Multi-attribute utility Deep Reinforcement Learning method for Sequential Multi-Criteria Decision problems: Application to human resource planning

Mohammadreza Nematollahi, Adel Guitouni, Nafiseh Izadyar, Nabil Belacel, & Andrew Park

2026

Original citation:
Nematollahi, M., Guitouni, A., Izadyar, N., Belacel, N., & Park, A. (2026). Multi-attribute utility Deep Reinforcement Learning method for Sequential Multi-Criteria Decision problems: Application to human resource planning. *Computers & Operations Research*, *190*, Article 107426.
https://doi.org/10.1016/j.cor.2026.107426

Computers & Operations Research 190 (2026) 107426

Contents lists available at ScienceDirect

Computers and Operations Research

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

Multi-Attribute Utility Deep Reinforcement Learning method for Sequential Multi-Criteria Decision problems: Application to human resource planning

Mohammadreza Nematollahi [a,*], Adel Guitouni [b], Nafiseh Izadyar [c], Nabil Belacel [d], Andrew Park [b]

[a] Edwards School of Business, University of Saskatchewan, SK, Canada
[b] Gustavson School of Business, University of Victoria, BC, Canada
[c] Department of Computer Science, University of Victoria, BC, Canada
[d] Digital Technologies Research Centre, National Research Council of Canada, ON, Canada

ARTICLE INFO

Keywords:
Deep reinforcement learning
Sequential multi-criteria decision
Multi-attribute utility theory
Risk preference
Human resource planning
Sustainable farming

ABSTRACT

Problem-solving and decision-making can be complex. There are often conflicting criteria, and decisions must take into account both immediate and long-term impacts, which define Sequential Multi-Criteria Decision (SMCD). Deep Reinforcement Learning (DRL) has emerged by integrating traditional Reinforcement Learning with Deep Learning to tackle intricate sequential decision-making problems. Although DRL has seen significant progress recently, there has been limited focus on developing DRL algorithms specifically for SMCD problems, which usually involve conflicting and non-commensurable attributes. To bridge this gap, we introduce a novel algorithm called Multi-Attribute Utility DRL (MAUDRL), which combines DRL with Multi-Criteria Decision Analysis (MCDA). This innovative approach provides a clear and transparent DRL model that can address the intricacies of SMCD problems while integrating the risk attitudes and preferences of the decision-maker. We showcase the potential of MAUDRL in promoting sustainable decision-making for human resource planning for blueberry farming in British Columbia, Canada. We evaluate the performance of MAUDRL in comparison with two benchmark algorithms—Oracle Discrete Multi-Attribute Utility Theory (MAUT) and the Single Reward Aggregation Approach—using three metrics: policy quality, goal achievement, and run times. The numerical analysis and benchmarks validate that MAUDRL offers practical solutions for SMCD problems by assisting in exploring diverse solution spaces efficiently. The theoretical implications and practical applications of these results are discussed, underscoring the capability of MAUDRL in tackling complex SMCD problem domains and advancing sustainable and socially responsible decision-making while considering the risk preferences of decision-makers.

1. Introduction

In most operational settings, decisions must balance several, often conflicting, criteria at once. Supplier selection, for example, weighs price, quality, delivery lead time/distance, reliability, and reputation. The multi-criteria decision-making (MCDM) literature formalizes such trade-offs and offers tools to identify preferred alternatives under explicit preferences and constraints (Thakkar, 2021).

A second, equally common, class of problems is sequential decision-making, where actions unfold over time and today's choice shapes tomorrow's options and payoffs. In these settings, a decision maker acts, observes the state transition and realized rewards, and updates subsequent actions—classic territory for reinforcement learning and dynamic optimization (Powell, 2021). Airline revenue management is a canonical example: prices are continuously adjusted as demand information arrives and departure approaches.

Most real applications are sequential choices under multiple criteria. In sequential multi-criteria decision-making (SMCD), the decision maker must evaluate several attributes at each decision point while accounting for the system's temporal dynamics. SMCD integrates ideas from MCDM (preference modelling, trade-off analysis) and sequential decision-making (state dynamics, policies), aiming to produce time-consistent policies that respect preferences and feasibility over the entire horizon. This paper contributes to that agenda by introducing a modular framework that learns criterion-wise value functions and aggregates them via risk-sensitive multi-attribute utility to select

* Corresponding author.
E-mail addresses: nematollahi@edwards.usask.ca (M. Nematollahi), adelg@uvic.ca (A. Guitouni), nizadyar@uvic.ca (N. Izadyar), nabil.belacel@nrc-cnrc.gc.ca (N. Belacel), apark1@uvic.ca (A. Park).

https://doi.org/10.1016/j.cor.2026.107426
Received 8 January 2025; Received in revised form 1 November 2025; Accepted 6 February 2026
Available online 17 February 2026

feasible policies over time. In today's complex and interconnected world, most modern decision problems can be characterized as SMCD problems. This is especially true in sustainability, where decisions must balance short-term operational efficiencies with long-term sustainable outcomes, thereby increasingly considering multiple conflicting criteria. These decisions, often sequential and interconnected, range from selecting sustainable transportation modes and optimizing resource allocation to implementing ethical labour practices and minimizing environmental footprints. The need to make such balanced decisions is evident in various fields, including logistics, supply chain management, and environmental planning, where each decision impacts subsequent choices and outcomes, creating a web of interconnected decisions that must satisfy diverse stakeholders, adhere to regulatory standards, achieve economic objectives, and contribute positively to social and environmental well-being (Seuring and Müller, 2008; Tang, 2006).

SMCD problems appear whenever decisions unfold over time under multiple, often conflicting, criteria. In revenue management, firms set dynamic prices for limited, perishable capacity (airline seats, hotel rooms) as inventory dwindles and departure nears, trading off profit against customer satisfaction, perceived fairness (Cohen et al., 2025), and spoilage risk. In inventory systems—especially for perishables—ordering and stocking policies evolve with demand and shelf life, balancing holding and ordering costs against waste, service levels, and environmental impact (Ghafour, 2024). Traffic management requires sequential choices on signal timing, lane assignment, and tolling while weighing travel time, fuel use, emissions, and equity among road users (Zhang et al., 2024). Healthcare operations (e.g., bed allocation, surgical scheduling) similarly optimize day-to-day assignments under criteria including patient urgency, resource availability, staff workload, and outcomes (Diaby et al., 2013). In energy systems, operators dispatch resources over time to satisfy demand while promoting renewable utilization and maintaining grid stability (Zhou, 2023). These cases illustrate the breadth of SMCD—time-linked decisions that must reconcile heterogeneous objectives under operational constraints.

The methodological challenges associated with SMCD problems are substantial. Traditional methods, such as multi-criteria Markov decision processes and dynamic programming, face significant difficulties due to the curse of dimensionality and the complexity of integrating decision-makers' preferences and risk attitudes (Roijers et al., 2013). As a result, existing SMCD methods struggle to capture the complexity and uncertainty inherent in real-world decision-making environments. This highlights the need for more advanced models that can integrate preference-sensitive decision-making with scalable learning approaches. Deep Reinforcement Learning (DRL) combines Deep Learning (DL) and traditional Reinforcement Learning (RL) to solve a wide range of previously thought complex and intractable decision-making problems. RL enables an agent to successively learn more effective actions to maximize cumulative reward over time (Sutton and Barto, 2018). DL exploits artificial neural networks through representation learning, and DRL scholars leverage big data, high-performance computing, and increased financial investments to push the frontiers of DRL further concerning its core capabilities, mechanisms, and applications (Li, 2017; Mnih et al., 2015). The deep neural networks combat the curse of dimensionality, making previously intractable tasks feasible (Goodfellow et al., 2016).

Our research integrates DRL with Multi-Criteria Decision Analysis (MCDA) to solve complex SMCD problems. By combining DRL with MCDA, we can map each decision and action to the decision-makers distinct risk preferences and the existing world state, establishing a transparent link from inputs to outputs. We introduce a novel Multi-Attribute Utility Deep Reinforcement Learning (MAUDRL) model to guide a decision agent in learning successful policies for solving SMCD problems. The MAUDRL model is validated through its application to a human resource planning (HRP) problem in sustainable blueberry farming in British Columbia, Canada, considering the uncertainty associated with workers' performance and availability. The performance of MAUDRL is compared to two benchmark algorithms—Oracle Discrete Multi-Attribute Utility Theory (MAUT) and the Single Reward Aggregation Approach—based on three key metrics: policy quality, goal achievement, and run times.

This paper makes three contributions. (1) To operations research, it introduces a scalable decision-support approach for high-dimensional SMCD by combining deep reinforcement learning with multi-criteria preference modelling, yielding time-consistent policies under explicit trade-offs. (2) To AI, it proposes MAUDRL, an explainable and computationally efficient framework that learns criterion-wise value functions, maps them to risk-sensitive partial utilities, and aggregates them via multi-attribute utility to select feasible sequential actions. (3) To the human-resource planning literature, it demonstrates MAUDRL on sustainable HR planning in blueberry farming (British Columbia), evaluating multiple decision-maker profiles to reflect heterogeneous sustainability goals and risk preferences. The findings from our numerical analysis and benchmarks provide robust validation that MAUDRL delivers effective solutions for SMCD problems, enhancing the interpretability and trustworthiness of analytics and AI. We also highlight how an agent's sustainable decision-making behaviour significantly impacts the efficiency of HRP policies in the context of the blueberry farming industry.

The structure of the paper is organized as follows. In Section 2, a detailed review of the relevant literature is presented. Section 3 introduces and elaborates on the SMCD problem. In Section 4, we propose the MAUDRL methodology. Section 5 details the results obtained from applying MAUDRL to sustainable HRP in blueberry farming. Section 6 examines the implications of the results, while Section 7 discusses the conclusions and future research opportunities. The Appendix includes detailed pseudo-code for the methodology.

2. Literature review

This study intersects three areas of research—SMCD problems, DRL models, and HRP in farming—which are reviewed below to highlight the study's contributions and positioning.

2.1. Sequential multi-criteria decision (SMCD)

SMCD problems are predicated on making sequential decisions over time, while considering a set of heterogeneous and conflicting criteria. Kornbluth (1985) propose simple programming techniques for solving SMCD problems, where the decision-maker must evaluate and either accept or reject a series of outcomes characterized by multiple attributes. Campanella and Ribeiro (2011) present a versatile framework for addressing SMCD problems, demonstrated through a small-scale example involving helicopter landing scenarios. Liu et al. (2018) develop the study of Campanella and Ribeiro (2011) by incorporating a fuzzy environment, where decision-makers' evaluations are expressed using a bipolar linguistic scale comprising negative, neutral, and positive assessments.

Lang et al. (2021) propose a decomposition process to address the multi-criteria dynamic slotting problem, breaking the optimization into two interconnected components: predictive routing and revenue management. Dahooie et al. (2021) integrate Multi-Attribute Decision-Making with Data Envelopment Analysis (DEA) to address the credit risk evaluation problem.

Usual approaches to solving SMCD problems involve multi-criteria decision trees and approximate dynamic programming, which suffer from the curse of dimensionality (Frini and Amor, 2019; Chikalov et al., 2018; Frini et al., 2012). Thus, the literature on effective solutions to SMCD problems is still scarce, though notably, several studies have identified the inordinate complexity of SMCD problems. For example, Wood and Khosravanian (2015) observe that mutable conditions, such as the changing risk attitudes of the stakeholder, are rarely

accounted for in this problem domain, which can critically impact effective decision-making.

In recent years, DRL has been increasingly applied to a wide range of complex sequential decision-making problems, including mobile robot navigation, personalized targeting strategies on digital platforms (Wang et al., 2023b), dynamic order picking (Mahmoudinazlou et al., 2025), resource allocation (Chen et al., 2025), dynamic job-shop scheduling (Zhang et al., 2026), and multi-depot electric vehicle routing in dynamic environments (Shahbazian et al., 2025). Although there have been substantial advancements in DRL, its application to SMCD has been largely overlooked. However, we contend that this approach can be highly effective. One of the few examples we found of DRL being applied to SMCD was a study by He et al. (2021). They applied DQN to generate a decision support framework, and determined that this combined model outperformed traditional methods.

2.2. Deep reinforcement learning (DRL)

RL algorithms can tackle simple sequential decision-making problems by informing an agent, which interacts with its environment, what action to take at each state to optimize its cumulative reward (Sutton and Barto, 2018). Yan et al. (2022) review the RL applications in addressing challenges within logistics and supply chain management. While RL was novel in advancing experience-driven autonomous learning (Sutton, 1995), it suffers from a lack of scalability and cannot manage high-dimension state spaces, resulting in inordinately high computation times for complex problems. To address these limitations of traditional RL, Mnih et al. (2015) proposed Deep Reinforcement Learning (DRL), which leverages recent advancements in deep neural networks.

Liu et al. (2020), Alcaraz et al. (2022), and Liu et al. (2022) show DRL applications in enhancing efficiency and precision in the context of real-world route planning. DRL has been also widely adopted to tackle dynamic scheduling problem (Liu et al., 2023a; Wu and Yan, 2023; Wang et al., 2023a). Xie et al. (2022) tackle the challenge of optimizing dynamic information dissemination within a transportation network by developing a double deep Q-learning. Ding et al. (2023) present an integrated airline recovery framework, leveraging a decision recommendation system that incorporates DRL for efficient decision-making in the face of disruptions. Liu et al. (2023b) apply DRL to the context of taxi cruising route planning. Wang et al. (2023c) propose an innovative ride-sharing model incorporating passenger transfers. Their approach integrates DRL with Integer-Linear Programming (ILP) in a unified decision framework to manage dispatching, transfer coordination, and vehicle rebalancing.

In the operations research field, the utility of DRL on its own has typically been limited to a single reward function. Some recent studies have extended DRL models to problems with multiple sources of reward such as those utilizing multi-dimensional distributional DQNs, which captures correlated randomness from multiple sources of reward (Zhang et al., 2021).

Recent DRL advances tackle multi-objective optimization by combining decomposition with preference conditioning. Li et al. (2020) introduce DRL-MOA, an end-to-end scheme that decomposes an MOP into scalar subproblems solved by neural policies with neighbourhood-based parameter transfer, producing strong Pareto sets on combinatorial instances (e.g., multi-objective TSP). In aero-engine maintenance, Wei et al. (2024) similarly scalarize cost–reliability trade-offs and accelerate convergence via neighbourhood transfer; for EV charging, Wang et al. (2021) pair an actor–critic architecture with a modified Pointer Network to schedule service and energy allocations under scalarized objectives. For design and routing, Wang et al. (2023d) propose two DDPG variants for wind-turbine blade optimization—MO-DDPG (Pareto design generation) and MO-DSPG (enhanced exploration via a restricted Boltzmann machine)—while Wu et al. (2024) decompose multi-objective VRPs into preference-driven single-objective tasks learned in parallel with a preference-based attention network and improved through a collaborative agent system that shares knowledge via imitation. In residential energy management, Lu et al. (2022b) develop a tunable multi-objective DQN that learns a single policy adaptable to user-specified weights (cost, peak demand, punctuality), and Jiang et al. (2024) extend the setting to multi-objective games with vector-valued payoffs. Despite impressive results, most approaches (i) rely on scalarization during training or inference, (ii) treat feasibility and constraints episodically rather than as time-coupled requirements, and (iii) offer limited, interpretable links to decision-maker preferences (e.g., risk attitudes and attribute-wise utilities); in contrast, we learn criterion-wise value functions, map them to risk-sensitive partial utilities, and aggregate via MAUT while enforcing feasibility over time, yielding preference-aware sequential policies with auditable trade-offs.

Another body of research on DRL with multi-objective scenarios has begun to emerge (Nguyen et al., 2020a; Wang et al., 2020; Lu et al., 2022a). These studies use DRL to improve multi-objective optimization techniques and meta-heuristics. Our study does not contribute to this specific stream of literature, which is focused on DRL-itself and is agnostic to the decision analysis context. Rather, we contribute to the operations research field by leveraging highly novel DRL methods, and combining them with multi-criteria decision analysis (MCDA) to solve SMCD problems.

2.3. Sustainable human resource planning (HRP) in farming

We review key literature on sustainable human resource planning (HRP) within the farming sector, providing context for the application of MAUDRL in this domain.

Macke and Genari (2019) offer a systematic and comprehensive review of the sustainable human resource management literature, underlining its significance in various industries. Kramar (2022) examine six core features of sustainable HRM which are applicable across multiple sectors. For farming-specific applications, Stup et al. (2006) investigate how HRM practices in dairy farm businesses influence productivity and profitability, illustrating the direct relationship between effective HR management and operational success. Moreover, Pandey et al. (2024) examine the role of training effectiveness, along with farmers' psychological and demographic characteristics, in influencing the adoption of sustainable farming practices.

Human resource planning (HRP) refers to the process of ensuring that an organization has the right quality and quantity of personnel to achieve its goals effectively (Price et al., 1980). Research in HRP often focuses on staffing and recruitment, which remain critical and widely studied areas. The HRP literature applies various methodologies to address workforce management challenges. These approaches generally fall into three categories: optimization models, Markovian models, and computer simulation techniques (Karimi-Majd et al., 2017). For example, Bordoloi and Matsuo (2001) develop an optimization model that considers workforce learning, promotion, turnover rates, and stochastic demand to allocate human resources efficiently. Similarly, Yalçındağ et al. (2016) propose pattern-based decomposition methods for HRP in home healthcare services, focusing on scheduling, assignment, and routing decisions.

Other researchers introduce stochastic dynamic, and decomposition methods. Sadeghi-Dastaki and Afrazeh (2018) use stochastic programming to address manpower planning in production environments. Their model incorporates workforce diversity, hierarchical skills, skill substitution, and uncertain demand. In higher education, Aviso et al. (2019) apply a P-graph methodology to support decision-making on staffing level expansions. Additionally, Amorim-Lopes et al. (2021) introduce a multi-methodology that integrates foresight and scenario planning to assist strategic planners in predicting workforce needs over the long term, particularly in health resource planning. Annear et al. (2023) propose an approximate dynamic programming framework to manage the assignment of multi-skilled workers in job shops, addressing uncertainty in demand and variability in resource availability.

While mathematical and computational models dominate HRP research, the use of artificial intelligence (AI) remains limited. The closest study we found to ours was (Karimi-Majd et al., 2017), who use the Q-learning technique for HRP in production systems, but do not combine this technique with MCDA methods, and the findings do not appear to be generalizable to SMCD problems. The current study focuses on sequential multi-criteria HRP decision-making, expanding the application of AI tools in HRP. Besides, the proposed Multi-Attribute Utility DRL (MAUDRL) in this study represents a breakthrough improvement over previously proposed machine learning models, such as those in Wei and Jin (2021) and Yuan et al. (2022), which rely on historical data and struggle to generalize beyond their training data.

This review reveals that most studies rely on optimization (Bordoloi and Matsuo, 2001), stochastic programming (Sadeghi-Dastaki and Afrazeh, 2018), or simulation models (Amorim-Lopes et al., 2021). Even when reinforcement learning has been explored (Karimi-Majd et al., 2017), formulations are typically single-objective and do not integrate MCDA principles. Consequently, existing HRP models underrepresent the *sequential, multi-criteria* nature of workforce planning in dynamic, uncertain settings—especially in sustainable agriculture, where trade-offs among efficiency, cost, environmental stewardship, and social outcomes are pronounced.

2.4. Research gaps and contributions

Recent DRL approaches address sequential multi-criteria decision-making (SMCD), but most adopt *decomposition* strategies that split a multi-objective problem into scalar subproblems and solve them collaboratively via neighbourhood-based parameter transfer and Actor–Critic training (Li et al., 2020; Wang et al., 2021, 2023b). Others employ *policy-gradient* methods to learn Pareto policies in continuous state–action spaces (Jiang et al., 2024; Wu et al., 2024). While methodologically effective, these lines of work (i) emphasize algorithmic performance over decision-analytic integration, (ii) rarely encode decision-maker *preferences, utilities, or risk attitudes* explicitly, and (iii) are tailored to continuous environments that differ from the *discrete* decision contexts common in operations.

We propose a *Multi-Attribute Utility Deep Reinforcement Learning* (MAUDRL) method that (a) trains policies with respect to each criterion separately and then (b) aggregates outcomes using decision-maker preferences and risk attitudes, yielding final policies that reflect explicit cross-criterion trade-offs. Methodologically, MAUDRL builds on a *Deep Q-Network* architecture suited to discrete state–action spaces, in contrast to prior work that leans on Actor–Critic schemes (Li et al., 2020; Wang et al., 2021, 2023b) or policy-gradient methods (Jiang et al., 2024; Wu et al., 2024). Conceptually, it bridges DRL with MCDA, improving interpretability and practical relevance for decision support.

From an application perspective, most HRP studies employ optimization or simulation, and the few that use RL (e.g., Karimi-Majd et al., 2017) do not extend to multi-criteria settings. By applying MAUDRL to sustainable HRP in blueberry farming, we illustrate how this integration can reduce costs, capture sustainability and workforce considerations, and maintain computational efficiency.

3. Sequential multi-criteria decision problem

Consider a decision-agent who makes sequential decisions over time and at each decision epoch evaluates the effects of their decisions based on multiple criteria. This is known as a SMCD problem, characterized by a series of decisions made over time, where each decision affects subsequent choices under multiple and often conflicting criteria (Olson, 1997). Unlike traditional multi-criteria decision-making, SMCD problems are inherently dynamic, requiring the decision-maker to evaluate the immediate consequences of a decision and its long-term implications across various dimensions (Pomerol and Barba-Romero, 2000). Furthermore, SMCD problems often involve high uncertainty and unpredictability, particularly in the context of rapidly changing environments. The resolution of SMCD problems requires the ability to balance conflicting criteria and forecast and incorporate the potential evolution of these criteria over time. As such, SMCD poses unique challenges and demands sophisticated decision-making strategies that can effectively navigate the temporal and multi-dimensional nature of these problems (Bouyssou et al., 2006).

A Multi-Criteria Markov Decision Process can be used to model a SMCD problem, which is represented by a set of states $\mathcal{S} = \{s_l | l = 1, \dots, L\}$, a set of actions $\mathcal{A} = \{a_j | j = 1, \dots, J\}$, a transition probability $\mathcal{P} = \{p_{l-1,l} | l = 1, \dots, L\}$ from state $l-1$ to state l, a set of attributes-elasticity coefficients $(\mathcal{G}, \Psi) = \{(g_i, \psi_i) | i = 1, \dots, I\}$, where ψ_i represent the elasticity coefficient associated with attribute g_i, and a reward function $\mathcal{R} = \{r_i(g_i) | i = 1, \dots, I\}$.

We define $\mathcal{S} = \bigcup_{t=1}^{T} \mathcal{S}^t$, $\mathcal{A} = \bigcup_{t=1}^{T} \mathcal{A}^t$, and $\mathcal{R} = \bigcup_{t=1}^{T} \mathcal{R}^t$ for a discrete SMCD problem over the planning horizon T, with $t = 1, \dots, T$. At time t, the agent chooses action $a_j^t \in \mathcal{A}^t$ considering their observations of the current state $s_l^t \in \mathcal{S}^t$, and receives a set of rewards $r_i^t(g_i) \in \mathcal{R}^t$. The transition from state s_l^t to s_{l+1}^{t+1} is random with a probability $\mathcal{P}((s_l^t, s_{l+1}^{t+1})|_{a_j^t}) \sim \mathcal{P}$. We use a single criterion synthesizing multiple criteria decision method (Guitouni and Martel, 1998). The large number of actions that need to be compared at each time period prevents using pairwise comparison or interactive methods as suggested by Guitouni and Martel (1998). We also consider decision-making under uncertainty. We, therefore, use Multi-Attribute Utility Theory (MAUT) (Keeney et al., 1993). First, we consider u_i^t as a partial utility function constructed for each reward $r_i^t(g_i)$ corresponding to attribute g_i. Second, we obtain a total utility function $U^t(u_1^t, \dots, u_i^t, \dots, u_I^t)$ by aggregating the partial utilities u_i^t. The total utility over the planning horizon is given by $U = \bigcup_{t=1}^{T} U^t$ (Yang et al., 2019).

In our model, each attribute g_i refers to a measurable characteristic or performance dimension of a decision alternative that reflects one aspect of the decision maker's objectives. The corresponding partial utility function u_i^t captures preferences over the outcomes of that attribute, and the total utility function aggregates these partial utilities into a single measure of performance.

For example, in agriculture, a farmer choosing among crop varieties may need to weigh attributes such as expected yield, resistance to pests and diseases, water and fertilizer requirements, input costs, and market price stability. In irrigation management, attributes such as water-use efficiency, energy consumption, installation cost, and long-term soil health become central to the decision of which irrigation technology to adopt. In supply chain management, selecting a supplier typically involves comparing purchase cost, delivery reliability, product quality, lead time, environmental footprint, and flexibility. In transportation, the decision may involve choosing between alternative modes of transport (e.g., road, rail, air, or sea) or routing strategies, where important attributes include fuel consumption, travel time, on-time delivery rate, emissions, and overall cost. In healthcare, a common decision is the selection of treatment plans, which may require balancing attributes such as treatment effectiveness, side effects, patient satisfaction, and financial cost. Each of these examples demonstrates how the definition of attributes depends on the problem setting, while the general structure of MAUT provides a systematic way of combining them into a single decision framework.

In this study, we apply MAUDRL to sustainable human-resource planning in farming, defining three attributes aligned with the sustainability pillars: environmental—pruning quality (proxy for environmental performance), social—job creation/turnover, and economic—operational efficiency as a measure of financial performance.

4. Multi-attribute utility deep reinforcement learning (MAUDRL)

DRL was introduced by Mnih et al. (2015) to enhance traditional RL. By combining RL with deep neural networks, Mnih et al. (2015)

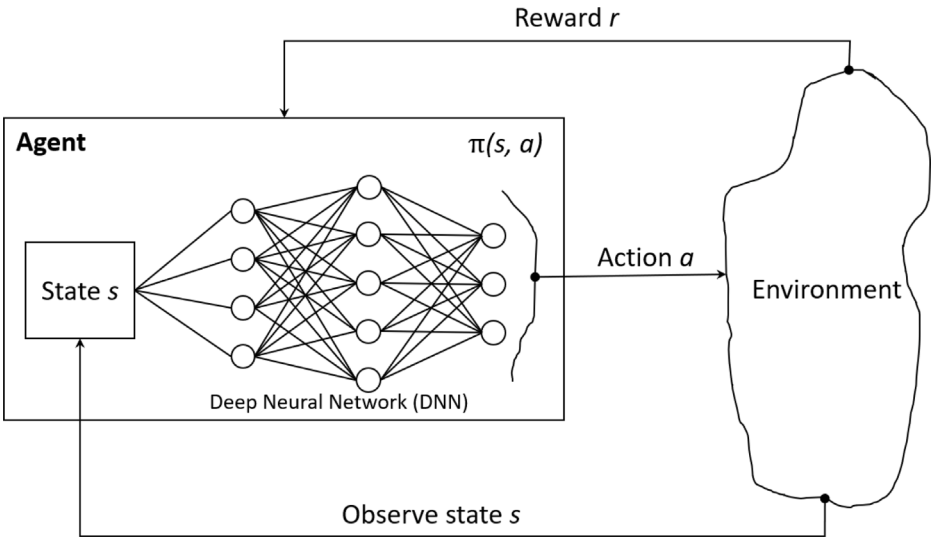

Fig. 1. Deep reinforcement learning structure.

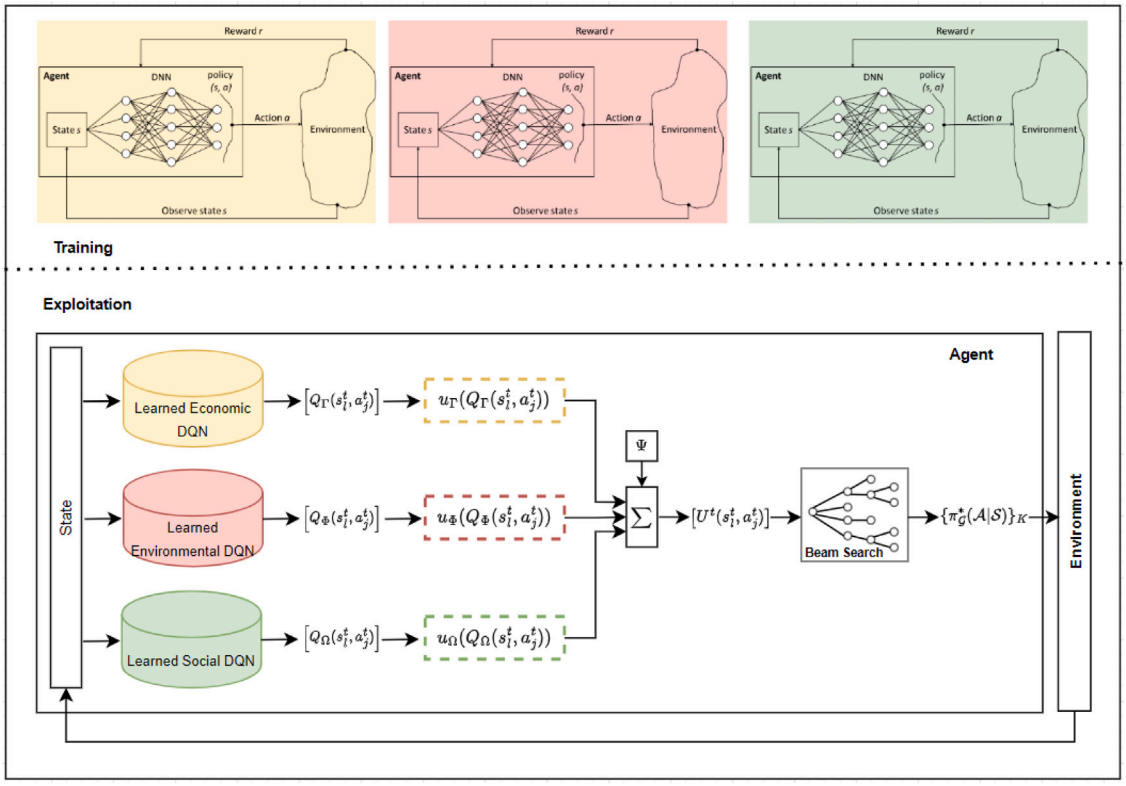

Fig. 2. MAUDRL conceptualization.

made a significant advance, and demonstrated the ability for an agent to complete a diverse array of tasks of high dimensionality (see Fig. 1).

Within the framework of a single reward function, a decision-making agent seeks to identify a policy $\pi_i^*(\mathcal{A}|S)$ that optimizes their anticipated cumulative reward $\sum_{t=1}^{T} \gamma^{t-1} r_i^t(g_i)$ for a specific attribute g_i, with future rewards being discounted by a factor $\gamma \in [0, 1]$. In a multi-criteria setting, the agent considers receiving rewards associated with various attributes, and aims to find a policy $\pi_{\mathcal{G}}^*(\mathcal{A}|S)$ which maximizes their cumulative utility function $\sum_{t=1}^{T} \gamma^{t-1} U^t$ calculated based on the coefficient set of attributes Ψ. Therefore, we develop MAUDRL to aid an agent in making sequential multi-criteria decisions (see Fig. 2).

MAUDRL includes two stages: (1) training, and (2) exploitation. In the training stage, we use a Deep Q-Network (DQN) regarding each attribute $g_i \in \mathcal{G}$ to find the policy that achieves the highest expected cumulative reward: $\sum_{t=1}^{T} \gamma^{t-1} r_i^t(g_i)$. We designed MAUDRL for parallel learning of the DQN along each attribute g_i, which significantly reduces the computation time in complex multi-dimensional problems. In the exploitation stage, we calculate the partial utility for each action-state $u_i^t(a_j^t, s_l^t)$ and policy $u_i[\pi(\mathcal{A}, S)]$ based on the decision-agent's preferences and risk attitude. Then, we apply a MAUT aggregation model (e.g., additive, multiplicative or hybrid) to calculate an aggregated utility $U[\pi(\mathcal{A}, S)]$. We develop the MAUDRL model in Algorithm 1.

Algorithm 1: MAUDRL

Input : $(\mathcal{A}, S, \mathcal{R}, \mathcal{P}, \mathcal{G}, \Psi, \rho)$

for i=1:I **do**
  PROCEDURE LEARNING-DQN$_i$ $(\mathcal{A}, S, \mathcal{R}, \mathcal{P}, \mathcal{G})$
    $Q_i(S, A) \leftarrow$ Calculate Q-values of all the state-actions (S, A) for each attribute g_i
  return $Q_i(S, A)$
  end PROCEDURE
end
PROCEDURE MAUT $(Q_i(S, A),\ \forall g_i \in \mathcal{G})$
for t=1:T **do**
  SUBPROCEDURE NORMALIZATION-DQN$_i$ $(Q_i(S^t, \mathcal{A}^t))$
    $\overline{Q}_i(S^t, \mathcal{A}^t) \leftarrow$ Normalize Q-values regarding attribute g_i at time period t
  return $\overline{Q}_i(S^t, \mathcal{A}^t)$
  end SUBPROCEDURE
  SUBPROCEDURE PARTIAL-UTILITY$_i$ $(\overline{Q}_i(S^t, A^t), \rho)$
    $u_i(\overline{Q}_i(S^t, \mathcal{A}^t)) \leftarrow$ Calculate partial utility of Q-values for attribute g_i at time period t considering risk attitude coefficient ρ of the agent
  return $u_i(\overline{Q}_i(S^t, \mathcal{A}^t))$
  end SUBPROCEDURE
  SUBPROCEDURE TOTAL-UTILITY $(u_i(\overline{Q}_i(S^t, \mathcal{A}^t)), \Psi)$
    $U^t(S^t, \mathcal{A}^t) \leftarrow$ Calculate total utility for each state-action at time period t
  return $U(S^t, \mathcal{A}^t)$
  end SUBPROCEDURE
end
return $U(S, \mathcal{A})$
end PROCEDURE
PROCEDURE BEAM-SEARCH $(U(S, \mathcal{A}), \kappa)$
  $\{\pi^*_{\mathcal{G}}(\mathcal{A}|S)\}_\kappa \leftarrow$ Select the best κ policies
return $\{\pi^*_{\mathcal{G}}(\mathcal{A}|S)\}_\kappa$
end PROCEDURE

Output: $\{\pi^*_{\mathcal{G}}(\mathcal{A}|S)\}_\kappa$

We evaluate the effectiveness of MAUDRL through the lens of a sustainable HRP problem on a mid-sized farm (30 acres) spanning a pruning horizon of eight weeks. With three tiers of pruner skills (beginner, intermediate, and advanced) and a maximum of 12 pruners employable simultaneously, the SMCD problem formulation gives rise to $15.62 * 10^3$ potential actions at every state, and $3.55 * 10^{33}$ possible policy paths. The sheer scale of these paths renders the problem virtually unsolvable, particularly when factoring in each farmer's varying decision-making preferences and risk attitudes (or personas). Nevertheless, MAUDRL effectively solves this problem within a manageable computation time and consistently converges on optimal policies in line with each farmer's persona. These significant findings underscore the benefits of fusing MCDA and DRL to create multi-dimensional, explainable AI methods capable of navigating intricate SMCD problems. Furthermore, our study highlights the role of sustainability in enhancing the quality and diversity of HRP strategies in farming.

4.1. DQN learning procedure for each attribute g_i

We use LEARNING-DQN procedure to learn the best decision policy when considering a single reward function along the attribute g_i. We calculate the Q-values of all the state-actions along each attribute at each time period t using a DQN-learning algorithm, which obtains an optimal policy. The DQN-learning algorithm repeatedly updates the Q-values using the update rule below to find the best policy as follows:

$$Q_i(s^t, a^t) \leftarrow (1-\alpha)Q_i(s^t, a^t) + \alpha\left(r^t(g_i) + \gamma\ Max_a Q_i(s^{t+1}, a)\right), \quad \forall t = 1, \ldots, T \tag{1}$$

where the learning rate is shown by α, the obtained reward when taking action a^t at state s^t is shown by $r^t(g_i)$, and $Max_a Q_i(s^{t+1}, a)$ is the maximum reward that can be obtained at state s^{t+1}. An ϵ-greedy algorithm is used to choose an action at each state. After calculating the Q-table, the best policy can be obtained using Eq. (2). A thorough review of RL is proposed in the seminal paper by Sutton and Barto (2018).

$$\pi^*_i(\mathcal{A}|S) = arg\ Max_a\ Q^*_i(S, \mathcal{A}) \tag{2}$$

While RL usually offers a viable solution for simple sequential problems, as problem dimensionality increases, the number of actions and states grows exponentially, and it no longer becomes feasible (or sometimes impossible) to handle all the potential state-actions. High memory consumption is one of the major drawbacks to using DQN-learning for highly dimensional problems, which makes it impossible to fill out the Q-table accurately in a reasonable time. A way to tackle this problem is to use function approximators such as deep Q-networks (DQNs), which leverages neural networks to approximate and solve complex sequential decision problems and multi-dimensional problems where traditional Q-learning and dynamic programming methods are insufficient (Gijsbrechts et al., 2022). When facing states which have never been encountered before, the approximate model can generalize the corresponding state using previous encounters with similar states. Approximation models use parameterized functions to estimate the value for a given input state. These models eliminate the need for tables that record the actual value of each state–action pair. By replacing the tables with approximation functions, we can decrease representational capacity, thus reducing the need for large computational resources. By taking examples from a value function, function approximation methods strive to generalize them to construct an approximation of the entire function (Sutton and Barto, 2018).

DRL-based problems can be addressed using OpenAI Gym, an open-source Python library designed for RL algorithm development, where various customized environments are either already available or can be developed (Brockman et al., 2016). We create the environment in OpenAI Gym, then we train our models using Stable Baselines 3 (SB3), which is a set of reliable PyTorch implementations of RL environments (Raffin et al., 2021). The pseudo-code developed for the DQNs associated with each attribute considered in sustainable farming can be found in the Appendix.

In DQN, a neural network is used to approximate the Q-value function, $Q(s, a)$, which represents the expected cumulative reward for taking action a^t in state s^t and then following the optimal policy (Mnih et al., 2015) This use of neural networks allows DQN to generalize across different states, making it possible to efficiently estimate Q-values even for states that have not been previously encountered.

A neural network is a parameterized non-linear function constructed from layers of computational units, or neurons. Each layer first applies an affine transformation of the form $xW + b$, then a non-linear activation function. In this work, we use the rectified linear unit (ReLU), $f(x) = \max(0, x)$, which allows the network to represent complex mappings that cannot be captured by linear transformations alone. The network parameters θ are optimized during training by minimizing a loss function through gradient descent with backpropagation. Once trained, inference consists of a forward pass that maps input states to outputs.

In DQN specifically, the neural network approximates a mapping $f_\theta : \mathcal{S} \rightarrow \mathbb{R}^J$, where $\mathcal{S}$ represents the state space, J denotes the number of possible actions, and θ corresponds to the network parameters. Given an input state, the network produces a vector of Q-values, with each entry corresponding to a specific action. This output provides the agent with the necessary information to select actions that maximize the expected cumulative reward.

In this implementation, the neural network is structured as a multi-layer perceptron (MLP) consisting of two fully connected hidden layers with 64 neurons each. The hidden layers employ ReLU activations, while the output layer is linear and produces one Q-value for each

possible action. During training, the network parameters are optimized to minimize the discrepancy between the predicted Q-values and the target values computed via the Bellman equation. The learning rate is set to 0.0001, and exploration is guided by an ϵ-greedy strategy in which ϵ decays from 1.0 to 0.001 over 80% of the training steps. The model is trained for 400,000 timesteps, with performance evaluated through episodic reward. Through this iterative process, the network progressively improves its approximation of the true Q-function, enabling DQN to learn effective policies in environments where tabular DQN-learning is computationally impractical.

4.2. Calculation of partial utility and total utility functions using MAUT procedure

Consider $Q_i(S, \mathcal{A})$ as the learned Q-function associated with the attribute g_i, which is obtained using the corresponding DQN procedure. In practice, the Q-functions generated by different DQNs are not necessarily normalized. Different methods of normalization can be applied to normalize the Q-functions within $[0, 1]$. We propose a NORMALIZATION-DQN subprocedure. We then construct a partial utility function u_i for each normalized $Q_i(S, \mathcal{A})$, represented by $\overline{Q}_i(S, \mathcal{A})$. We propose a PARTIAL-UTILITY subprocedure to build the partial utility functions in accordance with the utility theory axioms for utility functions (Quirk and Saposnik, 1962). Specifically, in this study, the normalization rescales the Q-values to a unit range through a min–max transformation:

$$\overline{Q}_i(s^t, a^t) = \frac{Q_i(s^t, a^t) - Q_i^{\min}}{Q_i^{\max} - Q_i^{\min}}, \tag{3}$$

where $Q_i^{\max}$ and $Q_i^{\min}$ represent, respectively, the best and worst observed Q-values for attribute g_i. This step ensures that all value functions are comparable across heterogeneous dimensions and remain bounded in $[0, 1]$. Next, the PARTIAL-UTILITY subprocedure converts each normalized $\overline{Q}_i(s^t, a^t)$ into a partial utility u_i that reflects the decision-agent's risk preference. In this study, the formulation proposed by Wood and Khosravanian (2015) is used to construct each partial utility function on normalized Q-values as follows:

$$u_i(\overline{Q}_i(s_l^t, a_j^t)) = \frac{e^{-\rho \overline{Q}_i^{worst}(S^t, \mathcal{A}^t)}}{e^{-\rho \overline{Q}_i^{worst}(S^t, \mathcal{A}^t)} - e^{-\rho \overline{Q}_i^{best}(S^t, \mathcal{A}^t)}} - \frac{e^{-\rho \overline{Q}_i(s_l^t, a_j^t)}}{e^{-\rho \overline{Q}_i^{worst}(S^t, \mathcal{A}^t)} - e^{-\rho \overline{Q}_i^{best}(S^t, \mathcal{A}^t)}} \tag{4}$$

where $\overline{Q}_i^{worst}(S^t, \mathcal{A}^t)$ and $\overline{Q}_i^{best}(S^t, \mathcal{A}^t)$, respectively, show the worst and best amounts of normalized Q-values constructed on attribute g_i at time period t. In addition, ρ represents the risk attitude coefficient of the farmer. The more positive the amount of ρ, the more the risk-averse the farmer. The more the negative the amount of ρ, the more risk-prone the farmer. When $\rho = 0$ means that the farmer tends to be risk-neutral, where Eq. (12) does not apply and the utility of risk-neutral farmer can be simply calculated as $u_i(\overline{Q}_i(s_l^t, a_j^t)) = \overline{Q}_i(s_l^t, a_j^t)$. Finally, we obtain the multi-attribute utility of all actions-states with respect to each attribute using a TOTAL-UTILITY subprocedure. For instance, the multi-attribute utility of agent for taking action a_j^t at state s_l^t can be formulated as:

$$U^t(s_l^t, a_j^t) = U^t\Big(u_1(\overline{Q}_1(s_l^t, a_j^t)), \ldots, u_i(\overline{Q}_i(s_l^t, a_j^t)), \ldots, u_I(\overline{Q}_I(s_l^t, a_j^t))\Big) \tag{5}$$

where $u_i(s_l^t, a_j^t)$ represents the utility of action a_j^t at state s_l^t with respect to attribute $g_i \in \mathcal{G}$. Different MAUT aggregation models can be adopted in order to calculate the total utility function depending the preference systems and risk attitude of the farmer.

4.3. Beam search procedure

The total utility obtained in Eq. (5) is used to choose the best policy at each time period considering multiple attributes. Considering the problem's sequential nature, we need to acquire the best policy over the whole planning horizon. Moreover, we need to ensure that the aggregated policy is feasible after aggregating the utility values. To do so, we first filter all actions that are not feasible at a given state and time period. Then, we apply the beam search technique using a BEAM-SEARCH procedure, which searches decision trees in order to obtain the best policy when considering all attributes (Sabuncuoglu and Bayiz, 1999; Akeb et al., 2009). A beam search systematically keeps a set of κ solutions for each time period in order to find a good policy with minimal search efforts (Ow and Morton, 1988). When searching the tree, a beam search chooses a subset of actions with a reasonably high aggregated utility value for further branching at each level. This approach ultimately leads to the best, or close to the best policy, and discards the other actions. The beam search is provided in algorithm 2. The broad set of actions in our model is evaluated using the utility value of these actions, which are influenced by the coefficients of the different attributes. We set the beam width to $\kappa = 7$. At each period $t = 1, \ldots, T$, starting from the current state, we consider a_t actions, apply feasibility *pre-aggregation*, and discard those that violate applicability/availability or resource/capacity limits (e.g., exceeding the budget). From the remaining feasible set, we compute aggregated utilities and retain the top $\kappa = 7$ successors for the next period.

Algorithm 2: Beam search

input : $(U^1(S, \mathcal{A}), \kappa)$

b ← 0 (variable to count the number of found policies)
t ← 1
// Initializing κ policies by picking best κ actions
$\tilde{U}^t(S, A) \leftarrow$ *sort* $U^t(S, \mathcal{A})$ *in decreasing order*
$Y^1_{\mathcal{G}}(\mathcal{A}|S)_\kappa \leftarrow$ *select the best* κ *actions corresponding to the highest* $\tilde{U}^t(S, A)$ $t \leftarrow t + 1$
// Selecting the best κ actions for each $Y_{\mathcal{G}}(\mathcal{A}|S)_\kappa$
for *d=1:*κ **do**
  $U^t_d(S, A) \leftarrow$ *MAUT algorithm for each* $Y_{\mathcal{G}}(\mathcal{A}|S)_d$
  $\tilde{U}^t_d(S, A) \leftarrow$ *sort* $U^t_d(S, \mathcal{A})$ *in decreasing order*
  $\tilde{U}^t_d(S, A) \leftarrow$ *select the best* κ *actions corresponding to the* κ *highest* $\tilde{U}^t_d(S, A)$
end
// Aggregating utilities of all the selected κ policies
$\tilde{U}^t(S, A) \leftarrow \tilde{U}^t_1(S, A) \cup \tilde{U}^t_2(S, A) \cup \ldots \cup \tilde{U}^t_\kappa(S, A)$, *getting union* $\tilde{U}^t_i$ *which leads to* κ^2 *actions*
while $b < \kappa$ **do**
  $\tilde{U}^t(S, A) \leftarrow$ *sort* $U^t(S, \mathcal{A})$ *in decreasing order*
  $Y^t_{\mathcal{G}}(\mathcal{A}|S)_\kappa \leftarrow$ *select the best* κ *actions corresponding to the* κ *highest* $\tilde{U}^t(S, A)$
  $t \leftarrow t + 1$
  for *d=1:* κ **do**
    $U^t_d(S, A) \leftarrow$ *MAUT algorithm for each* $Y_{\mathcal{G}}(\mathcal{A}|S)_d$
    $\tilde{U}^t_d(S, A) \leftarrow$ *sort* $U^t_d(S, \mathcal{A})$ *in decreasing order*
    $\tilde{U}^t_d(S, A) \leftarrow$ *select the best* κ *actions corresponding to the* κ *highest* $\tilde{U}^t_d(S, A)$
  end
  $\tilde{U}^t(S, A) \leftarrow \tilde{U}^t_1(S, A) \cup \tilde{U}^t_2(S, A) \cup \ldots \cup \tilde{U}^t_\kappa(S, A)$, *getting union* $\tilde{U}^t_i$ *which leads to* κ^2 *actions*
  // Checking if we found any policies reaching the goal
  for *z=1:*κ **do**
    if $Y^*_{\mathcal{G}}(\mathcal{A}|S)_z$ *reached objective* **then**
      $\pi^*_{\mathcal{G}}(\mathcal{A}|S)_b \leftarrow Y^*_{\mathcal{G}}(\mathcal{A}|S)_z$
      $b \leftarrow b + 1$
    end
  end
end

output: $\{\pi^*_{\mathcal{G}}(\mathcal{A}|S)\}_\kappa \leftarrow$ Set best policies

Table 1
Notations.

Indices	
t	Time period index ($t = 1, 2, \ldots, T$)
k	Index for worker's skill level $k \in \{1, 2, 3\}$, where 1, 2, and 3, respectively, represent the beginner, intermediate, and advanced levels.
Parameters	
B	Available budget for pruning
N	Farm size (acre)
λ	Number of blueberry plants per acre
h	Fixed cost of hiring a worker
f	Fixed cost of firing a worker
w_k	Wage per hour for worker with skill level k
δ	Working hours per time period
q_k	Quality of pruned plants by worker with skill level k
$\tilde{\xi}_k$	Number of plants pruned by worker with skill level k per hour, which follows a normal distribution
$\tilde{\theta}_k$	Count of workers possessing skill level k in the labour pool, which follows a normal distribution
Θ	Limit of the number of employed workers
Decision variables	
x_k^t	Count of workers possessing skill level k to be hired/fired at the beginning of period t. When $x_k^t > 0$, the farmer hires x_k^t workers and when $x_k^t < 0$, the farmer fires $-x_k^t$ workers with skill level k at time period t

5. Application of MAUDRL to sustainable farming

5.1. Problem formulation

The sustainable human resource planning (HRP) challenge in farming serves as an excellent representation of SMCD problems. This problem is predicated on decisions related to hiring and firing agricultural workers while considering the economic, social and environmental consequences of these decisions. In this section, we focus on managing pruning workers during the blueberry pruning season, an activity that is critical for controlling insects and diseases, and thus maintaining blueberry quality, size, and yields (BC-Government, 2022). The notations are given in Table 1.

The blueberry farm is divided into a certain number of blocks, each containing a homogeneous set of blueberry plants. We assume the farm is divided in N one-acre blocks. The sustainable sequential HRP process is executed by a farmer who decides to hire or fire farming workers (pruners) over the pruning season T to achieve his goals. The working hours δ during each time period are fixed. Drawing from the HRP literature (Bordoloi and Matsuo, 2001), We classify pruners based on their skill level k, where $(k = 1)$, $(k = 2)$, and $(k = 3)$ correspond to beginner, intermediate, and advanced, respectively. The pruning quantity $\tilde{\xi}_k$ and quality q_k differ among workers with different skill levels. For example, higher skilled workers are more careful pruners, and thus contribute to regenerative farming while less skilled workers may damage the plants, contributing to negative environmental impacts. However, skilled pruners are more expensive and use up more of the farmer's budget. To be socially responsible, a farmer may choose to hire more people and keep them for a longer period to positively contribute to their community (i.e., maximizing job creation and minimizing turnover).

The farmer pays wages, but also incurs hiring and firing costs. In practice, farmers usually allocate a specific budget B for the pruning season. In British Columbia, farm workers are usually employed on an hourly rate. More skilled pruners are paid a higher hourly rate, with the hourly rate for each skill level denoted by w_k. The unit costs for hiring and firing each worker are h and f, respectively.

The pruning season usually coincides with a blueberry plant's dormant period, which can be divided into a set of time periods (e.g., weeks) denoted by t, where $t = 1, 2, \ldots, T$. In the sequential HRP setting, the farmer decides to hire and fire pruners with different skill levels at the start of each time period t, and then observes the consequences of his decisions as they are carried out by pruners (actions) on the farm as illustrated in Fig. 3. The farmer observes the current state (i.e., farm, workforce, budget and outcomes) at the start of each time period t, and decides on the number of pruners to hire or fire, based on their skill levels. Then, the farmer evaluates the results of his decisions by assessing the corresponding rewards. In this study, we consider economic, social and environmental rewards, which are measured by pruning efficiency, pruning quality, and job creation, respectively. The farmer learns from previous periods and continuously improves his hiring and firing decisions over time. The decisions made over the pruning season is thus called a human resource planning policy. The farmer's objective is to learn the best sustainable policy, which is a rule (i.e., system of rules) used to decide what actions to take next based on the state observations and set of rewards obtained from the environment. The sustainable sequential HRP problem in the blueberry sector is a SMCD problem because the attributes associated with each dimension of sustainability are conflicting and non-commensurable.

State representation. The state space $\mathcal{S}$ includes all the possible states during the pruning season. The state at time step t is represented by a vector $s_t = [m_1^t, m_2^t, m_3^t, z_B^t, z_P^t] \in \mathcal{S}$, where m_1^t, m_2^t, m_3^t are the number of beginner, intermediate, and advanced pruners working on the farm at the beginning of period t before hiring or firing. At the beginning of the pruning season, these values are set to zero. z_B^t is a binary variable representing running out budget. z_P^t is a binary variable representing the completion of pruning all plants. Over the pruning season, the states are updated repeatedly at every time step, considering the selected action and the feedback from the environment.

Action representation. The action space $\mathcal{A}$ includes all possible decisions the farmer can make at each time step t. We represent the decision at t as a vector $a_t = [x_1^t, x_2^t, x_3^t] \in \mathcal{A}$, where x_1^t, x_2^t, x_3^t are the variations in the number of pruners working on the farm. If $x_k^t > 0$, then x_k^t shows the number of hires and if $x_k^t < 0$, then $-x_k^t$ shows the number of fired workers with skill level k. $x_k^t = 0$ when the farmer neither hires nor fires workers. In the agricultural sector, farmers typically face hiring challenges (Fletcher, 2020). Accordingly, we assume the available number of pruners for each skill level at each time period to be uncertain and constrained ($\tilde{\theta}_k$). Moreover, the maximum number of pruners working at the same time is capped (Θ). The number of firings at each time period cannot exceed the total number of pruners from the previous time period. Thus, the number of hirings and firings at each state and time period t is restricted as given in Eqs. (6)–(8).

$$[-x_k^t]^+ \le m_k^t, k = 1, 2, 3 \tag{6}$$

$$\sum_{k=1}^{3} (m_k^t + x_k^t) \le \Theta \tag{7}$$

$$[x_k^t]^+ \le \tilde{\theta}_k \ , \ k = 1, 2, 3 \tag{8}$$

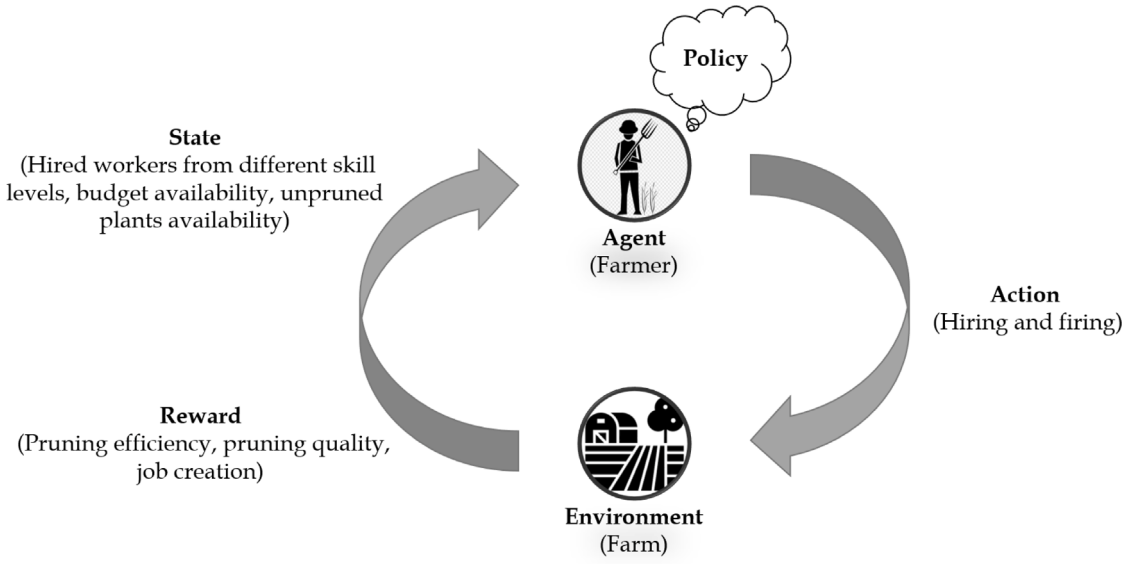

Fig. 3. Problem: Farmer's action and learning process.

where $[.]^+ = Max(0,.)$ and $\tilde{\theta}_k$ represents the availability pools of pruners for each skill level k.

Reward functions. The farmer evaluates the consequence of his decision-action from different perspectives. In this study, we consider a multi-dimensional reward function $r^t = \{r^t_I, r^t_\Phi, r^t_\Omega\}$ representing economic, environmental, and social attributes, as follows.

Economically-viable pruning: We use pruning efficiency as a proxy to represent the economic dimension of sustainability. This attribute characterizes the farmer's natural tendency to prune as many plants as he can at the lowest cost, which is given by Eq. (9).

$$r^t_I = \frac{\sum_{k=1}^{3}(m^t_k + x^t_k)\tilde{\xi}_k}{h\sum_{k=1}^{3}[x^t_k]^+ + f\sum_{k=1}^{3}[-x^t_k]^+ + \sum_{k=1}^{3} w_k(m^t_k + x^t_k)} \tag{9}$$

where r^t_I is obtained by dividing the number of pruned plants by budget spent at time period t. The number of pruned plants $\sum_{k=1}^{3}(m^t_k + x^t_k)\tilde{\xi}_k$ is obtained based on the number of workers and their productivity. The budget spent $h\sum_{k=1}^{3}[x^t_k]^+ + f\sum_{k=1}^{3}[-x^t_k]^+ + \sum_{k=1}^{3} w_k(m^t_k + x^t_k)$ includes the hiring and firing costs, and the wages paid to the pruners.

Environmentally-responsible pruning: We use the quality of pruning as a proxy for the environmental dimension of sustainability. Previous studies established the effects of pruning quality and intensity on the quality and yield of blueberry fruits and plants (Strik and Davis, 2022). We use Eq. (10) to calculate the environmental reward at each time period t.

$$r^t_\Phi = \frac{\sum_{k=1}^{3} q_k\tilde{\xi}_k(m^t_k + x^t_k)}{q_3\sum_{k=1}^{3}\tilde{\xi}_k(m^t_k + x^t_k)} \tag{10}$$

where $\sum_{k=1}^{3} q_k\tilde{\xi}_k(m^t_k + x^t_k)$ represents the achieved pruning quality and $q_3\sum_{k=1}^{3}\tilde{\xi}_k(m^t_k + x^t_k)$ represents the highest possible quality at time period t.

Socially-responsible pruning: We use job creation and turnover as a proxy for the social dimension of sustainability. Job creation and turnover have been widely used as a measure of social responsibility in the agriculture sector (Chávez et al., 2018; Allaoui et al., 2018). At each time period t, the farmer measures their socially-responsible attribute as follows:

$$r^t_\Omega = \frac{\sum_{k=1}^{3}(m^t_k + x^t_k)}{\Theta} \tag{11}$$

State transitions. At the start of each state, the farmer assesses the current state $s_t = [m^t_1, m^t_2, m^t_3, z^t_B, z^t_P] \in S$. Assuming there are still unpruned plants ($i.e.$, $z^t_P = False$) and available pruning budget ($i.e.$, $z^t_B = False$), the farmer decides on the hiring or firing of beginner x^t_1, intermediate x^t_2, and advanced x^t_3, level workers. This action is then assessed against three constraints: (1) the maximum workforce size, (2) the maximum firing limit, and (3) the available workforce to be hired. If the chosen action fails any constraint, the farmer receives a negative reward, ending the current episode. If not, pruning begins. At the end of each state, the farmer evaluates the economic, environmental, and social rewards (based on Eqs. (9), (10), and (11)) that depend on the workers' productivity (which is uncertain) and pruning results. If the budget expended exceeds the available amount, the farmer receives a negative reward, and z^{t+1}_B is set to True. Similarly, z^{t+1}_P is set to True if all the plants have been pruned by the end of the state. The next state s^{t+1} will be obtained based on z^{t+1}_B, z^{t+1}_P, and the number of workers are set to $m^{t+1}_1 = m^t_1 + x^t_1$, $m^{t+1}_2 = m^t_2 + x^t_2$, and $m^{t+1}_3 = m^t_3 + x^t_3$. For a detailed breakdown of state transitions and reward calculations, refer to Algorithms 6, 7, and 8 provided in the online companion.

5.2. Numerical example

We investigate the MAUDRL algorithm's performance using data from a 30-acre blueberry farm. The data for this application is given in Table 2. In the province of British Columbia, Canada, standard blueberry plant spacing is 3 ft. × 10 ft., and there are typically around 1450 plants per acre (Building business success, 2016). Each acre is referred to as a "block". We limit the number of time periods to $T = 8$ weeks, as the pruning season is usually limited to two months. Pruners are classified into three main skill level groups: beginner, intermediate, and advanced based on the quality and speed of their work. The maximum number of workers that can simultaneously work on the farm is limited to 12 due to operational constraints. We consider five working days per week with eights hours per day. The number of beginner, intermediate, and advanced workers in the labour pool follows normal distributions: $N_1(20, 3)$, $N_2(15, 2)$, and $N_3(12, 1)$, respectively. According to blueberry farmers, typical beginner, intermediate, and advanced agricultural workers usually prune 350, 550, and 750 blueberry plants per week, which is roughly equivalent to 24.13, 37.93, and 51.72 percentage of a block, respectively. Typical productivity rates, wages, and pruning quality (Moura et al., 2017) corresponding to each skill level along with the unit hiring and firing costs are provided in Table 2.

Table 2
Input data for the blueberry pruning HRP problem.

Parameter	Value
Farm size	30 acres
Number of plants	1450 per acre
Planning season	2 months (8 weeks)
Decision steps	Every week
Skill Levels	3 (1: beginner, 2: intermediary, 3: advanced)
Productivity rate for each skill level (block/week)	$N_1(0.241, 0.034)$, $N_2(0.379, 0.026)$, *and* $N_3(0.517, 0.017)$
Quality rate for each skill level	1: 506.88, 2: 790.76, and 3: 1074.10
Wage per worker for each skill level ($/week)	1: 626, 2: 800, and 3: 1200
Hiring cost per worker ($ per worker)	750
Firing cost per worker ($ per worker)	450
Worker limit on the farmland	12
Pool of workers	$N_1(20, 3)$, $N_2(15, 2)$, and $N_3(12, 1)$
Budget ($)	100,000

5.3. Specification of MAUDRL in the context of sustainable blueberry farming

We provide in Algorithm 3 the specification of MAUDRL for the sustainable sequential HRP problem in the context of blueberry farming. The farmer's multi-attribute utility function can be formulated as $U(S,\mathcal{A}) = U\big(u_\Gamma(\overline{Q}_\Gamma(S,\mathcal{A})), u_\Phi(\overline{Q}_\Phi(S,\mathcal{A})), u_\Omega(\overline{Q}_\Omega(S,\mathcal{A}))\big)$, where Γ, Φ, and Ω represent the economic, environmental, and social attributes, respectively. The formulation proposed by Wood and Khosravanian (2015) is used to represent the preferences and risk attitudes of different individual farmers to construct each partial utility function on normalized Q-values as follows:

$$u_i(\overline{Q}_i(s_l^t, a_j^t)) = \frac{e^{-\rho\overline{Q}_i^{worst}(S^t,\mathcal{A}^t)}}{e^{-\rho\overline{Q}_i^{worst}(S^t,\mathcal{A}^t)} - e^{-\rho\overline{Q}_i^{best}(S^t,\mathcal{A}^t)}} - \frac{e^{-\rho\overline{Q}_i(s_l^t,a_j^t)}}{e^{-\rho\overline{Q}_i^{worst}(S^t,\mathcal{A}^t)} - e^{-\rho\overline{Q}_i^{best}(S^t,\mathcal{A}^t)}} \tag{12}$$

where $\overline{Q}_i^{worst}(S^t,\mathcal{A}^t)$ and $\overline{Q}_i^{best}(S^t,\mathcal{A}^t)$, respectively, show the worst and best amounts of normalized Q-values constructed on attribute g_i at time period t. In addition, ρ represents the risk attitude coefficient of the farmer. The more positive the amount of ρ, the more the risk-averse the farmer. The more the negative the amount of ρ, the more risk-prone the farmer. When $\rho = 0$ means that the farmer tends to be risk-neutral, where Eq. (12) does not apply and the utility of risk-neutral farmer can be simply calculated as $u_i(\overline{Q}_i(s_l^t, a_j^t)) = \overline{Q}_i(s_l^t, a_j^t)$.

5.4. Total utility

In this application, the preferences of the decision-agents along each attribute are mutually independent. We, therefore, use a simple MAUT additive utility aggregation model (Keeney et al., 1993) as shown in Eq. (13). We can estimate these elasticity coefficients using elicitation procedures based on comparisons of payoff lotteries (Keeney et al., 1993).

$$U(S,\mathcal{A}) = \psi_\Gamma.u_\Gamma(\overline{Q}_\Gamma(S,\mathcal{A})) + \psi_\Phi.u_\Phi(\overline{Q}_\Phi(S,\mathcal{A})) + \psi_\Omega.u_\Omega(\overline{Q}_\Omega(S,\mathcal{A})) \tag{13}$$

where $\psi_\Gamma + \psi_\Phi + \psi_\Omega = 1$. The parameters ψ_Γ, ψ_Φ, and ψ_Ω are elasticity coefficients associated with the pruning efficiency, pruning quality, and job creation attributes, respectively.

5.5. Fine tuning of DQN hyperparameters

In the MAUDRL model, the hyperparameters associated with the DQN models need to be adjusted. We trained and evaluated our models with various sets of hyperparameters while tracking their performances. The most effective collection of hyperparameters was then selected, as explained below.

To tackle the exploration–exploitation trade-off when choosing actions, the initial exploration rate is set to 1 so that the model can first fully explore the environment. Then the exploration rate decreases based on the exploration fraction value, which is the proportion of the total training time during which the exploration rate is annealed. Based on our experiments, the best exploration fraction value for our

Algorithm 3: MAUDRL for sustainable farming

Input : $(\mathcal{A}, S, \mathcal{R}, \mathcal{P}, \mathcal{G}, \Psi, \rho)$

PROCEDURE LEARNING-DQN$_i$ $(\mathcal{A}, S, \mathcal{R}, \mathcal{P}, \mathcal{G})$
  $Q_\Gamma(S,A) \leftarrow$ Calculate Q-values for the economic attribute Γ using algorithm 6
  $Q_\Phi(S,\mathcal{A}) \leftarrow$ Calculate Q-values for the environmental attribute Φ using algorithm 7
  $Q_\Omega(S,\mathcal{A}) \leftarrow$ Calculate Q-values for the social attribute Ω using algorithm 8
return $Q_\Gamma(S,\mathcal{A}), Q_\Phi(S,\mathcal{A}), Q_\Omega(S,A)$
end PROCEDURE

PROCEDURE MAUT $(Q_i(S,A),\ \forall g_i \in \mathcal{G} = \{\Gamma, \Phi, \Omega\})$
for *t=1:8* **do**
  SUBPROCEDURE NORMALIZATION-DQN$_i$ $(Q_i(S^t,\mathcal{A}^t))$
    $\overline{Q}_\Gamma(S^t,\mathcal{A}^t) \leftarrow$ Normalize Q-values for the economic attribute Γ at time period t
    $\overline{Q}_\Phi(S^t,\mathcal{A}^t) \leftarrow$ Normalize Q-values for the environmental attribute Φ at time period t
    $\overline{Q}_\Omega(S^t,\mathcal{A}^t) \leftarrow$ Normalize Q-values for the social attribute Ω at time period t
  return $\overline{Q}_\Gamma(S^t,\mathcal{A}^t), \overline{Q}_\Phi(S^t,A^t), \overline{Q}_\Omega(S^t,\mathcal{A}^t)$
  end SUBPROCEDURE
  SUBPROCEDURE PARTIAL-UTILITY$_i$ $(\overline{Q}_\Gamma(S^t,\mathcal{A}^t), \overline{Q}_\Phi(S^t,A^t), \overline{Q}_\Omega(S^t,\mathcal{A}^t), \rho)$
    $u_\Gamma(\overline{Q}_\Gamma(S^t,\mathcal{A}^t)) \leftarrow$ Calculate partial utility of Q-values for economic attribute Γ using equation (12)
    $u_\Phi(\overline{Q}_\Phi(S^t,\mathcal{A}^t)) \leftarrow$ Calculate partial utility of Q-values for environmental attribute Φ using equation (12)
    $u_\Omega(\overline{Q}_\Omega(S^t,\mathcal{A}^t)) \leftarrow$ Calculate partial utility of Q-values for social attribute Ω using equation (12)
  return $u_\Gamma(\overline{Q}_\Gamma(S^t,\mathcal{A}^t), u_\Phi(\overline{Q}_\Phi(S^t,\mathcal{A}^t))$, and $u_\Omega(\overline{Q}_\Omega(S^t,\mathcal{A}^t))$
  end SUBPROCEDURE
  SUBPROCEDURE TOTAL-UTILITY $(u_\Gamma(\overline{Q}_\Gamma(S^t,\mathcal{A}^t)), u_\Phi(\overline{Q}_\Phi(S^t,\mathcal{A}^t)), u_\Omega(\overline{Q}_\Omega(S^t,\mathcal{A}^t)), \Psi)$
    $U(S^t,\mathcal{A}^t) \leftarrow$ Calculate total utility for each state-action at time period t using equation (13)
  return $U(S^t,\mathcal{A}^t)$
  end SUBPROCEDURE
end
return $U(S,\mathcal{A})$
end SUBPROCEDURE

PROCEDURE BEAM-SEARCH $(U(S,A),\kappa)$
  $\{\pi^*_\mathcal{G}(\mathcal{A}|S)\}_\kappa \leftarrow$ Select the best κ policies using algorithm 2
return $\{\pi^*_\mathcal{G}(\mathcal{A}|S)\}_\kappa$
end PROCEDURE

Output: $\{\pi^*_\mathcal{G}(\mathcal{A}|S)\}_\kappa$

problem is 0.8, as a lower rate forces our model to exploit too early before obtaining a good policy. On the other hand, a higher value keeps the model in exploration mode for too long, when good policies have already been observed. We also set the learning rate in our model to 0.0001.

We train each of the economic, environmental, and social DQN models separately until full convergence of each model is achieved. The number of required iterations (timesteps) to fully converge is around 400,000 (Fig. 4). All three models start to converge at around 100,000 iterations. After 400,000 iterations, the increased performance in comparison to the spent time is negligible. Also, there is a penalty/bonus hyperparameter (M) considered in the MAUDRL model. When the agent reaches the goal (i.e., pruning all the plants), it receives a bonus (M) in addition to the usual reward. On the other hand, its reward is decreased by (M) when any of the constraints are violated (e.g., hiring workers over capacity).

The behaviour illustrated in Fig. 4 corresponds specifically to the training phase of the DQN model, where the agent explores the environment and learns an optimal policy. These are computational units of learning rather than real-world time periods. This process does indeed require many interactions to converge (on the order of tens of thousands of steps). However, this is an offline, one-time process that is performed during the model development stage. Once the policy has been learned, it can be easily deployed to generate Pareto-optimal solutions that capture different preferences and risk attitudes of decision makers. This deployment stage is computationally inexpensive and does not involve further training. In deployment and test time, end-users (e.g., farmers) would not repeat the learning process. Instead, they would directly query the trained model, which produces decisions at negligible computational cost (on the order of milliseconds). This separation between an offline training stage and a fast, online inference stage is standard practice in real-world machine learning systems and therefore does not limit the applicability of our approach. In summary, although training requires many steps to reach convergence, this does not impact real-world usability: farmers and practitioners interact only with the trained policy, which operates in real time and is readily applicable in practice.

The other hyperparameters of the MAUDRL model include buffer size (50,000), batch size (32) and discount factor (0.99), which are set according to the default values offered by Stable Baselines 3 (SB3) (Raffin et al., 2021). For the activation function in our neural network architecture, ReLU (Rectified Linear Unit) is used, as it is the default activation function recommended for most feed-forward neural networks (Goodfellow et al., 2016).

5.6. Results

The sustainable sequential HRP problem described above leads to 15,622 potential decision-actions at each s_i^t, with approximately potential 3.55 10^{33} policies. At each period t, we calculate Q-values for each action and then use the normalization and partial utility procedures to calculate their corresponding partial utility values. For example, Fig. 5 shows the partial utility values for actions at $t = 1$ along each attribute. Fig. 6 shows partial utility values for ordered Q-values from low to high at period $t = 5$. These figures illustrate different decision-making risk attitudes. The results show the ability of MAUDRL to explain the results based on the decision-agent's preference profile.

MAUDRL's efficiency is evidenced by its ability to determine policies within a very short computational timeframe. On average, training each of our DQN models over 400,000 iterations takes approximately $2927s$. Importantly, we only need to conduct this training once, in parallel, after which the trained models can be employed repeatedly to identify the optimal policy based on decision-making preferences and risk attitudes. The MAUT process is remarkably swift, taking around 0.12 s to execute. These results underscore MAUDRL's performance in calculating the most suitable policy for each case within an impressively short computational run time. Furthermore, our approach facilitates sensitivity analysis and offers a straightforward path to converge on the most relevant policy, responsive to changing decision-maker preferences (e.g., modifications in investment strategies or human resource management policies).

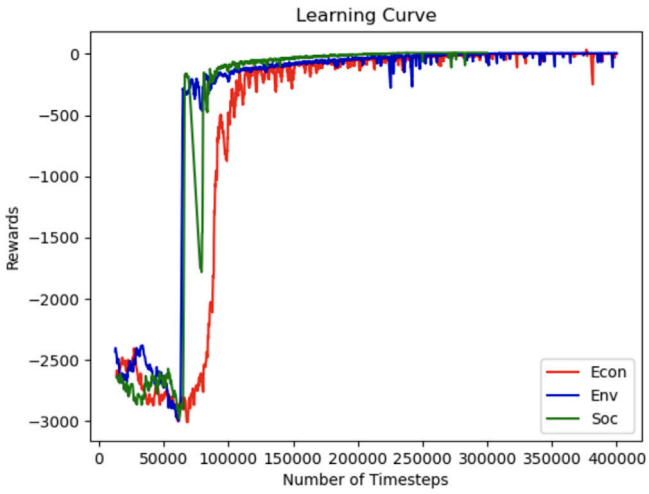

Fig. 4. Average expected rewards over simulation timesteps (iterations) during the offline training of the Economic (Econ), Environmental (Env), and Social (Soc) DQN models.

5.7. Benchmarking

We compare the performance of MAUDRL against two benchmark algorithms: Benchmark 1 and Benchmark 2. We note that we choose these two benchmarking algorithms that are salient to the prevailing research on multiple criteria decision analysis (MCDA), and not DRL. This is because our contribution is to the scholarly tradition around SMCD, incorporating DRL as a candidate addendum solution to those problems, and not vice versa. In other words, the novelty of our paper lies not in advancing DRL algorithms themselves, but in developing a framework that applies DRL to individual criteria within SMCD problems to address their attendant complexities. Traditional MCDA methods struggle with dimensionality in sequential settings, where an array of criteria must be balanced over multiple time periods, and in our context, under uncertainty and risk preferences. MAUDRL decomposes SMCD problems by training separate DRL agents for each criterion, which not only makes large-scale SMCD problems more computationally tractable by enabling parallel training, but also enhances explainability, a persistent challenge in the deep learning field. This process repeats across time periods, allowing for the inclusion of dynamic adaptation in solving SMCD problems.

Our study uses DQN, which serves as a specific DRL method to solve the decomposed subproblems. This is not our central contribution. We selected DQN because it is a well-established technique that has demonstrated stability and effectiveness across numerous applications since its introduction. Thus, using DQN, a foundational method in DRL, allows for a reliable baseline for illustrating our integration of DRL into SMCD problems. Incidentally, we fully acknowledge the evolution of DRL beyond DQN with several modern techniques addressing small samples (e.g., Rainbow DQN (Hessel et al., 2018) for sample efficiency), overestimation bias (e.g., Twin Delayed Deep Deterministic Policy Gradient (Fujimoto et al., 2018) for incorporating double Q-learning and taking the minimum of two Q-value estimates thus countering overestimation from function approximations), or exploration–exploitation trade-offs (e.g., Proximal Policy Optimization (Schulman et al., 2017)

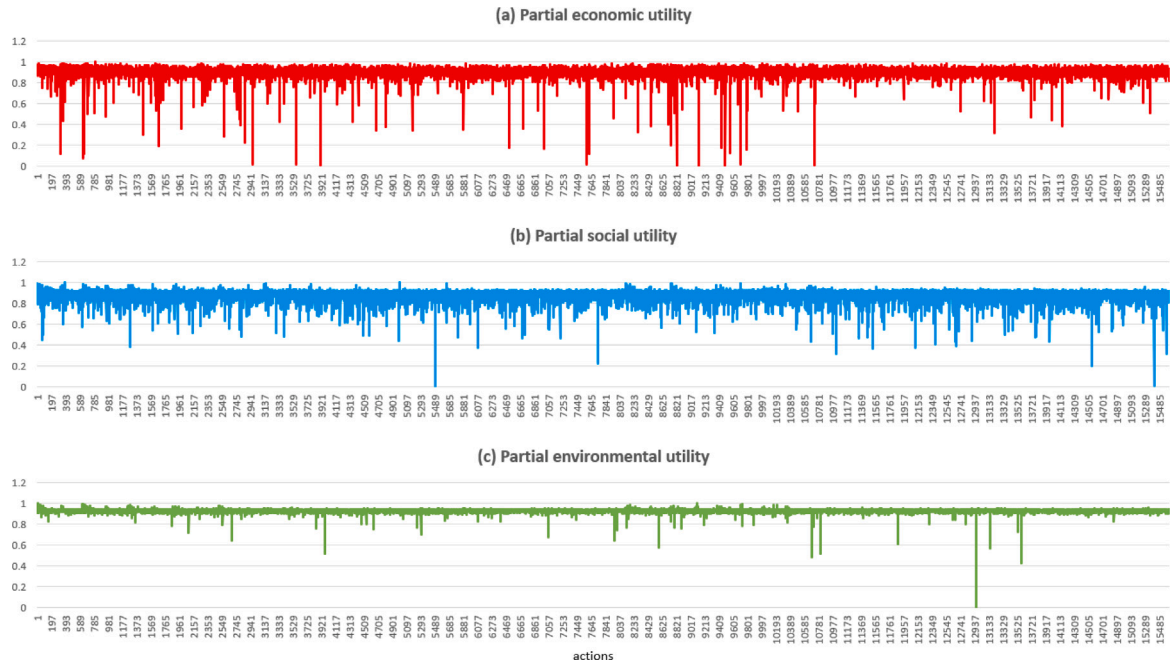

Fig. 5. The partial utility value for each decision-action at a given time period $t = 1$: **(a)** partial economic utility values, **(b)** partial social utility values, **(c)** partial environmental utility values.

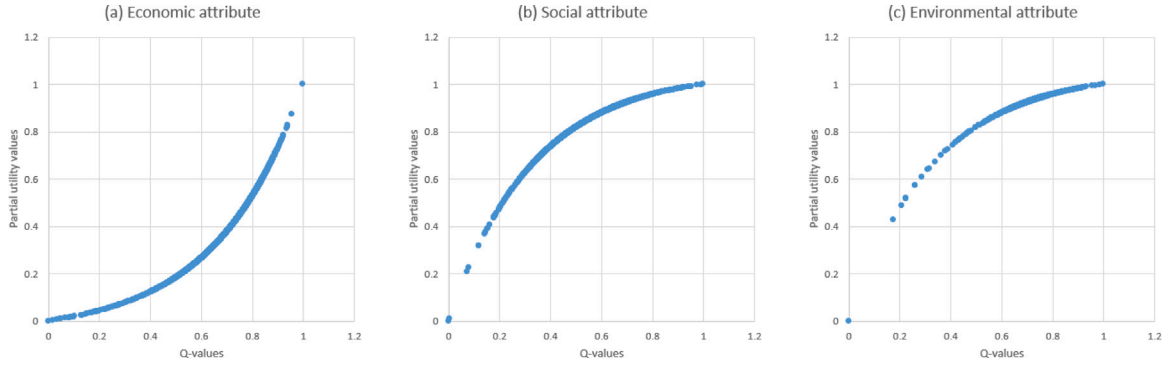

Fig. 6. Illustration of mapping the partial utility functions and Q-values at period $t = 5$ for different decision-agent preference profiles: **(a)** partial economic utility values for risk-prone farmer, and **(b)** partial social, and **(c)** environmental utility values for risk-prone farmer constructed on Q-values for each action.

for clipping policy ratios to allow for controlled exploration) in single objective settings.

These modern DRL methods can certainly be valuable for optimizing DRL in isolation, but they are orthogonal to the core contribution of our work. MAUDRL is modular by design, and the DRL method chosen is replaceable and substitutable with any of the aforementioned modern DRL methods. Choosing to benchmark against individual DRL techniques would only highlight differences in subproblem solving efficiency but would not be relevant to our effort in addressing tractability and scalability issues in complex SMCD problems, nor would it improve explainability given that MAUDRL's novelty in decomposing criterion into individual training regimes would not be affected. Despite our choice not to benchmark against individual DRL methods, extending benchmarks to include advanced DRL techniques could be a promising avenue for future research, particularly context-specific ones where observing marginal performance gains in specific scenarios are of high interest.

Benchmark 1 (Oracle Discrete MAUT): Benchmark 1 employs an oracle policy, detailed in algorithm 4. Creating a straightforward policy that outperforms random behaviour is often possible in many settings using methods other than RL, where one can manually program a policy to exhibit basic behaviour which is termed an oracle policy (Tai et al., 2022). If the farmer can adopt an oracle policy based on total availability of the information as opposed to relying on taking actions under uncertain environments, it would, in theory, improve the farmer's performance. Thus, the oracle policy under such a hypothetical deterministic setting can be considered as a quasi upper benchmark. In this case, the farmer makes decisions at each state based on the expected value of the information at hand, considering the aggregated economic, environmental, and social utility values. The most favourable action is the one which has the greatest total utility value for that state. The CALCULATING-Q procedure in the algorithm supersedes the LEARNING-DQN procedure, while the BEAM-SEARCH procedure is omitted. Notably, the CALCULATING-Q procedure merely computes the rewards linked to each criterion without learning.

Benchmark 2 (Single Reward Aggregation Approach): We propose a second benchmark algorithm based on MCDA single criterion aggregation approach (Guitouni and Martel, 1998), detailed in algorithm 5. In this approach, DRL uses the MAUT aggregated utility to calculate the reward function at each state to determine the best action. This

Algorithm 4: Benchmark 1 algorithm: Oracle Discrete MAUT

Input : $(\mathcal{A}, S, \mathcal{R}, \mathcal{P}, \mathcal{G}, \Psi, \rho)$

for i=1:I **do**
  PROCEDURE CALCULATING-Q$_i$ $(\mathcal{A}, S, \mathcal{R}, \mathcal{P}, \mathcal{G})$
    $Q_i(S, A) \leftarrow$ Calculate Q-values of all the state-actions (S, A) for each attribute g_i
  return $Q_i(S, A)$
  end PROCEDURE
end
PROCEDURE MAUT $(Q_i(S, A),\ \forall g_i \in \mathcal{G})$
for t=1:T **do**
  SUBPROCEDURE NORMALIZATION-Q$_i$ $(Q_i(S^t, \mathcal{A}^t))$
    $\overline{Q}_i(S^t, \mathcal{A}^t) \leftarrow$ Normalize Q-values regarding attribute g_i at time period t
  return $\overline{Q}_i(S^t, \mathcal{A}^t)$
  end SUBPROCEDURE
  SUBPROCEDURE PARTIAL-UTILITY$_i$ $(\overline{Q}_i(S^t, A^t), \rho)$
    $u_i(\overline{Q}_i(S^t, \mathcal{A}^t)) \leftarrow$ Calculate partial utility of Q-values for attribute g_i at time period t considering risk attitude coefficient ρ of the agent
  return $u_i(\overline{Q}_i(S^t, \mathcal{A}^t))$
  end SUBPROCEDURE
  SUBPROCEDURE TOTAL-UTILITY $(u_i(\overline{Q}_i(S^t, \mathcal{A}^t)), \Psi)$
    $U^t(S^t, \mathcal{A}^t) \leftarrow$ Calculate total utility for each state-action at time period t
  return $U(S^t, \mathcal{A}^t)$
  end SUBPROCEDURE
end
return $U(S, \mathcal{A})$
end PROCEDURE
PROCEDURE ACTION-SELECTION $(U(S, \mathcal{A}))$
  $\{\pi^*_{\mathcal{G}}(\mathcal{A}|S)\} \leftarrow$ Select the best policy
return $\{\pi^*_{\mathcal{G}}(\mathcal{A}|S)\}$
end PROCEDURE

Output: $\{\pi^*_{\mathcal{G}}(\mathcal{A}|S)\}$

approach entails deriving a single policy by applying linear scalarization to the criteria with a predetermined set of weights (Nguyen et al., 2020b). Consequently, any change in the preference modelling parameters necessitates re-learning (retraining) a new policy. As described in algorithm 1, the MAUDRL model first learns separate Q-functions for each reward criterion, followed by aggregating Q-values based on the utility functions to identify the optimal action.

We employ performance metrics in evaluating MAUDRL against the two benchmark algorithms, including policy quality, goal achievement, and run times. To gauge policy quality, we first generate normalized economic, social, and environmental rewards for risk-neutral farmers. We assess the goal achievement by comparing the percentage of plants pruned and percentage of budget used over planning horizon. We also consider the average computation time required for convergence.

The policy quality is evaluated using normalized economic, social, and environmental rewards for risk-neutral farmers, as illustrated in Fig. 7. Benchmark 1 is based on an oracle that is deterministic and not influenced by uncertainty and thus it provides a non-dominated solution that remains unaffected by preferences or uncertainty. Therefore, the policy of Benchmark 1 may be viewed as a quasi upper bound for our comparison, considering that it is not derived under perfect information. The impact of risk attitudes on MAUDRL's performance is contrasted with Benchmark 2, as shown in Fig. 8. Table 3 provides a broad comparison of the various algorithms, highlighting goal accomplishment (percentage of plants pruned and percentage of the budget utilized by the end of planning horizon), the average normalized reward for the best policy retained and the average computation time necessary for convergence.

Algorithm 5: Benchmark 2 algorithm: Single Reward Aggregation Approach

Input : $(\mathcal{A}, S, \mathcal{R}, \mathcal{P}, \mathcal{G}, \Psi, \rho)$

PROCEDURE LEARNING-DQN-MAUT $(\mathcal{A}, S, \mathcal{R}, \mathcal{P}, \mathcal{G})$
SUBPROCEDURE CALCULATING-Q$_i$
  $Q_i(S, A) \leftarrow$ Calculate Q-values of all the state-actions (S, A) for each attribute g_i
return $Q_i(S, A)$
end SUBPROCEDURE
SUBPROCEDURE NORMALIZATION-DQN$_i$ $(Q_i(S, \mathcal{A}))$
  $\overline{Q}_i(S, \mathcal{A}) \leftarrow$ Normalize Q-values regarding attribute g_i
return $\overline{Q}_i(S, \mathcal{A})$
end SUBPROCEDURE
SUBPROCEDURE PARTIAL-UTILITY$_i$ $(\overline{Q}_i(S, A), \rho)$
  $u_i(\overline{Q}_i(S^t, \mathcal{A})) \leftarrow$ Calculate partial utility of Q-values for attribute g_i considering risk attitude coefficient ρ of the agent
return $u_i(\overline{Q}_i(S, \mathcal{A}))$
end SUBPROCEDURE
SUBPROCEDURE TOTAL-UTILITY $(u_i(\overline{Q}_i(S, \mathcal{A})), \Psi)$
  $U^t(S, \mathcal{A}) \leftarrow$ Calculate total utility for each state-action
return $U(S, \mathcal{A})$
end SUBPROCEDURE
return $U(S, \mathcal{A})$
end PROCEDURE
  $\{\pi^*_{\mathcal{G}}(\mathcal{A}|S)\} \leftarrow$ Select the best policy
return $\{\pi^*_{\mathcal{G}}(\mathcal{A}|S)\}$
end PROCEDURE

Output: $\{\pi^*_{\mathcal{G}}(\mathcal{A}|S)\}$

The results unmistakably demonstrate MAUDRL's superior performance over the Benchmark 2 algorithm in most instances. Compared to Benchmark 1, which provides upper bounds, the MAUDRL algorithm consistently delivers high economic, social, and environmental rewards. Beyond its explainability, the comparison of MAUDRL's runtime with that of Benchmarks 1 and 2, as presented in Table 3, reinforces MAUDRL's practical utility in real-world scenarios. These typically necessitate real-time decision-making over a planning horizon.

Distinct from Benchmark 1, Benchmark 2 aggregates Q-values based on utility functions and identifies the optimal action according to the anticipated cumulative reward. As outlined in algorithm 5, this method substitutes the LEARNING-DQN and MAUT steps in the MAUDRL model with the LEARNING-DQN-MAUT step in the benchmark algorithm. This allows for concurrent Q-learning and multi-attribute aggregation.

6. Discussion

The numerical results and benchmarking suggest that MAUDRL model has been proven to be highly effective in solving complex SMCD problems and supporting sustainable decision-making. Our application of the MAUDRL to the sustainable sequential HRP case in blueberry farming has demonstrated its ability to handle social, environmental, and economic aspects of sustainability with remarkable efficiency. This model has far-reaching potential and can significantly contribute to sustainable decision-making across industries and sectors.

The MAUDRL operates in two stages: training and exploitation. During training, each dimension of sustainability is addressed in parallel, reducing the model's sensitivity to hyperparameters and minimizing the required tuning time (Gijsbrechts et al., 2022). Moreover, it allows for parallel computation, with each processor handling a distinct DQN during the training phase. This accelerates the learning process, enhances scalability for handling multiple objectives, and helps balance the trade-off between exploitation and exploration.

Table 3
Comparison of MAUDRL to the Benchmark 1 and Benchmark 2 algorithms.

Case	Algorithm	Pruned plants	Used budget	Average % rewards			Run-time
		(%)	(%)	Econ.	Soci.	Envi.	(s)
1–9	Benchmark 1 (Oracle)	100.000	78.231	0.923	1.000	1.000	1566
1	MAUDRL	100.000	75.216	0.787	0.964	0.854	93
	Benchmark 2	100.000	77.943	0.793	0.824	0.764	6172
2	MAUDRL	100.000	75.650	0.894	0.750	0.998	51
	Benchmark 2	100.000	76.671	0.867	0.833	1.000	6369
3	MAUDRL	100.000	79.758	0.894	0.750	0.998	85
	Benchmark 2	100.000	76.851	0.816	0.857	0.933	5853
4	MAUDRL	100.000	82.829	0.826	0.939	0.960	64
	Benchmark 2	100.000	79.438	0.828	0.833	1.000	3525
5	MAUDRL	100.000	77.191	0.908	1.000	1.000	50
	Benchmark 2	100.000	76.293	0.558	0.857	0.897	5344
6	MAUDRL	100.000	77.856	0.894	0.909	0.948	62
	Benchmark 2	100.000	74.926	0.855	0.785	0.925	5355
7	MAUDRL	100.000	87.090	0.829	0.979	0.857	62
	Benchmark 2	100.000	80.698	0.844	0.857	0.933	3594
8	MAUDRL	100.000	86.994	0.829	0.979	0.857	60
	Benchmark 2	100.000	80.630	0.858	0.791	1.000	5233
9	MAUDRL	100.000	81.070	0.894	1.000	0.976	75
	Benchmark 2	100.000	85.264	0.891	1.000	0.778	6221

The reported run times for MAUDRL exclude the individual attribute training times, given that the DRL model in MAUDRL does not necessitate retraining for each case. This training is conducted once for all decision-making personas. The run times for the one-time training of the DRL algorithm are as follows: 2910 s for Economic, 2091 s for Environmental, and 3779 s for Social criteria.

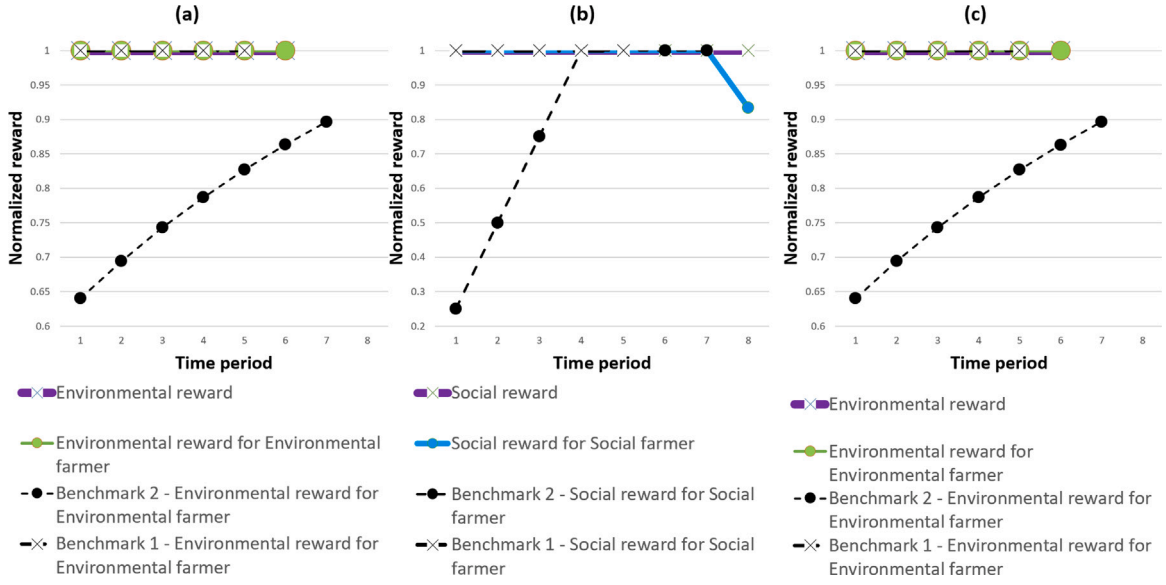

Fig. 7. Comparison of the performance of MAUDRL, Benchmark 1 and Benchmark 2 algorithms using the normalized reward profiles of the best policies for the same decision-agent with a neutral risk attitude: **(a)** economic rewards, **(b)** social rewards, and **(c)** environmental rewards.

Within blueberry farming, our results revealed that prioritizing environmental and social outcomes can be achieved with minimal economic trade-offs. This challenges the common belief that sustainability inevitably compromises economic performance, suggesting that initial investments in skilled labour can lead to more efficient operations, offsetting the initial cost increase (Berk et al., 2019). The MAUDRL model considers the decision-maker's risk attitude and sustainability preferences during the decision-making process and accommodates changes in these factors over time, which is essential in dynamic logistical environments. For instance, as the farming season progresses and budgetary constraints become more pressing, farmers might shift their focus to prioritize economic outcomes, reflecting an adaptive change in risk attitude.

Furthermore, our study simplifies DRL's often complex and time-consuming hyperparameter tuning process, as exemplified in our research (Gijsbrechts et al., 2022). The integration of MCDA and DRL opens up new opportunities for explainable AI, underscoring the potential of sustainability in human resource planning and its capability to generate economic value, not just in agriculture but in broader logistics operations.

7. Conclusion

Sequential Multi-Criteria Decision (SMCD) problems involve making choices based on several conflicting factors, a challenge that is becoming more important as we aim for sustainability in logistics. Our study

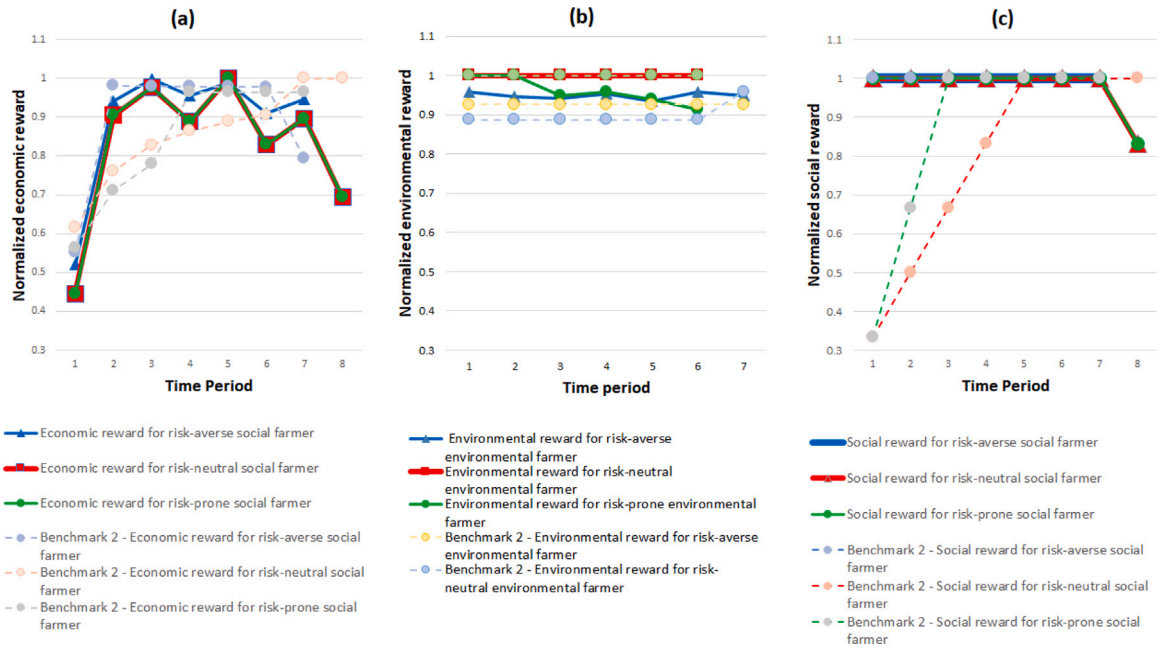

Fig. 8. Comparison of the performance of MAUDRL and Benchmark 2 algorithms using the normalized reward profiles of the best policies for the same decision-agent with different risk attitudes: **(a)** economic rewards, **(b)** social rewards, and **(c)** environmental rewards.

contributes to this area by introducing the Multi-Attribute Utility Deep Reinforcement Learning (MAUDRL) algorithm, which combines DRL with Multi-Criteria Decision Analysis (MCDA). This is a new step in applying DRL to help decision-makers choose the best action based on their preferences and risk attitudes across different criteria. MAUDRL is one of the first models to combine DRL and MCDA, offering a new way to solve SMCD problems. Our work also adds to the field of operational research by providing a practical and easier-to-understand model in terms of analytics and AI.

MAUDRL improves the clarity and understandability of DRL models by requiring specific criteria and parameters for their development and training. This approach helps manage the complexity of training large datasets, mainly when these datasets cover economic, environmental, and social aspects. Our method starts by training different Deep Q-Networks separately for each criterion and then combines them. This makes the results more transparent, especially when understanding how each trained model affects the overall outcome. Moreover, MAUDRL is less affected by hyperparameter changes and can be scaled up to include more attributes or objectives. Training models separately for each dimension speeds up the learning process and helps balance exploration and exploitation.

Implementing MAUDRL in sustainable human resource planning for farming has provided important insights. Our model indicates it can create effective policies promptly. Interestingly, focusing on environmental and social aspects in decision-making results in only a tiny loss in economic performance. Additionally, the MAUDRL model allows for integrating a decision agent's preferences and risk attitude, which is essential as these preferences can change over time.

There are areas for future research. The use of the additive model in Multi-Attribute Utility Theory (MAUT) component of the MAUDRL might be a limitation in situations where preferences across attributes are interconnected. Exploring other methods for combining criteria in MAUDRL could extend its use to different situations and sustainability challenges. Also, investigating other methods for exploring decision trees in multi-dimensional scenarios is worth exploring. It is also essential to see if sustainability-driven decisions always lead to better outcomes in SMCD problems. Finally, our model does not consider competition among different decision-makers or the individual choices of workers, as it assumes they follow the farmer's decisions. Extending MAUDRL to include these aspects would be an exciting area for future research.

CRediT authorship contribution statement

Mohammadreza Nematollahi: Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Adel Guitouni:** Writing – review & editing, Writing – original draft, Resources, Methodology, Funding acquisition, Conceptualization. **Nafiseh Izadyar:** Writing – original draft, Visualization, Software, Methodology, Data curation. **Nabil Belacel:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Conceptualization. **Andrew Park:** Writing – review & editing, Validation, Investigation.

Acknowledgements

The authors would like to express their sincere appreciation to the Area Editor for handling the manuscript and to the two anonymous reviewers for their insightful comments and constructive suggestions, which have significantly improved the quality of this work. Additionally, the authors acknowledge the financial support received from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the National Research Council Canada (NRC) (operating grant CDTS-101-1).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cor.2026.107426.

Data availability

Data will be made available on request.

References

Akeb, Hakim, Hifi, Mhand, M'Hallah, Rym, 2009. A beam search algorithm for the circular packing problem. Comput. Oper. Res. 36 (5), 1513–1528.

Alcaraz, Juan J., Losilla, Fernando, Caballero-Arnaldos, Luis, 2022. Online model-based reinforcement learning for decision-making in long distance routes. Transp. Res. E: Logist. Transp. Rev. 164, 102790.

Allaoui, Hamid, Guo, Yuhan, Choudhary, Alok, Bloemhof, Jacqueline, 2018. Sustainable agro-food supply chain design using two-stage hybrid multi-objective decision-making approach. Comput. Oper. Res. 89, 369–384.

Amorim-Lopes, Mário, Oliveira, Monica, Raposo, Mariana, Cardoso-Grilo, Teresa, Alvarenga, António, Barbas, Marta, Alves, Marco, Vieira, Ana, Barbosa-Póvoa, Ana, 2021. Enhancing optimization planning models for health human resources management with foresight. Omega 103, 102384.

Annear, Luis Mauricio, Akhavan-Tabatabaei, Raha, Schmid, Verena, 2023. Dynamic assignment of a multi-skilled workforce in job shops: An approximate dynamic programming approach. European J. Oper. Res. 306 (3), 1109–1125.

Aviso, Kathleen B, Chiu, Anthony SF, Demeterio III, Feorillo PA, Lucas, Rochelle Irene G, Tseng, Ming-Lang, Tan, Raymond R, 2019. Optimal human resource planning with P-graph for universities undergoing transition. J. Clean. Prod. 224, 811–822.

BC-Government, 2022. Growing blueberries. https://www2.gov.bc.ca/gov/content/industry/agriservice-bc/production-guides/berries/blueberries,

Berk, Lauren, Bertsimas, Dimitris, Weinstein, Alexander M, Yan, Julia, 2019. Prescriptive analytics for human resource planning in the professional services industry. European J. Oper. Res. 272 (2), 636–641.

Bordoloi, Sanjeev K., Matsuo, Hirofumi, 2001. Human resource planning in knowledge-intensive operations: A model for learning with stochastic turnover. European J. Oper. Res. 130 (1), 169–189.

Bouyssou, Denis, Marchant, Thierry, Pirlot, Marc, Tsoukias, Alexis, Vincke, Philippe, 2006. Evaluation and Decision Models with Multiple Criteria: Stepping Stones for the Analyst, vol. 86, Springer Science & Business Media.

Brockman, Greg, Cheung, Vicki, Pettersson, Ludwig, Schneider, Jonas, Schulman, John, Tang, Jie, Zaremba, Wojciech, 2016. Openai gym. arXiv preprint arXiv:1606.01540.

Campanella, Gianluca, Ribeiro, Rita A., 2011. A framework for dynamic multiple-criteria decision making. Decis. Support Syst. 52 (1), 52–60.

Chávez, Marcela María Morales, Sarache, William, Costa, Yasel, 2018. Towards a comprehensive model of a biofuel supply chain optimization from coffee crop residues. Transp. Res. E: Logist. Transp. Rev. 116, 136–162.

Chen, Xiaoyu, Tian, Tian, Dai, Guangming, Wang, Maocai, Song, Zhiming, Xing, Lining, 2025. Deep reinforcement learning-based resource allocation method for multi-satellite scheduling. Comput. Oper. Res. 107088.

Chikalov, Igor, Hussain, Shahid, Moshkov, Mikhail, 2018. Bi-criteria optimization of decision trees with applications to data analysis. European J. Oper. Res. 266 (2), 689–701.

Cohen, Maxime C., Miao, Sentao, Wang, Yining, 2025. Dynamic pricing with fairness constraints. Oper. Res.

Dahooie, Jalil Heidary, Hajiagha, Seyed Hossein Razavi, Farazmehr, Shima, Zavadskas, Edmundas Kazimieras, Antucheviciene, Jurgita, 2021. A novel dynamic credit risk evaluation method using data envelopment analysis with common weights and combination of multi-attribute decision-making methods. Comput. Oper. Res. 129, 105223.

Diaby, Vakaramoko, Campbell, Kaitryn, Goeree, Ron, 2013. Multi-criteria decision analysis (MCDA) in health care: A bibliometric analysis. Oper. Res. Health Care 2 (1–2), 20–24.

Ding, Yida, Wandelt, Sebastian, Wu, Guohua, Xu, Yifan, Sun, Xiaoqian, 2023. Towards efficient airline disruption recovery with reinforcement learning. Transp. Res. E: Logist. Transp. Rev. 179, 103295.

Fletcher, Tanya, 2020. B.C. agriculture sector bracing for 8,000 worker shortfall this year. https://www2.gov.bc.ca/gov/content/employment-business/employment-standards-advice/employment-standards/forms-resources/igm/esr-part-4-section-18,

Frini, Anissa, Amor, Sarah Ben, 2019. MUPOM: A multi-criteria multi-period outranking method for decision-making in sustainable development context. Environ. Impact Assess. Rev. 76, 10–25.

Frini, Anissa, Guitouni, Adel, Martel, Jean-Marc, 2012. A general decomposition approach for multi-criteria decision trees. European J. Oper. Res. 220 (2), 452–460.

Fujimoto, Scott, Hoof, Herke, Meger, David, 2018. Addressing function approximation error in actor-critic methods. In: International Conference on Machine Learning. PMLR, pp. 1587–1596.

Ghafour, Karzan, 2024. Multi-objective continuous review inventory policy using MOPSO and TOPSIS methods. Comput. Oper. Res. 163, 106512.

Gijsbrechts, Joren, Boute, Robert N, Van Mieghem, Jan A, Zhang, Dennis J, 2022. Can deep reinforcement learning improve inventory management? performance on lost sales, dual-sourcing, and multi-echelon problems. Manuf. Serv. Oper. Manag. 24, 1349–1368.

Goodfellow, Ian, Bengio, Yoshua, Courville, Aaron, 2016. Deep Learning. MIT Press.

Guitouni, Adel, Martel, Jean-Marc, 1998. Tentative guidelines to help choosing an appropriate MCDA method. European J. Oper. Res. 109 (2), 501–521.

He, Zhenglei, Tran, Kim-Phuc, Thomassey, Sebastien, Zeng, Xianyi, Xu, Jie, Yi, Changhai, 2021. A deep reinforcement learning based multi-criteria decision support system for optimizing textile chemical process. Comput. Ind. 125, 103373. http://dx.doi.org/10.1016/j.compind.2020.103373.

Hessel, Matteo, Modayil, Joseph, Van Hasselt, Hado, Schaul, Tom, Ostrovski, Georg, Dabney, Will, Horgan, Dan, Piot, Bilal, Azar, Mohammad, Silver, David, 2018. Rainbow: Combining improvements in deep reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. 32, (1).

Jiang, Ruhao, Deng, Yanchen, Chen, Yingying, Luo, He, An, Bo, 2024. Deep reinforcement learning for multi-objective game strategy selection. Comput. Oper. Res. 168, 106683.

Karimi-Majd, Amir-Mohsen, Mahootchi, Masoud, Zakery, Amir, 2017. A reinforcement learning methodology for a human resource planning problem considering knowledge-based promotion. Simul. Model. Pr. Theory 79, 87–99.

Keeney, Ralph L., Raiffa, Howard, Meyer, Richard F., 1993. Decisions with Multiple Objectives: Preferences and Value Trade-Offs. Cambridge University Press, ISBN: 0-521-43883-7.

Kornbluth, JSH, 1985. Sequential multi-criterion decision making. Omega 13 (6), 569–574.

Kramar, Robin, 2022. Sustainable human resource management: Six defining characteristics. Asia Pac. J. Hum. Resour. 60 (1), 146–170.

Lang, Magdalena A.K., Cleophas, Catherine, Ehmke, Jan Fabian, 2021. Multi-criteria decision making in dynamic slotting for attended home deliveries. Omega 102, 102305.

Li, Yuxi, 2017. Deep reinforcement learning: An overview. arXiv preprint arXiv:1701.07274.

Li, Kaiwen, Zhang, Tao, Wang, Rui, 2020. Deep reinforcement learning for multiobjective optimization. IEEE Trans. Cybern. 51 (6), 3103–3114.

Liu, Shan, Jiang, Hai, Chen, Shuiping, Ye, Jing, He, Renqing, Sun, Zhizhao, 2020. Integrating Dijkstra's algorithm into deep inverse reinforcement learning for food delivery route planning. Transp. Res. E: Logist. Transp. Rev. 142, 102070.

Liu, Hongbin, Jiang, Le, Martínez, Luis, 2018. A dynamic multi-criteria decision making model with bipolar linguistic term sets. Expert Syst. Appl. 95, 104–112.

Liu, Zeyu, Li, Xueping, Khojandi, Anahita, 2022. The flying sidekick traveling salesman problem with stochastic travel time: A reinforcement learning approach. Transp. Res. E: Logist. Transp. Rev. 164, 102816.

Liu, Renke, Piplani, Rajesh, Toro, Carlos, 2023a. A deep multi-agent reinforcement learning approach to solve dynamic job shop scheduling problem. Comput. Oper. Res. 159, 106294.

Liu, Shan, Zhang, Ya, Wang, Zhengli, Gu, Shiyi, 2023b. AdaBoost-Bagging deep inverse reinforcement learning for autonomous taxi cruising route and speed planning. Transp. Res. E: Logist. Transp. Rev. 177, 103232.

Lu, Jiawei, Jiang, Jielin, Balasubramanian, Venki, Khosravi, Mohammad R, Xu, Xiaolong, 2022a. Deep reinforcement learning-based multi-objective edge server placement in Internet of Vehicles. Comput. Commun. 187, 172–180.

Lu, Junlin, Mannion, Patrick, Mason, Karl, 2022b. A multi-objective multi-agent deep reinforcement learning approach to residential appliance scheduling. IET Smart Grid 5 (4), 260–280.

Macke, Janaina, Genari, Denise, 2019. Systematic literature review on sustainable human resource management. J. Clean. Prod. 208, 806–815.

Mahmoudinazlou, Sasan, Sobhanan, Abhay, Charkhgard, Hadi, Eshragh, Ali, Dunn, George, 2025. Deep reinforcement learning for dynamic order picking in warehouse operations. Comput. Oper. Res. 107112.

Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Rusu, Andrei A, Veness, Joel, Bellemare, Marc G, Graves, Alex, Riedmiller, Martin, Fidjeland, Andreas K, Ostrovski, Georg, et al., 2015. Human-level control through deep reinforcement learning. Nature 518 (7540), 529–533.

Moura, Gisely Correa De, Vizzotto, Marcia, Picolotto, Luciano, Antunes, Luis Eduardo Corrêa, 2017. Production, physical-chemical quality and bioative compounds of misty blueberry fruit under different pruning intensities. Rev. Bras. Frutic. http://dx.doi.org/10.1590/0100-29452017158.

Nguyen, Thanh Thi, Nguyen, Ngoc Duy, Nahavandi, Saeid, 2020a. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. IEEE Trans. Cybern. 50 (9), 3826–3839.

Nguyen, Thanh Thi, Nguyen, Ngoc Duy, Vamplew, Peter, Nahavandi, Saeid, Dazeley, Richard, Lim, Chee Peng, 2020b. A multi-objective deep reinforcement learning framework. Eng. Appl. Artif. Intell. 96, 103915.

Olson, David L., 1997. Decision aids for selection problems. J. Oper. Res. Soc. 48 (5), 541–542.

Ow, Peng Si, Morton, Thomas E., 1988. Filtered beam search in scheduling. Int. J. Prod. Res. 26 (1), 35–62.

Pandey, Satyendra C, Modi, Pratik, Pereira, Vijay, Fosso Wamba, Samuel, 2024. Empowering small farmers for sustainable agriculture: a human resource approach to SDG-driven training and innovation. Int. J. Manpow.

Pomerol, Jean-Charles, Barba-Romero, Sergio, 2000. Multicriterion Decision in Management: Principles and Practice, vol. 25, Springer Science & Business Media.

Powell, Warren B., 2021. From reinforcement learning to optimal control: A unified framework for sequential decisions. In: Handbook of Reinforcement Learning and Control. Springer, pp. 29–74.

Price, Wilson L., Martel, Alain, Lewis, K.A., 1980. A review of mathematical models in human resource planning. Omega 8 (6), 639–645.

Quirk, James P., Saposnik, Rubin, 1962. Admissibility and measurable utility functions. Rev. Econ. Stud. 29 (2), 140–146.
Raffin, Antonin, Hill, Ashley, Gleave, Adam, Kanervisto, Anssi, Ernestus, Maximilian, Dormann, Noah, 2021. Stable-baselines3: Reliable reinforcement learning implementations. J. Mach. Learn. Res. 22, 1–8.
Roijers, Diederik M, Vamplew, Peter, Whiteson, Shimon, Dazeley, Richard, 2013. A survey of multi-objective sequential decision-making. J. Artificial Intelligence Res. 48, 67–113.
Sabuncuoglu, Ihsan, Bayiz, M., 1999. Job shop scheduling with beam search. European J. Oper. Res. 118 (2), 390–412.
Sadeghi-Dastaki, Mohsen, Afrazeh, Abbas, 2018. A two-stage skilled manpower planning model with demand uncertainty. Int. J. Intell. Comput. Cybern. 11 (4), 526–551.
Schulman, John, Wolski, Filip, Dhariwal, Prafulla, Radford, Alec, Klimov, Oleg, 2017. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.
Seuring, Stefan, Müller, Martin, 2008. From a literature review to a conceptual framework for sustainable supply chain management. J. Clean. Prod. 16 (15), 1699–1710.
Shahbazian, Reza, Ciacco, Alessia, Macrina, Giusy, Guerriero, Francesca, 2025. Knowledge-guided hybrid deep reinforcement learning for the dynamic multi-depot electric vehicle routing problem. Comput. Oper. Res. 107217.
Strik, Bernadine C., Davis, Amanda J., 2022. Pruning method and trellising impact hand-and machine-harvested yield and costs of production in 'Legacy'highbush blueberry. HortScience 57 (7), 811–817.
Stup, R.E., Hyde, J., Holden, L.A., 2006. Relationships between selected human resource management practices and dairy farm performance. J. Dairy Sci. 89 (3), 1116–1120.
Building business success, 2016. Cost of producing fresh and processing blueberries in the fraser valley of british columbia. https://www2.gov.bc.ca/assets/gov/farming-natural-resources-and-industry/agriculture-and-seafood/farm-management/farm-business-management/enterprise-budgets/enterprise_budget_fresh_and_processing_blueberries_fraser_valley_2016.pdf,
Sutton, Richard S., 1995. Generalization in reinforcement learning: Successful examples using sparse coarse coding. Adv. Neural Inf. Process. Syst. 8.
Sutton, Richard S., Barto, Andrew G., 2018. Reinforcement Learning: An Introduction. MIT Press.
Tai, Jun Jet, Terry, Jordan K, Innocente, Mauro S, Brusey, James, Horri, Nadjim, 2022. Some supervision required: Incorporating oracle policies in reinforcement learning via epistemic uncertainty metrics. arXiv preprint arXiv:2208.10533.
Tang, Christopher S., 2006. Perspectives in supply chain risk management. Int. J. Prod. Econ. 103 (2), 451–488.
Thakkar, Jitesh J., 2021. Multi-Criteria Decision Making, vol. 336, Springer.
Wang, Zhenyu, Cai, Bin, Li, Jun, Yang, Deheng, Zhao, Yang, Xie, Huan, 2023a. Solving non-permutation flow-shop scheduling problem via a novel deep reinforcement learning approach. Comput. Oper. Res. 151, 106095.
Wang, Wen, Li, Beibei, Luo, Xueming, Wang, Xiaoyi, 2023b. Deep reinforcement learning for sequential targeting. Manag. Sci. 69 (9), 5439–5460.
Wang, Hao-nan, Liu, Ning, Zhang, Yi-yun, Feng, Da-wei, Huang, Feng, Li, Dong-sheng, Zhang, Yi-ming, 2020. Deep reinforcement learning: a survey. Front. Inf. Technol. Electron. Eng. 21 (12), 1726–1744.
Wang, Hui, Wang, Ran, Xu, Hu, Kun, Zhu, Yi, Changyan, Niyato, Dusit, 2021. Multi-objective mobile charging scheduling on the internet of electric vehicles: a DRL approach. In: 2021 IEEE Global Communications Conference. GLOBECOM, IEEE, pp. 01–06.
Wang, Dujuan, Wang, Qi, Yin, Yunqiang, Cheng, TCE, 2023c. Optimization of ride-sharing with passenger transfer via deep reinforcement learning. Transp. Res. E: Logist. Transp. Rev. 172, 103080.
Wang, Zheng, Zeng, Tiansheng, Chu, Xuening, Xue, Deyi, 2023d. Multi-objective deep reinforcement learning for optimal design of wind turbine blade. Renew. Energy 203, 854–869.
Wei, Guohua, Jin, Yi, 2021. Human resource management model based on three-layer BP neural network and machine learning. J. Intell. Fuzzy Systems 40 (2), 2289–2300.
Wei, Zeqi, Zhao, Zhibin, Zhou, Zheng, Ren, Jiaxin, Tang, Yajun, Yan, Ruqiang, 2024. A deep reinforcement learning-driven multi-objective optimization and its applications on aero-engine maintenance strategy. J. Manuf. Syst. 74, 316–328.
Wood, David A., Khosravanian, Rassoul, 2015. Exponential utility functions aid upstream decision making. J. Nat. Gas Sci. Eng. 27, 1482–1494.
Wu, Yaoxin, Fan, Mingfeng, Cao, Zhiguang, Gao, Ruobin, Hou, Yaqing, Sartoretti, Guillaume, 2024. Collaborative deep reinforcement learning for solving multi-objective vehicle routing problems. In: 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), pp. 1956–1965.
Wu, Xinquan, Yan, Xuefeng, 2023. A spatial pyramid pooling-based deep reinforcement learning model for dynamic job-shop scheduling problem. Comput. Oper. Res. 160, 106401.
Xie, Jiaohong, Yang, Zhenyu, Lai, Xiongfei, Liu, Yang, Yang, Xiao Bo, Teng, Teck-Hou, Tham, Chen-Khong, 2022. Deep reinforcement learning for dynamic incident-responsive traffic information dissemination. Transp. Res. E: Logist. Transp. Rev. 166, 102871.
Yalçındağ, Semih, Cappanera, Paola, Scutella, Maria Grazia, Şahin, Evren, Matta, Andrea, 2016. Pattern-based decompositions for human resource planning in home health care services. Comput. Oper. Res. 73, 12–26.
Yan, Yimo, Chow, Andy HF, Ho, Chin Pang, Kuo, Yong-Hong, Wu, Qihao, Ying, Chengshuo, 2022. Reinforcement learning for logistics and supply chain management: Methodologies, state of the art, and future opportunities. Transp. Res. E: Logist. Transp. Rev. 162, 102712.
Yang, Runzhe, Sun, Xingyuan, Narasimhan, Karthik, 2019. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. Adv. Neural Inf. Process. Syst. 32.
Yuan, Shuai, Qi, Qian, Dai, Enliang, Liang, Yongfeng, 2022. Human resource planning and configuration based on machine learning. Comput. Intell. Neurosci. http://dx.doi.org/10.1155/2022/3605722.
Zhang, Gongquan, Chang, Fangrong, Jin, Jieling, Yang, Fan, Huang, Helai, 2024. Multi-objective deep reinforcement learning approach for adaptive traffic signal control system with concurrent optimization of safety, efficiency, and decarbonization at intersections. Accid. Anal. Prev. 199, 107451.
Zhang, Pushi, Chen, Xiaoyu, Zhao, Li, Xiong, Wei, Qin, Tao, Liu, Tie-Yan, 2021. Distributional reinforcement learning for multi-dimensional reward functions. Adv. Neural Inf. Process. Syst. 34, 1519–1529.
Zhang, Wenquan, Peng, Zhaoxian, Zhao, Fei, Feng, Bo, Mei, Xuesong, 2026. A novel deep reinforcement learning framework based on digital twins for dynamic job shop scheduling problems. Expert Syst. Appl. 296, 128708.
Zhou, Yuekuan, 2023. A dynamic self-learning grid-responsive strategy for battery sharing economy—multi-objective optimisation and posteriori multi-criteria decision making. Energy 266, 126397.