Reward function design in reinforcement learning for HVAC Control: A review of thermal comfort and energy efficiency Trade-offs

Article · September 2025

DOI: 10.1016/j.enbuild.2025.116439

1 author:

Eisuke Togashi
Kogakuin University
36 PUBLICATIONS **85** CITATIONS

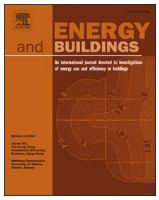
SEE PROFILE

Energy & Buildings 348 (2025) 116439

Contents lists available at ScienceDirect

Energy & Buildings

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



Reward function design in reinforcement learning for HVAC Control: A review of thermal comfort and energy efficiency Trade-offs

Eisuke Togashi *

Kogakuin University, Tokyo, Japan

ARTICLE INFO

Keywords:
Building Automation
Smart Buildings
Multi-objective Optimization
Occupant Comfort
PMV (Predicted Mean Vote)
Energy-Comfort Trade-off
Reward Shaping

ABSTRACT

Reinforcement learning is increasingly applied to Heating, Ventilation, and Air Conditioning control to balance energy efficiency with occupant comfort. However, the design of the reward function, crucial for managing this trade-off, remains a relatively underexplored topic in existing review literature.

This paper addresses this gap through a systematic review of 79 studies published since 2020. We introduce a novel standardization methodology to enable a systematic comparison of the diverse reward formulations reported in the literature.

The analysis reveals two primary findings: (1) a substantial heterogeneity in reward function structures, with 68 unique designs identified, a factor that severely impedes research comparability; and (2) a prevalent reliance on empirically derived weighting coefficients that often lack a clear theoretical basis. Furthermore, this study identifies and quantifies the usage patterns of four common design techniques: occupancy consideration, comfort deadbands, error exponentiation, and acceptable limits.

Based on this comprehensive analysis, we propose a typical piecewise reward function structure that synthesizes common best practices and is grounded in established Heating, Ventilation, and Air Conditioning domain knowledge. This proposed structure is intended to serve as a foundational baseline, addressing the identified limitations and aiming to improve the comparability of future research in Reinforcement learning driven Heating, Ventilation, and Air Conditioning control.

1. Introduction

The building sector is a major contributor to global energy consumption. An IEA report[1]indicates that in 2022, this sector accounted for approximately 33 % of global CO2 emissions. The same report notes that 79 % of those emissions (approximately 26 % of total global emissions) occur during the operational phase of buildings. Consequently, enhancing operational strategies, in addition to construction practices, can significantly impact global energy efficiency.

For this reason, numerous studies have long focused on technologies for optimizing building operations, treating it as a multi-objective optimization problem with multiple performance metrics such as occupant comfort, energy efficiency, and indoor air quality[2].

A key challenge in optimizing HVAC systems lies in the inherent uniqueness of each building. Because each building's location, structure, and occupant characteristics differ, and these factors lead to building-specific HVAC system configurations. An optimal control strategy derived for one building cannot be directly applied to another; tuning must be repeated for each building, which in turn incurs significant labor costs.

In recent years, the application of machine learning has been explored as a promising approach to address these challenges. Unlike traditional physics-based models, machine learning techniques are characterized by their ability to build data-driven models using real-world operational data. Several review studies have surveyed applications of machine learning in the HVAC domain, each covering hundreds of publications (e.g., [3,4,5]. Automating parameter tuning through machine learning using field-measured data offers the potential for cost-effective optimization tailored to individual building characteristics (e.g., [6,7,8].

Among machine learning methods, reinforcement learning (RL) is particularly well-suited for control optimization problems. In RL, an agent (the decision-making entity) interacts with an environment (the system to be controlled) in a simulated setting; through trial-and-error, the agent learns to take optimal actions (in this context, HVAC control setpoints or adjustments). The agent receives feedback in the form of a reward from the environment for each action, and it adjusts its strategy

* Corresponding author.
E-mail address: e.togashi@gmail.com.

https://doi.org/10.1016/j.enbuild.2025.116439
Received 8 June 2025; Received in revised form 21 August 2025; Accepted 10 September 2025
Available online 16 September 2025

Nomenclature

A	Action. [-]
$APAQ$	Acceptability of perceived air quality. [-]
c_{co2}	CO_2 concentration. [ppm]
$C_{penalty}$	Penalty constant. [-]
D	Depreciation. [USD]
dp	Dumper position. [-]
E	Energy consumption. [GJ]
E_{Bat}	Battery charging/discharging energy. [GJ]
E_{lux}	Illuminance. [lx]
f_{PD} (•)	performance-decrement function. [-]
GTS	Group thermal sensation. [-]
J_{ALV} (•)	Cost function for comfort when outside acceptable limits (Acceptable Limit Violation). [-]
J_C (•)	Cost function for comfort within acceptable limits. [-]
N (•)	Normalization function. [-]
N_{oc}	Number of occupants. [person]
P	Power for exponentiation. [-]
PMV	Predicted mean vote. [-]
PPD	Predicted percentage of dissatisfied. [-]
Q	Heat load. [GJ]
r	Reward. [-]
S	Status. [-]
S_{alm}	Monthly salary of occupants. [USD/month]
SOC	State of charge (battery). [GJ]
T	Temperature. [°C]
T_{op}	Operating temperature. [°C]
u_{epi}	Epistemic uncertainty. [-]
V	Voltage at common bus. [kW]
vp	Valve position. [-]
w	Weight coefficient. [-]
WH_m	Monthly working hours. [hours/month]
x	Comfort indicator. [-]
α	Coefficient. [-]
β	Coefficient. [-]
$\in$ (•)	Function determining energy consumption. [-]
λ_E	Energy purchase price. [USD/GJ]
$\lambda_{sell,E}$	Energy selling price. [USD/GJ]
ϕ	Relative humidity. [%]
χ (•)	Function determining the comfort indicator. [-]

Subscript:

ALL	acceptable lower limit
AUL	acceptable upper limit
ahu	air handling unit
ave	average
chw	chilled water
C	comfort
CLL	comfortable lower limit
CUL	comfortable upper limit
E	energy
ext	exterior
fan	fan
$hvac$	HVAC
N	nominal
$pm10$	PM10
$pm25$	PM2.5
sp	setpoint
$sply$	supply air
t	time
tp	thermal preference

to maximize the cumulative reward. Numerous studies have applied RL to HVAC control optimization, and several review papers have summarized these efforts from various perspectives, [9,10,11,12,13,14,15,16] (see Table 1). While these reviews organize the applicability and challenges of reinforcement learning from different viewpoints, none of them have made the design methodology of the reward function their primary focus. The contribution of this paper, therefore, is to provide a review specializing in reward functions. We aim to interpret the design intent behind the various proposed reward functions and organize their respective strengths and weaknesses for comparative analysis (Fig. 1).

In the field of RL itself, reward function design is widely recognized as a fundamental issue that influences agent learning outcomes[18]. For example, it has been pointed out that learning becomes remarkably difficult in sparse reward environments where rewards are given only upon task completion, and it is challenging to prevent *reward hacking*, where agents find unintended shortcuts (loopholes)[19]. To address such issues, techniques like *reward shaping*, which provide auxiliary rewards to promote learning, have also been proposed[20].

However, merely applying general reward design theories from information science research is insufficient; reward functions require designs informed by domain-specific knowledge[21]. This is particularly true in situations where concepts with different units, such as energy efficiency and comfort, must be balanced for optimization. An effective reward function cannot be designed without a thorough understanding of both concepts. Therefore, design principles for reward functions tailored to the HVAC control domain warrant thorough consideration by experts within the field.

Based on the preceding discussion, this study reviews research that applies reinforcement learning to the optimization of HVAC systems. By standardizing and comparing the reward functions from 79 papers, this study quantitatively demonstrates for the first time that a remarkable diversity of designs exists, which significantly hinders inter-study comparability. It then extracts common techniques found within this diversity and examines their significance from the perspective of our domain expertise. Finally, based on the insights gained from this analysis, this paper proposes a new "typical piecewise reward function structure" to address the limitations of existing research. The unique contribution of this study lies in its potential to accelerate progress in the field by enhancing the comparability of research applying reinforcement learning to HVAC optimization.

2. Methods for literature search and Selection

Papers for this review were selected through a systematic procedure, following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram shown in Fig. 2.

A Scopus search conducted on February 1, 2025, using the keywords *reinforcement learning* AND *hvac* yielded 391 documents. From this set, 191 academic articles written in English were selected. To focus on recent trends that reflect the rapid increase in RL-related publications, we excluded papers published before 2020 (10 instances).

The 181 extracted articles were subsequently screened, and those meeting the following exclusion criteria, which were based on the study's objectives, were removed:

1) Studies with a broad research scope (e.g., autonomous driving) where HVAC control was merely illustrative (e.g., [22,23].
2) Studies where the objective is a comparison of multiple optimization methods, and RL is only one example (e.g., [24,25].
3) Studies focused on the development of RL testbeds rather than the application of RL itself (e.g., [26,27].

Table 1
Comparison with Prior Papers.

Paper	Main Focus	Coverage of Reward Function Design
Ajifowowe et al. [9] (Review)	Comparison of traditional control with RL; demonstrating RL's effectiveness.	Conceptually introduces the reward function as one component of RL, without detailed analysis.
Al Sayed et al. [10] (Review)	Challenges in sim-to-real transfer for RL agents.	Lists the objectives included in the reward (e.g., energy saving, comfort), but does not delve into the specific design of the mathematical formulations that integrate them.
Chatterjee et al. [11] (Review)	Possibilities for creating a dynamic indoor thermal environment.	Focuses on the system-level discussion of how RL control creates a dynamic environment, rather than on the reward design itself.
[12,13] (Review)	Improving occupant comfort and modeling occupant behavior.	Points out that reward design is a difficult issue but does not provide a systematic comparative analysis of how specific formulations are constructed.
Sierla et al. [14] (Review)	Analysis of the RL action space and its impact on control abstraction.	Does not focus on reward analysis; only uses comfort objectives (temperature, humidity, etc.) for a high-level categorization of studies.
Wang & Hong [15] (Review)	Comprehensive review of the five key components of RL (algorithms, states, actions, rewards, environment).	Discusses the main strategies for integrating multiple objectives (weighted sum, constrained optimization) but does not refer to specific formula shapes or design patterns.
Yu et al. [16] (Review)	Review of RL applications from the perspective of occupant-building interaction.	Noted challenges in reward design due to delayed or sparse feedback, but didn't compare reward design methods in depth.
Liu et al. [17] (Article)	Proposing an occupant-centric HVAC & window controller.	Compiles a list of reward functions from several prior studies but provides no comparative analysis of their features, advantages, or disadvantages.

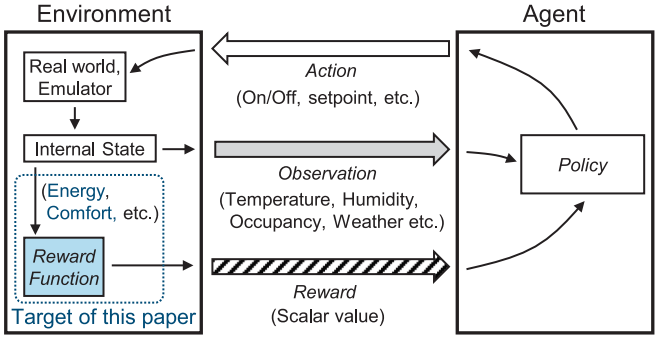


Fig. 1. Basic Structure of Reinforcement Learning for HVAC Control and the Focus of This Review.

4) Studies not incorporating occupant comfort in their evaluation metrics, such as those focused on plant equipment optimization or low-level VAV controller tuning (e.g., [28,29].
5) Studies where energy is not included in the reward function (e.g., [30,31,32].
6) Studies where transfer learning is the main theme (e.g., [33,34].
7) Studies in which the same authors reused a reward function from a previous publication. To avoid duplication, only the primary study was included in the analysis (e.g., [35,36].
8) Studies where the reward function is not explicitly stated or is ambiguous (e.g., [37,38].
9) Studies aimed at fault diagnosis (e.g., [39,40].

Following this screening process, 79 papers were selected for in-depth review. For reference, a comparison of this review's scope with other recent review papers on RL for HVAC is shown in Table 2.

3. Results: summary table of reward functions

Table 3 summarizes the reward functions from the selected literature to facilitate their comparison. Time-dependent state variables are consistently denoted with the subscript *t*. To enable systematic comparison of these reward functions, the original formulations were abstracted and simplified, as detailed below, rather than being directly transcribed.

It should be noted that these abstractions and simplifications involve subjective judgments and limitations. A simplification that is permissible for the purpose of comparing reward functions may be inappropriate in the specific context where reinforcement learning was applied in each paper. For example, the omission of some multiplicative coefficients and constant terms, while making the core mathematical structures easier to compare, does not fully reproduce the readability or specific behaviors intended by the original authors. Therefore, there is no guarantee that applying these simplified reward functions directly to the RL agents from the original papers would result in identical optimization outcomes. In addition, while we have endeavored to use simplification methods that could be judged as objectively as possible, cases where the original designer's intent was not fully explained required some conjecture on our part. In this sense, subjective judgment could not be completely eliminated.

3.1. Standardization of variables and unit Notations

3.1.1. Unification of variable symbols

Variable symbols that differed between papers were unified. In this process, all related physical quantities, regardless of their original units, were represented by common symbols. For example, Wh, kWh, kJ, and MJ were all expressed as E_t [GJ]. When the time step is fixed, W and kW also effectively represent the same concept, so these were also expressed as E_t [GJ].

3.1.2. Integration of Energy-Related terms

In some cases, power consumption for heat source equipment, pumps, fans, etc., was calculated individually, and different weighting coefficients were applied to each (e.g., [45,48]. However, assuming no differential pricing for electricity purchase and sale, the economic value per unit of electricity is constant; therefore, these power consumption terms were aggregated. Cases using heat load in the reward were also considered as energy.

3.1.3. Unified representation of electricity purchase and sale

Some studies evaluated electricity purchase and sale by introducing power generation facilities such as solar panels (e.g., [106,110]. Instead of assigning separate symbols for power generation and energy consumption, the total energy consumption was denoted as E_t, with positive values representing electricity purchase and negative values representing electricity sale.

3.1.4. Taxonomy and standardization of comfort indicators

Given the wide variety of comfort indicators used in reward functions, we first introduce a taxonomy to structure our analysis, as shown in Fig. 3. The overarching goal of Occupant Comfort is divided into two

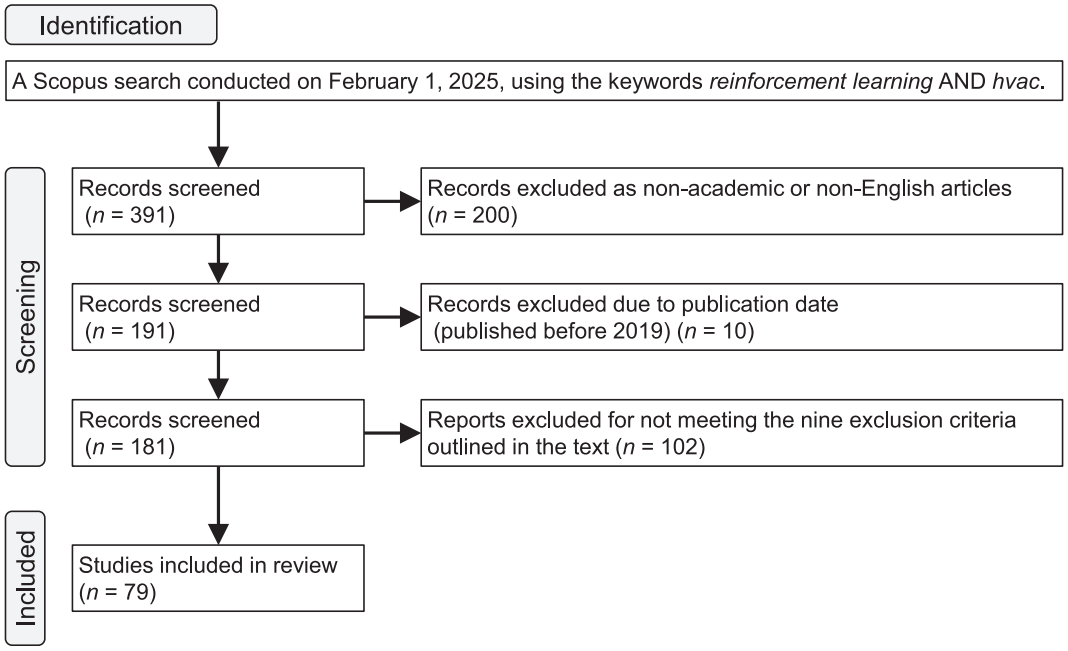


Fig. 2. PRISMA 2020 flow diagram for the systematic review.

Table 2
Comparison of Review Papers on the Application of RL to HVAC Optimization.

Paper	Initial hits	Selected articles	Database†1	Timeframe of Reviewed Papers†2	Search Date†2	Topic structure
This paper	391	79	SC	from 2020	February 2025	"reinforcement learning" AND hvac
Ajifowowe et al. [9]	821	120	SC, GS, WS, IX	Not restricted	Not specified	building AND HVAC AND "reinforcement learning" AND "indoor air quality" AND energy AND comfort
Al Sayed et al. [10]	135	48	SC	Not specified	August 2023	TITLE-AVS-KEY("reinforcement learning") AND TITLE-AVS-KEY("building") AND TITLE-AVS-KEY("HVAC systems")
Chatterjee et al. [11]	108	63	SD	Not specified	in 2022	"reinforcement learning" AND {(building OR house OR home) AND control} OR "smart thermostat"
Han et al. [12]	–	33	WS, SD, GS	Not restricted	Not specified	{building(s) AND ("reinforcement learning" OR "Markov decision processes" OR "Q-learning") AND (comfort OR "thermal comfort" OR "visual comfort" OR "indoor air quality" OR occupant OR "indoor environment")} OR "model free control" OR "intelligent control"
Han et al. [13]	40	32	SC	Not restricted	Not specified	("reinforcement learning" OR "Q-learning" OR "policy gradient" OR "A3C" OR "actor-critic" OR "SARSA*") AND "occupant*"
Sierla et al. [14]	278	83	WS	from 2013	Not specified	"reinforcement learning" AND (heating OR ventilation OR "air conditioning" OR cooling OR HVAC)
Wang & Hong [15]	77	77	WS	Not specified	December 2019	"reinforcement learning" AND (building OR house OR home OR residential) AND control
Yu et al. [16]	795	68	WS, GS	Not specified	September 2023	"reinforcement learning" AND "occupant behavior" AND "building" AND "energy"

†1 SC: Scopus, GS: Google Scholar, WS: Web of Science, SD: Science Direct, IX: IEEE Xplore.
†2 Not specified: not mentioned in the paper; Not restricted: explicitly stated as unrestricted.

main categories: Thermal Comfort and Indoor Air Quality (IAQ). Thermal Comfort is further subdivided into Physical Metrics, such as temperature, and Integrated Indices, which combine multiple factors, such as PMV, PPD, and TSV.

Several studies estimated Thermal Sensation Vote (TSV) and used this as an element of the reward function. For example, studies by Lan et al. [7], Li et al. [77], and Lim et al. [79] estimated these values using Kernel Density Estimation model, Takagi-Sugeno fuzzy model, and Machine learning, respectively. Although these approaches are strictly different from the definition of PMV [116], their common purpose of representing human thermal preference on a 7-point scale led us to represent them with the same variable, PMV.

3.2. Simplification of reward calculation Formulae

3.2.1. Omission of redundant multipliers and Non-essential terms

Some studies involved multiple multiplicative steps in calculating the reward value (e.g., [59,69]. The initial multiplication might serve to convert comfort and energy performance terms to roughly the same order of magnitude, but mathematically, the same output can be obtained by appropriately adjusting the final weighting coefficient. Consequently, to enhance comparability, multipliers other than the

Table 3
Standardized Reward Functions from Selected Literature on RL for HVAC Control.

Paper	The reward at timestep t (r_t)
[41]	$\begin{cases} -E_t & (c_{co2,ave,t} \le 1000) \\ -1.5E_t & (1000 < c_{co2,ave,t}) \end{cases}$
[42]	$-w_1 E_t - w_2 D_t - w_3 f_{err(CL)}(T_t)$
[43]	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CL)}(T_t)$
[44]	$-w_1 N_{max}(\lvert E_t \rvert) - w_2 N_{minmax}\left(f_{err(CL)}(PMV_t)\right)$
[45]	$-w_1 N_{max}(E_{ahu,t}) - w_2 \begin{cases} 1 & (0 < E_{fan,t}) \\ 0 & (0 = E_{fan,t}) \end{cases} - w_3 \begin{cases} 95 - 95\exp\left(-0.029PMV^4 - 0.205PMV^2\right) + \frac{PMV_t}{4} & (\lvert PMV_t \rvert \le 0.5) \\ 20 & (0.5 < \lvert PMV_t \rvert) \end{cases} - w_4 f_{err(CL,AL)}\left(N_{minmax}(c_{co2,t})\right)$
[46]	$-w_1 E_t \lambda_{E,t} + w_2 \left\{ \exp\left(-\alpha f_{err(P)}(T_t)\right) - \beta f_{err(CL)}(T_t) \right\}$
[47]	$-w_1 \frac{Q_t}{T_{sp,t} - T_{ext}} - w_2 f_{(OC,P)}(T_t)$
[30]	$-w_1 E_t - w_2 f_{err(CL)}(T_t) - w_3 f_{err,tp}(T_t) - w_4 u_{epi}$
[48]	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CL)}(E_t) - w_3 f_{err(OC,CL)}(T_t) - w_4 f_{err(CL,OC)}\left(c_{co2,t}\right)$
[49]	$-w_1 E_t - w_2 f_{err(OC,CL,P)}(T_t)$
[50]	$-w_1 E_t - w_2 f_{err(CL)}(T_t) - w_3 f_{err(CL)}\left(c_{co2,t}\right)$
[51]	$-w_1 \exp\left(vp_{chw,t}\right) - w_2 \left(f_{err}(T_t) + dp_t f_{err}\left(x_{co2,t}\right)\right)$
[35]	$-w_1 N_{minmax}(E_t) - w_2 f_{err(OC,AL)}(PPD_t)$
[52]	$-w_1 N_{minmax}(E_t) - w_2 N_{minmax}\left\{f_{err(CL)}(PMV_t)\right\} - w_3 N_{minmax}\left\{f_{err(CL)}\left(E_{lux,t}\right)\right\} - w_4 N_{minmax}\left\{f_{err(CL)}\left(c_{co2,t}\right)\right\}$
[53]	$-w_1 N_{minmax}(E_t) - \begin{cases} w_2 & (N_{oc} > 0) \\ w_3 & (N_{oc} = 0) \end{cases} N_{minmax}\left\{f_{err}(PMV_t)\right\}$
[54]	$\begin{cases} -w_1(\lvert PMV \rvert + PPD) - w_2 E_t \lambda_{E,t} & (PMV, PPD meet requirements) \\ -C_{penalty} & (else) \end{cases}$
[55]	$-w_1 N_{linear}(E_t) - w_2 N_{linear}\left(f_{err(CL)}(T_t)\right)$
[56]	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CAL)}(T_t)$
[57]	$-w_1 E_t - w_2 f_{err(OC,P)}\left(T_{op,t}\right)$
[58]	$-w_1 N_{minmax}(E_t) - w_2 f_{err(OC)}(N_{minmax}(T_t))$
[59]	$E_t \begin{cases} w_1 & (0 < E_t) \\ w_2 & (E_t \le 0) \end{cases} - w_3 f_{err(P)}(T_t)$
[60]	$-w_1 E_t - w_2 (T_t - T_{ext})^2$
[61]	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CL)}(T_t)$
[62]	$-w_1 E_t - w_2 f_{err(CL)}(PMV)$
[63]	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CL)}(T_t)$
[64]	$-w_1 E_t - w_2 f_{err(CL)}(T_t) - w_3 f_{err(CL)}(\phi_t) - w_4 f_{err(CAL)}(c_{co2}) - w_5 f_{err(CAL)}\left(\rho_{pm25}\right)$
[65]	$-w_1 N_{minmax}(E_t \lambda_E) - w_2 N_{minmax}\left(f_{err(P)}(T_t)\right)$
[66]	$-w_1 (T_t - T_{t-1}) - w_2 f_{err(OC,CAL)}(T_t)$
[67]	$-w_1 N_{max}(E_t)\lambda_E - w_2 1_{NOC}{}^* \begin{cases} 0 & (T_{CLL} < T_t < T_{CUL}) \\ \exp(\max(0, T_{ALL} - T_t, T_t - T_{AUL})) & (T_{ALL} < T_t < T_{CLL} or T_{CUL} < T_t < T_{AUL}) \\ 0.5 f_{err(CL)}(T_t) & (otherwise) \end{cases}$
[68]	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CL)} N_{minmax}(E_t) - w_3 f_{err(CL)}(T_t)$
[69]	$-w_1 E_t - w_2 \exp\left(f_{err(CL)}(T_t)\right)$
[70]	$-E_t \left(w_1 + w_2 f_{err(CL,P)}(T_t)\right)$
[71]	$-w_1 E_t - w_2 f_{err(CL)}(T_t)$
[72]	$-w_1 E_t - w_2 f_{err(CL)}(T_t)$
[73]	$-w_1 E_t \lambda_{E,t} - w_2 \left(\rho_{pm25} + \rho_{pm10}\right)$
[7]	$-w_1 E_t^{\alpha} - w_2 \lvert PMV_t \rvert^{\beta}$
[74]	$-E_t \begin{cases} w_1 & (N_{oc} > 0) \\ w_2 & (N_{oc} = 0) \end{cases} - w_3 (\overline{APAQ} + \alpha)^P$

(continued on next page)

Table 3 (*continued*)

Paper	The reward at timestep t (r_t)
[75]	$-w_1 E_t - w_2 f_{err(OC)}(PPD_t)$
[76]	$-w_1 \begin{cases} f_{err(CL)}(\lvert T_{sp,t+1} - T_t \rvert) & (GTS > 0) \\ f_{err(CL)}(\lvert T_{sp,t+1} - T_{AUL} \rvert) & (GTS < 0) \end{cases} - w_2 f_{err(AL)}(GTS)$
[77]	$-w_1 E_t - \begin{cases} w_2 \lvert PMV_t \rvert^{2.5} & (\lvert PMV_t \rvert \le 0.5) \\ w_3 \lvert PMV_t \rvert^{1.5} & (0.5 < \lvert PMV_t \rvert) \end{cases}$
[78]	$-w_1 N_{max}(E_t) - w_2 f_{err(CL)}(N_{max}(\lvert PMV_t \rvert))$
[79]	$-w_1 E_t - w_2 1_{NOC} \begin{cases} \lvert PMV_t \rvert & (\lvert PMV_t \rvert \le 1.0) \\ \lvert PMV_t \rvert^2 & (1.0 < \lvert PMV_t \rvert) \end{cases}$
[80]	$-w_1 E_t - w_2 f_{err(CL)}(T_t)$
[81]	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(P)}(\lvert T_t - T_{sp,t} \rvert)$
[82]	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CL)}(T_t)$
[17]	$-\begin{cases} 1 & (workinghours) \\ 0 & (otherwise) \end{cases} \bullet \left[E_t \begin{cases} w_1 & (windowopened) \\ w_2 & (windowclosed) \end{cases} + w_3 f_{err(CL,P)}(T_t) \right]$
[83]	$-w_1 E_t - w_2 PPD$
[84]	$-w_1 E_t - w_2 f_{err(CL)}(T_t)$
[85]	$w_1 \exp\left(-(\alpha E_t)^2\right) - w_2 \left\{ f_{err(CL)}(T_t) + \exp\left(-\beta (T_t - T_{sp})^2\right) \right\}$
[86]	$-w_1 E_t - w_2 f_{err(CL)}(T_t) - w_3 1(A_t \ne A_{t-1})$
[87]	$-w_{1,t} E_t - w_{2,t} f_{err(CL)}(T_t)$
[88]	$-w_1 E_t - w_2 \begin{cases} 50(T_t - T_{sp})^2 - 100 & (\lvert T_t - T_{sp} \rvert \le 1) \\ 6.25(T_t - T_{sp})^2 - 56.25 & (1 < \lvert T_t - T_{sp} \rvert \le 3) \\ 3.125(T_t - T_{sp})^2 - 28.125 & (3 < \lvert T_t - T_{sp} \rvert \le 5) \\ 500 & (5 < \lvert T_t - T_{sp} \rvert) \end{cases}$
[89]	$-w_1 E_t - w_2 f_{err(CL)}(PMV_t)$
[90]	$-w_1 E_t - w_2 \begin{cases} f_{err(CL)}(T_t) & (T_t < T_{ALL}) \\ f_{err(CL,P)}(T_t) & (T_{AUL} < T_t) \end{cases} - w_3 f_{err(CL,P)}(T_{sply,t})$
[91]	$-w_1 N_{max}(E_t) - w_2 N_{max}(f_{err}(T_t)) - w_3 1(T_t < T_{min} \vee T_{max} < T_t) - w_4 \begin{cases} 1 & (occupantinteractswithFCU) \\ 0 & (otherwise) \end{cases}$
[92]	$-w_1 E_t - w_2 \begin{cases} \lvert T_t - T_{sp} \rvert^2 & (T_t < T_{ALL}) \\ 0 & (T_{ALL} \le T_t < T_{sp}) \\ \lvert T_t - T_{sp} \rvert & (T_{sp} \le T_t < T_{AUL}) \\ \lvert T_t - T_{sp} \rvert^3 & (T_{AUL} \le T_t) \end{cases}$
[93]	$-w_1 E_t - w_2 f_{err(CL)}(T_t)$
[94]	$-w_1 E_t - w_2 f_{err(OC,CL,P)}(T_t)$
[95]	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CAL)}(T_t)$
[96]	$-w_1 \begin{cases} 2E_t & ((S_{hvac,t-1} = off) \wedge (S_{hvac,t} = on)) \\ E_t & (otherwise) \end{cases} - w_2 \begin{cases} -500 & (22 < T_t < 26) \\ -300 & ((20 < T_t \le 22) \wedge (5 \le h < 9) \wedge (S_{hvac,t-1} = off)) \\ 500 & ((20 < T_t \le 22) \wedge (5 \le h < 9) \wedge (S_{hvac,t-1} = on)) \\ 200 & (otherwise) \end{cases} - w_3 \begin{cases} 1000 & (28 < T_t) \\ 0 & (otherwise) \end{cases}$
[97]	$-w_1 E_t \lambda_{E,t} - w_2 \left\{ -\exp\left(-0.5(T_t - T_{sp})^2\right) + f_{err(CL)}(T_t) \right\} - w_3 \begin{cases} 20 & ((E_{Bat,t} < 0) \wedge (SOC_t < SOC_{min})) \\ 20 & ((E_{Bat,t} > 0) \wedge (SOC_t > SOC_{max})) \\ 0 & (otherwise) \end{cases}$
[98]	$-w_1 E_t - w_2 f_{err(CL,P)}(T_t) - w_3 f_{err(CL)}(\phi_t) - w_4 1(A_t \ne A_{t-1})$
[99]	$-w_1 \max(0, E_t - E_{min,t}) - w_2 f_{err(P)}(T_t)$
[100]	$-w_{1,t} \begin{cases} 1 - N_{minmax}(E_t) & (T_{ALL} < T_t < T_{AUL}) \\ -N_{minmax}(E_t) & (T_{AUL} < T_t (winter) \vee T_t < T_{ALL} (summer)) \\ N_{minmax}(E_t) - 1 & (T_t < T_{ALL} (winter) \vee T_{AUL} < T_t (summer)) \end{cases} - (1 - w_{1,t}) \begin{cases} 2 - \dfrac{2}{1 + \exp(-\lvert T_t - T_{SP,t} \rvert)} & (T_{ALL} < T_t < T_{AUL}) \\ 1 - \dfrac{2}{1 + \exp(-\lvert T_t - T_{SP,t} \rvert)} & (otherwise) \end{cases}$
[101]	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CL)}(T_t)$
[102]	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CL)}(T_t)$
[103]	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(OC,CL)}(T_t) - w_3 f_{err(OC,CL)}(c_{co2,t}) - w_4 f_{err(OC,CL)}(\phi_t)$
[104]	$-w_1 E_t - w_2 f_{err(CL)}(T_t)$
[105]	$-w_1 E_t - w_2 \begin{cases} \lvert PMV_t \rvert & (PMV_{CLL} < PMV_t < PMV_{CUL}) \\ -\lvert PMV_t \rvert & (otherwise) \end{cases}$

(*continued on next page*)

Table 3 (*continued*)

Paper	The reward at timestep t (r_t)
[106]	$-w_1 E_t \begin{cases} \lambda_{E,t} & (0 < E_t) \\ \lambda_{sell,E,t} & (E_t \le 0) \end{cases} - w_2 f_{err(CL)}(T_t) - w_3 \Delta E_{Bat,t}$
[107]	$-w_1 E_t - 1_{NOC}\left\{ w_2 f_{err(CL)}(T_t) + w_3 f_{err(CL)}(x_{co2,t}) \right\}$
[108]	$-w_1 E_t - w_2 f_{err(OC,CL)}(PMV_t) - w_3 f_{err(OC,CL)}(T_t)$
[109]	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CL,P)}(T_t)$
[110]	$-w_1 E_t \begin{cases} \lambda_{E,t} & (0 < E_t) \\ \lambda_{sell,E,t} & (E_t \le 0) \end{cases} - w_2 f_{err(CL)}(T_t) - w_3 P_{SoC,t}$ $P_{SoC,t} = \begin{cases} 0 & (E_{Bat,min} < E_{Bat,t} < E_{Bat,max}) \\ \beta + (1-\beta) P_{SoC,t-1} & (otherwise) \end{cases}$
[111]	$-w_1 E_t - w_2 f_{err(CL,P)}(T_t) - w_3 \begin{cases} \exp\left(-\frac{E_{bat,t}}{E_{Bat,max}}\right) & (E_{Bat,t} < E_{Bat,min}) \\ \exp\left(\frac{E_{bat,t}}{E_{Bat,max}} - 1\right) & (E_{Bat,max} < E_{Bat,t}) \\ 0 & (otherwise) \end{cases} - w_4 f_{err(P)}(V_t)$
[112]	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CL)}(T_t) - w_3 f_{err(CL)}(A_t)$
[113]	$-w_1 N_{minmax}(E_t) - w_2 f_{err(OC,AL)}(PPD_t)$
[114]	$-w_1 N_{minmax}(E_t) - w_2 f_{err(OC,CL,AL)}(PMV_t)$
[115]	$-w_1 E_t - w_2 f_{err(OC,P)}(PPD_t)$

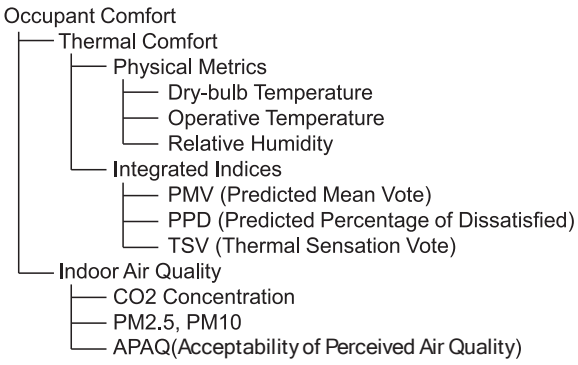


Fig. 3. Taxonomy of the comfort indicators discussed in this review.

primary weighting coefficients were omitted.

Some reward functions included simple additive or subtractive terms (e.g., [59,77]. This was likely intended to make the reward function more understandable to humans by placing desirable states in a positive range and undesirable states in a negative range. However, for RL learning algorithms, the sign of the value is irrelevant; the relative magnitude is what matters. Thus, these additive/subtractive terms were omitted.

3.2.2. Simplification of summation for Multi-Object evaluation

In cases with multiple similar evaluation targets, such as separate rooms, where the same evaluation formula was applied to each to obtain a cumulative reward (e.g., [50,53], this summation does not aid in comparing the fundamental shapes of the reward functions. Therefore, the summation symbols were omitted.

3.2.3. Introduction of normalization functions for data scaling

As scaling of reward elements was common practice, we defined symbols for typical normalization formulas that use the maximum value (x_{max}) and minimum value (x_{min}) of a state variable:

$$N_{max}(x_t) = \frac{x_t}{x_{max}} \tag{1}$$

$$N_{minmax}(x_t) = \frac{x_t - x_{min}}{x_{max} - x_{min}} \tag{2}$$

Note that linear scaling via multiplication by a coefficient k (Eq. (3) can be integrated into the main weighting coefficient w as previously mentioned, and was therefore omitted from our standardized representation.

$$N_{linear}(x_t) = kx_t \tag{3}$$

3.2.4. Standardization of weighting coefficient representation

In some reward formulations, the total reward is expressed as a weighted sum of two component terms (e.g., a comfort term and an energy term):

$$w_1 r_1 + w_2 r_2 \tag{4}$$

As a simplification, one could define a new weight $w_3 = w_1/w_2$ and rewrite the two-term reward as:

$$w_3 r_1 + r_2 \tag{5}$$

However, to facilitate comparison across studies, we chose to retain the original weighting form (Eq. (4) in our standardized representation. In other words, even if a reward function could be algebraically simplified by combining weights, we represented it with separate weight coefficients for each component, to make the presence of each term's weight clear for comparison.

In cases like Kurte et al. [72], where the reward function was formulated as a plain summation of different components without showing any weighting coefficients, we treated it as equivalent to having all weights set to 1 and included explicit w values in our standardized representation.

3.2.5. Representation of core reward components in shaping

In some cases, the reward function was defined as the difference in an evaluation index between one time step, possibly intended for agent learning promotion (so-called shaping, etc.) (e.g., [61,73]. In such cases, to maintain comparability with immediate rewards, the reward was represented by the evaluation index at the current timestep, which is the basis for calculating this difference.

3.3. Introduction of a common error function f_{err} and typical adjustments

The most straightforward method to incorporate comfort into a reward function is to treat an error function (Eq. (6) as a cost term and use its negation as the reward:

$$f_{err}(x_t) = |x_t - x_{sp}| \tag{6}$$

Here, x is a comfort-related state variable, such as dry-bulb temperature, relative humidity, or PMV. Hereafter, x is referred to as the *comfort indicator*, and x_{sp} is its the ideal value from a comfort perspective. The error is thus the absolute deviation from this ideal value. Note that this formula can also be applied to indicators where 0 is optimal, such as carbon dioxide concentration, PM2.5, PMV, and PPD, by setting $x_{sp} = 0$.

However, many prior studies did not use such a simple formula; several typical improvements were added. Therefore, to facilitate the comparison of reward functions, a common error function symbol is introduced. The formula incorporating four typical adjustments is shown below:

$$f_{err(OC,CL,P,AL)}(x_t) = 1^{\varepsilon}_{NOC} \begin{cases} f_{penalty}(x_t) & (x_t < x_{ALL} or x_{AUL} < x_t) \\ \left([x_t - x_{CUL}]_+ + [x_{CLL} - x_t]_+\right)^P & (otherwise) \end{cases} \tag{7}$$

Each of the four modifications is explained below.

3.3.1. Incorporating occupancy information into rewards

1^{ε}_{NOC} is a generalized indicator function shown in Eq. (8), taking a value of 1 or an extremely small value. It is 1 when occupants are present and an extremely small value when absent.

$$1^{\varepsilon}_{NOC} = \begin{cases} 1 & (N_{oc} > 0) \\ \varepsilon & (N_{oc} = 0) \end{cases} \tag{8}$$

Here, N_{oc} [person] is the number of occupants. Note that ε may be 0, in which case it becomes a normal indicator function. This aims to nullify or substantially diminish the impact of indoor comfort on the reward function during occupant absence.

3.3.2. Defining comfort range

For temperature and humidity, slight deviations from the ideal value do not cause significant discomfort. Therefore, for example, ASHRAE 55–2017 (2017) recommends a certain range for PMV. In Eq. (7), this comfortable upper limit (CUL) and comfortable lower limit (CLL) are denoted as x_{CUL} and x_{CLL}, respectively. In condition (2) (otherwise case) of Eq. (7), a positive value is generated only when the state value deviates from this comfortable range, and it is 0 within the comfortable range. That is, no negative reward is given. Note that there are cases where a positive reward is actively given within this range [50], but these are also represented by this symbol.

3.3.3. Non-linear error Transformation

To accelerate learning, it is better to significantly increase the penalty as the state moves away from the optimal value, thereby encouraging a return to the appropriate range. For this purpose, Eq. (7) raises the error to the power of P. However, there were exceptions, such as Friansa et al. [59], where P was less than 1 to moderate the increase in penalty.

3.3.4. Defining acceptable range

If temperature or humidity deviates significantly from the comfort range, it can harm health or lead to occupant complaints, becoming unacceptable. The acceptable upper limit (AUL) and acceptable lower limit (ALL) are denoted as x_{AUL} and x_{ALL}, respectively. In condition (1) (first case) of Eq. (7), a large penalty $f_{penalty}$ is applied when these upper and lower limits are exceeded. $f_{penalty}$ can be a large constant or calculated using the comfort indicator x , but in any case, it is made larger than in condition (2) (otherwise case).

3.3.5. Notation of error function

As explained above, Eq. (7) is an error function that adds four modifications—Occupancy (OC), Comfortable Limits (CL), Raising the error to the power (P), and Acceptable Limits (AL)—to the simple error function shown in Eq. (6). Many prior studies adopted only some of these modifications. Therefore, in Table 3, the four symbols (*OC*, *CL*, *P*, *AL*) are added as subscripts to the f_{err} function to indicate which modifications were adopted. Instances where the cost remained zero within specific upper and lower bounds but substantial penalty imposed immediately upon exceeding these limits (e.g., [56,64,66], were interpreted as having coinciding comfort and acceptable ranges, denoted by the subscript *CAL*.

4. Discussion

4.1. Diversity of reward functions and challenges in research comparability

A primary observation is the remarkable diversity among the reward functions presented in Table 3. The equations with nearly identical structures, found in 12 papers, combined energy performance with comfort represented by temperature; six of these[71,72,80,84,93,104] represented energy performance as energy consumption, while the other six[43,61,63,82,101,102] used energy cost. This means that 68 distinct reward functions were designed across the 79 papers reviewed. This diversity exists despite the efforts described in Section 3 to standardize variable symbols and introduce abstract error functions to facilitate mutual comparison. Furthermore, the hyperparameters, i.e., weighting coefficients w_n, used in each formula usually take different values, so even formulas with seemingly identical shapes are not strictly the same.

Although the flexibility in reward function design is a key strength of reinforcement learning, this inherent diversity poses a significant challenge to the comparability of research outcomes. For example, studies by Heidari et al. [67,88,92], and [96] complicate their formulas by finely segmenting the reward function's domain and applying different non-linearities to each region. However, it should be verified whether the learning improvements obtained through this complexity outweigh the cost of reduced research comparability. Furthermore, the detailed conditional branches in the works of Liu and Gou [17], Qin et al. [88], and Sun et al. [96] can be interpreted as suggesting how the action should be performed. Such an approach risks inducing reward hacking by contradicting a well-established principle of reward function design:

> "*The reward signal is your way of communicating to the robot what you want it to achieve, not how you want it achieved.*[18]"

In the case of Liu and Gou [17], for example, instead of altering the reward based on the window's state, the window's open/closed status should be provided as part of the agent's observation. This would allow the agent to learn the relationship between the window's state and energy consumption on its own. With the current design, the agent might learn a short-sighted policy of always keeping the window closed to receive an immediate reward (by avoiding a penalty), and consequently fail to learn beneficial strategies such as using free cooling when the outdoor air temperature is low.

This suggests a potential misapplication of RL principles in some HVAC research, wherein the reward function embeds aspects of the solution rather than solely defining the problem. Such an approach might lead to agents performing well in simulation by exploiting these "hints" but failing to generalize or discover truly novel strategies.

Of course, it may not be possible to use completely identical, simple reward function shapes in various studies. However, to ensure comparability with other research, an effort should be made to adopt typical patterns as much as possible. Otherwise, the field risks generating a large number of case studies that are difficult to compare and are merely

standalone products. The discussion in the following sections primarily explores these *typical patterns*.

4.2. Integrating Energy Performance and Comfort: Current Approaches and Limitations

The integration of energy performance and comfort in nearly all reviewed studies is achieved via a weighted sum, using coefficients (w_n). However, a theoretical basis for the specific values of these weighting coefficients is often absent. Many studies either do not specify the values (e.g., [93,104], provide them without justification (e.g., [56,62,71,80,101], or state they were determined through trial and error (e.g., [57,61,82]. While some studies used sensitivity analysis or grid search (e.g., [43,49,58], potentially exploring a Pareto front, they ultimately selected a single operating point. Simply assigning equal weights (e.g., 0.5) to both terms (e.g., [55,84,87]does not guarantee a balanced evaluation, as the units and scales of energy and comfort metrics are different.

To illustrate how different weighting coefficients can produce significantly different optimization outcomes, we present a conceptual example. The scatter plot in Fig. 4 shows the results of virtually testing various building operations using an emulator [117]. The horizontal axis represents energy consumption, while the vertical axis represents the average dissatisfied rate for the thermal environment. Generally, these two objectives are in a trade-off, forming a relationship that is convex toward the origin, as depicted in the figure. The dotted line represents the Pareto front, and the area to its upper right is the feasible region. The optimal point lies on this Pareto front and is determined by the point of tangency with a line (shown in red) whose slope is defined by the ratio of the weighting factors for energy (w_E) and dissatisfaction (w_D). This demonstrates that different weighting coefficients can select entirely different optimal points, and without a theoretical basis, the choice is entirely arbitrary.

A more serious issue is that the shape of this Pareto front is unknown a priori. If the chosen weighting coefficients correspond to the steep "knee" of the front, minor changes in the weights might not significantly affect the resulting optimal point. However, if the weights correspond to a "flat" region with a gentle slope, even a slight difference in the weighting can dramatically shift the optimal point, carrying the risk of producing unintended and extreme control strategies.

A few exceptions to the linear sum exist. One is the work of Ahn and Park [41], which applied a discontinuous penalty to energy consumption based on whether the CO2 concentration exceeded a threshold. However, the rationale for the 1.5-fold penalty was not provided, leaving the theoretical integration of energy performance and comfort unresolved. Another example is from Kannari et al. [70], where the reward was expressed by multiplying energy and comfort performance. This type of formulation ensures that a change in one performance metric influences the evaluation of the other. Consequently, it can prevent the learning failures often seen in simple linear sums, where inappropriate weighting coefficients cause one performance metric to be overly prioritized[118]. Nevertheless, this method has a drawback: the loss of linearity makes it even more difficult than with a linear sum to theoretically justify which combinations of the two performance metrics are of equivalent value.

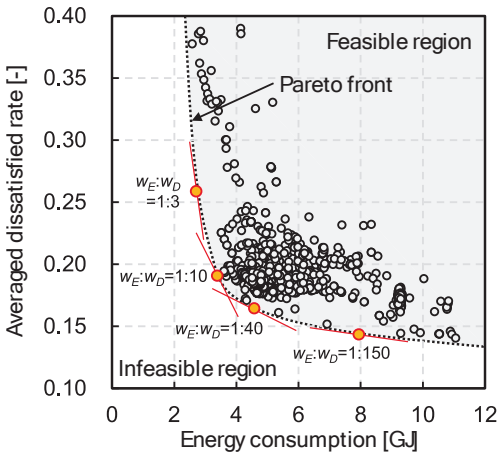


Fig. 4. Illustration of How Weighting Coefficients Determine the Optimal Point on the Energy-Comfort Pareto Front.

Some studies indirectly represented energy consumption using other state variables (e.g., room temperature, damper position) rather than direct energy consumption, E_t [47,51,66,76]. The reward function in Heidari and Khovalyg [66] represents energy as the change in room temperature over a single time step (T_t-T_{t-1}), allowing both the energy and comfort terms to be expressed in the same unit of temperature (°C). Thus, it is an interesting approach in that it avoids the need to sum state variables with different units, unlike other reward functions. However, when the formula is rearranged, the optimal room temperature ultimately becomes a weighted average of the previous time step's temperature (T_{t-1}) and the setpoint temperature (T_{sp}) using the weighting coefficients (w_1, w_2). This means the fundamental problem remains unsolved.

Monetizing the terms is one way to unify units. Eighteen studies multiplied energy consumption (E) by its cost (λ) to convert the term to a monetary value. While some did this to optimize electricity trading [97,106,110], others (e.g., [68,107,108]attempted to integrate it with comfort by using coefficients with units like $/°C, °C/$, °C/ppm, implying an effort to unify the final reward unit. However, the theoretical basis for the chosen conversion rates (e.g., 0.1$ /°C in [68]remains unexplained.

There is a history of research, such as that by Fisk [119] and Seppänen et al. [120], that aims to economically evaluate the indoor thermal environment from the perspective of *productivity*, which may offer a solution to this integration problem. If the impact of a unit change in temperature on occupant work efficiency can be quantified, it could provide a strong rationale for converting temperature into a monetary cost via occupant wages. However, among the reviewed reinforcement learning papers, none appeared to determine weighting coefficients by deductively translating the indoor thermal environment into economic terms based on such research findings.

In summary, a theoretically sound and widely accepted method for integrating the disparate metrics of energy and comfort within a reward function has yet to be established.

4.3. State variables for comfort Assessment: Selection and Implications

To represent comfort in a reward function, the first issue is which comfort indicator to use. It is important to note that since the literature search was conducted using "HVAC" as a keyword, the focus is predominantly on the thermal environment and air quality. An aggregation of the comfort indicators used in the reward functions summarized in Fig. 5 shows that the most frequently used, in descending order, were: dry-bulb temperature (57 cases), PMV (14 cases), CO_2 concentration (8 cases), PPD (6 cases), relative humidity (2 cases), $PM_{2.5}$ (2 cases), and others (5 cases). The 'others' category consisted of one case each of operative temperature, PM_{10}, supply air temperature, illuminance, and APAQ[121]. It should be noted that the total count of 94 does not match the number of papers reviewed (79), as some reward functions incorporate multiple comfort indicators. The combinations of the top four types of comfort indicators are shown in Fig. 3.

A related issue is whether to use a single comfort indicator (e.g., only dry-bulb temperature or only PMV) or multiple indicators to represent comfort. Sixty-eight studies (86 %) used a single comfort indicator, while eight (10 %) combined two indicators, two used three, and one used four.

The use of only dry-bulb temperature to represent comfort was the

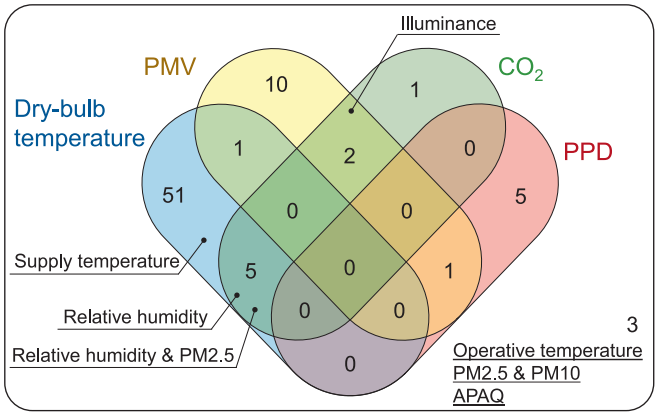


Fig. 5. Classification of Comfort Indicators Used in Reward Functions. The diagram illustrates the usage frequency and co-occurrence of the four primary indicators (Dry-bulb temperature, PMV, PPD, and CO_2 concentration). Less frequent indicators, other than these four, are listed directly within their respective regions.

most typical approach, found in 50 cases (63 %). As dry-bulb temperature is the state variable most directly controlled and abundantly sensed by HVAC systems, its use to represent comfort is a natural choice. Conversely, the most complex example was the reward function by Guo et al. [64], which combined dry-bulb temperature, relative humidity, PM2.5, and CO2 concentration. Although this approach has the advantage of evaluating thermal comfort and indoor air quality simultaneously, it raises the same concern as integrating energy and comfort: the difficulty of determining the weighting coefficients.

When evaluating comfort using multiple indicators, it may be possible to integrate them based on theory rather than weighting coefficients. For example, if both dry-bulb and radiant temperatures are observed, their combined effect on the human body can be expressed using convective and radiative heat transfer coefficients. This allows them to be integrated into a single state variable like operative temperature, as demonstrated in the reward function of Esrafilian-Najafabadi and Haghighat [57]. In particular, indices like PMV and PPD integrate the non-linear relationships of multiple thermal factors using theoretical formulas, making them potentially more accurate in evaluating thermal comfort than a simple linear sum with weighting coefficients. However, combining PMV and PPD within the same reward function, as was done by Ding et al. [54], is arguably meaningless for the purpose of identifying an optimal point. Because both indices are determined by the same six thermal factors and have a one-to-one correspondence, the information they provide is redundant. That said, as one moves away from the comfort zone, PPD increases more rapidly than PMV, so it might have value as a form of reward shaping to accelerate learning in the initial stages.

The decision to use integrated indices like operative temperature or PMV cannot be based on reward function design alone; it is also contingent on the observability of the required input variables within the environment. The subject of reinforcement learning generally needs to follow a Markov Decision Process (MDP), which requires observability. Therefore, if some elements needed to calculate these integrated indices cannot be observed or predicted and merely fluctuate uncertainly, learning can become unstable. For this reason, for example, Zhuang et al. [114] used PMV in the reward function but treated only dry-bulb temperature and relative humidity as variables when calculating it. The other four factors (radiant temperature, air velocity, clothing value, and metabolic rate) were set to some estimated or fixed values because only dry-bulb temperature and relative humidity were observable. Thus, when using an integrated index as a reward, care must be taken to ensure it aligns with the observable state variables.

Using these comfort indicators in reward design poses a fundamental problem: the potential to induce *reward hacking*. The ultimate goal is the comfort of individual occupants, for which indices like temperature or PMV are merely proxies. Therefore, rewarding an agent for adjusting such indices is a case of specifying the *means* rather than the *end*. Previously, directly observing an individual's comfort was difficult, so specifying this *means* has been accepted as a practical compromise. However, this situation is beginning to change. Recent efforts reviewed in this paper—such as those by Lan et al. [7], Li et al. [77], and Lim et al. [79]—use machine learning to predict individual thermal sensations, aiming for reward designs that are closer to the true *end*. As technologies for directly estimating individual comfort (e.g., physiological measurements from wearable devices) continue to advance, reward designs based on indirect, representative comfort indices will likely become less justifiable.

4.4. Structuring comfort in Rewards: Common techniques and Considerations

Fig. 6 summarizes how the four typical techniques discussed in Section 3.3—Occupancy Consideration (OC), Comfort Limits (CL), error exponentiation (P), and Acceptable Limits (AL)—were combined in the reward functions of the reviewed literature. While CL was the most frequent technique used in isolation (in 30 cases), no other particular combination was notably prevalent. This suggests that researchers are often designing reward functions on a trial-and-error basis and that a standard methodology for combining these techniques has not yet been established.

As a premise for the following discussion, these four techniques can be grouped into two categories based on their objectives. The primary objective of P and AL is to improve learning speed by encouraging the agent to quickly abandon exploration outside the acceptable range and return to a viable region. It is likely for this reason that, as shown in Fig. 6, there are relatively few cases where P and AL are used in combination. In contrast, the primary objective of OC and CL is to define the final optimal state that the agent should achieve. Therefore, compared to P and AL, OC and CL are techniques that should more strongly reflect our domain expertise.

4.4.1. Incorporating occupancy information into rewards

The purpose of air conditioning is to make occupants present in a room feel comfortable. Therefore, achieving a comfortable indoor environment when occupants are absent is of no direct value. For this

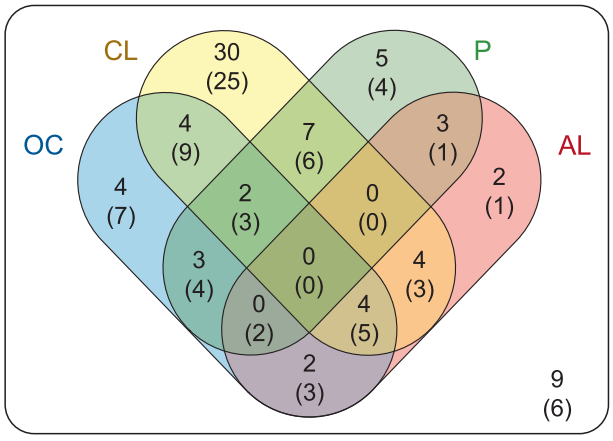


Fig. 6. Venn diagram showing the adoption and combination of the four typical reward function design techniques (OC, CL, P, and AL) across the reviewed literature. The numbers indicate the count of papers corresponding to each intersection for explicitly defined techniques. The numbers in parentheses represent the results of a re-aggregation that includes 14 studies in which occupancy (OC) was considered indirectly (e.g., through time-based schedules).

reason, HVAC systems are typically controlled by changing their on/off state or temperature setpoints based on a schedule. A more advanced method is CO_2 control, which estimates occupant presence to vary the ventilation rate. However, the reward function should not be designed using the HVAC's operational state or the ventilation rate. Instead, the reward signal should reflect the comfort of the occupants who are actually present; the agent itself should then learn what equipment control strategy is necessary to achieve that outcome.

In addition to the 19 cases that explicitly represented occupant presence in the reward function, there were also three types of indirect methods, as illustrated below.

The first method was to apply different reward functions for different time periods, effectively creating a time-based schedule (e.g., [54,17,92,97,99]. This is likely based on the idea that since occupant presence can be predicted by the time of day, the comfort-related reward weight can be adjusted in advance.

The second was to change the comfort range according to occupant presence (e.g., [56,61,63]. For instance, in the example from Gao and Wang [61], the comfort range was set to 21–24 °C during occupancy and 15–30 °C during absence.

The third was to evaluate comfort using TSV or PPD and aggregate the values only for the occupants present[79],Liu et al. [83]. With this calculation method, the comfort-related reward becomes zero during periods of absence.

In total, there were 14 cases presumed to indirectly represent occupant presence in the ways exemplified above. Combined with the direct methods, this amounts to 33 cases (42 %), meaning that just under half of all studies incorporated occupancy information into the reward function in some way.

However, the first and second methods described above risk specifying the *means* to the agent, rather than the *end*, and thus may lead to reward hacking.

In the first method, occupant presence is predicted based on the time of day and embedded into the reward function; however, this is arguably a pattern the agent itself should learn. Since the *end* for reinforcement learning is to improve the comfort of the occupants currently present, the *means* of pre-adjusting the comfort weight based on a prediction of how many people are likely to be present should not be reflected in the reward function. The agent should be given information to predict the number of occupants (e.g., day of the week, time) and be made to learn the model of how many people are likely to be present on.

The second method is based on the idea that maintaining a certain temperature during absence allows for a quick adjustment to the comfort zone when occupants arrive; this is clearly a *means*. The appropriate temperature to maintain during absence is influenced by factors such as the building's thermal capacity and the frequency of occupants entering and leaving. Learning the *means* of what the temperature should be and at what time is the agent's role, and we should not pre-emptively build it into the reward function. If the objective—to achieve an appropriate temperature range during occupancy—is correctly communicated, the agent will learn the method of maintaining a sufficient temperature as needed, even during periods of absence.

4.4.2. Defining comfort Zones (Deadbands) and their significance

This technique was the most frequently adopted, likely due to its consistency with established control methods such as deadband control [122]and with Key Performance Indicators (KPIs) used in well-known testbeds like BOPTEST[123].

Although it is presumed to have been adopted empirically in many papers, a more objective basis can be established by leveraging the theoretical framework provided by standards such as ASHRAE 55–2017 (2017). Underpinning this standard is the pragmatic goal of defining a thermal environment acceptable to a substantial majority (80 %) of occupants. Specifically, it sets a quantitative target of keeping the Predicted Percentage of Dissatisfied (PPD) below 10 % for general thermal comfort, from which the PMV range of −0.5 to + 0.5 is derived. Thus, the standard's concept of minimizing a concrete, human-centric metric like the dissatisfaction rate provides a strong theoretical grounding for setting the deadband's boundaries. This would encourage a departure from empirical parameter setting.

This technique also offers a significant advantage for designing reward functions in reinforcement learning. As previously discussed, an unresolved problem in our field is how to integrate energy performance and comfort. By establishing a deadband, the influence of comfort can be disregarded within this range, thereby circumventing the integration problem and allowing the reward function to be designed based solely on energy performance.

A general advantage of deadband control is the improvement of control stability. This is an effect where, by intentionally avoiding a single optimal point, the HVAC system is prevented from cycling on and off excessively. However, it should be noted that this effect cannot be expected simply by introducing a comfort deadband in the reward function design. If the energy performance, which is paired with comfort, is not also constant, an optimal point will still exist in the integrated reward function. Therefore, to ensure stability, the reward can be structured differently. For example, some studies have attempted to achieve this by introducing a penalty for changes in action (e.g., [51,86,98]. However, few of the reward functions listed in Table 3 included a term aimed at such control stabilization. This is likely for two reasons. First, in simulation-based studies, the phenomenon of equipment degradation from frequent on and off, as occurs in real buildings, is not reproduced, so the problem does not become apparent. Second, the simulations used are often not at the minute or second level where cycling can be discussed, but rather discrete, static simulations are still common. Consequently, it will likely be necessary to incorporate such control stabilization into the reward function in the future.

4.4.3. Non-linear error Transformation and penalty design for exceeding acceptable limits

Error exponentiation is primarily aimed at accelerating and stabilizing the learning process rather than precisely indicating the final optimal solution. Whether this technique improves learning depends on the overall shape of the reward function and the specific learning algorithm used; its introduction is not always effective. Therefore, the appropriateness of its setup should be evaluated pragmatically, as it is difficult to determine theoretically in advance. Thus, if it proves effective for improving learning, using a non-linear function like the natural exponential function, as exemplified in the reward functions of Biemann et al. [46], Kadamala et al. [69], and Miao et al. [85], is acceptable.

However, it is crucial to note that these techniques have a negative impact on the objective we have been discussing: the integration of energy performance and comfort. Applying transformations that increase the non-linearity of the target makes the concept more difficult to explain. Therefore, unless there is a theory—for example, that the value of one unit of energy loss is proportional to the square of the temperature difference—such transformations typically make the integration with energy performance even more challenging. Consequently, such non-linear transformations should likely be avoided within the operational region where the trade-off between energy performance and comfort is being evaluated. Conversely, this is not an issue in ranges that can never be optimal from a comfort perspective—that is, the region beyond the acceptable limits. In this region, an optimal point is not expected to exist; rather, the requirement is to return to the acceptable range as quickly as possible.

For instance, Biemann et al. [46] aimed to improve learning by exponentiating the error in one term of the reward function. However, this also provides a strong reward as the room temperature approaches the setpoint, which is inconsistent with another term that treats all comfort levels within the comfortable range as equivalent. Similarly, in studies like Gupta et al. [65] and Li et al. [77], the error is exponentiated across all regions. While this may increase learning speed, it complicates the discussion of the optimal point (i.e., integration with energy

performance). Conversely, the work of Lim et al. [79] presents a case where these two aims—improving learning and integrating with energy performance—could potentially be balanced. In that study, the absolute value of PMV was used as the reward penalty, but this penalty was squared when the PMV exceeded a threshold of 1.0, which can be considered the acceptable limit (as ± 0.5 is the typical comfort range). The intention here is likely to encourage a rapid exit from the region once the value exceeds this limit. The reward function in Shi et al. [92] also progressively magnifies the error, presumably with a similar goal. However, because their reward functions are discontinuous at the conditional branching points, they risk causing learning instability for gradient-based RL algorithms.

4.5. Proposal of a typical reward function structure based on literature review

This section proposes a specific reward function structure designed to mitigate some of the potential risks discussed previously. The intention, however, is not to present a standard function for universal adoption, as such a prescription would undermine a major strength of reinforcement learning: its flexibility in reward function design. On the other hand, as has been shown, current reward functions are so diverse that they suffer from a lack of comparability. Therefore, in designing reward functions, a balance must be struck between two conflicting requirements: tailoring the function to the specific problem being solved, and adhering as much as possible to a common pattern. The goal of this section is to present a typical structure that can be used as a reference in such situations.

Let A be the controllable variable vector (i.e., the action), and let the energy consumption E and the comfort indicator x be represented by functions $\in(A)$ and $\chi(A)$, respectively. Here, x is treated as a scalar, though it can be easily extended to a vector. However, for the purpose of integration with energy performance, it is desirable to use as few indicators as possible. Furthermore, while energy consumption might in reality be measured separately for individual pieces of equipment, it should be evaluated in the reward function as a single, aggregated scalar.

The reward function $r(A)$ is defined piecewise in three regions based on the value of the comfort-related state variable $\chi(A)$, as expressed in the following equation:

$$r(\boldsymbol{A}) = \begin{cases} -w_E \in(\boldsymbol{A}) & (x_{CLL} < \chi(\boldsymbol{A}) < x_{CUL}) \\ -w_E \in(\boldsymbol{A}) - w_C J_C(\chi(\boldsymbol{A})) & ((x_{ALL} < \chi(\boldsymbol{A}) < x_{CLL}) \bigvee (x_{CUL} < \chi(\boldsymbol{A}) < x_{AUL})) \\ -J_{ALV}(\chi(\boldsymbol{A})) & ((\chi(\boldsymbol{A}) < x_{ALL}) \bigvee (x_{AUL} < \chi(\boldsymbol{A}))) \end{cases} \tag{9}$$

Here, $J_C(\bullet)$ is the cost function for the comfort indicator, $J_{ALV}(\bullet)$ is the cost function for when the comfort indicator violates the acceptable limits, and w_E and w_C are the weighting coefficients for energy performance and comfort, respectively (Fig. 7).

As shown in the first case of the equation, when the comfort indicator is within the comfortable range, the agent seeks the optimal point based solely on energy performance. In the second case, when the comfort indicator is within the acceptable limits but outside the comfort zone, the agent seeks the optimal point by evaluating the trade-off between energy performance and comfort. Although they are integrated here using weights w_E and w_C, as previously discussed, a theoretically sound method for this integration has not yet been established. Occupancy information should be incorporated into the reward function within this region. In the third case, if the comfort indicator moves outside the acceptable limits, the $J_{ALV}(\bullet)$ term is used to quickly guide the agent back into the acceptable range. In this range, since the objective is no longer to find an optimal point, the energy performance term is unnecessary.

The function $J_{ALV}(\bullet)$ must impose a large penalty that increases as the indicator moves further away from the acceptable limits. This can be expressed by the following condition:

$$\begin{cases} \dfrac{dJ_{ALV}(\chi)}{d\chi} < 0 & if \quad \chi(\boldsymbol{A}) < x_{ALL} \\ 0 < \dfrac{dJ_{ALV}(\chi)}{d\chi} & if \quad x_{AUL} < \chi(\boldsymbol{A}) \end{cases} \tag{10}$$

Generally, the cost function $J_C(\cdot)$ should also increase as the indicator moves away from the comfort limits, leading to the following condition:

$$\begin{cases} \dfrac{dJ_C(\chi)}{d\chi} < 0 & if \quad \chi(\boldsymbol{A}) < x_{CLL} \\ 0 < \dfrac{dJ_C(\chi)}{d\chi} & if \quad x_{CUL} < \chi(\boldsymbol{A}) \end{cases} \tag{11}$$

For learning stability, the reward function should preferably be continuous. To satisfy this requirement at the comfort limits, the following condition is needed:

$$J_C(\chi(\boldsymbol{A})) = 0 \quad if \quad (\chi(\boldsymbol{A}) = x_{CLL}) \bigvee (\chi(\boldsymbol{A}) = x_{CUL}) \tag{12}$$

Similarly, continuity at the acceptable limits requires:

$$w_E \in(\boldsymbol{A}) + w_C J_C(\chi(\boldsymbol{A})) = J_{ALV}(\chin(\boldsymbol{A})) \quad if \quad (\chi(\boldsymbol{A}) = x_{ALL}) \bigvee (\chi(\boldsymbol{A}) = x_{AUL}) \tag{13}$$

However, because $\in(A)$ can typically be determined only through simulation, it is difficult to pre-design the relationship between $J_C(\bullet)$ and $J_{ALV}(\bullet)$ to guarantee that Eq. 13 holds. Therefore, in practice, the

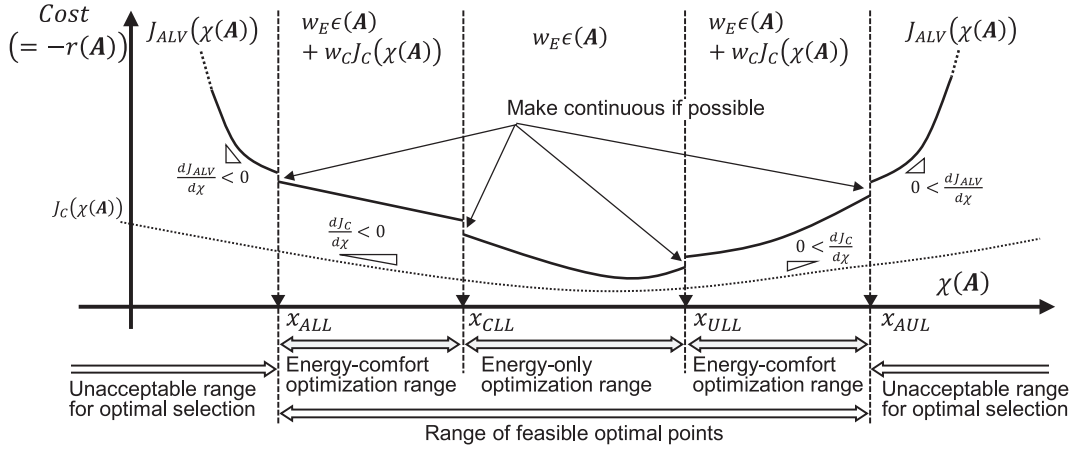


Fig. 7. Conceptual diagram of the proposed typical piecewise reward function structure, illustrating how the optimization objective and reward function shape vary depending on the value of the comfort indicator $\chi(\boldsymbol{A})$.

most that can be done is likely to impose the minimum condition that the penalty at the acceptable limit is no less than the sum of the energy and comfort costs at that boundary. This can be expressed by the following equation:

$$w_E \in (\boldsymbol{A}) + w_C J_C(\chi(\boldsymbol{A})) \leq J_{ALV}(\chi(\boldsymbol{A})) \quad if \quad (\chi(\boldsymbol{A}) = x_{ALL}) \bigvee (\chi(\boldsymbol{A}) = x_{AUL}) \tag{14}$$

For example, if a nominal energy consumption, E_N [GJ], is assumed to be the maximum possible energy consumption, a reward function that satisfies a condition of this nature can be readily designed.

The piecewise definition described above merely offers a solution to the structural problem of determining in which region the balance between energy performance and comfort should be considered. The fundamental challenge of determining specific values for the weighting coefficients w_E and w_C in the intermediate region—whether theoretically or pragmatically—persists. Addressing this issue represents the core challenge in reward function design for this field and should be a primary focus of future research.

As the above remains a conceptual design guideline, a concrete example that satisfies it is presented below. The reward function and its weighting coefficients are defined by Equations 15–17.

$$r(\boldsymbol{A}) = \begin{cases} -w_E \in (\boldsymbol{A}) - w_C J_C(0.5) & (|PMV| \leq 0.5) \\ -w_E \in (\boldsymbol{A}) - w_C J_C(PMV) & (0.5 < |PMV| \leq 1.0) \\ -w_E E_N - w_C J_{ALV}(PMV) & (1.0 < PMV) \end{cases} \tag{15}$$

$$w_E = \lambda_{sell,E} \tag{16}$$

$$w_C = \frac{N_{oc} Sal_m}{WH_m} \tag{17}$$

Here, S_{alm} [USD/month] is the monthly salary of occupants, and WH_m [hours/month] is the monthly working hours. By defining the weighting coefficients in this manner, the first and second terms in each case of Eq. 15 can be aligned to the same unit, such as cost per floor area (USD/m^2). This is a crucial point; if the weights can be defined by such physically meaningful concepts rather than by arbitrary values from trial-and-error, the comparability with other studies is greatly enhanced.

The cost functions for comfort are based on the performance-decrement function f_{PD} (•) from Lan et al. [124] and are defined as:

$$J_C(PMV) = f_{PD}(PMV) \tag{18}$$

$$J_{ALV}(PMV) = f_{PD}(PMV^2) \tag{19}$$

$$f_{PD}(PMV) = 0.00135 + 0.000351 PMV^3 + 0.005294 PMV^2 + 0.00215 PMV \tag{20}$$

It should be noted that while a comfort cost term appears in the first case of Eq. 15, its purpose is not to account for comfort (which is already satisfied), but rather to ensure mathematical continuity with the equation in the second case. Similarly, the purpose of the nominal energy cost term in the third case is not to impose a penalty based on energy use, but to guarantee that the reward in this violation region is always lower than that of the intermediate region. However, a limitation of this specific formulation is that it introduces a discontinuity at the boundary between the acceptable and unacceptable ranges, which can cause learning instability for some reinforcement learning algorithms that rely on gradient-based optimization methods. Furthermore, it is important to note that on the right-hand side of Eq. 19, only the PMV term is squared, not the entire function. Since PMV always exceeds 1.0 in this region, this ensures that the increase in cost accelerates relative to the increase in PMV in the third case of Eq. 15. However, in this specific example, this manipulation is not strictly necessary, as f_{PD} (•) already increases non-linearly with PMV.

While the typical reward function structure has been presented above, it is one that has been logically constructed by interpreting the intent of the patterns shown in many of the reward functions from prior research. Whether this reward function structure is truly effective should be validated through practical problems, which will require future studies.

5. Conclusions

This study investigated the critical impact of reward function design on performance in the application of reinforcement learning (RL) to HVAC control. We conducted a detailed review of 79 academic papers published since 2020, focusing on how the trade-off between thermal comfort and energy performance is managed within the reward function. To enable a meaningful comparison of these functions, we introduced a methodology for their abstraction and standardization, which involved unifying variable symbols and defining common error and normalization functions.

The review revealed significant diversity in reward function definitions, which severely hinders the comparability of research outcomes. The integration of energy performance and comfort, in the vast majority of studies, relies on a linear weighted sum, but the weighting factors employed often lack a strong theoretical basis. Consequently, a robust and theoretically sound method for integrating these two disparate concepts has yet to be established. Dry-bulb temperature was the most frequently used state variable for comfort assessment, while the use of integrated indices like PMV and PPD was limited, with their application facing challenges related to observability. Furthermore, we analyzed common techniques used to shape the comfort component of the reward—occupant consideration (OC), comfort limits (CL), error exponentiation (P), and acceptable limits (AL). While each serves a distinct purpose, these techniques, if not applied judiciously, can risk contradicting foundational principles of reward design or further complicating the integration of energy and comfort.

Drawing on the findings of this review, a structure for a piecewise reward function is presented, which applies distinct evaluations across three regions: the comfort range, the acceptable range, and outside the acceptable range. This represents an attempt to systematically incorporate varying situational demands into the reward function—such as the pursuit of energy efficiency, the balance with comfort, and rapid recovery from deviation states—and can serve as a guideline for future reward function design.

A key future challenge for this field lies in establishing a more theoretically grounded and quantitative methodology for integrating the disparate scales of energy and comfort. Related to this is the fundamental question of whether aggregating individual occupant comfort—typically by summing or averaging—to form a single collective index is a valid approach. The problem of how to aggregate, or perhaps individually address, the comfort requirements of different individuals may demand discussion that extends to philosophical and ethical perspectives, such as fairness and respect for the individual in HVAC control. In the field of information science, research has already begun on how to incorporate such “fairness” into reinforcement learning (S. [125].

To tackle this problem, it is first necessary to know the individual preferences related to comfort, and as exemplified in this paper, some studies have begun to identify the thermal preferences of each occupant using machine learning. Since occupants enter and leave rooms, and their thermal sensations are stochastic, a reward function would need to have highly dynamic characteristics, varying significantly over time. To achieve successful learning in real-time under such conditions, Adaptive Reward Shaping[126]may offer a technical solution.

On the other hand, a perfect integration of energy performance and comfort as described above is unlikely to be achieved in the short term. Given this reality, advancing multi-objective optimization (MOO) techniques presents an effective interim path forward. If MOO can be used to reveal the Pareto front, it would enable operators to later select a single operating point once a theoretically sound integration formula becomes available. This is, at the very least, more valuable than the

current approach of using weighting coefficients with an unclear basis, which arbitrarily designates a single point on the front as the sole candidate for the optimum. However, as has been pointed out, it must be accurately recognized that exploring the Pareto front does not, in itself, lead to the identification of a specific operating point. Ultimately, the integration of both performance indicators remains the most critical need for the field.

Enhancing inter-study comparability will necessitate fostering a common understanding of reward function design patterns and developing frameworks for benchmark or standardized reward functions. Additionally, factors that were only mentioned briefly in this review, such as control stability and applicability to real systems, should be more deeply considered in future reward function design.

Regarding applicability to real systems, several important issues remain that extend beyond the scope of the present study. For instance, while almost all the reviewed studies were premised on immediate rewards given at each time step, real-world scenarios often involve sparse and delayed rewards, such as those provided only upon task completion. In terms of energy performance, this applies to cases affected by a building's thermal lag or the presence of energy-shifting equipment like thermal or battery storage. For comfort, it will likely take considerable time before immediate physiological measurements from wearable devices become feasible; until then, discrete surveys will be the most that can be done. Effective reward function design under such conditions is an area awaiting future research.

Furthermore, this study focused on the structure of the reward function itself; however, an optimal reward function is inherently intertwined with and mutually influences the design of the agent's available choices (action space) and observable information (observation space). The development of a comprehensive design theory that encompasses these interdependencies presents a significant future challenge. The persistence of these unresolved issues strongly indicates that the discourse surrounding reward function design in this domain is still in its developmental stages, necessitating continuous and multifaceted investigation.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author used *ChatGPT-4o* and *gemini* in order to proofread the English text. After using this service, the author reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Eisuke Togashi: Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The author has no other known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was partially supported by JSPS KAKENHI Grant Numbers JP 23K04148.

Data availability

Data will be made available on request.

References

[1] International Energy Agency (IEA) (2023). *The Breakthrough Agenda Report 2023, Accelerating transition across the world's most emitting sectors*, https://www.iea.org/reports/breakthrough-agenda-report-2023.
[2] T. Al Mindeel, E. Spentzou, M. Eftekhari, Energy, thermal comfort, and indoor air quality: Multi-objective optimization review, Renew. Sustain. Energy Rev. 202 (2024) 114682, https://doi.org/10.1016/j.rser.2024.114682.
[3] Ala'raj, M., Radi, M., Abbod, M. F., Majdalawieh, M., & Parodi, M. (2022). Data-driven based HVAC optimisation approaches: A systematic literature review. Journal of Building Engineering, 46, 103678. https://doi.org/10.1016/j.jobe.2021.103678.
[4] X. Xin, Z. Zhang, Y. Zhou, Y. Liu, D. Wang, S. Nan, A comprehensive review of predictive control strategies in heating, ventilation, and air-conditioning (HVAC): Model-free vs. model, Journal of Building Engineering 94 (2024) 110013, https://doi.org/10.1016/j.jobe.2024.110013.
[5] S.L. Zhou, A.A. Shah, P.K. Leung, X. Zhu, Q. Liao, A comprehensive review of the applications of machine learning for HVAC, DeCarbon 2 (2023) 100023, https://doi.org/10.1016/j.decarb.2023.100023.
[6] S. Brandi, M. Fiorentini, A. Capozzoli, Comparison of online and offline deep reinforcement learning with model predictive control for thermal energy management, Autom. Constr. 135 (2022) 104128, https://doi.org/10.1016/j.autcon.2022.104128.
[7] H. Lan, H. Huiying, G. Zhonghua, Z. Gou, User-centric approach to optimizing thermal comfort in university classrooms: Utilizing computer vision and Q-XGBoost reinforcement learning, Energ. Buildings 323 (2024) 114808, https://doi.org/10.1016/j.enbuild.2024.114808.
[8] H.-A. Park, G. Byeon, W. Son, J. Kim, S. Kim, Data-Driven Modeling of HVAC Systems for Operation of Virtual Power Plants using a Digital Twin, Energies 16 (20) (2023) 7032.
[9] I. Ajifowowe, H. Chang, C.S. Lee, S. Chang, Prospects and challenges of reinforcement learning- based HVAC control, Journal of Building Engineering 98 (2024) 111080, https://doi.org/10.1016/j.jobe.2024.111080.
[10] K. Al Sayed, A. Boodi, R. Sadeghian Broujeny, K. Beddiar, Reinforcement learning for HVAC control in intelligent buildings: a technical and conceptual review, Journal of Building Engineering 95 (2024) 110085, https://doi.org/10.1016/j.jobe.2024.110085.
[11] A. Chatterjee, D. Khovalyg, Dynamic indoor thermal environment using Reinforcement Learning-based controls: Opportunities and challenges, Build. Environ. 244 (2023) 110766, https://doi.org/10.1016/j.buildenv.2023.110766.
[12] Han, M., May, R., Zhang, X., Wang, X., Pan, S., Yan, D.,…Xu, L. (2019). A review of reinforcement learning methodologies for controlling occupant comfort in buildings. Sustainable Cities and Society, 51, 101748. https://doi.org/10.1016/j.scs.2019.101748.
[13] M. Han, J. Zhao, X. Zhang, J. Shen, Y. Li, The reinforcement learning method for occupant behavior in building control: a review, Energy Built Environ. 2 (2) (2021) 137–148, https://doi.org/10.1016/j.enbenv.2020.08.005.
[14] S. Sierla, H. Ihasalo, V. Vyatkin, A review of reinforcement learning applications to control of heating, ventilation and air conditioning systems, Energies 15 (10) (2022) 3526.
[15] Z. Wang, T. Hong, Reinforcement learning for building controls: the opportunities and challenges, Appl. Energy 269 (2020) 115036, https://doi.org/10.1016/j.apenergy.2020.115036.
[16] H. Yu, V.W.Y. Tam, X. Xu, A systematic review of reinforcement learning application in building energy-related occupant behavior simulation, Energ. Buildings 312 (2024) 114189, https://doi.org/10.1016/j.enbuild.2024.114189.
[17] X. Liu, Z. Gou, Occupant-centric HVAC and window control: a reinforcement learning model for enhancing indoor thermal comfort and energy efficiency, Build. Environ. 250 (2024) 111197, https://doi.org/10.1016/j.buildenv.2024.111197.
[18] R.S. Sutton, A.G. Barto, Reinforcement learning: an introduction, (2nd ed.)., MIT Press, 2018.
[19] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. arXiv:1606.06565. Retrieved June 01, 2016, from https://ui.adsabs.harvard.edu/abs/2016arXiv160606565A.
[20] A. Ng, D. Harada, S.J. Russell, Policy Invariance under Reward Transformations: Theory and Application to Reward Shaping, in: In *Proceedings of the Sixteenth International Conference on Machine Learning*, 1999, pp. 278–287.
[21] R. Devidze, Reward Design for Reinforcement Learning Agents, ArXiv abs/2503.21949 (2025).
[22] E. Andrés, M.P. Cuéllar, G. Navarro, On the use of Quantum Reinforcement Learning in Energy-Efficiency Scenarios, Energies 15 (16) (2022) 6034.
[23] M. Devasenan, S. Madhavan, Thermal intelligence: exploring AI's role in optimizing thermal systems – a review, Interactions 245 (1) (2024) 282, https://doi.org/10.1007/s10751-024-02122-6.
[24] Z.K. Ding, Q.M. Fu, J.P. Chen, H.J. Wu, Y. Lu, F.Y. Hu, Energy-efficient control of thermal comfort in multi-zone residential HVAC via reinforcement learning, Connect. Sci. 34 (1) (2022) 2364–2394, https://doi.org/10.1080/09540091.2022.2120598.
[25] H.T. Dinh, D. Kim, MILP-Based Imitation Learning for HVAC Control, IEEE Internet Things J. 9 (8) (2022) 6107–6120, https://doi.org/10.1109/JIOT.2021.3111454.
[26] A. Campoy-Nieves, A. Manjavacas, J. Jiménez-Raboso, M. Molina-Solana, J. Gómez-Romero, Sinergym – a virtual testbed for building energy optimization with Reinforcement Learning, Energ. Buildings 327 (2025) 115075, https://doi.org/10.1016/j.enbuild.2024.115075.

[27] T. Marzullo, S. Dey, N. Long, J. Leiva Vilaplana, G. Henze, A high-fidelity building performance simulation test bed for the development and evaluation of advanced controls, J. Build. Perform. Simul. 15 (3) (2022) 379–397, https://doi.org/10.1080/19401493.2022.2058091.

[28] Q. Fu, X. Chen, S. Ma, N. Fang, B. Xing, J. Chen, Optimal control method of HVAC based on multi-agent deep reinforcement learning, Energ. Buildings 270 (2022) 112284, https://doi.org/10.1016/j.enbuild.2022.112284.

[29] K. He, Q. Fu, Y. Lu, J. Ma, Y. Zheng, Y. Wang, J. Chen, Efficient model-free control of chiller plants via cluster-based deep reinforcement learning, Journal of Building Engineering 82 (2024) 108345, https://doi.org/10.1016/j.jobe.2023.108345.

[30] Chen, C., An, J., Wang, C., Duan, X., Lu, S., Che, H.,…Yan, D. (2023). Deep Reinforcement Learning-Based Joint Optimization Control of Indoor Temperature and Relative Humidity in Office Buildings. Buildings, 13(2), 438.

[31] L. Chen, F. Meng, Y. Zhang, Fast Human-in-The-Loop Control for HVAC Systems via Meta-Learning and Model-based Offline Reinforcement Learning, IEEE Trans. Sustainable Comput. 8 (3) (2023) 504–521, https://doi.org/10.1109/TSUSC.2023.3251302.

[32] Y. Fan, Q. Fu, J. Chen, Y. Wang, Y. Lu, K. Liu, A deep reinforcement learning control method for multi-zone precooling in commercial buildings, Appl. Therm. Eng. 260 (2025) 124987, https://doi.org/10.1016/j.applthermaleng.2024.124987.

[33] M. Esrafilian-Najafabadi, F. Haghighat, Transfer learning for occupancy-based HVAC control: a data-driven approach using unsupervised learning of occupancy profiles and deep reinforcement learning, Energ. Buildings 300 (2023) 113637, https://doi.org/10.1016/j.enbuild.2023.113637.

[34] X. Fang, G. Gong, G. Li, L. Chun, P. Peng, W. Li, X. Shi, Cross temporal-spatial transferability investigation of deep reinforcement learning control strategy in the building HVAC system level, Energy 263 (2023) 125679, https://doi.org/10.1016/j.energy.2022.125679.

[35] X. Deng, Y. Zhang, H. Qi, Toward smart multizone HVAC control by combining context-aware system and deep reinforcement learning, IEEE Internet Things J. 9 (21) (2022) 21010–21024, https://doi.org/10.1109/JIOT.2022.3175728.

[36] X. Deng, Y. Zhang, H. Qi, Towards optimal HVAC control in non-stationary building environments combining active change detection and deep reinforcement learning, Build. Environ. 211 (2022) 108680, https://doi.org/10.1016/j.buildenv.2021.108680.

[37] R. Shen, S. Zhong, R. Zheng, D. Yang, B. Xu, Y. Li, J. Zhao, Advanced control framework of regenerative electric heating with renewable energy based on multi-agent cooperation, Energ. Buildings 281 (2023) 112779, https://doi.org/10.1016/j.enbuild.2023.112779.

[38] J. Yang, J. Yu, S. Wang, Heating ventilation air-conditioner system for multi-regional commercial buildings based on deep reinforcement learning, Adv. Control Appl. 6 (4) (2024) e190.

[39] Y. Masdoua, M. Boukhnifer, K.H. Adjallah, Active fault-tolerant control based on DDQN architecture applied to HVAC system, Trans. Inst. Meas. Control 01423312241273767 (2024), https://doi.org/10.1177/01423312241273767.

[40] K. Yan, C. Lu, X. Ma, Z. Ji, J. Huang, Intelligent fault diagnosis for air handing units based on improved generative adversarial network and deep reinforcement learning, Expert Syst. Appl. 240 (2024) 122545, https://doi.org/10.1016/j.eswa.2023.122545.

[41] K.U. Ahn, C.S. Park, Application of deep Q-networks for model-free optimal control balancing between different HVAC systems, Sci. Technol. Built Environ. 26 (1) (2020) 61–74, https://doi.org/10.1080/23744731.2019.1680234.

[42] N.S. Alsharafa, R. Suguna, R.J. Krishna, V.K. Sonthi, S.M. Padmaja, P. Mariaraja, Optimizing Building Energy Management with Deep Reinforcement Learning for Smart and Sustainable Infrastructure, Journal of Machine and Computing 4 (2) (2024) 381–391. https://doi.org/10.53759/7669/jmc202404036.

[43] A. Azimi, O. Akbari, A deep reinforcement learning-based method for dynamic quality of service aware energy and occupant comfort management in intelligent buildings. e-Prime - advances in Electrical Engineering, Electronics and Energy 9 (2024) 100700, https://doi.org/10.1016/j.prime.2024.100700.

[44] D. Azuatalam, W.L. Lee, F. de Nijs, A. Liebman, Reinforcement learning for whole-building HVAC control and demand response, Energy AI 2 (2020) 100020, https://doi.org/10.1016/j.egyai.2020.100020.

[45] L. Bai, Z. Tan, Optimizing energy efficiency, thermal comfort, and indoor air quality in HVAC systems using a robust DRL algorithm, Journal of Building Engineering 98 (2024) 111493, https://doi.org/10.1016/j.jobe.2024.111493.

[46] M. Biemann, P.A. Gunkel, F. Scheller, L. Huang, X. Liu, Data center HVAC control harnessing flexibility potential via real-time pricing cost optimization using reinforcement learning, IEEE Internet Things J. 10 (15) (2023) 13876–13894, https://doi.org/10.1109/JIOT.2023.3263261.

[47] S. Brandi, M.S. Piscitelli, M. Martellacci, A. Capozzoli, Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings, Energ. Buildings 224 (2020) 110225, https://doi.org/10.1016/j.enbuild.2020.110225.

[48] Z. Chen, L. Yu, S. Zhang, S. Hu, C. Shen, Multiagent Hierarchical Deep Reinforcement Learning for operation Optimization of Grid-Interactive Efficient Commercial buildings, IEEE Trans. Artif. Intell. 5 (8) (2024) 4280–4292, https://doi.org/10.1109/TAI.2024.3366869.

[49] D. Coraci, S. Brandi, M.S. Piscitelli, A. Capozzoli, Online implementation of a soft actor-critic agent to enhance indoor temperature control and energy efficiency in buildings, Energies 14 (4) (2021), https://doi.org/10.3390/en14040997.

[50] C. Cui, J. Xue, Energy and comfort aware operation of multi-zone HVAC system through preference-inspired deep reinforcement learning, Energy 292 (2024) 130505, https://doi.org/10.1016/j.energy.2024.130505.

[51] S.M. Dawood, A. Hatami, R.Z. Homod, Trade-off decisions in a novel deep reinforcement learning for energy savings in HVAC systems, J. Build. Perform. Simul. 15 (6) (2022) 809–831, https://doi.org/10.1080/19401493.2022.2099465.

[52] X. Ding, A. Cerpa, W. Du, Exploring Deep Reinforcement Learning for Holistic Smart Building Control, ACM Trans. Sens. Netw. 20 (3) (2024), https://doi.org/10.1145/3656043.

[53] X. Ding, A. Cerpa, W. Du, Multi-Zone HVAC Control with Model-based Deep Reinforcement Learning, IEEE Trans. Autom. Sci. Eng. 1–19 (2024), https://doi.org/10.1109/TASE.2024.3410951.

[54] Z. Ding, Q. Fu, J. Chen, Y. Lu, H. Wu, N. Fang, B. Xing, MAQMC: Multi-Agent Deep Q-Network for Multi-Zone Residential HVAC Control, CMES - Computer Modeling in Engineering and Sciences 136 (3) (2023) 2759–2785. https://doi.org/10.32604/cmes.2023.026091.

[55] A. Dmitrewski, M. Molina-Solana, R. Arcucci, CNTRLDA: a building energy management control system with real-time adjustments. Application to indoor temperature, Build. Environ. 215 (2022) 108938, https://doi.org/10.1016/j.buildenv.2022.108938.

[56] Du, Y., Zandi, H., Kotevska, O., Kurte, K., Munk, J., Amasyali, K.,…Li, F. (2021). Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning. Applied Energy, 281, Article 116117. https://doi.org/10.1016/j.apenergy.2020.116117.

[57] M. Esrafilian-Najafabadi, F. Haghighat, Towards self-learning control of HVAC systems with the consideration of dynamic occupancy patterns: Application of model-free deep reinforcement learning, Build. Environ. 226 (2022) 109747, https://doi.org/10.1016/j.buildenv.2022.109747.

[58] Fang, X., Gong, G., Li, G., Chun, L., Peng, P., Li, W.,…Chen, X. (2022). Deep reinforcement learning optimal control strategy for temperature setpoint real-time reset in multi-zone building HVAC system. Applied Thermal Engineering, 212, Article 118552. https://doi.org/10.1016/j.applthermaleng.2022.118552.

[59] Friansa, K., Pradipta, J., Mahesa Nanda, R., Nashirul Haq, I., Armanto Mangkuto, R., Fauzi Iskandar, R.,…Leksono, E. (2024). Enhancing University Building Energy Flexibility Performance Using Reinforcement Learning Control. IEEE Access, 12, 192377-192395. https://doi.org/10.1109/ACCESS.2024.3512543.

[60] C. Fu, Y. Zhang, Research and Application of Predictive Control Method based on Deep Reinforcement Learning for HVAC Systems, IEEE Access 9 (2021) 130845–130852, https://doi.org/10.1109/ACCESS.2021.3114161.

[61] C. Gao, D. Wang, Comparative study of model-based and model-free reinforcement learning control performance in HVAC systems, Journal of Building Engineering 74 (2023) 106852, https://doi.org/10.1016/j.jobe.2023.106852.

[62] G. Gao, J. Li, Y. Wen, DeepComfort: Energy-Efficient thermal Comfort Control in buildings Via Reinforcement Learning, IEEE Internet Things J. 7 (9) (2020) 8472–8484, https://doi.org/10.1109/JIOT.2020.2992117.

[63] Y. Gao, S. Shi, S. Miyata, Y. Akashi, Successful application of predictive information in deep reinforcement learning control: a case study based on an office building HVAC system, Energy 291 (2024) 130344, https://doi.org/10.1016/j.energy.2024.130344.

[64] F. Guo, S.W. Ham, D. Kim, H.J. Moon, Deep reinforcement learning control for co-optimizing energy consumption, thermal comfort, and indoor air quality in an office building, Appl. Energy 377 (2025) 124467, https://doi.org/10.1016/j.apenergy.2024.124467.

[65] A. Gupta, Y. Badr, A. Negahban, R.G. Qiu, Energy-efficient heating control for smart buildings with deep reinforcement learning, Journal of Building Engineering 34 (2021) 101739, https://doi.org/10.1016/j.jobe.2020.101739.

[66] A. Heidari, D. Khovalyg, DeepValve: Development and experimental testing of a Reinforcement Learning control framework for occupant-centric heating in offices, Eng. Appl. Artif. Intel. 123 (2023) 106310, https://doi.org/10.1016/j.engappai.2023.106310.

[67] A. Heidari, L. Girardin, C. Dorsaz, F. Maréchal, A trustworthy reinforcement learning framework for autonomous control of a large-scale complex heating system: simulation and field implementation, Appl. Energy 378 (2025) 124815, https://doi.org/10.1016/j.apenergy.2024.124815.

[68] Z. Jiang, M.J. Risbeck, V. Ramamurti, S. Murugesan, J. Amores, C. Zhang, K. H. Drees, Building HVAC control with reinforcement learning for reduction of energy cost and demand charge, Energ. Buildings 239 (2021) 110833, https://doi.org/10.1016/j.enbuild.2021.110833.

[69] K. Kadamala, D. Chambers, E. Barrett, Enhancing HVAC control systems through transfer learning with deep reinforcement learning agents, Smart Energy 13 (2024) 100131, https://doi.org/10.1016/j.segy.2024.100131.

[70] L. Kannari, J. Kantorovitch, K. Piira, J. Piippo, Energy cost Driven heating Control with Reinforcement Learning, Buildings 13 (2) (2023), https://doi.org/10.3390/buildings13020427.

[71] N. Kodama, T. Harada, K. Miyazaki, Home energy management algorithm based on deep reinforcement learning using multistep prediction, IEEE Access 9 (2021) 153108–153115, https://doi.org/10.1109/ACCESS.2021.3126365.

[72] Kurte, K., Munk, J., Kotevska, O., Amasyali, K., Smith, R., McKee, E.,…Zandi, H. (2020). Evaluating the adaptability of reinforcement learning based HVAC control for residential houses. Sustainability (Switzerland), 12(18), Article 7727. https://doi.org/10.3390/su12187727.

[73] K.B. Kwon, J.Y. Park, S.M. Hong, J.H. Heo, H. Jung, Development of machine learning-based energy management agent to control fine dust concentration in railway stations, J. Electr. Eng. Technol. 19 (4) (2024) 2757–2766, https://doi.org/10.1007/s42835-023-01730-6.

[74] Y. Lei, S. Zhan, E. Ono, Y. Peng, Z. Zhang, T. Hasama, A. Chong, A practical deep reinforcement learning framework for multivariate occupant-centric control in

buildings, Appl. Energy 324 (2022) 119742, https://doi.org/10.1016/j.apenergy.2022.119742.
[75] R. Li, Z. Zou, How far back shall we peer? Optimal air handling unit control leveraging extensive past observations, Build. Environ. 269 (2025) 112347, https://doi.org/10.1016/j.buildenv.2024.112347.
[76] W. Li, Y. Zhao, J. Zhang, C. Jiang, S. Chen, L. Lin, Y. Wang, Indoor temperature preference setting control method for thermal comfort and energy saving based on reinforcement learning, Journal of Building Engineering 73 (2023) 106805, https://doi.org/10.1016/j.jobe.2023.106805.
[77] W. Li, H. Wu, Y. Zhao, C. Jiang, J. Zhang, Study on indoor temperature optimal control of air-conditioning based on Twin delayed Deep Deterministic policy gradient algorithm, Energ. Buildings 317 (2024) 114420, https://doi.org/10.1016/j.enbuild.2024.114420.
[78] Z. Li, Z. Sun, Q. Meng, Y. Wang, Y. Li, Reinforcement learning of room temperature set-point of thermal storage air-conditioning system with demand response, Energ. Buildings 259 (2022) 111903, https://doi.org/10.1016/j.enbuild.2022.111903.
[79] S.H. Lim, T.G. Kim, D.J. Yeom, S.G. Yoon, Robust deep reinforcement learning for personalized HVAC system, Energ. Buildings 319 (2024) 114551, https://doi.org/10.1016/j.enbuild.2024.114551.
[80] X. Lin, D. Yuan, X. Li, Reinforcement Learning with dual Safety policies for Energy Savings in Building Energy Systems, Buildings 13 (3) (2023), https://doi.org/10.3390/buildings13030580.
[81] B. Liu, M. Akcakaya, T.E. McDermott, Automated Control of Transactive HVACs in Energy distribution Systems, IEEE Trans. Smart Grid 12 (3) (2021) 2462–2471, https://doi.org/10.1109/TSG.2020.3042498.
[82] X. Liu, M. Ren, Z. Yang, G. Yan, Y. Guo, L. Cheng, C. Wu, A multi-step predictive deep reinforcement learning algorithm for HVAC control systems in smart buildings, Energy 259 (2022) 124857, https://doi.org/10.1016/j.energy.2022.124857.
[83] X. Liu, Y. Wu, H. Wu, Enhancing HVAC energy management through multi-zone occupant-centric approach: a multi-agent deep reinforcement learning solution, Energ. Buildings 303 (2024) 113770, https://doi.org/10.1016/j.enbuild.2023.113770.
[84] A. Manjavacas, A. Campoy-Nieves, J. Jiménez-Raboso, M. Molina-Solana, J. Gómez-Romero, An experimental evaluation of deep reinforcement learning algorithms for HVAC control, Artif. Intell. Rev. 57 (7) (2024), https://doi.org/10.1007/s10462-024-10819-x.
[85] C. Miao, Y. Cui, H. Li, X. Wu, Efficient multi-agent reinforcement learning HVAC power consumption optimization, Energy Rep. 12 (2024) 5420–5431, https://doi.org/10.1016/j.egyr.2024.11.011.
[86] A. Naug, M. Quinones-Grueiro, G. Biswas, Deep reinforcement learning control for non-stationary building energy management, Energ. Buildings 277 (2022) 112584, https://doi.org/10.1016/j.enbuild.2022.112584.
[87] A.T. Nguyen, D.H. Pham, B.L. Oo, M. Santamouris, Y. Ahn, B.T.H. Lim, Modelling building HVAC control strategies using a deep reinforcement learning approach, Energ. Buildings 310 (2024) 114065, https://doi.org/10.1016/j.enbuild.2024.114065.
[88] H. Qin, Z. Yu, T. Li, X. Liu, L. Li, Energy-efficient heating control for nearly zero energy residential buildings with deep reinforcement learning, Energy 264 (2023) 126209, https://doi.org/10.1016/j.energy.2022.126209.
[89] T.V. Quang, N.L. Phuong, Using Deep Learning to Optimize HVAC Systems in Residential Buildings, Journal of Green Building 19 (1) (2024) 29–50, https://doi.org/10.3992/jgb.19.1.29.
[90] G. Razzano, S. Brandi, M.S. Piscitelli, A. Capozzoli, Rule extraction from deep reinforcement learning controller and comparative analysis with ASHRAE control sequences for the optimal management of heating, Ventilation, and Air Conditioning (HVAC) systems in multizone buildings, Appl. Energy 381 (2025) 125046, https://doi.org/10.1016/j.apenergy.2024.125046.
[91] L. Scarcello, F. Cicirelli, A. Guerrieri, C. Mastroianni, G. Spezzano, A. Vinci, Pursuing Energy Saving and thermal Comfort with a Human-Driven DRL Approach, IEEE Trans. Hum.-Mach. Syst. 53 (4) (2023) 707–719, https://doi.org/10.1109/THMS.2022.3216365.
[92] Shi, Z., Zheng, R., Zhao, J., Shen, R., Gu, L., Liu, Y.,…Wang, G. (2024). Towards various occupants with different thermal comfort requirements: A deep reinforcement learning approach combined with a dynamic PMV model for HVAC control in buildings. Energy Conversion and Management, 320, Article 118995. https://doi.org/10.1016/j.enconman.2024.118995.
[93] M. Shin, S. Kim, Y. Kim, A. Song, H.Y. Kim, Development of an HVAC system control method using weather forecasting data with deep reinforcement learning algorithms, Build. Environ. 248 (2024) 111069, https://doi.org/10.1016/j.buildenv.2023.111069.
[94] Silvestri, A., Coraci, D., Brandi, S., Capozzoli, A., Borkowski, E., Köhler, J.,… Schlueter, A. (2024). Real building implementation of a deep reinforcement learning controller to enhance energy efficiency and indoor temperature control. Applied Energy, 368, Article 123447. https://doi.org/10.1016/j.apenergy.2024.123447.
[95] Y. Su, X. Zou, M. Tan, H. Peng, J. Chen, Integrating few-shot personalized thermal comfort model and reinforcement learning for HVAC demand response optimization, Journal of Building Engineering 91 (2024) 109509, https://doi.org/10.1016/j.jobe.2024.109509.
[96] L. Sun, Z. Hu, M. Mae, T. Imaizumi, Individual room air-conditioning control in high-insulation residential building during winter: a deep reinforcement learning-based control model for reducing energy consumption, Energ. Buildings 323 (2024) 114799, https://doi.org/10.1016/j.enbuild.2024.114799.
[97] Touzani, S., Prakash, A. K., Wang, Z., Agarwal, S., Pritoni, M., Kiran, M.,… Granderson, J. (2021). Controlling distributed energy resources via deep reinforcement learning for load flexibility and energy efficiency. Applied Energy, 304, Article 117733. https://doi.org/10.1016/j.apenergy.2021.117733.
[98] H. Wang, X. Chen, N. Vital, E. Duffy, A. Razi, Energy optimization for HVAC systems in multi-VAV open offices: a deep reinforcement learning approach, Appl. Energy 356 (2024) 122354, https://doi.org/10.1016/j.apenergy.2023.122354.
[99] M. Wang, B. Lin, MF^2: Model-free reinforcement learning for modeling-free building HVAC control with data-driven environment construction in a residential building, Build. Environ. 244 (2023) 110816, https://doi.org/10.1016/j.buildenv.2023.110816.
[100] X. Wang, N. Mahdavi, S. Sethuvenkatraman, S. West, An environment-adaptive SAC-based HVAC control of single-zone residential and office buildings, Data-Centric Eng. 6 (2025), https://doi.org/10.1017/dce.2024.57.
[101] T. Wei, S. Ren, Q. Zhu, Deep Reinforcement Learning for Joint Datacenter and HVAC load Control in distributed Mixed-Use buildings, IEEE Trans. Sustainable Comput. 6 (3) (2021) 370–384, https://doi.org/10.1109/TSUSC.2019.2910533.
[102] M. Xia, F. Chen, Q. Chen, S. Liu, Y. Song, T. Wang, Optimal Scheduling of Residential heating, Ventilation and Air Conditioning based on Deep Reinforcement Learning, J. Mod Power Syst. Clean Energy 11 (5) (2023) 1596–1605. https://doi.org/10.35833/MPCE.2022.000249.
[103] Y. Xia, X. Wang, X. Yin, W. Bo, L. Wang, S. Li, K. Li, Federated Accelerated Deep Reinforcement Learning for Multi-Zone HVAC Control in Commercial buildings, IEEE Trans. Smart Grid (2024), https://doi.org/10.1109/TSG.2024.3524756.
[104] Xu, D. (2022). Learning Efficient Dynamic Controller for HVAC System. Mobile Information Systems, 2022, Article 4157511. https://doi.org/10.1155/2022/4157511.
[105] W. Xue, N. Jia, M. Zhao, Multi-agent deep reinforcement learning based HVAC control for multi-zone buildings considering zone-energy-allocation optimization, Energ. Buildings 329 (2025) 115241, https://doi.org/10.1016/j.enbuild.2024.115241.
[106] Yu, L., Xie, W., Xie, D., Zou, Y., Zhang, D., Sun, Z.,…Jiang, T. (2020). Deep Reinforcement Learning for Smart Home Energy Management. IEEE Internet of Things Journal, 7(4), 2751-2762. https://doi.org/10.1109/JIOT.2019.2957289.
[107] L. Yu, Y. Sun, Z. Xu, C. Shen, D. Yue, T. Jiang, X. Guan, Multi-Agent Deep Reinforcement Learning for HVAC Control in Commercial buildings, IEEE Trans. Smart Grid 12 (1) (2021) 407–419, https://doi.org/10.1109/TSG.2020.3011739.
[108] L. Yu, Z. Xu, T. Zhang, X. Guan, D. Yue, Energy-efficient personalized thermal comfort control in office buildings based on multi-agent deep reinforcement learning, Build. Environ. 223 (2022) 109458, https://doi.org/10.1016/j.buildenv.2022.109458.
[109] X. Yuan, Y. Pan, J. Yang, W. Wang, Z. Huang, Study on the application of reinforcement learning in the operation optimization of HVAC system, Build. Simul. 14 (1) (2021) 75–87, https://doi.org/10.1007/s12273-020-0602-9.
[110] I. Zenginis, J. Vardakas, N.E. Koltsaklis, C. Verikoukis, Smart Home's Energy Management through a Clustering-based Reinforcement Learning Approach, IEEE Internet Things J. 9 (17) (2022) 16363–16371, https://doi.org/10.1109/JIOT.2022.3152586.
[111] B. Zhang, W. Hu, A.M.Y.M. Ghias, X. Xu, Z. Chen, Multi-agent deep reinforcement learning-based coordination control for grid-aware multi-buildings, Appl. Energy 328 (2022) 120215, https://doi.org/10.1016/j.apenergy.2022.120215.
[112] H. Zhao, J. Zhao, T. Shu, Z. Pan, Hybrid-Model-based Deep Reinforcement Learning for heating, Ventilation, and Air-Conditioning Control, Front. Energy Res. 8 (2021) 610518, https://doi.org/10.3389/fenrg.2020.610518.
[113] X. Zhong, Z. Zhang, R. Zhang, C. Zhang, End-to-End Deep Reinforcement Learning Control for HVAC Systems in Office Buildings, Designs 6 (3) (2022), https://doi.org/10.3390/designs6030052.
[114] D. Zhuang, V.J.L. Gan, Z. Duygu Tekler, A. Chong, S. Tian, X. Shi, Data-driven predictive control for smart HVAC system in IoT-integrated buildings with time-series forecasting and reinforcement learning, Appl. Energy 338 (2023) 120936, https://doi.org/10.1016/j.apenergy.2023.120936.
[115] Z. Zou, X. Yu, S. Ergan, Towards optimal control of air handling units using deep reinforcement learning and recurrent neural network, Build. Environ. 168 (2020) 106535, https://doi.org/10.1016/j.buildenv.2019.106535.
[116] American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE). (2017). *ANSI/ASHRAE Standard 55-2017: Thermal environmental conditions for human occupancy*.
[117] Togashi, E., Ogata, H., Ayame, H., Nakatsuka, K., Satoh, M., Ukai, M.,…Iio, Y. A benchmarking framework for HVAC optimization via competitive evaluation: insights from the 2nd wccbo. Journal of Building Performance Simulation, 1-14. https://doi.org/10.1080/19401493.2025.2539356.
[118] E. Togashi, M. Miyata, Y. Yamamoto, The first world championship in cybernetic building optimization, J. Build. Perform. Simul. 13 (3) (2020) 391–408, https://doi.org/10.1080/19401493.2020.1741685.
[119] W.J. Fisk, Estimates of potential nationwide productivity and health benefits from better indoor environments, in: J.D. Spengler, J.M. Samet, J.F. McCarthy (Eds.), *Indoor Air Quality Handbook*, McGraw-Hill, 2000.

[120] Seppänen, O., Fisk, W. J., & Lei, Q., H. (2006). Room temperature and productivity in office work. Healthy Buildings: Creating a Healthy Indoor Environment for People,.
[121] L. Gunnarsen, P. Ole Fanger, Adaptation to indoor air pollution, Environ. Int. 18 (1) (1992) 43–54, https://doi.org/10.1016/0160-4120(92)90209-M.
[122] Paoluccio, J. P. (1978). Dead band controls guide (CR 79.002).
[123] Blum, D., Arroyo, J., Huang, S., Drgoňa, J., Jorissen, F., Walnum, H. T.,…Helsen, L. (2021). Building optimization testing framework (BOPTEST) for simulation-based benchmarking of control strategies in buildings. Journal of Building Performance Simulation, 14(5), 586-610. https://doi.org/10.1080/19401493.2021.1986574.
[124] L. Lan, P. Wargocki, Z. Lian, Quantitative measurement of productivity loss due to thermal discomfort, Energ. Buildings 43 (5) (2011) 1057–1062, https://doi.org/10.1016/j.enbuild.2010.09.001.
[125] Zhang, S., Bai, J., Guan, M., Zhang, Y., Sun, J., Huang, Y.,…Pu, G. (2024). CFP: A Reinforcement Learning Framework for Comprehensive Fairness-Performance Trade-Off in Machine Learning. Artificial Neural Networks and Machine Learning – ICANN 2024, Cham.
[126] M. Chahoud, H. Sami, R. Mizouni, J. Bentahar, A. Mourad, H. Otrok, C. Talhi, Reward shaping in DRL: a novel framework for adaptive resource management in dynamic environments, Inf. Sci. 715 (2025) 122238, https://doi.org/10.1016/j.ins.2025.122238.