studies.

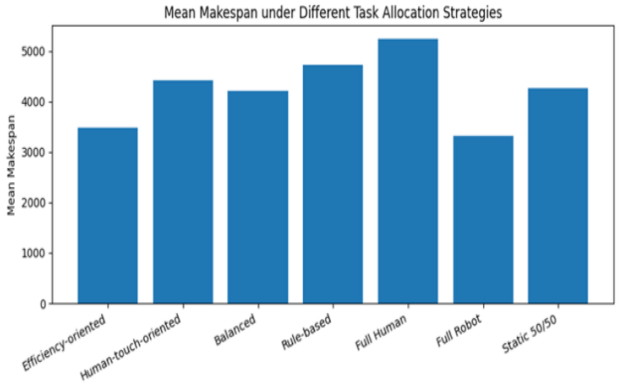

Fig. 5. Mean makespan comparison across preference-controllable strategies and baseline methods. Lower values indicate higher efficiency.

TABLE II. COMPARISON OF PREFERENCE-CONTROLLABLE AND BASELINE TASK ALLOCATION STRATEGIES

Category	Strategy	Mean Makespan	Avg Task Time (s)	Robot %	Human %
PreCo+PPO	Efficiency-oriented	3476.439	69.529	0.856	0.144
	Human-touch-oriented	4414.963	88.299	0.366	0.634
	Balanced	4207.118	84.142	0.547	0.453
Baseline	Rule-based	4730.312	94.606	0.417	0.583
	Full Human	5240.619	104.812	0.000	1.000
	Full Robot	3320.709	66.414	1.000	0.000
	Static 50/50	4268.116	85.362	0.510	0.490

The balanced configuration provides a practical trade-off between efficiency and human involvement. It achieves lower makespan values than the rule-based and static baselines while maintaining a near-equal human–robot allocation ratio. The slightly lower makespan compared to the static 50/50 strategy suggests that adaptive preference-aware control can offer modest efficiency gains over fixed allocation ratios under similar human–robot balance targets.

Overall, these results demonstrate that the proposed preference-controllable framework enables adaptive and interpretable task allocation across diverse operational priorities. Unlike static or heuristic-based methods, the proposed approach allows explicit control over trade-offs between efficiency and human-centric considerations without retraining or redesigning the policy, highlighting its practical applicability in real-world service-oriented environments.

B. External Validation Results

After establishing internal validity through controlled simulated experiments, we further evaluate the external validity of the proposed framework using real-world transactional data. The objective of this evaluation is to assess the robustness of the learned policy when exposed to realistic, heterogeneous, and noisy demand distributions that differ substantially from the synthetic order streams used during training. Importantly, the external evaluation does not modify the environment dynamics, reward formulation, or learning architecture. The trained PreCo+PPO policy is kept fixed throughout all experiments, and only the incoming order sequences are replaced with those observed in real transaction logs. This design allows the effect of realistic demand patterns on policy behavior to be isolated without introducing confounding factors related to retraining or architectural changes.

To assess external validity across different operational conditions, three real-world coffee sales datasets are employed [57-59], as summarized in Table III. These datasets differ in scale, structure, and data source, ranging from a large-scale transactional dataset comprising 149,116 transactions to smaller datasets with a few thousand records each. All datasets contain complete transactional information and exhibit demand characteristics commonly observed in real coffee shop operations.

Exploratory analysis shows that customer demand follows a skewed, long-tailed distribution, with coffee and tea products accounting for most orders, followed by bakery and specialty items. Most transactions include one or two items, indicating that the datasets largely reflect individual service requests rather than bulk orders. These characteristics are consistent with the task-level abstraction used in the simulated café environment. To integrate the real-world data, each transaction is mapped to the predefined service task categories using a simple rule-based mapping that preserves the original order sequence and task characteristics.

External evaluation follows the same controlled protocol used during internal validation. Three fixed and interpretable preference configurations are evaluated under deterministic execution, and performance is assessed using preference alignment (Ψ), Pareto hypervolume (HV), and human–robot utilization ratios. A quantitative summary of preference-aligned performance across both internal and external datasets is reported in Table IV, while Fig. 6 visualizes the results across datasets.

Across all external datasets and preference configurations, the proposed framework exhibits consistent, robust, and interpretable behavior. Under efficiency-oriented preferences, the policy consistently favors robotic execution, achieving robot utilization rates of approximately 83–85% across all datasets. Preference alignment and Pareto quality remain comparable to, and in several cases higher than, internal validation results, indicating effective prioritization of service efficiency under realistic and noisy demand conditions. In contrast, human-touch-oriented preferences result in a pronounced shift toward human execution, with robot utilization decreasing to approximately 35–38% across all datasets. This behavior remains stable across datasets of different sizes and sources, confirming that the policy responds to preference information rather than overfitting to specific order distributions.

The balanced preference configuration consistently yields the highest preference alignment ($\Psi \approx 0.97$–0.99) and competitive hypervolume values across all datasets, producing stable and interpretable human–robot workload distributions. Minor variations between internal and external evaluations are expected given the skewed and noisy nature of real-world transactional data and do not indicate performance degradation.

Overall, the external validation results demonstrate that the learned policy does not overfit to synthetic task distributions and maintains stable, interpretable, and preference-aligned behavior when deployed under realistic service demand compositions. The consistency of behavioral trends across datasets of different scales and sources provides strong evidence of external validity and supports the applicability of the proposed preference-

controllable framework to real-world human–robot collaborative service environments.

C. Discussion

Building on the experimental results, this section discusses the implications of preference-controllable reinforcement learning for human–robot task allocation in service environments. The findings demonstrate that a single preference-conditioned policy can adapt its allocation behavior in response to varying operational priorities while maintaining stable learning dynamics and interpretable decision patterns.

A key observation is that balanced preference configurations consistently yield stronger preference alignment (Ψ) and higher Pareto quality (HV) compared to extreme or single-objective-dominant scenarios. This behavior aligns with established findings in preference-controllable reinforcement learning, where policies trained to respond across a distribution of preferences tend to achieve more stable and well-balanced trade-offs under moderate settings. In contrast, extreme preferences naturally constrain the solution space, reducing Pareto coverage and reflecting an inherent trade-off between specialization and controllability rather than a limitation of the proposed approach.

Importantly, these behavioral patterns are preserved under external validation using real-world transactional datasets. Despite differences in scale, structure, and demand variability between synthetic and real-world data, the learned policy exhibits consistent directional responses to preference changes. This indicates that the policy captures task allocation principles that remain stable under realistic distributional shifts, rather than relying on environment-specific patterns.

Beyond aggregate performance metrics, the results provide insight into how preference-controllable learning translates into interpretable operational behavior. Efficiency-oriented preferences lead to increased robot utilization and faster task execution, while human-touch-oriented preferences shift allocation toward human workers to support service interaction and personalization. Balanced preferences produce moderate robot utilization, reflecting a practical compromise between automation and human involvement. These systematic allocation patterns are observed consistently across internal and external evaluations, confirming that the policy responds coherently to preference inputs rather than relying on static heuristics or fixed allocation rules.

Notably, robot participation remains non-negligible across all preference scenarios. This outcome is driven by the task composition of service environments, where many tasks exhibit high automatability and favorable robot execution characteristics. Rather than suppressing automation under human-centric preferences, the proposed framework allows robotic participation whenever it aligns with task requirements and preference constraints. This design reflects realistic service operations, in which automation and human labor coexist dynamically rather than being treated as mutually exclusive.

From both a practical and conceptual perspective, these findings suggest that preference-controllable reinforcement learning provides a flexible mechanism for managing human–robot collaboration in service settings where objectives may shift over time or differ across stakeholders. Adjusting preference weights enables controlled changes in operational behavior without retraining or redesigning the policy, supporting adaptability to changing service priorities and contextual requirements.

Overall, the discussion highlights that the primary contribution of this work lies not in optimizing a single performance metric, but in demonstrating how preference-controllable reinforcement learning can support controllable, interpretable, and context-aware task allocation in human–robot collaborative services. By explicitly modeling human touch alongside efficiency and utilization, and validating behavior under realistic demand conditions, this work bridges multi-objective reinforcement learning with practical service-oriented human–robot collaboration, highlighting its applicability in real-world deployment scenarios.

TABLE III. SUMMARY OF DATASETS USED FOR INTERNAL AND EXTERNAL VALIDATION

Dataset	Size (Transactions)	Reference
Dataset 1	149,116	[57]
Dataset 2	3,636	[58]
Dataset 3	3,898	[59]

TABLE IV. SUMMARY OF PREFERENCE-ALIGNED PERFORMANCE ACROSS INTERNAL AND EXTERNAL DATASETS

Dataset	Scenario	Preference Alignment (Ψ)	Hypervolume (HV)	Robot Utilization
Simulated Dataset	Efficiency-oriented	0.513	0.305	0.85
	Human-touch-oriented	0.675	0.264	0.36
	Balanced	0.983	0.408	0.54
Dataset 1	Efficiency-oriented	0.597	0.426	0.84
	Human-touch-oriented	0.653	0.290	0.37
	Balanced	0.969	0.485	0.56
Dataset 2	Efficiency-oriented	0.521	0.355	0.84
	Human-touch-oriented	0.730	0.196	0.37
	Balanced	0.987	0.349	0.58
Dataset 3	Efficiency-oriented	0.515	0.357	0.83
	Human-touch-oriented	0.732	0.196	0.37
	Balanced	0.987	0.347	0.58

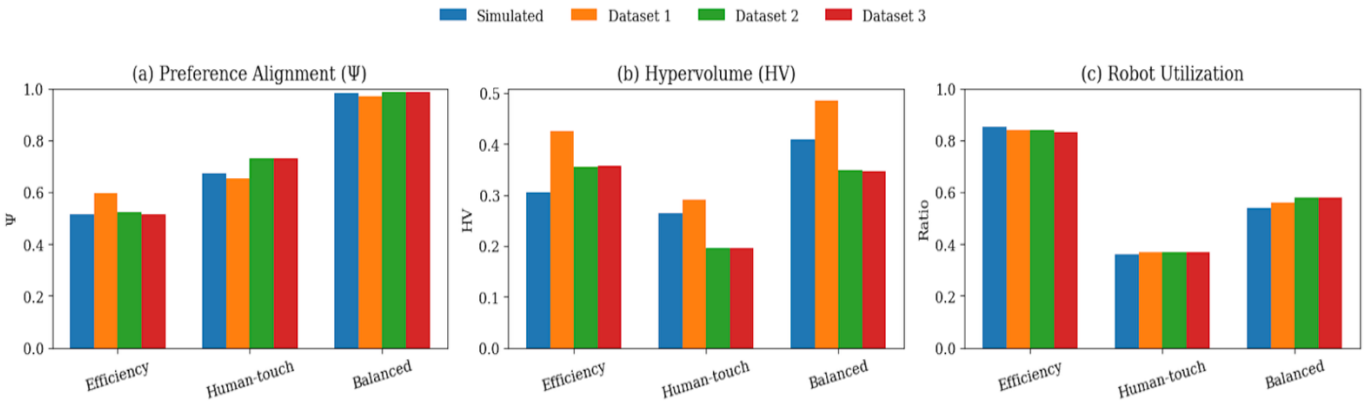

Fig. 6. Preference-controllable performance across datasets.

V. Limitations and Future Research

While the proposed framework demonstrates effective preference-controllable task allocation in a service-oriented environment, several limitations should be acknowledged. First, the task environment considered in this study contains a relatively high proportion of automatable tasks with favorable robot execution times. Although this reflects many contemporary service settings, environments with a larger share of human-exclusive or socially sensitive tasks may exhibit different allocation dynamics. Future work could explore alternative task compositions and service domains to assess the robustness of the proposed approach across a broader range of human-centered service contexts.

Second, the current study assumes explicitly specified and fixed preference configurations during evaluation, enabling controlled analysis of preference-dependent behavior. These configurations correspond to three common and practically relevant service scenarios in café environments, namely efficiency-oriented operation during peak hours, human-touch-oriented service during customer-focused or off-peak periods, and a balanced mode representing routine daily operation. In real service environments, however, preference priorities may shift gradually in response to operational conditions, customer feedback, or managerial decisions. Extending the framework to support smoothly varying or dynamically inferred preferences, while preserving policy stability and interpretability, represents an important direction

Third, although the simulation environment is designed to capture realistic service characteristics and demand patterns, the evaluation remains simulation-based. Real-world deployment may introduce additional sources of uncertainty, including human behavioral variability, execution delays, and operational constraints that are not explicitly modeled. In addition, task attributes such as automatability and human-touch importance are assumed to be known *a priori*. In this study, these attributes are derived from expert-informed estimates based on interviews with café domain specialists; however, in practice, such characteristics may be uncertain, noisy, or context-dependent. The simulated environment further assumes equal availability of human and robotic resources, adopting a 1:1 human-to-robot ratio to enable controlled analysis of task allocation behavior. Accordingly, future work will explore alternative human–robot resource compositions beyond the assumed 1:1 ratio, including configurations such as 1:2 or 2:1, to examine how varying workforce balance influences task allocation behavior and system performance.

Beyond these extensions, future research will focus on deploying the proposed framework in real or hybrid service settings involving human operators, learning task attributes online, and incorporating adaptive preference updates driven by human or customer feedback. Further extensions to more complex service scenarios, such as multiple human workers, heterogeneous robotic agents, or dynamically evolving task sets, would also enhance the applicability and scalability of preference-controllable human–robot collaboration beyond simulated environments.

VI. Conclusion

This paper introduced a preference-controllable reinforcement learning framework for human–robot task allocation in service-oriented environments. By integrating multi-objective reward design with preference-conditioned policy learning, the proposed framework enables a single trained policy to adapt task allocation behavior across varying efficiency, human-touch, and resource utilization priorities without requiring retraining. The experimental results demonstrate stable learning behavior, clear preference-dependent allocation patterns, and interpretable shifts between human and robotic execution under different operational priorities. Rather than enforcing fixed automation ratios, the framework supports adaptive task allocation that reflects both task characteristics and user intent, allowing human–robot collaboration strategies to be adjusted in response to changing service demands.

Validation across both simulated environments and real-world transactional datasets shows that the learned policy maintains consistent and interpretable behavior under realistic and heterogeneous demand conditions. These findings indicate that preference-controllable reinforcement learning provides a practical and flexible mechanism for managing trade-offs between efficiency and human-centric considerations in service settings. Overall, this work bridges multi-objective and preference-aware reinforcement learning with practical human–robot collaboration, demonstrating how preference-driven control can support adaptive, transparent, and context-aware

task allocation in real-world service environments, highlighting its potential for real-world deployment.

ACKNOWLEDGMENT

The project was funded by KAU Endowment (WAQF) at king Abdulaziz University, Jeddah, Saudi Arabia. The authors, therefore, acknowledge with thanks WAQF and the Deanship of Scientific Research (DSR) for technical and financial support.

REFERENCES

[1] M. Dhanda, B. A. Rogers, S. Hall, E. Dekoninck, and V. Dhokia, "Reviewing human-robot collaboration in manufacturing: Opportunities and challenges in the context of Industry 5.0," Robotics and Computer-Integrated Manufacturing, vol. 93, p. 102937, 2025.

[2] A. Baratta, A. Cimino, M. G. Gnoni, and F. Longo, "Human–robot collaboration in Industry 4.0: A literature review," Procedia Computer Science, vol. 217, pp. 1887–1895, 2023.

[3] S. K. Yadav and S. Shahi, "Safe human–robot collaboration in dynamic environments: An AI-powered situation awareness perspective," International Journal of Tropical Medicine, vol. 19, pp. 92–98, 2024.

[4] Y. Choi, M. Choi, M. Oh, and S. Kim, "Service robots in hotels: Understanding service quality perceptions of human–robot interaction," Journal of Hospitality Marketing & Management, vol. 29, no. 6, pp. 613–635, 2020.

[5] G. G. Liu, P. Benckendorff, and G. Walters, "Human–robot interaction research in hospitality and tourism: Trends and future directions," Tourism Review, 2024.

[6] T. Shimmura, R. Ichikari, T. Okuma, H. Ito, K. Okada, and T. Nonaka, "Service robot introduction to a restaurant enhances both labor productivity and service quality," Procedia CIRP, vol. 88, pp. 589–594, 2020.

[7] W. Grobbelaar, A. Verma, and V. K. Shukla, "Analyzing human–robot interaction in the food industry," Journal of Physics: Conference Series, vol. 1714, p. 012032, 2021.

[8] R. de Kervenoael, R. Hasan, A. Schwob, and E. Goh, "Leveraging human–robot interaction in hospitality services: Incorporating the role of perceived value, empathy, and information sharing into visitors' intentions to use social robots," Tourism Management, vol. 78, p. 104042, 2020.

[9] H. Qiu, M. Li, B. Shu, and B. Bai, "Enhancing hospitality experience with service robots: The mediating role of rapport building," Journal of Hospitality Marketing & Management, vol. 29, no. 3, pp. 247–268, 2020.

[10] I. Tussyadiah, "A review of research into automation in tourism: Launching the Annals of Tourism Research curated collection on artificial intelligence and robotics in tourism," Annals of Tourism Research, vol. 81, p. 102883, 2020.

[11] H.-W. Jang and S.-B. Lee, "Serving robots: Management and applications for restaurant business sustainability," Sustainability, vol. 12, no. 10, p. 3998, 2020.

[12] M. Xie and H.-B. Kim, "User acceptance of hotel service robots using the quantitative Kano model," Sustainability, vol. 14, no. 7, p. 3988, 2022.

[13] Y. Y. Liau and K. Ryu, "Task allocation in human–robot collaboration based on task characteristics and agent capability for mold assembly," Procedia Manufacturing, vol. 51, pp. 179–186, 2020.

[14] R. Mammadzada, "Various scheduling techniques used in human–robot collaborative system strategies for allocation and optimization of tasks," Journal of Modern Technology & Engineering, vol. 8, no. 3, 2023.

[15] J. Ding, M. Chen, T. Wang, J. Zhou, X. Fu, and K. Li, "A survey of AI-enabled dynamic manufacturing scheduling: From directed heuristics to autonomous learning," ACM Computing Surveys, vol. 55, no. 14s, pp. 1–36, 2023.

[16] M. L. Lee, X. Liang, B. Hu, G. Onel, S. Behdad, and M. Zheng, "A review of prospects and opportunities in disassembly with human–robot collaboration," Journal of Manufacturing Science and Engineering, vol. 146, no. 2, p. 020902, 2024.

[17] E. Merlo, E. Lamon, F. Fusaro, M. Lorenzini, A. Carfì, F. Mastrogiovanni, and A. Ajoudani, "An ergonomic role allocation framework for dynamic human–robot collaborative tasks," Journal of Manufacturing Systems, vol. 67, pp. 111–121, 2023.

[18] G. Bruno and D. Antonelli, "Dynamic task classification and assignment for the management of human–robot collaborative teams in workcells," The International Journal of Advanced Manufacturing Technology, vol. 98, no. 9, pp. 2415–2427, 2018.

[19] B. Tian, M. Janardhanan, and M. Marinelli, "A systematic investigation of the barriers to effective implementation of human–robot assembly lines: An integrated multi-criteria decision-making approach," International Journal of Computer Integrated Manufacturing, vol. 37, no. 1–2, pp. 198–223, 2024.

[20] W. Ren, X. Yang, Y. Yan, Y. Hu, and L. Zhang, "A digital twin-based framework for task planning and robot programming in human–robot collaboration," Procedia CIRP, vol. 104, pp. 370–375, 2021.

[21] Q. Lv, R. Zhang, X. Sun, Y. Lu, and J. Bao, "A digital twin-driven human–robot collaborative assembly approach in the wake of COVID-19," Journal of Manufacturing Systems, vol. 60, pp. 837–851, 2021.

[22] W. Hou, Z. Xiong, M. Yue, and H. Chen, "Human–robot collaborative assembly task planning for mobile cobots based on deep reinforcement learning," Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, vol. 238, no. 23, pp. 11097–11114, 2024.

[23] C. Urrea, "Hybrid deep learning–reinforcement learning for adaptive human–robot task allocation in Industry 5.0," Systems, vol. 13, no. 8, p. 631, 2025.

[24] C. Petzoldt, D. Niermann, E. Maack, M. Sontopski, B. Vur, and M. Freitag, "Implementation and evaluation of dynamic task allocation for human–robot collaboration in assembly," Applied Sciences, vol. 12, no. 24, p. 12645, 2022.

[25] A. Ali, H. Azevedo-Sa, D. M. Tilbury, and L. P. Robert Jr., "Heterogeneous human–robot task allocation based on artificial trust," Scientific Reports, vol. 12, no. 1, p. 15304, 2022.

[26] M. Faccio, I. Granata, and R. Minto, "Task allocation model for human–robot collaboration with variable cobot speed," Journal of Intelligent Manufacturing, vol. 35, no. 2, pp. 793–806, 2024.

[27] R. Wang, D. Zhao, A. Gupte, and B. C. Min, "Initial task allocation in multi-human multi-robot teams: An attention-enhanced hierarchical reinforcement learning approach," IEEE Robotics and Automation Letters, vol. 9, no. 4, pp. 3451–3458, 2024.

[28] S. Chand and Y. Lu, "Dual task scheduling strategy for personalized multi-objective optimization of cycle time and fatigue in human–robot collaboration," Manufacturing Letters, vol. 35, pp. 88–95, 2023.

[29] J. Xu, Y. Tian, P. Ma, D. Rus, S. Sueda, and W. Matusik, "Prediction-guided multi-objective reinforcement learning for continuous robot control," in Proc. International Conference on Machine Learning (ICML), PMLR, Nov. 2020, pp. 10607–10616.

[30] Y. Yang, T. Zhou, M. Pechenizkiy, and M. Fang, "Preference controllable reinforcement learning with advanced multi-objective optimization," in Proc. 42nd International Conference on Machine Learning (ICML), 2025.

[31] A. Alessio, K. Aliev, and D. Antonelli, "Multicriteria task classification in human–robot collaborative assembly through fuzzy inference," Journal of Intelligent Manufacturing, vol. 35, no. 5, pp. 1909–1927, 2024.

[32] C. Shyalika, T. Silva, and A. Karunananda, "Reinforcement learning in dynamic task scheduling: A review," SN Computer Science, vol. 1, no. 6, p. 306, 2020.

[33] E. Helms, R. D. Schraft, and M. Hägele, "rob@work: Robot assistant in industrial environments," in Proc. IEEE Int. Workshop on Robot and Human Interactive Communication, 2002, pp. 399–404.

[34] M. A. Goodrich and A. C. Schultz, "Human–robot interaction: A survey," Foundations and Trends in Human–Computer Interaction, vol. 1, pp. 203–275, 2007.

[35] D. Nguyen-Tuong and J. Peters, "Model learning for robot control: A survey," Cognitive Processing, vol. 12, no. 4, pp. 319–340, 2011.

[36] A. Khamis, A. Hussein, and A. Elmogy, "Multi-robot task allocation: A review of the state of the art," pp. 31–51, Springer International Publishing, 2015.

[37] C. Weckenborg, K. Kieckhäfer, C. Müller, M. Grunewald, and T. S. Spengler, "Balancing of assembly lines with collaborative robots," Business Research, vol. 13, no. 1, pp. 93–132, 2019, doi: 10.1007/s40685-019-0101-y.

[38] C. Ferreira, G. Figueira, and P. Amorim, "Scheduling human–robot teams in collaborative working cells," International Journal of Production Economics, vol. 235, p. 108094, 2021, doi: 10.1016/j.ijpe.2021.108094.

[39] F. Chen, K. Sekiyama, F. Cannella, and T. Fukuda, "Optimal subtask allocation for human and robot collaboration within hybrid assembly systems," IEEE Transactions on Automation Science and Engineering, vol. 11, pp. 1065–1075, 2014, doi: 10.1109/TASE.2013.2274099.

[40] A. Raatz, S. Blankemeyer, T. Recker, D. Pischke, and P. Nyhuis, "Task scheduling method for human–robot collaboration workplaces based on capabilities and execution time assumptions for robots," CIRP Annals, vol. 69, pp. 13–16, 2020, doi: 10.1016/j.cirp.2020.04.030.

[41] T. Yu, J. Huang, and Q. Chang, "Optimizing task scheduling in human–robot collaboration with deep multi-agent reinforcement learning," Journal of Manufacturing Systems, vol. 60, pp. 487–499, 2021, doi: 10.1016/j.jmsy.2021.07.015.

[42] M. Breque, L. De Nul, and A. Petridis, Industry 5.0: Towards a Sustainable, Human-Centric and Resilient European Industry, 2021.

[43] Y. Wang, H. S. Ma, J. H. Yang, and K. S. Wang, "Industry 4.0: A way from mass customization to mass personalization production," Advances in Manufacturing, vol. 5, pp. 311–320, 2017, doi: 10.1007/s40436-017-0204-7.

[44] Y. Lu, H. Zheng, S. Chand, W. Xia, Z. Liu, X. Xu, et al., "Outlook on human-centric manufacturing towards Industry 5.0," Journal of Manufacturing Systems, vol. 62, pp. 612–627, 2022, doi: 10.1016/j.jmsy.2022.02.001.

[45] E. Matheson, R. Minto, E. G. G. Zampieri, M. Faccio, and G. Rosati, "Human–robot collaboration in manufacturing applications: A review," Robotics, vol. 8, p. 100, 2019, doi: 10.3390/robotics8040100.

[46] A. Cherubini, R. Passama, A. Crosnier, A. Lasnier, and P. Fraisse, "Collaborative manufacturing with physical human–robot interaction," Robotics and Computer-Integrated Manufacturing, vol. 40, pp. 1–13, 2016, doi: 10.1016/j.rcim.2015.12.007.

[47] L. Wang, R. Gao, J. Váncza, J. Krüger, X. V. Wang, S. Makris, et al., "Symbiotic human–robot collaborative assembly," CIRP Annals, vol. 68, pp. 701–726, 2019, doi: 10.1016/j.cirp.2019.05.002.

[48] M. Pearce, B. Mutlu, J. Shah, and R. Radwin, "Optimizing makespan and ergonomics in integrating collaborative robots into manufacturing processes," IEEE Transactions on Automation Science and Engineering, vol. 15, pp. 1772–1784, 2018, doi: 10.1109/TASE.2018.2789820.

[49] M. Y. Jaber, Z. S. Givi, and W. P. Neumann, "Incorporating human fatigue and recovery into the learning–forgetting process," Applied Mathematical Modelling, vol. 37, pp. 7287–7299, 2013, doi: 10.1016/j.apm.2013.02.028.

[50] K. Li, Q. Liu, W. Xu, J. Liu, Z. Zhou, and H. Feng, "Sequence planning considering human fatigue for human–robot collaboration in disassembly," Procedia CIRP, vol. 83, pp. 95–104, 2019.

[51] J. Wang, Y. Yan, Y. Hu, X. Yang, and L. Zhang, "A transfer reinforcement learning and digital-twin based task allocation method for human–robot collaboration assembly," Engineering Applications of Artificial Intelligence, vol. 144, p. 110064, 2025.

[52] A. Abels, D. Roijers, T. Lenaerts, A. Nowé, and D. Steckelmacher, "Dynamic weights in multi-objective deep reinforcement learning," in Proc. International Conference on Machine Learning (ICML), PMLR, May 2019, pp. 11–20.

[53] M. D. Zhao, R. Simmons, and H. Admoni, "Learning human contribution preferences in collaborative human–robot tasks," in Proc. Conference on Robot Learning (CoRL), PMLR, Dec. 2023, pp. 3597–3618.

[54] F. Felten, L. N. Alegre, A. Nowé, A. Bazzan, E. G. Talbi, G. Danoy, and B. C. da Silva, "A toolkit for reliable benchmarking and research in multi-objective reinforcement learning," Advances in Neural Information Processing Systems, vol. 36, pp. 23671–23700, 2023.

[55] M. Towers, A. Kwiatkowski, J. Terry, J. U. Balis, G. De Cola, T. Deleu, et al., "Gymnasium: A Standard Interface for Reinforcement Learning Environments," in The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2026. [Online]. Available: https://openreview.net/forum?id=qPMLvJxtPK. Accessed: Jan. 2026.

[56] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," Journal of Machine Learning Research, vol. 22, no. 268, pp. 1–8, 2021.

[57] A. Abbas, "Coffee Sales Dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/ahmedabbas757/coffee-sales. Accessed: Jan. 2026.

[58] R. Richard, "Coffee Store Sales," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/reignrichard/coffee-store-sales. Accessed: Jan. 2026.

[59] I. Helon, "Coffee Sales Dataset," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/ihelon/coffee-sales. Accessed: Jan. 2026.