



Article

Multi-objective reinforcement-learning simulation for co-optimizing resource circularity and environmental performance in circular manufacturing

Ting Li*, ALogaiswari Indiran

Faculty of Management, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia

ARTICLE INFO

Article history:

Received 10 June 2026

Received in revised form

08 October 2026

Accepted 27 October 2026

Keywords:

Circular manufacturing, Discrete-event simulation, Multi-objective reinforcement learning, Resource circularity, Environmental performance

*Corresponding author

Email address:

l.ting@graduate.utm.my

DOI: 10.55670/fpll.futech.5.4.30

ABSTRACT

More regulations require manufacturers to enhance material circularity and reduce their environmental footprint simultaneously; however, it remains unclear whether these two objectives align, and if so, in which domain. A discrete event simulation model for a remanufacturing plant operating multiple production lines with environmental performance measures at the process level is defined as a multiobjective Markov decision process, which is then solved using a preference-conditioned actor-critic technique, employing feasibility masking for the precedence and capacity constraints and adaptive sampling between the surrogate model and the full-scale simulator. Two life cycle accounting methods have been used: Boundary A (cut-off) and Boundary B (avoided burden). The environmental burden reveals three regimes, synergistic, trade-off, and inverted, characterized by transition points in secondary-material input levels at 0.38 and 0.61 for Boundary A and at 0.47 and 0.78 for Boundary B; sweeps with fixed ratios validate these transition points. One conditionally generated network accounts for 87.4% of the non-dominated solutions of eleven scalarized agents using 9.1% of the simulation budget allocated to them, and the switching strategy reduces the number of simulation calls needed to reach the reference hypervolume by 27.7%. A screening based on the unit-cost criterion limits the joint synergistic and economically viable domain to input levels ranging from 0.21 to 0.38. All parameters are synthetic values from published ranges for a single facility, so the boundaries describe a mechanism rather than plant-specific forecasts.

1. Introduction

Manufacturers today need to prove their circular performance and environmental impact simultaneously. Instruments like EPR requirements, mandatory recycled material [1], and eco-modulated charges hold manufacturers accountable for end-of-life stages; therefore, reintroducing recovered cores and secondary raw materials becomes an operational task, not just reputation building. The challenge is most relevant for durable products like electric-vehicle batteries, which return in different states of maintenance and are hard to return for appropriate processing [2]. These metrics rely on monotonicity assumptions, under which a higher circularity rate yields better environmental performance. As the quality of recycled cores decreases, more energy is invested in inspecting, disassembling, and cleaning cores. Thus, the overall environmental efficiency of recycling depends on core quality and the recycling process used [3].

Life cycle analyses show that the arrangement providing the maximum circularity rate does not necessarily yield the minimum environmental burden [4]. Whether resource circularity and environmental performance can be co-optimized, and under what conditions, is therefore an open question rather than a settled premise. Current solutions have three limitations. First, remanufacturing system optimization is largely static and deterministic: metaheuristics optimize specific instances and treat return quantities and core-quality distributions as deterministic inputs [5]; thus, they cannot capture sequential decisions over time. A review of decision-making in remanufacturing operations revealed that only a fraction of studies focus on operational decisions and uncertainties [6]. Second, while some dynamic approaches handle multiple objectives by weighting the objective vector into a scalar function, this makes the process opaque, repetitive, and unable to reach solutions in non-convex

regions of the Pareto frontier, regardless of the weight choice. Finally, although reinforcement learning (RL) becomes more and more popular in solving production scheduling problems, most of the works utilize this approach in a single-objective formulation on tiny test instances far from real industry and use penalty reward functions to capture the process feasibility, not modeling it in a native way [7]. In addition, energy consumption is treated as a cost criterion in deep RL schedulers [8]. Another often-overlooked problem is that coupling reinforcement learning with a high-fidelity simulator entails sampling costs that depend on its complexity [9].

This work proposes an event-based simulation of a closed-loop manufacturing process with process-level environmental accounting, models it as a multi-objective Markov decision process (MOMDP), and solves it with a preference-conditioned actor-critic method that outputs infeasible action masks. To the best of our knowledge, no work identifies the range of secondary material input within a dynamic manufacturing plant model where both environmental performance and circularity improve, and investigates how this range is influenced by the accounting convention. The objectives of this paper are: (O1) to identify the environmental burden-against-secondary material-input regime boundaries under two different accounting conventions; (O2) to validate that a single preference-conditioned network learns the entire Pareto front and generalizes to unknown preferences at much reduced cost compared to per-weight training; and (O3) to validate whether the mask, preference condition, and adaptive fidelity are all necessary components and robust to perturbation. Section 2 summarizes the relevant literature and formalizes the problem; Section 3 explains the methodology used; Section 4 presents the results; Section 5 discusses them; and Section 6 concludes.

2. Related work and problem formulation

2.1 Response relationship between resource circularity and environmental performance: a three-regime conjecture to be tested

Let us denote by $\rho \in [0,1]$ the input rate of secondary material, that is, the fraction of material demand covered by the recovered cores and not virgin material, and let us denote by $E(\rho)$ the environmental load corresponding to such an input rate according to the particular operating strategy (carbon load per unit of production, see Section 3.2). Its sign divides the ρ -axis into three regions:

$$R_{gn} = \{\rho: dE/d\rho < -\varepsilon\}, \quad R_{nd} = \{\rho: |dE/d\rho| \leq \varepsilon\}, \quad R_{mv} = \{\rho: dE/d\rho > \varepsilon\} \quad (1)$$

where $\varepsilon > 0$ is a plateau tolerance specified in Section 3.4. This partitioning of the ρ -axis (Section 4.4) differs from the shape of the Pareto front in objective space, particularly its concave portion (Section 4.2). The two concepts complement each other since the circularity criterion corresponds to the material circularity indicator (MCI) and it depends on recovery output and disposal along with ρ (Section 3.2), hence, the front is not a reshaped $E(\rho)$.

Non-monotonicity of $E(\rho)$ is the conjecture here. At low ρ values, the process of recovery depends on the high quality of cores and on low levels of processing effort, so each unit of substitution provides a positive net effect; at higher ρ values, each extra core is of worse quality and the energy needed for its processing is increasing until it becomes greater than the marginal benefit of substitution. Points where $dE/d\rho$

reaches the tolerance band of Eq. (1) are called regime transition points. Because processing effort increases as core quality declines (see Section 3.1), the model should be non-monotonic; the remaining question is the location of transitions and the extent of the synergistic regime.

The relationship between both circularity and impact indicators has been proven to differ [10]. The mechanism analyzed in the paper, plant-level decreasing returns and burden-shifting, differs from the circular economy rebound effect, a market reaction (e.g., partial substitution [11]), and from the general concept of efficiency-driven offsets [12]. What this evidence does not supply is a boundary: existing accounts are retrospective and case-specific, and none formalizes $E(\rho)$ or locates where its gradient reverses.

2.2 Failure boundaries of static and single-objective optimization approaches

Closed-loop manufacturing optimization has used a static framework to date. Researchers have resolved disassembly sequences through precedence graphs and search, either exact or metaheuristic [13], and multi-objective frameworks that account for both disassembly sequences and line balancing produce a Pareto curve for a given problem instance [14]. Frameworks that incorporate uncertainty include simulations within meta-heuristics [15], and energy-aware remanufacturing schedulers that explicitly treat energy usage as an objective [16]. However, these frameworks provide one solution per problem instance. This approach poses difficulties in closed-loop manufacturing because returns become known over time. A second limitation concerns how multiple objectives are handled. Scalarization reduces the reward vector to $\omega \cdot R$, giving

$$\pi^*(\omega) = \arg \max_{\pi} \omega \cdot J(\pi) \quad (2)$$

The map in Eq. (2) $\omega \mapsto \pi^*(\omega)$ has no closed form, so weight adjustments cannot be anticipated in objective space; recovering a front of resolution N requires N independent runs; and over deterministic policies, exact maximizers of Eq. (2) are confined to the convex hull of the attainable set, leaving concave segments unreachable for every admissible ω ; only stochastic mixtures make the attainable set convex [17]. Rolling re-optimization, the traditional approach, confuses sampling noise with signals and may reverse previous decisions; state-based policies, not problem-based policies, are required [18].

2.3 Advances, open problems, and the positioning of this study

Multi-objective RL (MORL) distinguishes between multi-policy methods and single-policy methods, which parametrize an entire policy family with one neural network, using hypervolume, inverted generational distance, spacing, and coverage criteria [19]. Decomposition-based approaches are more suitable for stochastic combination of deterministic policies in concave parts of the front [20]. For single-policy methods, PD-MORL [21] shows that one can use a single neural network for any preference vector by conditioning policy and value networks on sampled preference vectors, and actor-critic methods can use nonlinear utility functions [22]. The following open questions still exist. Most results come from low-dimensional tasks, while industrial-scale verifications remain limited. Process constraints are mostly addressed using penalties for infeasible decisions, thus skewing the scale of rewards and providing no guarantees; other approaches build dynamic masks from a formalized system description [23]; use masks, containing domain-

specific knowledge, in operations research [24]; or incorporate precedences into states using graph encoders [25]. Lastly, preferences compete for capacity in a conditioned common network, but the coverage loss at the extremes remains unreported. Moreover, when the discrete-event model acts as a learning environment [26], the sampling costs increase with the fidelity, and the multi-fidelity approach, which reduces the variance via correlations between low- and high-fidelity rewards, uses a fixed sampling ratio [27], which is not applicable to the information requirements of the policy learning process.

The closest prior work uses RL to make disassembly decisions under core quality uncertainty, but addresses objective-function conflicts through weighted scalarization and treats delivery as an optimization goal, not an environmental constraint [28]. The present research is different from the prior research in three ways; namely, vector rewards for circularity and environmental performance, feasibility masking, and adaptive scheduling.

2.4 Problem statement and research questions

The closed-loop manufacturing system is modeled in a decision time interval T . Each point in time t is characterized by the state $st \in S$, which consists of buffer contents, the quality profile of work-in-process, material inventories, the equipment status, and a demand history; the action at $at \in A(st)$ determines the core routing, the ratio between the virgin and secondary material feeding, and the lot size and sequencing. The objective is the vector $J(\pi)$ in Eq. (3):

$$J(\pi) = E_{\pi} \left[\sum_{t=0}^T \gamma^t \mathbf{R}(s_t, a_t, s_{t+1}) \right], \quad \mathbf{R} = (r_{circ}, r_{env}) \quad (3)$$

and the task is to recover all Pareto-optimal policies with respect to J subject to $at \in A(st)$, where $A(st)$ encodes precedence and capacity feasibility. A policy π is Pareto optimal if no π' satisfies

$$J_i(\pi') \geq J_i(\pi) \quad \forall i, \quad \exists j : J_j(\pi') > J_j(\pi) \quad (4)$$

The following are the three research questions that emerge from Objectives O1 to O3 in Section 1: whether the three-regime structure exists and what its transition points are under each convention (RQ1); whether one preference-conditioned network is able to identify the front at substantially less cost than per-weight re-training (RQ2); and whether all three structural components are necessary and robust against increased return variance and reduced quality (RQ3). The analysis will be carried out for a single plant having multiple lines. The solution space is two-dimensional, and production cost per unit is considered a criterion, not an objective. The process of post-evaluation and its results are presented in Section 3.2 and Section 4.4, respectively.

3. Material and methods

3.1 Closed-loop simulation environment, accounting boundaries, and model calibration

A simulation model of the event-oriented type was created for the facility operating under closed material flow conditions across multiple lines. This model is made up of eight modules linked in a cyclically structured manner - forward production, return intake, inspection and grading, disassembly, cleaning, reprocessing, reassembly, and material recovery modules, while the disposal module serves as the sink where cores are disposed of if they are rejected in any of the other stages. The inspection and grading modules classify each core into one of five quality categories, after which routing is chosen so processing intensity depends on

core quality, not the number of cores. The model is illustrated in Figure 1.

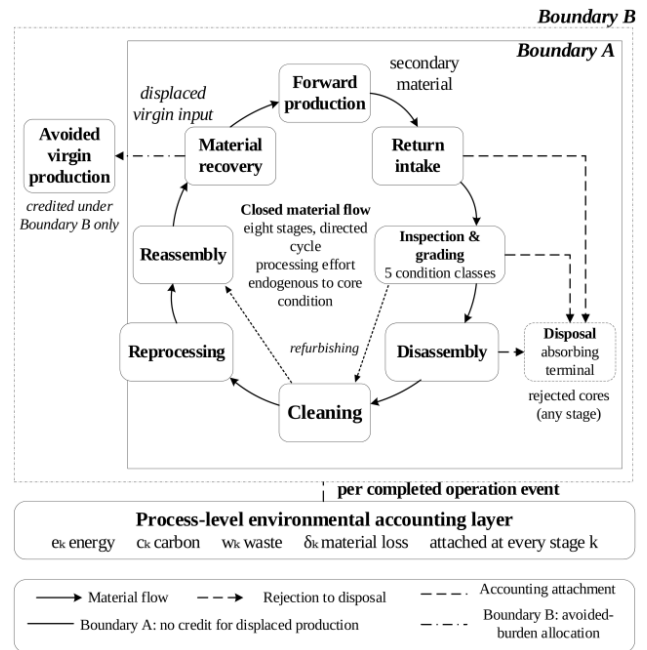


Figure 1. Closed-loop topology with environmental accounting attachment. Eight stages form a directed cycle and rejected cores leave to disposal from any stage; four accounting factors are attached to every completed operation, and the avoided-virgin-production credit is counted only under Boundary B.

The stochastic characteristics were defined as follows: core arrivals are modeled via a non-homogeneous Poisson process that fluctuates according to a weekly pattern; condition class for each arrival is randomly selected from a categorical distribution; processing times in each stage are distributed as lognormal with stage-dependent parameters weighted by condition class; and equipment availability is described by an alternating renewal process with exponential failures and lognormal repairs. Each kilogram of secondary input substitutes $s = 0.85$ kg of virgin input and determines m_{vir} . After discrete-event simulation augmented by life cycle assessment [29], four parameters are allocated to each completed activity k : energy e_k , direct emissions c_k , waste w_k and material loss δ_k ; background data sources rely on the ecoinvent v3.9.1 dataset (cut-off system approach, European datasets wherever possible) [30]. The cut-off approach applies to Boundary A: secondary input carries only its processing load and receives no credit for substituting virgin input production. The avoided burden approach applies to Boundary B: secondary input carries a virgin-like burden multiplied by the substitution factor $SF = 0.80$, and all materials extracted from the returned cores, including the reprocessing scrap delivered to material recovery, receive the corresponding credit. Two different policies are trained for each of these boundaries; Sections 4.1-4.3 consider Boundary A, while the two are compared in Section 4.4.

Parameters were not obtained from any particular industrial process individually. The processing times and failure times distribution were specified to be in the range of values that have been observed in other remanufacturing processes. Steady-state capacity is compared to existing data, mass balance is satisfied, and limits are examined (when the

number of returns is negligible and when all arrivals are of the lowest class). All parameter values along with the ranges from the literature used in the sensitivity analysis are presented in Section 3.4. This parameterization supports a controlled study of the mechanism rather than a prediction for any particular plant, so absolute values are not offered as forecasts.

3.2 Multi-objective MDP formulation and feasibility constraint embedding

Environment: It is modeled as the six-tuple $\langle S, A, P, R, \gamma, \Omega \rangle$ which extends the notation introduced in Section 2.4, where the preference simplex is represented by Ω . Transition Kernel P results from the stochastic processes mentioned above. Instead of being time-based, decision points are event-based. Specifically, the agent receives queries at each point in time when one of the processors completes a workpiece or a workstation becomes idle with pending workpieces.

The circularity reward follows the MCI methodology [31]:

$$MCI = 1 - LFI \cdot F(X), \quad LFI = \frac{V + W}{2M + (W_F - W_C)/2} \quad (5)$$

where LFI is the linear flow index, M the product mass, V the virgin feedstock, W the unrecoverable waste, W_F and W_C the waste from feedstock and component recycling, and $F(X) = 0.9$ for utility $X = 1$. J_{circ} is the episode MCI and $r_{circ,t}$ the per-event change of the running MCI, so J_{circ} equals the undiscounted episode return (training uses $\gamma = 0.99$). The environmental reward is the negative normalized carbon burden of the operations K_t completed at event t :

$$r_{env,t} = -\frac{1}{c} \left[\sum_{k \in K_t} (g e_k + c_k + \phi_w w_k + \phi_v \delta_k) + \phi_v m_t^{vir} + b SF \phi_v (m_t^{sec} - m_t^{rec}) \right] \quad (6)$$

where g is the grid emission factor, ϕ_w and ϕ_v the burdens of waste treatment and virgin material, m_t^{vir} , m_t^{sec} and m_t^{rec} the virgin input, secondary input and the mass recovered from returned cores (reused or sent to material recovery), $b = 0$ (Boundary A) or 1 (Boundary B), and $\bar{c}$ a normalizing constant. $E(\rho)$ is the burden per unit output indexed to 100 at $\rho = 0.05$, and J_{env} the normalized environmental performance (higher is better).

State, action, and notation are summarized in Table 1. The state concatenates five blocks, with the work-in-process (WIP) quality profile represented as a condition-class histogram rather than a mean, since the distribution rather than its first moment drives downstream processing effort. Precedence structure among stages is supplied to the encoder as a graph rather than flattened, preserving the topology that governs feasibility.

Hard constraints are imposed at the policy output rather than through the reward. Let $z_t(a)$ denote the logit assigned to action a before normalization. Masking modifies the logits directly:

$$\tilde{z}_t(a) = z_t(a) + \log \left[g_{prec}(a, s_t) \wedge g_{cap}(a, s_t) \right] \quad (7)$$

Under Eq. (7), infeasible actions receive $-\infty$ and carry zero probability after the softmax, while gradients for feasible actions remain unaffected. As opposed to a penalty that shares a scale with r_{circ} and r_{env} and is therefore tunable, masking eliminates infeasible actions from the support without involving the reward. The precedence flag g_{prec} is determined based on the adjacency matrix of the precedence graph with eight stages and the operations performed on the

work piece so far, while g_{cap} is determined by the buffer capacities and machine availabilities and recalculated at each decision epoch. If the feasible set is empty, the environment moves to the next event without making a decision; this was the case for 0.7% of decision epochs. Unit production cost is trajectory-dependent and cannot be evaluated when a decision is made, so it is ineligible for masking; instead, we apply it as a post-hoc filter over the archive. Unit cost is computed per episode as

$$c_u = \frac{1}{Q} \left[C_{fix} N_{sh} + \sum_k (\lambda_L \tau_k + \lambda_E e_k) + \lambda_V m^{vir} + \lambda_D m^{disp} \right] \quad (8)$$

where C_{fix} is the fixed recovery-line cost per shift, N_{sh} is the number of shifts, τ_k is the processing time of operation k , and λ_L , λ_E , λ_V , and λ_D are the labor, electricity, virgin-material, and disposal rates, m^{disp} is the disposed mass, and Q the units delivered. Quality-dependent effort enters through τ_k and e_k , which scale with each core's quality multiplier. A solution is cost-feasible if its unit cost does not exceed that of the threshold-dispatching reference policy described in Section 3.4, which represents current practice.

Table 1. State representation, action space, and notation

Symbol	Definition	Domain / dimension
s_t^{buf}	Station buffer occupancy	$\mathbb{Z}_{\geq 0}^3$
s_t^{qual}	WIP condition-class histogram	$\mathbb{Z}_{\geq 0}^5$
s_t^{inv}	Virgin and secondary material inventory	$\mathbb{R}_{\geq 0}^2$
s_t^{eq}	Equipment availability indicators	$\{0,1\}^8$
s_t^{dem}	Rolling demand window, $H=14$	$\mathbb{R}_{\geq 0}^{14}$
a_t^{route}	Core routing decision	4 options
a_t^{mix}	Virgin-to-secondary input ratio	$\{0,0.1,\dots,1.0\}$, 11 levels
a_t^{lot}	Lot size and sequencing index	$\{1,\dots,L\}$, $L=6$
g_{prec}	Precedence feasibility indicator	$\{0,1\}$
g_{cap}	Capacity feasibility indicator	$\{0,1\}$
$z_t(a)$	Pre-masking policy logit	$\mathbb{R}$
θ, ϕ	Actor and critic parameters	—

3.3 Preference-conditioned multi-objective actor-critic algorithm

A single network is trained to serve the whole preference space. Each epoch runs 16 parallel episodes, and each episode draws its own preference vector ω from the simplex Ω via Dirichlet(1) sampling, which is supplied alongside the state, so one parameter set encodes a family of policies indexed by ω rather than a single compromise. The architecture is shown in Figure 2.

The learning process uses vector-based utilities, and the critic objectives use the envelope operator [32], so bootstrapping selects the successor state with the highest utility for the selected preference. The overall objective function contains four components:

$$L(\theta, \phi) = L_{actor} + \lambda_1 L_{critic} + \lambda_2 L_{pref} - \lambda_3 H[\pi_\theta] \quad (9)$$

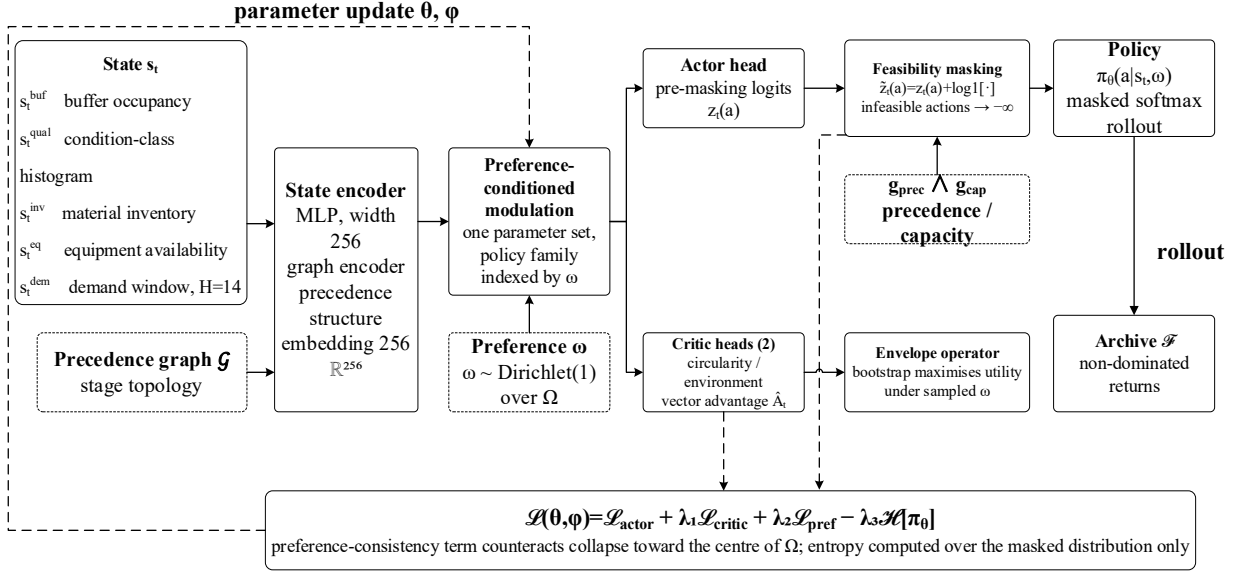


Figure 2. Preference-conditioned actor-critic architecture. State blocks and the stage-precedence graph are encoded and modulated by the sampled preference; a masked actor head and two critic heads with envelope targets are trained with the loss of Eq. (9).

The actor, critic, and entropy terms of Eq. (9) follow the vector-valued formulation. The preference-consistency term L_{pref} penalizes inconsistency between the utility ordering implied by ω and the ordering of returns actually achieved, counteracting the tendency of a shared network to collapse toward the center of preference space. It is a hinge loss over pairs of preferences in a batch:

$$L_{\text{pref}} = \mathbb{E}_{\omega, \omega'} [\max(0, \omega' \cdot \mathbf{Q}(s, \pi_{\theta}(\cdot|s, \omega')) - \omega \cdot \mathbf{Q}(s, \pi_{\theta}(\cdot|s, \omega)) + m)] \quad (10)$$

with margin $m = 0.01$, so that the policy conditioned on ω attains at least the ω -weighted value of policies conditioned on other preferences. The entropic bonus $H[\pi_{\theta}]$ is calculated using the masked distribution, so impossible actions do not affect it. Rewards are kept in a non-dominated archive, and duplicated solutions are removed in objective space.

A limitation is explicitly stated from the beginning: since all objectives fight for the same parameters, we should expect to see the worst results near the boundaries of Ω , which are analyzed in Section 4.2.

3.4 Adaptive-fidelity coupling mechanism and experimental protocol

Training involves two fidelity levels: the fluid surrogate, which models the throughput stage using mean rates but not specific events, and the discrete-event system. In contrast to the fixed proportions used in past multi-fidelity systems, this particular allocation is dependent on :

$$\xi_t = \alpha \frac{H[\pi_{\theta}(\cdot|s_t, \omega)]}{H_0} + (1 - \alpha) \frac{\text{Var}[\hat{A}_t]}{\sigma_0^2} \quad (11)$$

In Eq. (11), H_0 and σ_0^2 are baseline values taken as the running means of policy entropy and advantage variance over the first 5% of the training budget, so that ξ_t is dimensionless and comparable across scales. High entropy or high estimator variance indicates that the policy is still exploring coarsely, where surrogate rollouts suffice. For the surrogate-dominant regime, a fraction $\eta_{\text{lo}} = 0.05$ of the episodes are simulated using the high-fidelity model and correlated with the episodes

using the surrogate through common random number generation (CRN) to assess the return correlation. When ξ_t becomes less than τ , the algorithm goes into the high-fidelity dominant regime ($\eta_{\text{hi}} = 0.6$) because any further enhancement is dependent on the details that the surrogate doesn't capture; if ξ_t becomes greater than $\tau + \delta$ ($\delta = 0.1$), then it switches back to the surrogate-dominant regime. ξ_t is assessed every 10^4 simulation calls, and each transition in any regime counts as one switch. The switching algorithm is detailed in Algorithm 1.

The algorithm is benchmarked against seven alternative algorithms assuming the identical budget of 2×10^6 simulation runs except where indicated: the non-dominated sorting genetic algorithm II (NSGA-II), NSGA-II with feasibility repair operator, the MOEA/D algorithm with Tchebycheff aggregating function, multi-objective particle swarm optimization (MOPSO), PD-MORL (Section 2.3), threshold-based dispatching, and the control group of $N = 11$ independent scalarized PPO algorithms using evenly spaced weights. The threshold-based dispatching policy, which is the current practice, performs remanufacturing of cores of classes ≥ 4 , recertifies class 3 cores, directs cores from classes 1-2 to material recovery, and utilizes all secondary materials. The evolutionary baselines employ an integer encoding of the routing, mix ratio, and lot number, a population size of 100, and pymoo operators [33] with crossover and mutation probabilities set on the hold-out set and evaluated in the high-fidelity model; the repair operator transforms infeasible genes into the closest feasible actions. PD-MORL and PPO both use the masking and fidelity-adaptive interfaces from the proposed method, so the differences stem from the learning algorithm; the control group uses 11 times more budget because it retrains per preference. The wall-clock times are provided for all methods due to the low cost of surrogate evaluations.

The evaluation is performed using 21 preference vectors: the 11 weights used to train the scalarized control group plus the 10 interleaved weights $\omega_1 = 0.05, 0.15, \dots, 0.95$, for which no scalarized agent was trained. Comparison of front quality and cost involves all 21 for the conditioned case and 11 for the

weights trained for the control case, and the conditioned network is also evaluated on the 11 trained weights for a like-for-like test; zero-shot results use the 10 interleaved weights, for which the scalarized case is marked as not applicable; all policies are evaluated greedily. Quality of solution was evaluated using hypervolume (HV), inverted generational distance plus (IGD+), spacing, and C-metric; cost by number of simulation calls and elapsed wall-clock time; coverage by solution set overlap and gap statistics. The objectives are min-max normalized based on the pooled archive, while HV employs the reference point $(-0.05, -0.05)$, and IGD+ employs the pooled non-dominated set.

The recovery rate denotes the proportion of scalarized agents' pooled non-dominated solutions that are ε -dominated by the conditioned archive ($\varepsilon = 0.01$ additive). Since seeds are shared across the methods using CRN, statistical differences are assessed using the exact Wilcoxon signed-rank test with Holm correction for each table. Cost screening of Eq. (8) is conducted on the final archive. The $E(\rho)$ curve is fit to the archive using a penalized cubic smoothing spline (using a value for the degree of smoothing selected via generalized cross-validation).

Algorithm 1. Preference-conditioned multi-objective actor-critic with adaptive fidelity

Input: environments E_{lo}, E_{hi} ; preference space Ω ; parameters $\alpha, \tau, \delta, \eta_{lo}, \eta_{hi}$; budget B
Output: parameters θ, φ ; non-dominated archive F
1: initialize θ, φ ; $F \leftarrow \emptyset$; mode $\leftarrow$ surrogate-dominant ($\eta = \eta_{lo}$); 2: estimate H_0, σ_0^2 over the first $0.05B$ calls
3: while simulation budget $< B$ do
4: for each of 16 parallel episodes, sample $\omega \sim \text{Dirichlet}(1)$ over Ω
5: collect the episode in E_{hi} with probability η , otherwise in E_{lo} (CRN-paired), using $\pi_\theta(\cdot|s, \omega)$ masked by Eq. (7)
6: compute vector advantages $\hat{A}_t$; form envelope targets
7: update θ, φ by minimizing Eq. (9)
8: every 10^4 calls: compute ξ_t by Eq. (11)
9: if mode is surrogate-dominant and $\xi_t < \tau$ then $\eta \leftarrow \eta_{hi}$; reweight replay samples by fidelity
10: else if mode is high-fidelity-dominant and $\xi_t > \tau + \delta$ then $\eta \leftarrow \eta_{lo}$
11: update F with non-dominated returns; filter duplicates
12: end while; 13: return θ, φ, F

Table 2. Experimental configuration and simulation parameters

Item	Setting	Literature range
(a) Learning and evaluation		
Encoder / modulation width; critic heads	256 / 256; 2 (circularity, environment)	
Optimizer, learning rate; discount factor γ	Adam, 3×10^{-4} ; 0.99	
Loss weights $\lambda_1, \lambda_2, \lambda_3$; margin m	0.5, 0.1, 0.01; 0.01	
Switching parameters α, τ ; hysteresis δ	0.5, 0.2; 0.1	
High-fidelity shares η_{lo}, η_{hi} ; ξ_t interval	0.05, 0.6; every 10^4 calls	
Baseline estimation window; demand window H / lot index L	first $0.05B$ calls; 14 / 6	
Preferences sampled per epoch; evaluation preferences	16 (one per parallel episode); 21 (11 trained, 10 interleaved)	
Scalarized control agents N ; penalty weight λ_{pen}	11; 0.5	
Random seeds; simulation budget B ; scale levels	10; 2×10^6 calls; 3, 5, 8 production lines	
HV reference point; ε -dominance tolerance; plateau tolerance ε ; bootstrap resamples	$(-0.05, -0.05)$; 0.01; 5 per unit ρ ; 1,000	
(b) Simulation parameters		
Core arrivals (non-homogeneous Poisson)	42 cores day ⁻¹ , weekly amplitude $\pm 25\%$	30–60 cores day ⁻¹
Quality class probabilities, $q = 1$ (worst) to 5 (best)	0.13, 0.22, 0.30, 0.23, 0.12 (mean 2.99)	mean 2.4–3.6
Quality multiplier on time and energy; recovery yield ($q = 5$ to 1)	$m_q = 1 + 0.3(5 - q)$; 0.92, 0.81, 0.66, 0.49, 0.33	—
Median processing time (min), lognormal $\sigma_{log} = 0.3$	inspection 6, disassembly 18, cleaning 9, reprocessing 24, reassembly 15, material recovery 12, forward production 10	$\pm 40\%$
Energy per operation at $m_q = 1$ (kWh), same order	0.4, 2.1, 1.6, 3.8, 1.2, 2.9, 5.6	$\pm 30\%$
Direct emissions c_k ; waste w_k ; material loss δ_k per operation ($\times m_q$)	0.02–0.30 kg CO ₂ e; 0.05–0.80 kg; 1–6% of core mass	$\pm 30\%$
Buffer limit; demand; horizon T ; normalizer $\bar{c}$	12 per station; Poisson, 40 units day ⁻¹ ; 14 days; 25 kg CO ₂ e	—
Failures; repairs	exponential, mean time between failures 160 h; lognormal, median 2.5 h, $\sigma_{log} = 0.5$	100–250 h
Grid emission factor g	0.45 kg CO ₂ e kWh ⁻¹ , applied to process energy and to the electricity share (0.8) of φ_v	0.30–0.60
Virgin material burden φ_v ; waste burden φ_w ; product mass	3.6 kg CO ₂ e kg ⁻¹ ; 0.5 kg CO ₂ e kg ⁻¹ ; 6.5 kg	3.0–4.5; 0.3–0.8; —
Material substitution rate s (virgin displaced per kg secondary); SF (Boundary B)	0.85; 0.80	0.70–0.95; 0.70–0.95
Cost rates C_{fix} ; λ_L ; λ_E ; λ_V ; λ_D	1,850 € shift ⁻¹ ; 38 € h ⁻¹ ; 0.18 € kWh ⁻¹ ; 2.4 € kg ⁻¹ ; 0.12 € kg ⁻¹	—

With $\varepsilon = 5$ index points per unit of ρ , ρ_1 is the value at which the smoothed derivative falls into the range $|dE/d\rho| \leq \varepsilon$, and ρ_2 is the value at which it leaves that range, with 95% confidence intervals obtained via bootstrapping across 1,000 replications per seed. To break the link between archive values and policy selection, a controlled sweep is applied, overriding the 11-level grid and setting the mix action level to 0.05–0.95 in increments of 0.05 (with values of 0.02 near ± 0.06 of the transition point) while conditioning the network to select other actions at $\omega = (0.5, 0.5)$. The full parameterization is provided in Table 2; hyperparameters were tuned on a hold-out scenario set disjoint from the evaluation scenarios (Section 4.1).

4. Results

4.1 Training convergence, sample efficiency, and computational scalability

Training trajectories for ten seeds are displayed in Figure 3. The full method is compared with a fixed-ratio variant and a high-fidelity-only variant. With the terminal hypervolume of the high-fidelity-only variant ($HV = 0.742$) as the target, the full method converges after 8.6×10^5 simulation calls, the fixed-ratio variant requires 11.9×10^5 and the variant without variance reduction — that is, without common-random-number synchronization across the two fidelity levels — requires 10.4×10^5 . The high-fidelity-only run reaches this value only at the end of its 2.0×10^6 -call budget (17.4 h). Narrowing of the confidence bands after the first switching point indicates that the estimator variance also decreases.

Figure 3(b) indicates that the process of switching is not instant. Rollouts are largely carried out using the surrogate until the point when one-third of the budget (median first switch at 0.34B), whereupon the joint criterion value drops below the threshold, triggering the increased usage of the high-fidelity model. The number of switches on average per seed is 3.2 ± 0.9 , and 41.3% of rollouts take place in the high-fidelity environment, compared to 50% in the fixed ratio case. The switches involve reverting to the surrogate dominant case, while the value of 58% in Figure 3(b) involves η_{hi} and seeds not switched. Correlation between returns in paired episodes is 0.91 for circularity and 0.84 for burden. For the grid search over $\alpha \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$, $\tau \in \{0.10, 0.15, 0.20, 0.25, 0.30\}$ and windows of 2.5%, 5% and 10% of the budget in the hold-out scenario set, $\alpha=0.5$, $\tau=0.2$ and 5% were chosen (8.8×10^5 calls to target, against up to 10.4×10^5 elsewhere); a low τ delayed switching and left the policy underspecified, whereas a high τ switched before the surrogate had been exploited. Table 3 summarizes the costs at the reference scale of five lines and eleven mix-ratio levels. Deactivating adaptive switching increases the call cost by 38.4% and wall-clock time from 6.3 hours to 8.7 hours; deactivating variance reduction alone costs 20.9%; deactivating both costs 59.3%. In other words, adaptive switching makes the number of calls to the target HV 27.7% lower and wall-clock time 27.6% lower. The wall-clock time per 105 calls in the mixed fidelity setting is almost the same (0.73 hours), as updating the policy dominates the calculations in this range (9.4 milliseconds per high-fidelity and 0.8 milliseconds per surrogate call).

Table 3. Cost ablation of the adaptive-fidelity mechanism

Configuration	Calls to target HV ($\times 10^5$)	Wall-clock (h)	Switches	HF share (%)	Cost increase (%)
Full method	8.6 ± 0.7	6.3 ± 0.5	3.2 ± 0.9	41.3	—
w/o adaptive switching	11.9 ± 1.1	8.7 ± 0.8	n/a	50.0	+38.4
w/o variance reduction	10.4 ± 1.3	7.6 ± 0.9	3.4 ± 1.1	43.7	+20.9
w/o both	13.7 ± 1.6	9.9 ± 1.2	n/a	50.0	+59.3
High-fidelity only (reference)	20.0	17.4 ± 1.3	n/a	100	+132.6

Note: All configurations are measured at the reference scale (five lines, eleven mix-ratio levels). The target HV is the terminal value of the high-fidelity-only run (0.742), reached only at the end of its budget. Mean $\pm$ SD over ten seeds. "w/o variance reduction" denotes removal of common-random-number synchronization between the surrogate and the high-fidelity environment; "w/o adaptive switching" denotes the fixed 50/50 fidelity allocation. Cost increase is the relative increase in calls to target HV with respect to the full method; switches count mode changes in both directions; HF = high-fidelity.

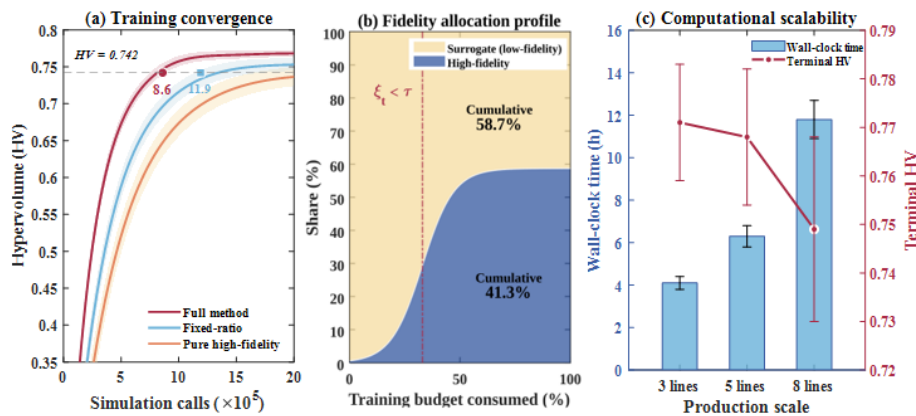


Figure 3. Training convergence, fidelity allocation, and scalability: (a) HV against simulation calls (ten-seed mean, 95% band); (b) rollout share per fidelity; (c) wall-clock time and terminal HV by scale.

Scalability can be seen in Figure 3(c): the wall clock time to target increases from 4.1 hours to 11.8 hours, whereas the terminal hypervolume decreases from 0.771 to 0.749 (-2.9%).

4.2 Pareto front coverage, zero-shot preference generalization, and baseline comparison

The conditioned network ϵ -dominates 87.4% of the pooled non-dominated solutions of the eleven separate linear scalarization agents, while using only 9.1% of their total training cost. Figure 4(a) illustrates the techniques in the objective space (PD-MORL and the corrected NSGA-II in Table 4 only). NSGA-II, MOEA/D, and MOPSO do not generate any non-dominated solutions in the shaded concave portion. This is a result of current experimental conditions, not a problem inherent to dominance or Tchebycheff-based search; the concave portion needs coordination between multiple epochs for successful solution generation, something that variation-based operators are incapable of. Nineteen point eight percent of non-dominated solutions generated by the conditioned network and twelve point five percent by the scalarized agents are located in the concave portion, where the scalarized agents are approximate maximizers of Eq (2), as opposed to being exact maximizers, which is why the latter can be close to a concave portion, and the points generated by the threshold heuristic are dominated. Performance across the preference space (Figure 4(b)) is uneven. Relative hypervolume contributions stay within 3% of the stable-region mean for $\omega_1 \in [0.10, 0.90]$ but drop at the extremes: by 14.9% and 13.8% at the vertices $\omega_1 = 0$ and 1, and by 7.4% and 10.6% at the interleaved weights 0.05 and 0.95. The decrease is roughly symmetric and occurs where the required behavior lies farthest from the population mean, as anticipated in Section 3.3; extreme preferences therefore benefit from tailored training. At the eight interior interleaved weights, the relative hypervolume deviates from the mean of the two adjacent grid weights by at most 2.8% ($\omega_1 = 0.75$), and by +1.5% and -3.4% at 0.05 and 0.95, so zero-shot transfer is nearly lossless except near the vertices.

Comparison of the techniques is shown in Table 4. The proposed technique yields 9.4% better HV than NSGA-II and 14.5% than MOPSO, and all differences between the two evolutionary techniques are statistically significant (Holm-adjusted $p=0.014$). The repair operator increases NSGA-II HV to 0.718, reducing the difference to 7.0%, and PD-MORL attains 0.749, which is 2.5% lower than the proposed method. Over the 11 common weights, the improvement on the scalarized benchmark is 3.1% (0.764 compared to 0.741, with nine out of ten seeds winning) and 3.6% in the case of using the conditioned front across all 21 weights. This improvement is obtained using just an eleventh of the budget and is the pertinent comparison. In case of spacing, the only metric where the other techniques excel, MOEA/D performs better (0.048 versus 0.052), due to uniformity of the decomposition weights, even though the difference is not statistically significant after adjustment. The superiority therefore consists in coverage and economy of effort and not in spacing. The evolutionary algorithms are quicker in wall-clock terms (5.3-5.6 h versus 14.6 h), since they require no network training, but no technique spans the concave part; surrogating them was not tried either.

4.3 Ablation of key components and scenario robustness

Results of the ablations are shown in Table 5. Feasibility masking produces the biggest decline (-10.3%, from HV = 0.768 to 0.689), while the penalty variant with the optimized penalty weight over $\lambda \text{ pen} \in \{0.1, 0.5, 1.0, 2.0\}$ on the hold-

out set still results in 6.8% of infeasible actions (4.1% precedence and 2.7% capacity constraints). Higher penalties decrease violations (3.1% at $\lambda \text{ pen} = 2.0$) but also decrease the HV (0.661), while lower penalties lead to higher violations (11.9% for 0.1) but higher HV (0.672). Thus, 10.3% violations is the best result among penalties. All infeasible actions are performed as no-operations, and the penalty is used exclusively in training rewards and not in the HV. The elimination of preference conditioning, which means learning one policy using scalarization per run, has a cost of 7.2%. The contribution of the graph encoder is the lowest (4.4%). This shows that precedence relations provide valuable but not critical information after observing buffer and equipment states (Holm-adjusted $p = 0.006$ for all three ablations). Perturbation does not cause failure of the algorithm but rather results in graceful degradation. For a 50% increase in return volatility, HV is reduced by 5.1%, and the fidelity transition becomes delayed since a larger advantage variance means that ξ_t stays above τ for a longer period (median switch to 0.43B; return correlation is now 0.78 and 0.71). Quality perturbation is more harmful (8.1%), which is intuitive since quality impacts effort in processing. The 30% reduction in grid carbon intensity leads to a modest but statistically significant 3.1% decrease in performance (Holm-corrected $p = 0.020$), as its effect is mostly on the environmental objective component related to energy. In all cases, ξ_t converged on the budget, and the archive was non-empty for all seeds; switch statistics are similar to the baseline, with the exception of return volatility (Table 5, Panel B).

4.4 Regime boundary identification and the cost-feasible decision region

Environmental burden is non-monotonic in secondary-material input under both accounting conventions, and the simulation evidence supports the three-regime conjecture of Section 2.1. Input was varied through the range $\rho \in [0.05, 0.95]$. Under Boundary A, burden falls by 17.2% as ρ increases from 0.05 to the first transition point $\rho_1 = 0.38$ [0.34, 0.42], so circularity and environmental performance improve together in this range. A plateau follows, and beyond the second transition point $\rho_2 = 0.61$ [0.56, 0.66] burden rises by 9.6% to the end of the sweep. The reversal is therefore present in the model but modest in magnitude relative to the gain available in R_{syn} . This process is a result of the quality of the marginal core. With $\rho = 0.10$ the quality class of this core is 4.1, while the processing cost amounts to 1.0 kg CO_{2e} per kg of secondary material used, much lower than the 3.6 kg CO_{2e} kg⁻¹ for the replaced virgin material. With $\rho = 0.60$, the class declines to 2.3. The effect of the quality multiplier increases the amount of energy required for inspection, disassembly, and reprocessing and distributes this energy over a smaller mass of secondary material, thus increasing the processing burden to 3.3 kg CO_{2e} kg⁻¹, with no gains from substitution left. According to Boundary B, however, the processing burden for the reused materials is balanced by the recycling credit, resulting in fewer substitution gains and a smaller initial reduction in the processing burden. Moreover, the scrap from low-quality cores, with their own recycling credit rather than processing burden, shifts both transition points to higher ρ . Figure 5(a) compares the two conventions. Under Boundary B the two transition points move to the right to $\rho_1=0.47$ [0.42, 0.52] and $\rho_2=0.78$ [0.71, 0.84], while R_{inv} shrinks from 0.34 to 0.17 within the swept interval.

Table 4. Front quality comparison against baselines

Method	HV	IGD+	Spacing	C-metric	Budget ($\times 10^6$)	Wall-clock (h)	Concave share (%)	Zero-shot
Ours (conditioned)	0.768 ± 0.014	0.041 ± 0.006	0.052 ± 0.009	—	2.0	14.6 ± 0.9	19.8	Available
PD-MORL	$0.749 \pm 0.017^*$	$0.052 \pm 0.007^*$	0.069 ± 0.011	0.38	2.0	15.9 ± 1.1	14.3	Available
Scalarized PPO (N = 11)	$0.741 \pm 0.019^*$	$0.058 \pm 0.008^*$	0.087 ± 0.014	0.34	22.0	154.0	12.5	n/a
NSGA-II	$0.702 \pm 0.023^*$	$0.079 \pm 0.011^*$	0.061 ± 0.010	0.58	2.0	5.4 ± 0.3	0.0	n/a
NSGA-II + repair	$0.718 \pm 0.020^*$	$0.072 \pm 0.010^*$	0.058 ± 0.010	0.52	2.0	5.6 ± 0.3	0.0	n/a
MOEA/D	$0.694 \pm 0.021^*$	$0.083 \pm 0.010^*$	0.048 ± 0.008	0.61	2.0	5.3 ± 0.3	0.0	n/a
MOPSO	$0.671 \pm 0.026^*$	$0.094 \pm 0.013^*$	0.073 ± 0.012	0.66	2.0	5.4 ± 0.3	0.0	n/a
Threshold heuristic	$0.583 \pm 0.008^*$	$0.146 \pm 0.009^*$	0.118 ± 0.015	0.89	—	0.2	28.6	n/a

Note: Mean $\pm$ SD over ten seeds at the reference scale. C-metric is the fraction of each method's solutions dominated by the proposed method. * denotes Holm-adjusted $p < 0.05$ against the proposed method (Wilcoxon signed-rank; scalarized PPO tested on the 11 shared weights, where the proposed method scores 0.764 ± 0.015). Tests cover HV, IGD+, and Spacing; lower is better for IGD+ and Spacing. PPO wall-clock is summed over 11 agents; concave share refers to the shaded segment of Figure 4(a).

Table 5. Component ablation and scenario robustness**Panel A — Component ablation**

Configuration	HV	IGD+	Spacing	Degradation (%)	Violations (%)
Full method	0.768 ± 0.014	0.041 ± 0.006	0.052 ± 0.009	—	0.0
w/o preference conditioning	0.713 ± 0.021	0.066 ± 0.009	0.079 ± 0.012	7.2	0.0
w/o feasibility masking	0.689 ± 0.028	0.081 ± 0.012	0.094 ± 0.016	10.3	6.8 (4.1 / 2.7)
w/o graph encoder	0.734 ± 0.017	0.055 ± 0.007	0.061 ± 0.010	4.4	0.0

Panel B — Scenario robustness

Scenario	HV	IGD+	Spacing	Degradation (%)	Switches / HF share (%)
Baseline	0.768 ± 0.014	0.041 ± 0.006	0.052 ± 0.009	—	$3.2 \pm 0.9 / 41.3$
Return volatility +50%	0.729 ± 0.022	0.053 ± 0.008	0.064 ± 0.011	5.1	$4.1 \pm 1.2 / 37.6$
Quality distribution shifted down	0.706 ± 0.025	0.061 ± 0.009	0.070 ± 0.013	8.1	$3.5 \pm 1.0 / 40.2$
Grid carbon intensity -30%	0.744 ± 0.016	0.047 ± 0.007	0.056 ± 0.010	3.1	$3.1 \pm 0.8 / 41.0$

Note: In Panel A, degradation is the relative HV loss of each ablated variant with respect to the full method; in Panel B, it is the relative HV loss of the full method in each perturbed scenario with respect to the baseline scenario. Violations: share of decisions breaking precedence / capacity constraints. All degradations are significant (Holm-adjusted $p \leq 0.020$); the quality shift lowers the mean class by 20%. Mean $\pm$ SD, ten seeds.

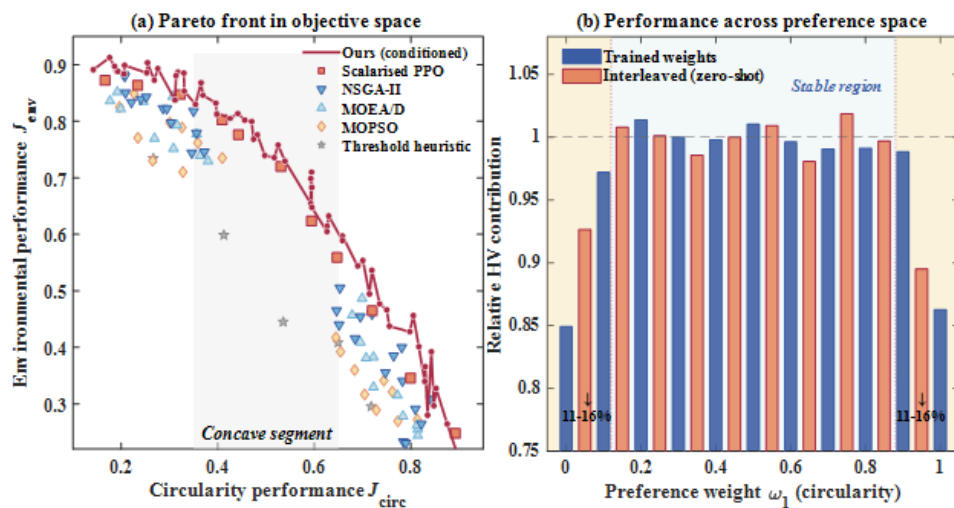


Figure 4. Pareto front coverage and preference-space performance: (a) non-dominated solutions, concave segment shaded; (b) relative HV contribution per weight, with the stable region shaded; the 11–16% labels bound the drops at the vertex weights $\omega_1 = 0$ and 1 (Section 4.2).

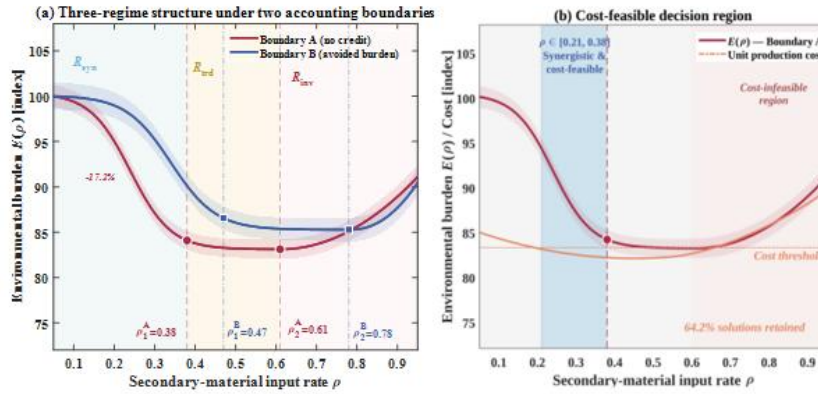


Figure 5. Regime structure under two accounting boundaries and the cost-feasible region: (a) $E(\rho)$ under Boundaries A and B, each indexed to 100 at $\rho = 0.05$, with 95% bootstrap bands; (b) $E(\rho)$ under Boundary A and unit cost on the same index scale, where the reference-policy threshold corresponds to 83.2.

The structure survives the change of convention, but the boundaries do not: a displacement of 0.09 in ρ_1 (paired bootstrap 95% interval of the difference [0.03, 0.15]) is sufficient to change an operational boundary.

The end of the synergistic regime must therefore be read relative to the convention, which has to be stated explicitly. The controlled fixed-ratio sweep delivers $\rho_1 = 0.37$ [0.33, 0.41] and $\rho_2 = 0.62$ [0.57, 0.67] under Boundary A and $\rho_1 = 0.48$ [0.43, 0.53] and $\rho_2 = 0.77$ [0.71, 0.83] under Boundary B, all within the confidence interval of the archived estimate; the burden is reduced by 16.8% until ρ_1 and increased by 9.9% above ρ_2 . For $\omega_1 = 0.25$ and 0.75, the critical points change by a maximum of 0.02, while the realized ρ differs from the targeted ratio by 0.018 on average (with a maximum of 0.041). Thus, the regime configuration is not an artifact of the policy's choice of ρ .

The one-factor sensitivity in Figure 6 shows that the mean of the core-quality distribution dominates ρ_1 : a $\pm 20\%$ perturbation shifts it by -0.11 to $+0.09$. The material substitution rate follows (-0.08 to $+0.07$), consistent with the substitution mechanism described above: incomplete displacement erodes the benefit that sustains R_{syn} . Reprocessing energy intensity, the grid emission factor, and the disassembly time coefficient each displace ρ_1 by no more than 0.06, and higher processing energy or longer disassembly moves ρ_1 to lower values, so the boundary is stable with respect to process and energy parameters but sensitive to what enters the loop and what it replaces. The core-quality effect is similar to the convention shift (0.09). $\pm 20\%$ variation bracket spans the range covered by the literature on core quality and substitution rate shown in Table 2; on the upper and lower limits of this wider range, ρ_1 varies from 0.32 to 0.46 in case of grid factor (0.30 and 0.60), respectively and varies between 0.31 and 0.47 for reprocessing energy ($\pm 30\%$) and disassembly time ($\pm 40\%$). The interaction between core quality and substitution rate is marginal; ρ_1 at -20% is 0.22 (additive 0.19) and $+20\%$, 0.53 (0.54). The cost constraint of Eq. (8) preserves 64.2% of the database with infeasible solutions at either end of the ρ range. In Boundary A, the cost-constraint intersection with R_{syn} results in $\rho \in [0.21, 0.38]$ (Figure 5(b)).

For values below 0.21, the reduction in virgin metal does not cover the recovery line's operational cost despite the low effort per core; values above 0.38 yield no additional synergy, while values above 0.63 make the solution cost-infeasible because reprocessing poor-quality cores is required. In Boundary B, the cost constraint keeps 59.7% of the database and yields $\rho \in [0.22, 0.47]$, with the upper boundary at the end of R_{syn} ; the lower boundary differs slightly from Boundary A because the archives are trained independently. This interval, rather than any single solution, is the analysis's operational output.

5. Discussion

The co-optimization concept is conditional in this study. The three regimes of Section 4.4 define the domain in which an increase in the amount of secondary material leads to an increase in environmental pressure. In this case, the cost-feasible synergistic range of $\rho \in [0.21, 0.38]$ under Boundary A and $\rho \in [0.22, 0.47]$ under Boundary B converts the co-optimization concept of a circle into an operating principle of the bounded type and not into a direction. It is a burden-shifting effect at the plant level that differs from the market-level circular economy rebound phenomenon in which the partial replacement makes the environmental benefit invalid due to demand reactions [34]. The contribution of this study lies in quantifying the reversal within a single plant and in showing that the result depends on the accounting convention.

Preference-trained network yields a range of compromises, instead of one, and costs less than one-tenth that of re-training for each weight for $\omega_1 \in [0.10, 0.90]$. It is in line with the change in RL applied to production planning from just scheduling to multi-criteria RL [35]. Inconsistency between the outer and inner spaces relates to sharing the resources by one parameter pool: at the center of the simplex, the behavior is very close to the average for the whole population, which means that the inner weight resembles a blend of its neighbors (up to 2.8% deviation). The outermost preferences will have profiles which deviate most from the mean values, but they all need to have identical parameters; therefore, the deterioration of 13.8-14.9% on the vertices will be due to parameter sharing, not due to poor learning [36].

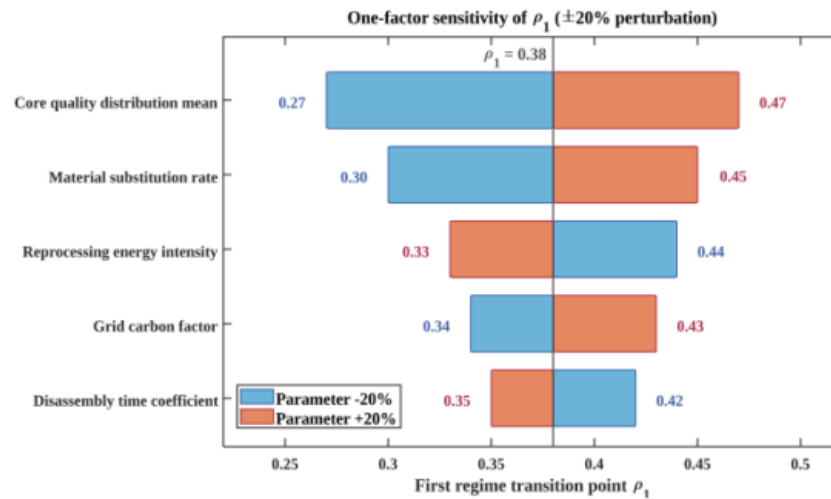


Figure 6. One-factor sensitivity of ρ_1 . Bars show ρ_1 when each parameter is lowered or raised by 20% from its baseline; the vertical line marks the baseline $\rho_1 = 0.38$ under Boundary A.

A solution to this problem would be prioritized sampling of underrepresented areas. The margin over PD-MORL (2.5% in HV) is much less than that over the evolutionary methods (7.0-14.5%), indicating that the main difference comes from the problem formulation, not from the architecture itself.

The benefit of adaptive switching depends on how strongly surrogate and high-fidelity returns are correlated: increased return variability kept the criterion above the threshold longer and lowered the correlations from 0.91 and 0.84 to 0.78 and 0.71, so noisier environments will benefit less. The removal of feasibility masking reduced HV by 10.3%, or equivalently, masking provides an increase of 11.5% relative to the best-tuned penalty version without any constraint violations. This is consistent with feasibility condition control based on reinforcement learning in hybrid disassembly systems [37]. Feasibility control through masking is confined to the information that can be carried by a mask; thus, quantities such as unit cost cannot be monitored at each time step and are only filtered a posteriori in this case, while constrained RL, applied to dynamic material handling [38], could have considered these quantities during training. Masks can be generalized to other action spaces through parametric masks [39] or learned from states [40].

As far as policy is concerned, the recycled content requirements in the EU batteries fall in the range of secondary inputs' percentage without giving credit for displaced production (near Boundary A), together with Extended Producer Responsibility (EPR) charges based on recycled content or recyclability, but not the Circular Footprint Formula in the context of the Product Environmental Footprint approach that divides impacts and credits between the suppliers and users of recycled products through an allocation factor [41] (between A and B). In terms of the limitations of the synthetic one-plant model, a system of incentives for circularity rates only can drive the process past the inversion point, and given that ρ_1 is sensitive to the core quality, it is more reasonable to set procurement targets on the quality distribution than on the input percentage. The application of intervals in terms of a one-plant case needs line calibration, core-quality distribution, regionalized inventories, and transfer to the industry, such as electric vehicle battery remanufacturing.

The adaptive preference-conditioned framework developed here also bears conceptual resemblance to AI-driven decision architectures in other operational domains. For instance, an omni-channel retail evaluation framework integrating big data and reinforcement learning for personalized marketing optimization [42] shares a similar logic of preference-guided multi-objective adaptation. Likewise, an AI-based risk management decision framework for chemical enterprises [43] demonstrates the broader applicability of adaptive multi-criteria decision paradigms under uncertainty. These cross-domain parallels suggest that the methodological contributions of this paper may generalize to sectors beyond remanufacturing where constrained, multi-objective decisions are made under stochastic conditions.

6. Conclusion

Modeled as an MOMDP with a discrete-event simulation, the multi-line closed-loop plant incorporating process-level environmental accounting was solved using a preference-conditioned actor-critic algorithm based on feasibility masking and adaptive-fidelity scheduling. The burden on the environment is a non-monotonic function of secondary materials used in both accounting methods. The critical transition points occur at $\rho_1 = 0.38$ and $\rho_2 = 0.61$ for the cut-off scheme, which grants no credit for displaced virgin production, and move to 0.47 and 0.78 under avoided-burden accounting: the structure persists, but its boundaries shift with the convention. A single conditioned network recovers most of the scalarized front at less than one tenth of its budget, adaptive switching reduces the simulation calls needed to reach the reference hypervolume by 27.7%, and the unit-cost screen confines the jointly synergistic and cost-feasible region to $\rho \in [0.21, 0.38]$ under Boundary A. Instead of assuming co-benefit, we have found an interval whose lower bound is the dominant sensitivity and upper bound is the cost feasibility, which has been presented in the form of a set of policies that can be accessed at any preference point. This research is constrained by the one-facility assumption, the use of literature-based artificial data instead of field data, and a two-objective analysis with limited coverage of the preference-space boundaries. A more robust policy recommendation requires, among other things, at least three

additional findings: generalizing the boundaries to multiple plants, real data on core quality in the returns, and a three-objective approach to optimizing unit cost directly.

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically regarding authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with research ethics policies. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere. All survey participants provided informed consent before participating, and the research team strictly maintained the anonymity and confidentiality of all respondents throughout the research process.

Data availability statement

The manuscript contains all the data. However, additional data will be provided by the corresponding author upon reasonable request.

Conflict of interest

The authors declare no potential conflict of interest.

References

- [1] European Parliament and Council of the European Union, "Regulation (EU) 2023/1542 of the European Parliament and of the Council of 12 July 2023 concerning batteries and waste batteries, amending Directive 2008/98/EC and Regulation (EU) 2019/1020 and repealing Directive 2006/66/EC," Official Journal of the European Union, vol. L 191, pp. 1–117, 28 Jul. 2023. [Online]. Available: <http://data.europa.eu/eli/reg/2023/1542/oj>
- [2] L. Li, F. Dababneh and J. Zhao, "Cost-effective supply chain for electric vehicle battery remanufacturing," *Applied Energy*, vol. 226, pp. 277–286, 2018. DOI: <https://doi.org/10.1016/j.apenergy.2018.05.115>
- [3] H. Liao, Q. Zhang, N. Shen and L. Li, "Stochastic analysis of quality uncertainty and optimal acquisition strategies for engine remanufacturing," *Journal of Cleaner Production*, vol. 261, Art. no. 121088, 2020. DOI: <https://doi.org/10.1016/j.jclepro.2020.121088>
- [4] L. Rigamonti and E. Mancini, "Life cycle assessment and circularity indicators," *The International Journal of Life Cycle Assessment*, vol. 26, no. 10, pp. 1937–1942, 2021. DOI: <https://doi.org/10.1007/s11367-021-01966-2>
- [5] W. Zhang, Y. Zheng and R. Ahmad, "The integrated process planning and scheduling of flexible job-shop-type remanufacturing systems using improved artificial bee colony algorithm," *Journal of Intelligent Manufacturing*, vol. 34, no. 7, pp. 2963–2988, 2023. DOI: <https://doi.org/10.1007/s10845-022-01969-2>
- [6] M. I. Rizova, T. C. Wong and W. Ijomah, "A systematic review of decision-making in remanufacturing," *Computers & Industrial Engineering*, vol. 147, Art. no. 106681, 2020. DOI: <https://doi.org/10.1016/j.cie.2020.106681>
- [7] M. Panzer and B. Bender, "Deep reinforcement learning in production systems: A systematic literature review," *International Journal of Production Research*, vol. 60, no. 13, pp. 4316–4341, 2022. DOI: <https://doi.org/10.1080/00207543.2021.1973138>
- [8] Y. Li, W. Gu, M. Yuan and Y. Tang, "Real-time data-driven dynamic scheduling for flexible job shop with insufficient transportation resources using hybrid deep Q network," *Robotics and Computer-Integrated Manufacturing*, vol. 74, Art. no. 102283, 2022. DOI: <https://doi.org/10.1016/j.rcim.2021.102283>
- [9] M. Cutler, T. J. Walsh and J. P. How, "Real-world reinforcement learning via multifidelity simulators," *IEEE Transactions on Robotics*, vol. 31, no. 3, pp. 655–671, 2015. DOI: <https://doi.org/10.1109/TRO.2015.2419431>
- [10] G. Lonca, R. Muggéo, H. Imbeault-Tétreault, S. Bernard and M. Margni, "Does material circularity rhyme with environmental efficiency? Case studies on used tires," *Journal of Cleaner Production*, vol. 183, pp. 424–435, 2018. DOI: <https://doi.org/10.1016/j.jclepro.2018.02.108>
- [11] C. G. Castro, A. H. Trevisan, D. C. A. Pigosso and J. Mascarenhas, "The rebound effect of circular economy: Definitions, mechanisms and a research agenda," *Journal of Cleaner Production*, vol. 345, Art. no. 131136, 2022. DOI: <https://doi.org/10.1016/j.jclepro.2022.131136>
- [12] D. Font Vivanco, J. Freire-González, R. Galvin, T. Santarius, H. J. Walnum, T. Makov et al., "Rebound effect and sustainability science: A review," *Journal of Industrial Ecology*, vol. 26, no. 4, pp. 1543–1563, 2022. DOI: <https://doi.org/10.1111/jiec.13295>
- [13] X. Guo, M. Zhou, A. Abusorrah, F. Alsokhry and K. Sedraoui, "Disassembly sequence planning: A survey," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 7, pp. 1308–1324, 2021. DOI: <https://doi.org/10.1109/JAS.2020.1003515>
- [14] X. Guo, Z. Zhang, L. Qi, S. Liu, Y. Tang and Z. Zhao, "Stochastic hybrid discrete grey wolf optimizer for multi-objective disassembly sequencing and line balancing planning in disassembling multiple products," *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 3, pp. 1744–1756, 2022. DOI: <https://doi.org/10.1109/TASE.2021.3133601>
- [15] Y. Fu, Z. Zhang, P. Liang, G. Tian and C. Zhang, "Integrated remanufacturing scheduling of disassembly, reprocessing and reassembly considering energy efficiency and stochasticity through group teaching optimization and simulation approaches," *Engineering Optimization*, vol. 56, no. 12, pp. 2018–2039, 2024. DOI: <https://doi.org/10.1080/0305215X.2023.2296538>
- [16] J. Zhao, S. Peng, T. Li, S. Lv, M. Li and H. Zhang, "Energy-aware fuzzy job-shop scheduling for engine remanufacturing at the multi-machine level," *Frontiers of Mechanical Engineering*, vol. 14, no. 4, pp. 474–488, 2019. DOI: <https://doi.org/10.1007/s11465-019-0560-z>
- [17] H. Lu, D. Herman and Y. Yu, "Multi-objective reinforcement learning: Convexity, stationarity and Pareto optimality," in *Proc. 11th International*

- Conference on Learning Representations (ICLR), 2023. [Online]. Available: <https://openreview.net/forum?id=TjEzIsyEsQ6>
- [18] T. Kegyes, Z. Süle and J. Abonyi, "Disassembly line optimization with reinforcement learning," *Central European Journal of Operations Research*, vol. 32, no. 4, pp. 1115–1142, 2024. DOI: <https://doi.org/10.1007/s10100-024-00906-3>
- [19] C. F. Hayes, R. Rădulescu, E. Bargiacchi, J. Källström, M. Macfarlane, M. Reymond et al., "A practical guide to multi-objective reinforcement learning and planning," *Autonomous Agents and Multi-Agent Systems*, vol. 36, no. 1, Art. no. 26, 2022. DOI: <https://doi.org/10.1007/s10458-022-09552-y>
- [20] F. Felten, E.-G. Talbi and G. Danoy, "Multi-objective reinforcement learning based on decomposition: A taxonomy and framework," *Journal of Artificial Intelligence Research*, vol. 79, pp. 679–723, 2024. DOI: <https://doi.org/10.1613/jair.1.15702>
- [21] T. Basaklar, S. Gumussoy and U. Y. Ogras, "PD-MORL: Preference-driven multi-objective reinforcement learning algorithm," in *Proc. 11th International Conference on Learning Representations (ICLR)*, 2023. [Online]. Available: <https://openreview.net/forum?id=zS9sRyaPFIJ>
- [22] M. Reymond, C. F. Hayes, D. Steckelmacher, D. M. Roijers and A. Nowé, "Actor-critic multi-objective reinforcement learning for non-linear utility functions," *Autonomous Agents and Multi-Agent Systems*, vol. 37, no. 2, Art. no. 23, 2023. DOI: <https://doi.org/10.1007/s10458-023-09604-x>
- [23] S. Lassoued, L. S. Baheti, N. Weiß-Borkowski, S. Lier and A. Schwung, "Flexible Manufacturing Systems intralogistics: Dynamic optimization of AGVs and tool sharing using Colored-Timed Petri Nets and actor-critic RL with actions masking," *Journal of Manufacturing Systems*, vol. 82, pp. 405–419, 2025. DOI: <https://doi.org/10.1016/j.jmsy.2025.06.017>
- [24] M. Stappert, B. Lutz, N. Goby and D. Neumann, "Integrating human knowledge through action masking in reinforcement learning for operations research," *Computers & Industrial Engineering*, vol. 222, Art. no. 112340, 2026. DOI: <https://doi.org/10.1016/j.cie.2026.112340>
- [25] W. Song, X. Chen, Q. Li and Z. Cao, "Flexible job-shop scheduling via graph neural network and deep reinforcement learning," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 1600–1610, 2023. DOI: <https://doi.org/10.1109/TH.2022.3189725>
- [26] A. Seipolt, R. Buschermöhle, V. Haag, W. Hasselbring, M. Höfinghoff, M. Schumacher et al., "Reinforcement learning and digital twin-driven optimization of production scheduling with the digital model playground," *Discover Internet of Things*, vol. 4, no. 1, Art. no. 34, 2024. DOI: <https://doi.org/10.1007/s43926-024-00087-0>
- [27] S. Khairy and P. Balaprakash, "Multi-fidelity reinforcement learning with control variates," *Neurocomputing*, vol. 597, Art. no. 127963, 2024. DOI: <https://doi.org/10.1016/j.neucom.2024.127963>
- [28] F. Paschko, A. Krini, M. Kemke and S. Knorn, "Reinforcement learning approach for decision making in disassembly processes," *Journal of Intelligent Manufacturing*, vol. 37, no. 5, pp. 1829–1855, 2026. DOI: <https://doi.org/10.1007/s10845-025-02622-4>
- [29] B. Löfgren and A.-M. Tillman, "Relating manufacturing system configuration to life-cycle environmental performance: discrete-event simulation supplemented with LCA," *Journal of Cleaner Production*, vol. 19, no. 17–18, pp. 2015–2024, 2011. DOI: <https://doi.org/10.1016/j.jclepro.2011.07.014>
- [30] G. Wernet, C. Bauer, B. Steubing, J. Reinhard, E. Moreno-Ruiz and B. Weidema, "The ecoinvent database version 3 (part I): overview and methodology," *The International Journal of Life Cycle Assessment*, vol. 21, no. 9, pp. 1218–1230, 2016. DOI: <https://doi.org/10.1007/s11367-016-1087-8>
- [31] Ellen MacArthur Foundation and Granta Design, "Circularity Indicators: An Approach to Measuring Circularity – Methodology," Ellen MacArthur Foundation, Cowes, UK, 2015 (adapted 2019). [Online]. Available: <https://ellenmacarthurfoundation.org/material-circularity-indicator>
- [32] R. Yang, X. Sun and K. Narasimhan, "A generalized algorithm for multi-objective reinforcement learning and policy adaptation," in *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019. [Online]. Available: <https://proceedings.neurips.cc/paper/2019/hash/4a46fbfca3f1465a27b210f4bdf6ab3-Abstract.html>
- [33] J. Blank and K. Deb, "Pymoo: Multi-objective optimization in Python," *IEEE Access*, vol. 8, pp. 89497–89509, 2020. DOI: <https://doi.org/10.1109/ACCESS.2020.2990567>
- [34] B. H. Lowe, A. Genovese, D. Font Vivanco and T. Zink, "Revisiting circular economy rebound: Market dynamics, policy implications, and future research directions," *Journal of Industrial Ecology*, vol. 29, no. 6, pp. 1936–1945, 2025. DOI: <https://doi.org/10.1111/jiec.70121>
- [35] A. Estes, D. Peidro, J. Mula and M. Díaz-Madroño, "Reinforcement learning applied to production planning and control," *International Journal of Production Research*, vol. 61, no. 16, pp. 5772–5789, 2023. DOI: <https://doi.org/10.1080/00207543.2022.2104180>
- [36] L. N. Alegre, A. L. C. Bazzan, D. M. Roijers, A. Nowé and B. C. da Silva, "Sample-efficient multi-objective learning via generalized policy improvement prioritization," in *Proc. 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, London, UK, 2023, pp. 2003–2012. [Online]. Available: <https://arxiv.org/abs/2301.07784>
- [37] M. Wurster, M. Michel, M. C. May, A. Kuhnle, N. Stricker and G. Lanza, "Modelling and condition-based control of a flexible and hybrid disassembly system with manual and autonomous workstations using reinforcement learning," *Journal of Intelligent*

- Manufacturing, vol. 33, no. 2, pp. 575–591, 2022. DOI: <https://doi.org/10.1007/s10845-021-01863-3>
- [38] C. Hu, Z. Wang, J. Liu, J. Wen, B. Mao and X. Yao, "Constrained reinforcement learning for dynamic material handling," in Proc. 2023 International Joint Conference on Neural Networks (IJCNN), 2023, pp. 1–9. DOI: <https://doi.org/10.1109/IJCNN54540.2023.10191999>
- [39] R. Stolz, H. Krasowski, J. Thumm, M. Eichelbeck, P. Gassert and M. Althoff, "Excluding the irrelevant: Focusing reinforcement learning through continuous action masking," in Advances in Neural Information Processing Systems (NeurIPS), vol. 37, pp. 95067–95094, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/hash/acf4a08f67724e9d2de34099f57a9c25-Abstract.html
- [40] Z. Wang, X. Li, L. Sun, H. Zhang, H. Liu and J. Wang, "Learning state-specific action masks for reinforcement learning," Algorithms, vol. 17, no. 2, Art. no. 60, 2024. DOI: <https://doi.org/10.3390/a17020060>
- [41] European Commission, "Commission Recommendation (EU) 2021/2279 of 15 December 2021 on the use of the Environmental Footprint methods to measure and communicate the life cycle environmental performance of products and organisations," Official Journal of the European Union, L 471, p. 1, 30 Dec. 2021. [Online]. Available: <https://eur-lex.europa.eu/eli/reco/2021/2279/oj>
- [42] Si Z, Ali D A, Rosli R B, Bhaumik A, Ghosh A. Omni-channel retail marketing effect evaluation framework integrating big data and artificial intelligence. Edelweiss Applied Science and Technology, 2025, 9(3): 568-583. DOI: <https://doi.org/10.55214/25768484.v9i3.5255>
- [43] Si Z, Ali D A, Rosli R B. AI-driven risk management decision framework and value creation in chemical enterprises. BEM Quarterly, 2025, 1(1): 1-32. Available: <https://system.wisacad-pub.com/index.php/bem/article/view/195>



This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).