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

e.togashi@gmail.com

Contents

Acknowledgements:

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

Abstract

Reinforcement learning (RL) is increasingly applied to Heating, Ventilation, and Air Conditioning (HVAC) control, aiming to optimize both building energy efficiency and occupant comfort. However, reward function design, crucial for balancing these often-conflicting objectives, has been underexplored in existing reviews. This paper presents a systematic review of 78 studies published since 2020, analyzing how reward functions integrate thermal comfort and energy efficiency in RL for HVAC control. This paper introduces a standardization methodology to enable systematic comparison of diverse reward formulations. Key findings reveal substantial heterogeneity in reward structures, which impedes research comparability. Furthermore, a prevalent reliance on empirically derived weighting factors for comfort-energy integration, often lacking a strong theoretical basis, is identified. Common techniques in shaping comfort-related rewards, such as occupancy considerations, comfort deadbands, error exponentiation, and acceptable limits, are identified and critically evaluated. Based on this comprehensive analysis, a typical piecewise reward function structure is proposed to address key identified limitations. This review clarifies current practices, highlights critical challenges, and suggests future research directions for the design of robust reward functions in RL-driven HVAC systems.

1. Introduction

建設部門のエネルギー消費は非常に大きく、IEA の報告書（IEA 2023）によれば、2022 年時点で世界の CO2 排出量の 33%程度を占めている。同報告書によればその内の 33%の内の 79%（全体の 26%）は運用段階に発生しており、建設方法だけではなく、建物を運用する方法を改善することは世界のエネルギー効率に大きな影響を持っている。

このため、建物の設備を最適化する技術に関しては昔から多くの研究があり、快適性、省エネルギー性、室内空気質など、複数の評価関数を持つ多目的最適化問題（Multi-objective optimization）として認識されている(Al Mindeel 2024)。

空調設備の最適化という課題は、世の中に同じ設備がないという点に難しさがある。建物の立地、構造、執務者の特性はすべて異なり、これらに影響されて空調設備の構成も建物ごとに異なる。このため、ある建物で成立した最適解は、そのまま別の建物で使うことはできない。建物ごとにチューニングを繰り返さねばならないために人権費は大きくなる。

近年では、このような課題の解決方法の一つとして、機械学習の応用が期待されている。機械学習は伝統的な物理モデルだけではなく、現場で収集されたデータにもとづいたデータドリブンのモデルを活用する点が特徴である。これらの技術の空調分野への応用方法については、数百もの文献にもとづいたレビュー研究がいくつも報告されている(Xin 2024; Zhou 2023; Ala'raj 2022)。機械学習と現場で計測されたデータを使うことで、最適化のためのパラメータチューニングが自動化できれば、それぞれの建物と設備の特性を踏まえた最適化を安価に実現できる可能性がある（e.g., Park et al. 2023; Haifeng, L. 2024; Silvio Brandi et al. 2022）。

機械学習の中で、特にシステムの最適化に応用しやすい技術としては強化学習がある。これは、コンピュータの中で最適化をする主体（Agent）と最適化の対象（Environment）を仮想的に構築し、Agent が行動（Action）を試行錯誤する中で最適な Action（本研究に即して言えば空調パラメータ）を獲得する手法である。Agent は Action に応じて Environment から報酬（Reward）を受け取り、この報酬が最大化するように Action を調整していく。空調設備の最適化に強化学習を適用する研究は多数あり、いくつかのレビュー論文がある（Ajifowowe et al. (2024); Al Sayed et al. (2024); Chatterjee and Khovalyg (2023); Han et al. (2019; 2021); Sierla et al. (2022); Wang and Hong (2020); Yu et al. (2024)）（Table 1）。これらのレビュー論文では、それぞれ異なる観点から強化学習の応用可能性や課題が整理されているが、いずれも報酬関数の設計方法を主眼としていない。本研究の貢献は、報酬関数に特化したレビューによって、様々な論文で提案された報酬関数の意図を読み取り、相互に比較することで長所や短所を整理することにある。

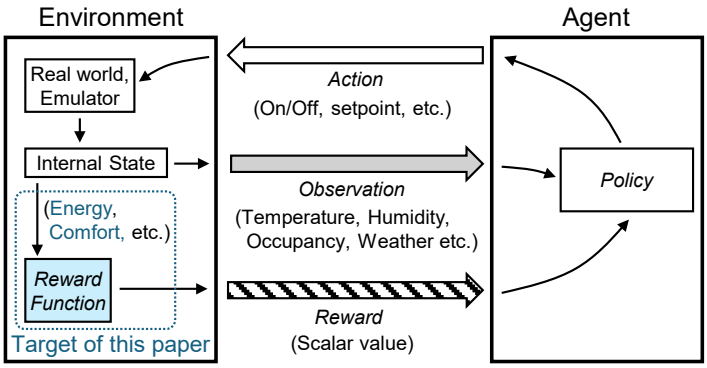

Figure 1 Basic Structure of Reinforcement Learning for HVAC Control and the Focus of This Review

Table 1 Comparison with Prior Papers

Paper	Main Focus	Coverage of Reward Function Design
Ajifowowe (2024) (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 (2024) (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 (2023) (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.
Han (2019; 2021) (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. (2022) (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 and Hong (2020) (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. (2024)	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. (2024a) (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.

そもそも強化学習の分野において、報酬関数の設計はエージェントの学習成果を左右する根源的かつ重要な課題として広く認識されている(Sutton, R. S., and Barto, A. G. 2018)。例えば、目的達成時にのみ報酬が与えられるような疎な報酬(Sparse Rewards)環境では学習が著しく困難になることや、設計者の意図しない抜け道を見つけてしまう Reward Hacking を防ぐことの難しさなどが指摘されている(Amodei et al. 2016)。このような課題に対応するため、学習を促進するための補助的な報酬を与える Reward Shaping といった技術も提案されている(Ng et al. 1999)。

しかし、情報分野の研究で例示された一般的な報酬設計論をそのまま使うだけでは不十分で、報酬関数はドメイン固有の専門知識(domain knowledge) を踏まえた設計が強く求められる (Devidze 2025)。特に省エネルギー性と快適性というような、異なる単位を持つ概念をバランスさせて最適化させるような場面では、両概念に精通しなければ適切な報酬関数は設計できない。従って、空調制御という応用領域に即した報酬関数の設計方針は、我々空調分野の専門家自身が十分に検討すべきであろう。

そこで本研究では、以上の議論を踏まえて、空調設備の最適化という目的に対して強化学習を応用した研究をレビューする。79 件の論文の報酬関数を標準化・比較し、その設計に驚くほどの多様性が存在し、研究間の比較可能性を著しく妨げているという現状を初めて定量的に明らかにする。その上で、多様な設計の中に共通して見られる典型的な工夫を抽出し、それらが我々の専門知識からはどのような意義を持つのかを検討する。最後に、これらの分析から得られた知見に基づき、既存研究の限界に対処するための新しい「典型的な区分的報酬関数構造」を提案する。これにより、強化学習を空調最適化に適用しようとする研究の相互の比較可能性が高まり、本分野の研究の進展を加速させる点に本研究の独自の貢献がある。

2. Methods for Literature Search and Selection

レビューの対象とする論文は PRISMA フローに従って抽出した（Figure 2）。

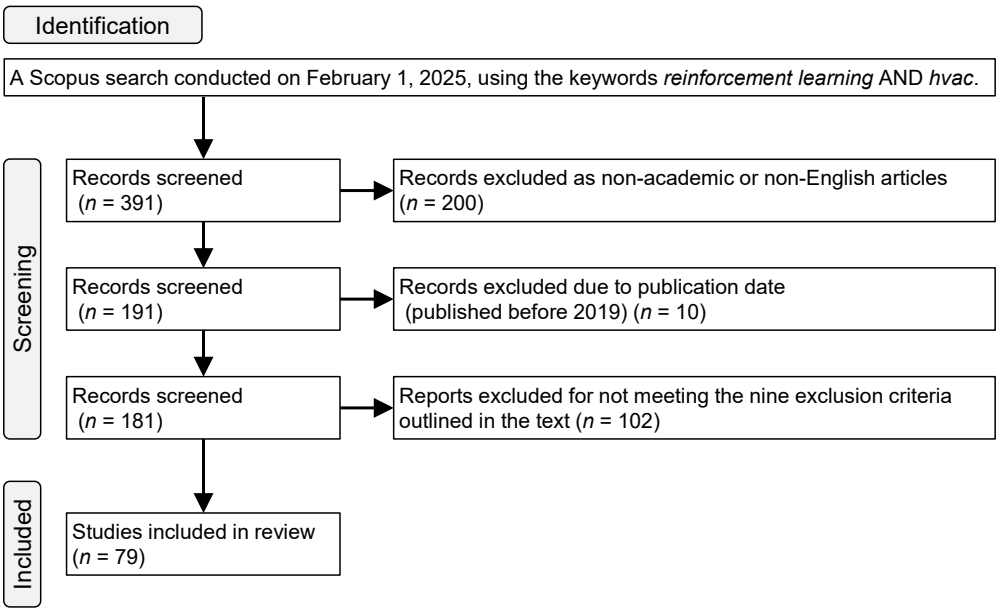

Figure 2 PRISMA 2020 flow diagram for the systematic review

Scopus を使い、2025 年 2 月 1 日時点で“reinforcement learning” AND hvac というキーワードで検索し、391 件の文献を得た。この中から英語（in English）で書かれた学術論文（Article）である 191 件を抽出した。強化学習を使った研究報告は 2020 年以降、急速に増えたため、近年の傾向を知るために 2019 年以前に発表された 10 件は除外した。

抽出された 181 件を読み、本研究の目的に照らして、以下に該当する文献は除外した。

1）自動運転などを含むなど、研究対象が広くて HVAC はその 1 例でしかない研究（Andrés, E. (2022)や Devasenan, M. (2024)など）
2）複数の最適化手法の比較が目的で、RL は手法の 1 例でしかない研究（Ding, Z. K. (2022)や Dinh, H. T. (2022)など）
3）RL 自体ではなく、RL のテストベッドの開発が目的の研究（Campoy-Nieves, A. (2025)や Marzullo, T. (2022)など）
4）プラント設備の最適化や VAV コントローラの制御など、執務者の快適性が評価に含まれていない研究（He, K.(2024)や Fu, Q.(2022)など）
5）エネルギーが報酬関数に含まれない研究（Chen, C.(2023)や Fan, Y.(2025)など）
6）転移学習が主題である研究（Esrafilian-Najafabadi, M. (2023)や Fang, X. (2023)など）
7）異なる論文で同一の報酬関数を再利用している研究。重複を避けるため、この場合には 1 件のみを本研究の事例として残した。（Deng, X.(2022a)と Deng, X.(2022b)など）
8）報酬関数が明示されていない、または曖昧なもの。（Yang, J. (2024)や Shen, R.(2023)など）
9）異常運転検知（Fault Diagnosis）が目的のもの（Masdoua, Y.(2024)や Yan, K. (2024)など）

以上の手続きの結果、本研究でレビュー対象とする論文は 79 件となった。参考に、本分野における他

のレビュー論文との比較を Table 2 に示す。

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	78	SC	from 2020	February 2025	“reinforcement learning” AND hvac
Ajifowowe (2024)	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 (2024)	135	48	SC	Not specified	August 2023	TITLE-AVS-KEY(“reinforcement learning”) AND TITLE-AVS-KEY(“building”) AND TITLE-AVS-KEY(“HVAC systems”)
Chatterjee (2023)	108	63	SD	Not specified	in 2022	“reinforcement learning” AND {(building OR house OR home) AND control} OR “smart thermostat”
Han (2019)	-	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 (2021)	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 (2022)	278	83	WS	from 2013	Not specified	“reinforcement learning” AND (heating OR ventilation OR “air conditioning” OR cooling OR HVAC)
Wang (2020)	77	77	WS	Not specified	December 2019	“reinforcement learning” AND (building OR house OR home OR residential) AND control
Yu (2024)	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.

3. Results: Summary Table of Reward Functions

抽出した文献に定義された報酬関数を比較するため、Table 3 に要約した。なお、時間依存の状態変数は共通して時刻 t を添字の t で表した。報酬関数相互の比較を容易にするため、原典の式をそのまま写すのではなく、以下に説明するように抽象化や簡略化を図った。

なお、これらの抽象化や簡略化には主観的な判断と限界が含まれることに注意されたい。報酬関数の相互比較という目的において許容される簡略化は、それぞれの論文で具体的に強化学習を適用する場面においては妥当ではない可能性は十分にある。例えば、一部の乗算係数や定数項の省略は数式の核心構造を比較しやすくする一方、元の著者が意図した可読性や特定の挙動を完全に再現するものではない。従ってこれらの簡略化した報酬関数をそのまま原論文の RL に適用したとしても、全く同じ最適化が保証されるわけではない。また、以下ではできる限り客観的に判定できる方法で簡略化するように務めたが、報酬関数を設計した者の意図が完全に解説されていない場合などには、筆者による予想が必要になり、その意味では主観的判断は完全には排除できない。

3.1 Standardization of Variables and Unit Notations

1) Unification of Variable Symbols

文献ごとに異なる変数記号は統一した。この際、スケーリングされた単位（Scaled Units）はすべて共通の記号とした。例えば Wh、kWh、kJ、MJ などはすべて E_t [GJ]と表現した。タイムステップが固定の場合には W や kW なども実質的には同じ概念となるため、これらも E_t [GJ]と表現した。

2) Integration of Energy-Related Terms

熱源機、ポンプ、ファンなどの消費電力を個別に計算して、それぞれに対して異なる重み係数を与える事例があった（Bai, L. (2024); Chen, Z. (2024)など）。しかし、異なる単価の売買電が無いとすれば 1 単位の電力の価値は等しいため、これらの電力は統合して表現した。また、報酬に熱負荷を使う事例もあったが、これもエネルギーとみなした。

3) Unified Representation of Electricity Purchase and Sale

太陽光パネルなどの発電設備を導入して売買電を評価する事例があった（Yu, L. (2020)や Zenginis, I.(2022)など）。発電量とエネルギー消費量に別々の記号を割り当てず、合計のエネルギー消費を E とし、これが正の場合に買電、負の場合に売電を表すこととした。

4) Taxonomy and Standardization of Comfort Indicators

報酬関数で使われる快適性指標は多いため、Figure 3 の通り分類した。包括的な目標である居住者快適性（Occupant Comfort）は、温熱快適性（Thermal Comfort）と室内空気質（IAQ）の 2 つのカテゴリに大別される。さらに、温熱快適性は、温度などの物理指標と、PMV や PPD、TSV といった複数の要素を統合した統合指標に細分化される。

いくつかの研究では Thermal Sensation Vote（TSV）を推定して、これを報酬関数の要素とした。例えば Haifeng, L.(2024)、Li, W.(2024)、Lim, S. H.(2024) はそれぞれ、KDE、Takagi–Sugeno fuzzy model、Machine learning を使って値を推定した。これらは厳密には PMV の定義（ASHRAE 2017）とは異なるが、人の熱的嗜好を 7 段階のスケールで表現するという目的は共通のため、同じ PMV という変数で表現した。

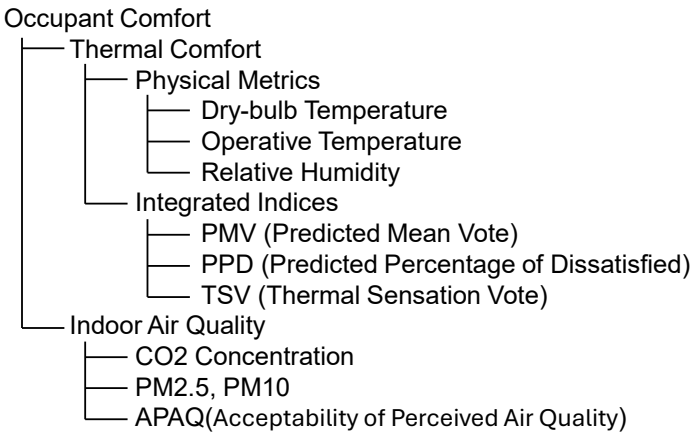

Figure 3 Taxonomy of the comfort indicators discussed in this review

3.2 Simplification of Reward Calculation Formulae

1) Omission of Redundant Multipliers and Non-essential Terms

報酬値を計算するまでに複数回の乗算を繰り返す事例があった（例えば Friansa, K. (2024)や Kadamala, K.(2024) など)。最初の乗算は快適性とエネルギー性能の項を概ね同じオーダーに変換するような意義を持っているが、数式の上では、最後に最後に乗じる重み係数を適切に調節すれば同じ出力が得られる。従って、比較可能性を高めるために重み係数以外の乗数は省略した。

報酬関数の中に単純な加減項を持つ事例があった（例えば Li. W (2024)や Friansa, K.(2024) など)。これは、望ましい状態を正の範囲、望ましくない状態を負の範囲にすることで、人間にとって報酬関数が理解しやすくすることなどが狙いだろう。しかし、強化学習の学習アルゴリズムにとっては、値の正負は意味を持たず、大小関係が重要である。このため、これらの加減項は省略した。

2) Simplification of Summation for Multi-Object Evaluation

部屋が分かれるなどして類似の評価対象が複数ある場合に、それぞれに同じ評価式を適用して積算値を報酬とする事例があった（Cui, C. (2024b)や Ding, X. (2024b)など)。報酬関数の形状を相互に比較するという目的においては、このような積算は意味を持たないため、積算記号を省略した。

3) Introduction of Normalization Functions for Data Scaling

報酬の要素をスケーリングする事例が多かったため、状態変数の最大値 x_{max} と最小値 x_{min} を使った典型的な標準化式を次の記号で表現した。

$$N_{max}(x_t) = \frac{x_t}{x_{max}} \quad \text{Eq. 1}$$

$$N_{minmax}(x_t) = \frac{x_t - x_{min}}{x_{max} - x_{min}} \quad \text{Eq. 2}$$

なお、次式のように適当な係数 k を乗じるスケーリングは、既に述べたように重み係数 w と統合できるため省略した。

$$N_{linear}(x_t) = kx_t \quad \text{Eq. 3}$$

4) Standardization of Weighting Coefficient Representation

例えば次式のように報酬関数が 2 つの要素（r_1、r_2）で表現されているとする。

$$w_1 r_1 + w_2 r_2 \quad \text{Eq. 4}$$

このとき、新しい重み係数 $w_3 = w_1/w_2$ を定義することで、次式のように簡略化する場合があった。

$$w_3 r_1 + r_2 \quad \text{Eq. 5}$$

この場合にも、3 以上の要素を持つ報酬関数との比較がしやすいように Eq. 4 で表現した。

また、Kurte, K.(2020)のように重み係数を表現せず、異なる報酬に関わる項を単に足し合わせた事例もあった。これは重みをすべて 1 にしたことと同じため、式の上では w を表示した。

5) Representation of Core Reward Components in Shaping

報酬関数が、エージェントの学習促進（いわゆるシェイピングなど）を意図してか、ある評価指標の 1 タイムステップ間の変化量（差分）として定義される事例があった（Gao, C.(2023); Kwon, K. B. (2024) など）。この場合には他の即時報酬との比較可能性を重視し、その変化量を計算する基となる現在のタイムステップにおける評価指標を報酬として表示した。

3.3 Introduction of a Common Error Function f_{err} and Typical Adjustments

快適性を報酬関数に反映するための最も単純な方法は、次式で表される誤差関数をコスト関数とみなし、この正負を逆転した値を報酬とする方法である。

$$f_{err}(x_t) = |x_t - x_{sp}| \quad \text{Eq. 6}$$

ここで、x_t は快適性に関わる状態変数（Comfort-related State Variable）で、乾球温度、相対湿度、PMV 値などが使える。以下、x_t を快適性指標（comfort indicator）、と呼ぶ。x_{sp} は快適性という観点からの理想値である。この理想値からの乖離の絶対値が誤差である。なお、本式は $x_{sp} = 0$ とすれば二酸化炭素濃度、PM2.5、PMV、PPD など、0 が最適となる指標にも援用できる点に注意されたい。

しかし、多くの既往研究ではこのような単純な式は使われず、いくつかの典型的な改良が加えられていた。そこで、報酬関数の比較を容易にするために、共通で表すことができる誤差関数記号を導入する。典型的な 4 つの改良を加えた式を次に示す。

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

以下、4 つのそれぞれの工夫について解説する。

1) Incorporating Occupancy Information into Rewards

1^{ε}_{NOC} は次式で示す 1 または極小値をとる一般化指示関数（generalized indicator function）である。執務者の在室時には 1、不在時には極めて小さい値 ε をとる。

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

ここで N_{oc} [person]は執務者数である。なお、ε は 0 となる可能性もあり、この場合には通常の指示関数となる。これは執務者の不在時に、室内の快適性が報酬関数に与える影響を無くす、または極めて小さくすることが目的である。

2) Defining Comfort Range

温度や湿度などは理想値から少し逸脱しても大きな不快は発生しない。このため、例えば ASHRAE 55-2017 (2017)では、PMV についてある程度の幅をとって推奨している。Eq. 7 では、この上限値（CUL: Comfortable Upper Limit）と下限値（Comfortable Lower Limit）をそれぞれ x_{CUL} と x_{CLL} と表した。Eq. 7 の条件 2（otherwise case）では、状態値がこの快適範囲から逸脱したときのみ正の値となり、快適範囲内の場合には 0 となる。つまり、負の報酬を与えない。なお、この範囲内において積極的に正の報酬をあたる

事例（Cui, C. 2024b）もあるが、これらも共通にこの記号で表現した。

3) Non-linear Error Transformation

学習を高速化させるためには、最適値から離れるとともにペナルティを大きく拡大させることで、適切な範囲に戻ることを促した方が良い。このために Eq. 7 では誤差を P 乗している。ただし、Friansa, K. (2024) のように、P を 1 未満としてペナルティの増加を緩やかにする例外もあった。

4) Defining Acceptable Range

温度や湿度が快適範囲から大きく離れると、健康を害したり、執務者のクレームが出たりして、許容できなくなる。この許容できる上限値（AUL: Acceptable Upper Limit）と下限値（ALL: Acceptable Lower Limit）をそれぞれ x_{AUL} と x_{ALL} と表す。Eq. 7 の条件 1（first case）ではこの上下の限界値を超えた場合に、大きなペナルティ $f_{penalty}$ を与えている。$f_{penalty}$ は大きな定数としたり、快適性指標 x_t を使って計算したりするが、いずれにせよ、条件 2（otherwise case）よりも大きな値にする。

5) Notation of error function

以上に解説したように、Eq. 6 で表される単純な誤差関数に対して、執務者の有無（Occupancy (OC)）、快適限界（Comfortable Limits (CL)）、誤差の累乗（Raising the error to the power (P)）、許容限界（Acceptable Limits (AL)）、という 4 つの工夫を加えた誤差関数が Eq. 7 である。既往研究ではこれらの工夫の全てではなく一部のみが採用されているものも多い。そこで、Table 2 では f_{err} 関数の添字に 4 つの記号（OC, CL, P, AL）を追加することで、どの工夫が採用されているのかを表すことにした。なお、ある上下限値まではコストを 0 とし、それらの限界値を超えた瞬間に極めて大きなペナルティを与える事例があった（Du, Y.(2021a), Guo, F.(2025), Heidari, A.(2023)）。これらは快適範囲と許容範囲が一致しているとみなせるため、添字の CAL で表現した。

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

Paper	The reward at timestep t (r_t)
Ahn & Park (2020)	$\begin{cases} -E_t & (c_{co2,ave,t} \le 1000) \\ -1.5E_t & (1000 < c_{co2,ave,t}) \end{cases}$
Alsharafa et al. (2024)	$-w_1 E_t - w_2 D_t - w_3 f_{err(CL)}(T_t)$
Azimi & Akbari (2024)	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CL)}(T_t)$
Azuatalam et al. (2020)	$-w_1 N_{max}(\lvert E_t \rvert) - w_2 N_{minmax}\left(f_{err(CL)}(PMV_t)\right)$
Bai & Tan (2024)	$-w_1 N_{max}(E_{ahu,t}) - w_2 \begin{cases} 1 & (0 < E_{fan,t}) \\ 0 & (0 = E_{fan,t}) \end{cases}$ $- w_3 \begin{cases} 95 - 95\exp(-0.029PMV^4 - 0.205PMV^2) + \frac{PMV_t}{4} & (\lvert PMV_t \rvert \le 0.5) \\ 20 & (0.5 < \lvert PMV_t \rvert) \end{cases}$ $- w_4 f_{err(CL,AL)}\left(N_{minmax}(c_{co2,t})\right)$
Biemann et al. (2023)	$-w_1 E_t \lambda_{E,t} + w_2 \left\{\exp\left(-\alpha f_{err(P)}(T_t)\right) - \beta f_{err(CL)}(T_t)\right\}$
Brandi et al. (2020)	$-w_1 \frac{Q_t}{T_{sp,t} - T_{ext}} - w_2 f_{(OC,P)}(T_t)$
L. Chen et al. (2023)	$-w_1 E_t - w_2 f_{err(CL)}(T_t) - w_3 f_{err,tp}(T_t) - w_4 u_{epi}$
Z. Chen et al. (2024)	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CL)}(E_t) - w_3 f_{err(OC,CL)}(T_t) - w_4 f_{err(CL,OC)}(c_{co2,t})$
Coraci et al. (2021)	$-w_1 E_t - w_2 f_{err(OC,CL,P)}(T_t)$
Cui & Xue (2024)	$-w_1 E_t - w_2 f_{err(CL)}(T_t) - w_3 f_{err(CL)}(c_{co2,t})$
Dawood et al. (2022)	$-w_1 \exp(vp_{chw,t}) - w_2 \left(f_{err}(T_t) + dp_t f_{err}(x_{co2,t})\right)$
Deng et al. (2022a)	$-w_1 N_{minmax}(E_t) - w_2 f_{err(OC,AL)}(PPD_t)$
X. Ding et al. (2024a)	$-w_1 N_{minmax}(E_t) - w_2 N_{minmax}\{f_{err(CL)}(PMV_t)\} - w_3 N_{minmax}\{f_{err(CL)}(E_{lux,t})\}$ $- w_4 N_{minmax}\{f_{err(CL)}(c_{co2,t})\}$
X. Ding et al. (2024b)	$-w_1 N_{minmax}(E_t) - \begin{cases} w_2 & (N_{oc} > 0) \\ w_3 & (N_{oc} = 0) \end{cases} N_{minmax}\{f_{err}(PMV_t)\}$
Z. Ding et al. (2023)	$\begin{cases} -w_1(\lvert PMV \rvert + PPD) - w_2 E_t \lambda_{E,t} & (PMV, PPD\ meet\ requirements) \\ -C_{penalty} & (else) \end{cases}$
Dmitrewski et al. (2022)	$-w_1 N_{linear}(E_t) - w_2 N_{linear}\left(f_{err(CL)}(T_t)\right)$
Du et al. (2021a)	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CAL)}(T_t)$
Esrafilian-Najafabadi & Haghighat (2022)	$-w_1 E_t - w_2 f_{err(OC,P)}(T_{op,t})$
Fang et al. (2022)	$-w_1 N_{minmax}(E_t) - w_2 f_{err(OC)}(N_{minmax}(T_t))$
Friansa et al. (2024)	$E_t \begin{cases} w_1 & (0 < E_t) \\ w_2 & (E_t \le 0) \end{cases} - w_3 f_{err(P)}(T_t)$
C. Fu & Zhang (2021)	$-w_1 E_t - w_2 (T_t - T_{ext})^2$
C. Gao & Wang (2023)	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CL)}(T_t)$
G. Gao et al. (2020)	$-w_1 E_t - w_2 f_{err(CL)}(PMV)$
Y. Gao et al. (2024)	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CL)}(T_t)$

Paper	The reward at timestep t (r_t)
Guo et al. (2025)	$-w_1 E_t - w_2 f_{err(CL)}(T_t) - w_3 f_{err(CL)}(\phi_t) - w_4 f_{err(CAL)}(c_{co2}) - w_5 f_{err(CAL)}(\rho_{pm25})$
Gupta et al. (2021)	$-w_1 N_{minmax}(E_t \lambda_E) - w_2 N_{minmax}\left(f_{err(P)}(T_t)\right)$
Heidari et al. (2023)	$-w_1 E_t^{\alpha} - w_2 \lvert PMV_t \rvert^{\beta}$
Heidari et al. (2025)	$-w_1(T_t - T_{t-1}) - w_2 f_{err(OC,CAL)}(T_t)$
Jiang et al. (2021)	$-w_1 N_{max}(E_t)\lambda_E - w_2 1_{NOC} * \begin{cases} 0 & (T_{CLL} < T_t < T_{CUL}) \\ \exp(\max(0, T_{ALL} - T_t, T_t - T_{AUL})) & (T_{ALL} < T_t < T_{CLL} \text{ or } T_{CUL} < T_t < T_{AUL}) \\ 0.5 f_{err(CL)}(T_t) & (otherwise) \end{cases}$
Kadamala et al. (2024)	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CL)} N_{minmax}(E_t) - w_3 f_{err(CL)}(T_t)$
Kannari et al. (2023)	$-w_1 E_t - w_2 \exp\left(f_{err(CL)}(T_t)\right)$
Kodama et al. (2021)	$-E_t\left(w_1 + w_2 f_{err(CL,P)}(T_t)\right)$
Kurte et al. (2020)	$-w_1 E_t - w_2 f_{err(CL)}(T_t)$
Kwon et al. (2024)	$-w_1 E_t - w_2 f_{err(CL)}(T_t)$
H. Lan et al. (2024)	$-w_1 E_t \lambda_{E,t} - w_2\left(\rho_{pm25} + \rho_{pm10}\right)$
Lei et al. (2022)	$-E_t \begin{cases} w_1 & (N_{oc} > 0) \\ w_2 & (N_{oc} = 0) \end{cases} - w_3\left(\overline{APAQ} + \alpha\right)^P$
R. Li & Zou (2025)	$-w_1 E_t - w_2 f_{err(OC)}(PPD_t)$
W. Li et al. (2023)	$-w_1 \begin{cases} f_{err(CL)}\left(\lvert T_{sp,t+1} - T_t \rvert\right) & (GTS > 0) \\ f_{err(CL)}\left(\lvert T_{sp,t+1} - T_{AUL} \rvert\right) & (GTS < 0) \end{cases} - w_2 f_{err(AL)}(GTS)$
W. Li et al. (2024)	$-w_1 E_t - \begin{cases} w_2 \lvert PMV_t \rvert^{2.5} & (\lvert PMV_t \rvert \le 0.5) \\ w_3 \lvert PMV_t \rvert^{1.5} & (0.5 < \lvert PMV_t \rvert) \end{cases}$
Z. Li et al. (2022)	$-w_1 N_{max}(E_t) - w_2 f_{err(CL)}\left(N_{max}(\lvert PMV_t \rvert)\right)$
Lim et al. (2024)	$-w_1 E_t - w_2 1_{NOC} \begin{cases} \lvert PMV_t \rvert & (\lvert PMV_t \rvert \le 1.0) \\ \lvert PMV_t \rvert^2 & (1.0 < \lvert PMV_t \rvert) \end{cases}$
Lin et al. (2023)	$-w_1 E_t - w_2 f_{err(CL)}(T_t)$
B. Liu et al. (2021)	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(P)}\left(\lvert T_t - T_{sp,t} \rvert\right)$
X. Liu et al. (2022)	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CL)}(T_t)$
X. Liu & Gou (2024a)	$-\begin{cases} 1 & (working\ hours) \\ 0 & (otherwise) \end{cases} \cdot \left[E_t \begin{cases} w_1 & (window\ opened) \\ w_2 & (window\ closed) \end{cases} + w_3 f_{err(CL,P)}(T_t)\right]$
X. Liu et al. (2024b)	$-w_1 E_t - w_2 PPD$
Manjavacas et al. (2024)	$-w_1 E_t - w_2 f_{err(CL)}(T_t)$
Miao et al. (2024)	$w_1 \exp(-(\alpha E_t)^2) - w_2\left\{f_{err(CL)}(T_t) + \exp\left(-\beta\left(T_t - T_{sp}\right)^2\right)\right\}$
Naug et al. (2022)	$-w_1 E_t - w_2 f_{err(CL)}(T_t) - w_3 1(A_t \neq A_{t-1})$
Nguyen et al. (2024)	$-w_{1,t} E_t - w_{2,t} f_{err(CL)}(T_t)$

Paper	The reward at timestep t (r_t)
Qin et al. (2023)	$-w_1 E_t - w_2 \begin{cases} 50(T_t - T_{sp})^2 - 100 & (\lvert T_t - T_{sp}\rvert \le 1) \\ 6.25(T_t - T_{sp})^2 - 56.25 & (1 < \lvert T_t - T_{sp}\rvert \le 3) \\ 3.125(T_t - T_{sp})^2 - 28.125 & (3 < \lvert T_t - T_{sp}\rvert \le 5) \\ 500 & (5 < \lvert T_t - T_{sp}\rvert) \end{cases}$
Quang & Phuong (2024)	$-w_1 E_t - w_2 f_{err(CL)}(PMV_t)$
Razzano et al. (2025)	$-w_1 E_t - w_2 \begin{cases} f_{err(CL)}(T_t) & (T_t < T_{ALL}) \\ f_{err(CL,P)}(T_t) & (T_{AUL} < T_t) \end{cases} - w_3 f_{err(CL,P)}(T_{sply,t})$
Scarcello et al. (2023)	$-w_1 N_{max}(E_t) - w_2 N_{max}(f_{err}(T_t)) - w_3 1(T_t < T_{min} \vee T_{max} < T_t)$ $- w_4 \begin{cases} 1 & (occupant\ interacts\ with\ FCU) \\ 0 & (otherwise) \end{cases}$
Shi et al. (2024)	$-w_1 E_t - w_2 \begin{cases} \lvert T_t - T_{sp}\rvert^2 & (T_t < T_{ALL}) \\ 0 & (T_{ALL} \le T_t < T_{sp}) \\ \lvert T_t - T_{sp}\rvert & (T_{sp} \le T_t < T_{AUL}) \\ \lvert T_t - T_{sp}\rvert^3 & (T_{AUL} \le T_t) \end{cases}$
Shin et al. (2024)	$-w_1 E_t - w_2 f_{err(CL)}(T_t)$
Silvestri et al. (2024)	$-w_1 E_t - w_2 f_{err(OC,CL,P)}(T_t)$
Su et al. (2024)	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CAL)}(T_t)$
Sun et al. (2024)	$-w_1 \begin{cases} 2E_t & \left((S_{hvac,t-1} = off) \wedge (S_{hvac,t} = on)\right) \\ E_t & (otherwise) \end{cases}$ $- w_2 \begin{cases} -500 & (22 < T_t < 26) \\ -300 & \left((20 < T_t \le 22) \wedge (5 \le h < 9) \wedge (S_{hvac,t-1} = off)\right) \\ 500 & \left((20 < T_t \le 22) \wedge (5 \le h < 9) \wedge (S_{hvac,t-1} = on)\right) \\ 200 & (otherwise) \end{cases}$ $- w_3 \begin{cases} 1000 & (28 < T_t) \\ 0 & (otherwise) \end{cases}$
Touzani et al. (2021)	$-w_1 E_t \lambda_{E,t} - w_2 \left\{-\exp\left(-0.5(T_t - T_{sp})^2\right) + f_{err(CL)}(T_t)\right\}$ $- w_3 \begin{cases} 20 & \left((E_{Bat,t} < 0) \wedge (SOC_t < SOC_{min})\right) \\ 20 & \left((E_{Bat,t} > 0) \wedge (SOC_t > SOC_{max})\right) \\ 0 & (otherwise) \end{cases}$
H. Wang et al. (2024)	$-w_1 E_t - w_2 f_{err(CL,P)}(T_t) - w_3 f_{err(CL)}(\phi_t) - w_4 1(A_t \ne A_{t-1})$
M. Wang & Lin (2023)	$-w_1 \max(0, E_t - E_{min,t}) - w_2 f_{err(P)}(T_t)$
X. Wang et al. (2025)	$-w_{1,t} \begin{cases} 1 - N_{minmax}(E_t) & (T_{ALL} < T_t < T_{AUL}) \\ -N_{minmax}(E_t) & \left(T_{AUL} < T_t(winter) \vee T_t < T_{ALL}(summer)\right) \\ N_{minmax}(E_t) - 1 & \left(T_t < T_{ALL}(winter) \vee T_{AUL} < T_t(summer)\right) \end{cases}$ $- (1 - w_{1,t}) \begin{cases} 2 - \dfrac{2}{1 + \exp(-\lvert T_t - T_{SP,t}\rvert)} & (T_{ALL} < T_t < T_{AUL}) \\ 1 - \dfrac{2}{1 + \exp(-\lvert T_t - T_{SP,t}\rvert)} & (otherwise) \end{cases}$

Paper	The reward at timestep t (r_t)
Wei et al. (2021)	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CL)}(T_t)$
M. Xia et al. (2023)	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CL)}(T_t)$
Y. Xia et al. (2024)	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(OC,CL)}(T_t) - w_3 f_{err(OC,CL)}(c_{co2,t}) - w_4 f_{err(OC,CL)}(\phi_t)$
Xu (2022)	$-w_1 E_t - w_2 f_{err(CL)}(T_t)$
Xue et al. (2025)	$-w_1 E_t - w_2 \begin{cases} \lvert PMV_t \rvert & (PMV_{CLL} < PMV_t < PMV_{CUL}) \\ -\lvert PMV_t \rvert & (otherwise) \end{cases}$
L. Yu et al. (2020)	$-w_1 E_t \begin{cases} \lambda_{E,t} & (0 < E_t) \\ \lambda_{sell,E,t} & (E_t \le 0) \end{cases} - w_2 f_{err(CL)}(T_t) - w_3 \Delta E_{Bat,t}$
L. Yu et al. (2021)	$-w_1 E_t - 1_{NOC}\{w_2 f_{err(CL)}(T_t) + w_3 f_{err(CL)}(x_{co2,t})\}$
L. Yu et al. (2022)	$-w_1 E_t - w_2 f_{err(OC,CL)}(PMV_t) - w_3 f_{err(OC,CL)}(T_t)$
Yuan et al. (2021)	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CL,P)}(T_t)$
Zenginis et al. (2022)	$-w_1 E_t \begin{cases} \lambda_{E,t} & (0 < E_t) \\ \lambda_{sell,E,t} & (E_t \le 0) \end{cases} - w_2 f_{err(CL)}(T_t) - w_3 P_{SoC,t}$ $P_{SoC,t} = \begin{cases} 0 & (E_{Bat,min} < E_{Bat,t} < E_{Bat,max}) \\ \beta + (1-\beta) P_{SoC,t-1} & (otherwise) \end{cases}$
B. Zhang et al. (2022)	$-w_1 E_t - w_2 f_{err(CL,P)}(T_t) - w_3 \begin{cases} \exp\left(-\frac{E_{bat,t}}{E_{Bat,max}}\right) & (E_{Bat,t} < E_{Bat,min}) \\ \exp\left(\frac{E_{bat,t}}{E_{Bat,max}} - 1\right) & (E_{Bat,max} < E_{Bat,t}) \\ 0 & (otherwise) \end{cases} - w_4 f_{err(P)}(V_t)$
Zhao et al. (2021)	$-w_1 E_t \lambda_{E,t} - w_2 f_{err(CL)}(T_t) - w_3 f_{err(CL)}(A_t)$
Zhong et al. (2022)	$-w_1 N_{minmax}(E_t) - w_2 f_{err(OC,AL)}(PPD_t)$
Zhuang et al. (2023)	$-w_1 N_{minmax}(E_t) - w_2 f_{err(OC,CL,AL)}(PMV_t)$
Zou et al. (2020)	$-w_1 E_t - w_2 f_{err(OC,P)}(PPD_t)$

4. Discussion

4.1 Diversity of Reward Functions and Challenges in Research Comparability

まず、Table 3 に示された報酬関数は極めて多様だという点に注意すべきである。同じ式となったのはエネルギーと乾球温度に関わる項を単純に積和した 9 件（Gao, C. (2023)や Gao, Y. (2024)など）と、これに類似の 3 件（Liu, X. (2022); Wei, T. (2021); Xia, M. (2022)）のみで、その他はすべて異なる式だった。つまり、78 件の論文に対して 68 種の異なる報酬関数が設計されたということである。第 3 節で示したように、報酬関数の相互比較を容易にするために、変数記号を共通化したり、抽象化した誤差関数を導入したりしたにも関わらず、これだけの多様性があった。さらに、それぞれの式で使われているハイパーパラメータである重み係数 w_n も通常は異なる値を取るため、一見、同じ式形状でも厳密には同じではない。

報酬関数を自由に設定できるという点は強化学習の一つの強みだが、一方で、その自由が誘発する多様性は研究成果の相互比較を非常に難しくしているという問題を認識すべきである。例えば Heidari, A. (2025)、Qin, H. (2023)、Shi, Z. (2024)、Sun, L. (2024)などは、報酬関数の領域を細かく区切った上で、それぞれの領域で異なる非線形性を与えるために式を複雑化させているが、これによって得られる学習の改善効果が、その対価として支払われる研究の相互比較性の低下に優るものなのかは検証されるべきだろう。また、Liu, X. (2024a)、Qin, H. (2023)、Sun, L. (2024)の細かい条件分岐は Action の方法を示唆しているように解釈できる。そうであれば、これらは以下のよく知られた報酬関数の設計に関わる原則に反しており、reward hacking を誘発する危険性がある。

“*The reward signal is your way of communicating to the robot what you want it to achieve, not how you want it achieved*. (Sutton, R. S., and Barto, A. G. 2018)”

例えば Liu, X. (2024a)について言えば、窓の開閉状態で報酬は変えず、むしろ観測できる情報として窓の開閉状態を与えることで、エージェントに窓の状態と消費エネルギーの関係を学習させるべきだろう。今のままでは窓を閉めることによる短絡的報酬を得ることを学習してしまい、例えば外気温度が低いときに外気冷房をするというような制御は学べない。

もちろん、様々な研究で完全に同じ単純な形の報酬関数を使うということはできないとしても、他の研究との比較可能性を保証するためには、できるだけ典型的なパターンを採用するという努力が必要だろう。そうでなければ相互比較が極めて困難な、一品生産品に過ぎないケーススタディ研究が大量に生み出されるという危険性がある。本節の以下の discussion は、この「典型的なパターン」を探ることを大きな目的としている。

4.2 Integrating Energy Performance and Comfort: Current Approaches and Limitations

ほぼすべての事例が重み係数 w_n を使うことで積和としてエネルギー性能と快適性を統合した。多くの事例は、重み係数の具体的な値が示されていない事例（Xu, D.(2022); Shin, M.(2024)など）、数値は示されているが根拠は示されていない事例（Du, Y.(2021a); Gao, G.(2020); Kodama, N.(2021); Lin, X.(2023); Wei, T.(2021)など）、試行錯誤で決めたとする事例（Gao, C.(2023); Liu, X.(2022); Esrafilian-Najafabadi, M.(2022)など）、などに属しており、具体的な値の理論的根拠が示されているとは言い難い。Azimi, A.(2024); Coraci, D.(2021); Fang, X.(2022)など、いくつかの事例は感度分析あるいはグリッドサーチを根拠としているため、多目的最適化とパレートフロントの探索を狙った可能性がある。しかし、最終的には何らかの根拠にもとづいて 1 つの運用を選ばなければならない。なお、Manjavacas, A.(2024)、Dmitrewski, A.(2022)、Nguyen,

A. T.(2024) などでは、エネルギー性能と快適性を等しく重み付けることを目的にデフォルトで 0.5 という重みを採用した。しかし、そもそも両性能を示す単位が異なる以上、0.5 は、それがたとえ 1.0 の半分であったとしても、両性能をバランスよく評価することを全く保証しない点には注意が必要だろう。

以下に、異なる重み係数が如何に異なる最適化結果を生み出すのかを例示しておく。Figure 4 の散布図はエミュレータを使って仮想的に様々な建物運用を試した結果である（Togashi 2025）。横軸がエネルギー消費量、縦軸が熱環境に対する平均不満足者率である。一般に両者はトレードオフで、図に示されるように原点に凸の関係性を持つ。点線がパレートフロントで、それよりも右上が選択可能（Feasible）な領域である。最適点はパレートフロント上にあり、エネルギー性能と快適性の重み係数（w_E と w_D）の比によって描かれる線が接する点になる。これらは図では赤線で示される。重み係数の設定によって全く異なる点を最適点として選ぶことができ、もしも理論的根拠がなければ全く恣意的であることがわかろう。

特に深刻なのは、このパレートフロントの形状が事前には未知であるという点だ。もし選択した重み係数が、フロントの勾配が急な「膝」の部分にあれば、重みの多少の変化は結果に大きな影響を与えないかもしれない。しかし、もし勾配が緩やかな「平坦な」部分にあれば、重みの僅かな違いが最適点を劇的に変化させ、意図しない極端な制御を生むリスクをはらんでいる。

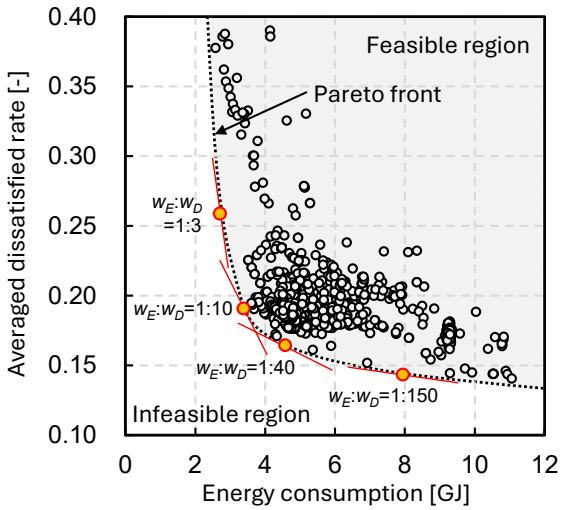

Figure 4 Illustration of How Weighting Factors Determine the Optimal Point on the Energy-Comfort Pareto Front

重み係数による線形和以外の事例はいくつかあり、1 つは Ahn, K. U. (2020)である。CO_2 濃度がしきい値を超えるか否かによってエネルギー消費のペナルティを非連続に変化させた。しかし、ペナルティである 1.5 倍という数値の根拠は示されておらず、依然としてエネルギー性能と快適性の統合は理論的に解けていない。もう 1 つの事例は Kannari, L. (2023) で、エネルギー性能と快適性を掛け合わせることで報酬を表現した。このような式にすれば、片方の性能変化がもう片方の性能の評価に影響を与えるようになる。従って、単純な線形和のように不適当な重み係数の設定によって片方ばかりの性能が重視されるという学習の失敗を抑制できる（Togashi et. al. (2020)）。しかし、線形性が失われるため、2 つの性能のどの組み合わせが、別のどの組み合わせと等しい価値を持つかという問題について理論的な根拠を持つことは、単純な線形和よりもさらに難しくなるという弱点もある。

いくつかの事例は、直接にエネルギー消費 E ではなく、その他の状態変数（部屋の温度やダンパの開度など）を使うことで実質的にエネルギー消費を表現するという工夫をした（Brandi, S. 2020; Dawood, S. M. 2022; Heidari, A. 2023; Li, W. 2023）。Heidari, A. (2023)の報酬関数は 1 タイプステップで室温がどれだけ変化したのか（T_t-T_{t-1}）でエネルギーを表現しており、これによってエネルギーの項と快適性の項が共に温度の単位（℃）で表現できている。従って、他の報酬関数のように単位の異なる状態量を加算しなくても良いという点では興味深いが、式を変形させると結局、前タイムステップの室温（T_{t-1}）と室温設定値（T_{sp}）を重み係数（w_1、w_2）で加重平均した温度が最適室温となり、問題は解決していない。

金銭換算は、エネルギー性能と快適性を統合するための 1 つの典型的な方法である。18 の事例がエネルギー消費 E にエネルギーコスト λ を乗じることで、エネルギー項の単位を$に変えた。ただし、その中のいくつかは太陽光発電などによる売買電の最適化を狙ったものだから、統合が目的ではなかろう（Touzani, S. 2021; Yu, L. 2020; Zenginis, I. 2022）。しかし、いくつかは快適性との統合を試みており、例えば Jiang, Z. (2021)、Yu, L. (2020)、Yu, L. (2021) はそれぞれ、重み係数の単位を$/°C、°C/$、°C/ppm とした。報酬関数の出力を単一の単位に揃えようとする意図が窺われる。特に Jiang, Z. (2021)は、0.1$/°C という具体的な値を示しており、この係数は、0.5°C の温度乖離が 1 日続いた場合に、通常の 1 日の電気料金とほぼ同等となるように定められた。しかし、その理論的な根拠は示されていない。

Fisk (2000)や Seppänen et al. (2006)など、知的生産性という観点から室内温熱環境を経済的に捉えようとする研究には歴史があり、この統合の問題に解決策を与えるかもしれない。1 単位の温度変化が執務者の作業効率に与える影響を定量的に捉えられれば、それは執務者の賃金を介して温度を費用に換算するための強い根拠となり得る。しかし、少なくとも我々が収集した強化学習関連の論文の中には、このような研究の知見を活かして演繹的に室内温熱環境を経済換算することで重み係数を求めたものは無かった。

以上をまとめると、ほぼすべての研究で重み係数による加重平均でエネルギー性能と快適性を統合しており、重み係数の設定には理論的な根拠は乏しい。そして我々はまだこれに代わる洗練された手法を生み出せていない、ということが現状であろう。

4.3 State Variables for Comfort Assessment: Selection and Implications

快適性を報酬関数に表現するために、まず、どのような快適指標を使うかという問題がある。なお、今回の文献検索は HVAC をキーワードに実施したため、おのずと温熱環境と空気質に関するものに焦点があたっている点には注意されたい。Table 3 にまとめた報酬関数で使われている快適指標を集計すると、多い順に、乾球温度が 57 件、PMV が 14 件、CO_2 濃度が 8 件、PPD が 6 件、相対湿度が 2 件、PM2.5 が 2 件、その他が 5 件だった。その他は 1 件ずつで、作用温度、PM10、給気温度、照度、APAQ (Gunnarsen and Fanger 1992)だった。1 つの報酬関数に複数の快適指標が使われる場合もあるため、合計数である 94 は論文の本数である 79 に一致しない点には注意されたい。上位 4 種類の快適指標の組み合わせを Figure 5 に示す。

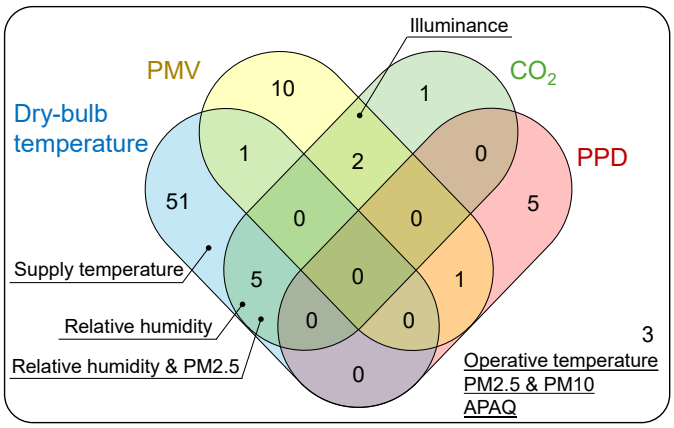

Figure 5 Classification of Comfort Indicators Used in Reward Functions. The diagram illustrates the usage frequency and co-occurrence of the four primary indicators (Dry-bulb temperature, PMV, PPD, and CO_2 concentration). Less frequent indicators, other than these four, are listed directly within their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

次に、快適性の表現のために単一の快適指標（例えば乾球温度のみ、あるいは PMV のみ、など）を使うか否かという問題がある。単一の快適指標を使ったものは 68 件（86%）で、2 つの快適指標を組み合わせたものが 8 件（10%）、3 つが 2 件、4 つが 1 件だった。

乾球温度のみを使って快適性を表現した事例は 50 件（63%）あり、最も典型的だった。一般に、空調設備で直接に制御でき、最も豊富にセンシングされている状態量は乾球温度だから、これを使って快適性を表現しようとすることは自然だろう。逆に最も複雑なものは、Guo, F. (2025)による報酬関数で、乾球温度、相対湿度、PM2.5、CO2 濃度が組み合わされていた。熱的快適性と室内空気質が同時に評価できるという利点はあるものの、エネルギー性能と快適性の統合と同じように、重み係数の決定が困難になるという問題も懸念される。

複数の快適指標を使って評価する場合には、重み係数を使わず、理論によって統合できる可能性がある。例えば乾球温度と放射温度が観測できた場合、両者が人体に及ぼす影響は対流熱伝達率と放射熱伝達率を使って表現できるはずだから、Esrafilian-Najafabadi, M. (2022)の報酬関数に示されるように単一の作用温度という状態量に統合できる。特に PMV や PPD などは、複数の熱的要素の非線形な関係性を理論式で統合しているため、重み係数による単純な線形和よりも熱的快適性を正しく評価できる可能性が高い。しかし Ding, Z.(2023) のように PMV と PPD を組み合わせても、最適点を指し示すという目的においては意味が無い。両者は共通の温熱 6 要素によって定まり、値は 1 対 1 に対応するため、情報が冗長である。もっとも、快適域から外れると PPD は PMV に比べて急速に値が大きくなるため、学習の初期において速度を高める報酬シェイピングとしての意義はあるかもしれない。

相当温度や PMV などの統合的指標を使うべきか否かの判断は、厳密には報酬関数の設計だけで決められず、どのような観測ができる環境なのかにも依存する。強化学習の対象は一般的には Markov Decision Process (MDP)に従う必要があり、可観測性が求められる。従って、もしもこれらの統合的指標を算出するための一部の要素が観測あるいは予測できず、不確実に変動するだけならば、学習が不安定になる。このため、例えば Zhuang, D.(2023)では PMV を報酬関数に使ったが、PMV を計算する際には、乾球温度と相対湿度のみを変数として扱い、その他の 4 つの要素（放射温度、風速、着衣量、代謝量）は、何らかの

推定値または固定値とした。乾球温度と相対湿度のみが観測可能だったためである。このように、報酬として統合的指標を使うためには観測できる状態量と整合させることに注意を払わなければならない。

快適指標を報酬設計に用いることには、reward hacking を誘発し得るという根本的な問題がある。最終目的は個々の執務者の快適性であり、温度や PMV といった指標はそれを知るための間接的な情報に過ぎないため、指標の調整を報酬とすることは「目的」ではなく「手段」の指定に他ならない。これまで個人の快適性を直接観測することは困難だったため、この「手段」の指定が現実的な妥協策として許容されてきた。しかし、本稿でレビューした Haifeng, L.(2024)、Li, W.(2024)、Lim, S. H.(2024)らのように、個人の温冷感を機械学習で予測し、より「目的」に近い報酬設計を目指す近年の試みは、この状況を大きく変えつつある。今後、ウェアラブルデバイスによる生理量計測など、個人の快適性を直接把握する技術がさらに発展すれば、代表的な快適指標を介した間接的な「手段」を報酬とする設計は、その正当性をいっそう失うことになると考えられる。

4.4 Structuring Comfort in Rewards: Common Techniques and Considerations

レビューした文献で使われた報酬関数について、前節で解説した 4 つの典型的な工夫（OC, CL, P, AL）の組み合わせの方法を Figure 6 にまとめた。単独で使う工夫としては CL が 30 件と多かったが、その他には特に数の多い組み合わせはなかった。それぞれの研究で試行錯誤的に報酬関数を設計しており、典型的な工夫を如何に組み合わせるかについては、現状では標準的な方法が確立していないと推察できる。

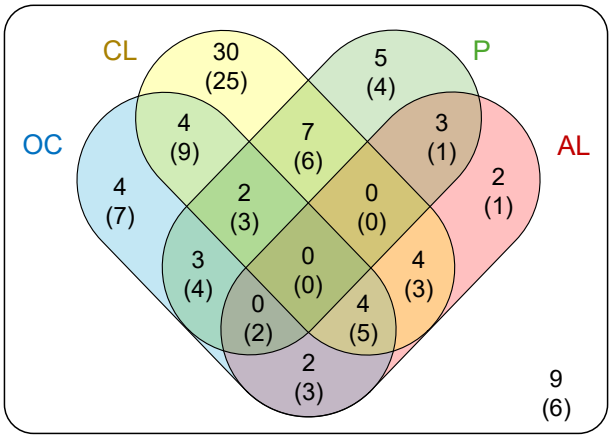

Figure 6 Venn diagram showing the adoption and combination of the four typical reward function design techniques (OC, CL, P, and AL) across the reviewed literature. The numbers indicate the count of papers corresponding to each intersection for explicitly defined techniques. The numbers in parentheses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

以降の議論の前提として、これらの 4 つの工夫が目的の違いに応じて 2 つにグルーピングできることを指摘すべきだろう。P と AL は選択可能範囲外の探索を素早く諦め、最適解にたどり着くまでの学習の速度を向上させることが主な目的である。Figure 6 において P と AL を組み合わせて使う事例がやや少ないことはこれが理由だろう。一方で、OC と CL は、最終的にたどり着くべき最適解を示すことが主な目的である。従って、P と AL に比較して OC と CL はより強く、我々の分野の専門知識が反映されるべき工夫である。

1) Incorporating Occupancy Information into Rewards

空調することの目的は、部屋に滞在する執務者に快適を感じさせることにある。従って、執務者が不在の場合に、室内が快適環境となっても直接的には意味がない。このため、空調設備は一般的にはスケジュールで On と Off や室温設定を変化させる。さらに高度な手法としては CO2 制御があり、執務者の在室状態を推定して換気量を変える。しかし報酬関数は、空調の発停状態や換気量を使って設計すべきではない。我々は現に滞在する執務者が快適を感じることに対して報酬を与えるべきであり、それがどのような設備の制御によって達成されるかはエージェントが学習すべきだからである。

執務者の在不在を報酬関数に明示的に表現した 19 件の事例の他にも、以下に例示するように、直接的にではなく間接的に表現する 3 種の方法もあった。

1 つ目として、時間帯別に異なる報酬関数を与えるという方法があった（Ding, Z. (2023)、Liu, X. (2024a)、Shi, Z. (2024)、Touzani, S. (2021)、Wang, M. (2023)など）。時間帯によって執務者の在室状況は予見できるのだから、それによって快適性に関わる報酬を予め調整しておけば良いという発想によるものだろう。

2 つ目として、執務者の在不在に応じて、快適性の範囲を変化させるという方法があった（Du, Y. (2021a)、Gao, C. (2023)、Gao, Y. (2024)など）。例えば Gao, C. (2023)の例では、快適性の範囲が在室時に 21-24 ℃、不在時に 15-30 ℃ とされた。

3 つ目として、TSV や PPD を使って快適性を評価し、その際に在室している執務者のみを集計するという方法があった（Lim, S. H. (2024)や Liu, X.(2024b)）。このような計算方法とすれば、不在時には快適性に関わる報酬は 0 となる。

上に例示した方法のように間接的に執務者の在不在を表現したと推測される事例は全部で 14 件あったため、直接的な方法と合計すると 33 件（42%）となり、全体の半数弱が何らかの方法で在不在情報を報酬関数に表現したことになる。

ただし、上記の 1 つ目と 2 つ目の方法は、エージェントに「目的」ではなく「手段」を示している可能性がある。

1 つ目の方法は、我々が時間帯別の執務者の在不在状態を予想し、それを報酬関数に組み込んでしまっているが、これはエージェントが学習すべきと考えられる。強化学習の「目的」は「現に滞在している」執務者の快適を向上させることなのだから、どれだけの人数が滞在しそうかという予見にもとづいて予め快適性の重みを調整するという「方法」を報酬関数に反映してはならない。エージェントには執務者の滞在人数を予想するための情報（例えば曜日や時刻）を与え、どれだけの人数が滞在するだろうかという問題も自ら学習させるべきである。

2 つ目の方法は、不在時にある程度の温度に維持しておけば、執務者が入手したときにも速やかに快適域まで温度を調整できるという考えにもとづいており、これは明らかに「方法」である。不在時に何度に維持すれば良いのかは、建物の熱容量や執務者の出入りの頻度にも影響を受ける。どの時刻に何度にすべきなのかという「方法」を学習するのはエージェントの役割であって、予め我々が報酬関数に反映すべきではない。執務者の在室時に適切な温度帯にするという目的が正しく伝われば、エージェントは不在時にも必要に応じて十分な温度を維持するという方法を学ぶだろう。

2) Defining Comfort Zones (Deadbands) and Their Significance

この工夫は、歴史ある制御方式であるデッドバンド制御（Paoluccio 1978）、よく知られたテストベッドである BOPTEST（Blum, D. 2021）に定義された Key Performance Indicator (KPI)、などとも整合しており、

もっとも採用数が多かった。

多くの論文では経験的に採用されたと予想されるが、ASHRAE 55 2017 (2017)のような基準が示す理論的枠組みを援用することで、より客観的な根拠を持たせることができる。同基準の根底にあるのは、居住者の大多数（80%）が許容できる温熱環境を定義するという実用的な目標である。具体的には、全身の快適性に対する予測不満足者率（PPD）を 10%未満に抑えるという定量的な目標を掲げ、これを達成する指標として PMV が-0.5 から+0.5 の範囲が導出されている。このように、「不満足者率」という具体的な人間中心の目標を最小化するという基準の考え方は、デッドバンドの境界値を設定する上での強力な理論的根拠となる。これは経験的なパラメータ設定からの脱却を促すだろう。

また、この工夫は強化学習の報酬関数の設計上も大きな利点がある。既に検討したように、我々の分野において報酬関数を設計する際の未解決の問題は、エネルギー性能と快適性をどのように統合するのかという点にある。従って、デッドバンドを設ければ、少なくともこの範囲については快適性の影響を無視できるため、両者をどのように統合するのかという問題が回避でき、エネルギー性能のみで報酬関数を設計すればよい。

また、一般的には、デッドバンド制御の一つの利点として、制御の安定性の向上が挙げられる。これは、あえて最適な一点を設けないことで、空調が過剰に発停を繰り返さなくなるという効果である。しかし、報酬関数の設計で快適性にデッドバンドを設けたとしても、この効果は期待できない点には注意すべきだろう。快適性と対になるエネルギー性能も一定にしないのであれば、依然として両者を統合した報酬関数には最適点が存在するためである。従って、安定性の確保のためには別の方法で報酬を用意すべきであり、例えば Dawood, S. M. (2022)、Naug, A. (2022)、Wang, H. (2024)はアクションの変更へのペナルティによってこれを実現しようとした。しかし Table 3 に挙げた中で、このような制御の安定化を目的とした項を持っている報酬関数は少なかった。おそらく以下の 2 つが理由だろう。第 1 に、シミュレーションを使った検討では現実の建物のように頻繁な操作によって機器が劣化するという現象は再現されず、問題が顕在化しない。第 2 に、そもそも発停を議論できるような分秒単位の動的設備シミュレーションではなく、まだ離散的な静的シミュレーションが使われている。従って、将来的には報酬関数にこのような制御の安定化を組み込むことも必要になるだろう。

3) Non-linear Error Transformation and Penalty Design for Exceeding Acceptable Limits

これらの 2 つの工夫は関連している。

まず、誤差の指数化は、最終的にたどり着くべき最適解を正しく示すことよりも、そこにたどり着くまでの学習の高速化や安定化を主な目的としている。そして、誤差の指数化によって学習が改善するか否かは報酬関数全体の形状や、適用した学習方法にも依存し、同じ工夫の導入が常に効果的とは限らない。従って、その設定が妥当だったか否かはプラグマティックに評価すべきであり、予め理論的に判断することは難しい。従って、学習の改善に効果的であるならば、Biemann, M. (2023)、Kadamala, K. (2024)、Miao, C.(2024) の報酬関数に例示されるように自然指数関数（exponential function）を使っても良い。

ただし、注意すべき点は、これらの工夫はこれまで議論してきた、エネルギー性能と快適性の統合という目的に対しては悪影響を持っているということである。このように議論する対象の非線形性を拡大させる変換を施すと、対象とする概念の説明を困難にするため、通常は（例えば 1 単位のエネルギー損失の価値は温度差の 2 乗と比例関係にある、というような理論が無い限りは）、エネルギー性能との統合をさらに難しくする。従って、エネルギー性能と快適性の統合を検討すべき領域においては、このような工夫

は避けるべきだろう。逆に、快適性の観点から決して最適値とはなり得ない範囲、つまり許容範囲を超える領域では問題がない。この領域においてはそもそも最適値が存在することが期待されておらず、むしろこの領域から速やかに許容範囲に戻ることが要請されているためである。

例えば Biemann, M. (2023)は報酬関数の第 2 項で誤差を指数化することで、学習改善を図った。しかし、これは室温が設定値に近づくことに強い報酬を与えることでもあるから、第 3 項において快適範囲内での快適性を同等と扱ったことと、態度が一貫しているとは言えない。Gupta, A. (2021) や Li, W. (2024) などもすべての領域で誤差が指数化されているため、学習速度は上がるだろうが、最適値の議論（エネルギー性能との統合）は難しくなる。逆に、学習の改善、エネルギー性能との統合、これら 2 つの狙いが両立できそうな事例として Lim, S. H. (2024) がある。Lim, S. H. (2024)は PMV の絶対値が 1.0、つまり許容限界（通常、PMV は±0.5 が快適範囲なので±1.0 は許容限界という位置づけだろう）を超えるまではそのまま報酬とし、このしきい値を超えると 2 乗を報酬とした。これは 1.0（許容限界）を超えた場合には速やかにその範囲から脱出するべきという意図だろう。Shi, Z. (2024) の報酬関数も段階的に誤差を拡大させており、同様の狙いがあると予想される。一方で、彼らの報酬関数では、条件分岐をした点で関数が不連続になるため、勾配を使う学習アルゴリズムが不安定になる危険性はある。

4.5 Proposal of a Typical Reward Function Structure Based on Literature Review

本節では上記で議論されたいくつかの潜在的な危険を回避できる、具体的な報酬関数の型を示す。ただし、その意図は共通に従うべき標準的（standard）な関数を提示することではない。そのような標準化は報酬関数を自由に設計できるという強化学習の大きな強みを奪うためである。一方で、既に見てきたように、現状の報酬関数はあまりに多様すぎて比較可能性が無いという問題を抱えている。従って、我々は報酬関数の設計において、解くべき問題に応じて特殊に作り込むことと、できるだけ典型的な型に倣うという、2 つの相反する要求のバランスを考えなければならない。このような場面で参照すべき典型的（typical）な型を示すことが本節の狙いである。

制御できる変数ベクトル（つまり action）を $\boldsymbol{A}$ とし、エネルギー消費量 E と快適指標 x はそれぞれ $\epsilon(\boldsymbol{A})$ と $\chi(\boldsymbol{A})$で表されるとする。x はスカラーとするが、容易にベクトルに拡張できる。しかし、上で議論したようにエネルギー性能との統合という目的においては、できるだけ少ない指標とすることが望ましいだろう。また、エネルギー消費量も現実には機器別に出力されるかも知れないが、報酬関数ではそれらを合算したスカラーで評価すべきである。

報酬関数 $r(\boldsymbol{A})$を快適性に関わる状態変数 $\chi(\boldsymbol{A})$の範囲に応じて 3 つの領域に分け、次式で表す（Figure 7）。

$$r(\boldsymbol{A}) = \begin{cases} -w_E\epsilon(\boldsymbol{A}) & (x_{CLL} < \chi(\boldsymbol{A}) < x_{CUL}) \\ -w_E\epsilon(\boldsymbol{A}) - w_C J_C\big(\chi(\boldsymbol{A})\big) & \big((x_{ALL} < \chi(\boldsymbol{A}) < x_{CLL}) \vee (x_{CUL} < \chi(\boldsymbol{A}) < x_{AUL})\big) \\ -J_{ALV}\big(\chi(\boldsymbol{A})\big) & \Big((\chi(\boldsymbol{A}) < x_{ALL}) \vee \big(x_{AUL} < \chi(\boldsymbol{A})\big)\Big) \end{cases} \quad \text{Eq. 9}$$

ここで $J_C(\cdot)$は快適指標のコスト関数、$J_{ALV}(\cdot)$は許容範囲外に快適指標が逸脱した場合のコスト関数、w_E と w_C はそれぞれエネルギー性能と快適性に関わる重み係数である。

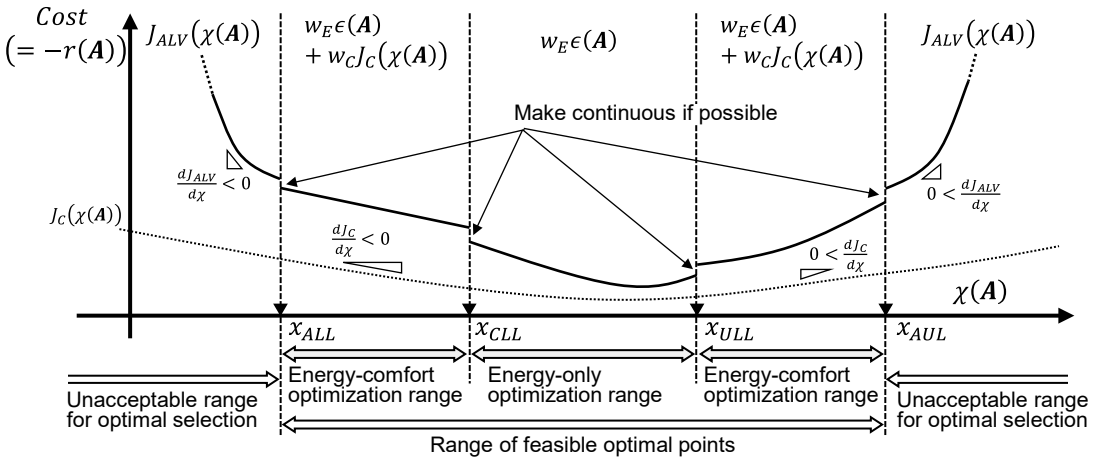

Figure 7 Conceptual diagram of the proposed typical piecewise reward function structure, illustrating how the optimization objective and reward function shape vary depending on the value of the comfort indicator $\chi(\boldsymbol{A})$.

上段に示されるように、快適指標が快適範囲内にある場合にはエネルギー性能のみで最適点を探索する。中段に示されるように、快適指標が許容範囲内にある場合にはエネルギー性能と快適性のバランスを評価して最適点を探索する。ここでは w_E と w_C を使って両者を統合したが、既に議論したように、この統合の方法は未だ確立していない。執務者の在不在は、この領域で報酬関数に組み込むべきである。下段に示されるように、快適指標が許容範囲を逸脱した場合には、$J_{ALV}(\cdot)$を使って速やかに許容範囲内に探索点を回復させる。この範囲では最適点の探索は目的ではなくなるため、エネルギー性能は報酬関数に含む必要がない。

$J_{ALV}(\cdot)$は許容限界から離れるとともに大きなペナルティを計上しなければならない。つまり、以下が条件となる。

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

$J_C(\cdot)$も一般的には快適限界から離れるとともにコストを大きくすべきだから、以下が条件となる。

$$\begin{cases} \dfrac{dJ_C(\chi)}{d\chi} < 0 \quad if \quad \chi(\boldsymbol{A}) < x_{CLL} \\ 0 < \dfrac{dJ_C(\chi)}{d\chi} \quad if \quad x_{CUL} < \chi(\boldsymbol{A}) \end{cases} \qquad \text{Eq. 11}$$

学習の安定化のためには報酬関数が連続的であることが望ましく、快適限界においてこれを満足するためには以下が条件となる。

$$J_C\big(\chi(\boldsymbol{A})\big) = 0 \quad if \quad (\chi(\boldsymbol{A}) = x_{CLL}) \vee (\chi(\boldsymbol{A}) = x_{CUL}) \qquad \text{Eq. 12}$$

許容限界においても同様に以下が条件となる。

$$w_E\epsilon(\boldsymbol{A}) + w_C J_C\big(\chi(\boldsymbol{A})\big) = J_{ALV}\big(\chi(\boldsymbol{A})\big) \quad if \quad (\chi(\boldsymbol{A}) = x_{ALL}) \vee (\chi(\boldsymbol{A}) = x_{AUL}) \qquad \text{Eq. 13}$$

しかし通常、$\epsilon(\boldsymbol{A})$はシミュレーションをしなければ求まらないため、予め上式が成り立つように $J_C(\cdot)$ と $J_{ALV}(\cdot)$の関係を設計することは困難である。従って現実には、許容限界点でペナルティがエネルギーコストと快適性コストの和を下回らないという、最低限の条件しか課せないだろう。これは具体的には次式で表される。

$$w_E\epsilon(\boldsymbol{A}) + w_C J_C\big(\chi(\boldsymbol{A})\big) \leq J_{ALV}\big(\chi(\boldsymbol{A})\big) \quad if \quad (\chi(\boldsymbol{A}) = x_{ALL}) \vee (\chi(\boldsymbol{A}) = x_{AUL}) \qquad \text{Eq. 14}$$

例えば定格のエネルギー消費 E_N [GJ]をエネルギー消費量の最大値と仮定すれば、この程度の条件を満たす報酬関数は容易く設計できる。

上記の区分は、あくまでエネルギー性能と快適性のバランスをどの領域で考慮すべきかという構造的な問題に解を与えるに過ぎない。中段の領域（許容範囲内）における重み係数 w_E と w_C の具体的な値をいかにして理論的に、あるいは実用的に決定するかという根源的な課題は依然として残されている。この問題の解決こそが、本分野における報酬関数設計の核心的な課題であり、今後の研究で重点的に取り組まれるべきだろう。

上記は概念上の設計指針に留まるため、この指針を満たす具体的な 1 つの例を以下に示す。報酬関数と重み係数をそれぞれ次式で定義する。

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

$$w_E = \lambda_{sell,E} \qquad \text{Eq. 16}$$

$$w_C = \frac{N_{oc} Sal_m}{WH_m} \qquad \text{Eq. 17}$$

ここで S_{alm} [USD/month]は執務者の月給（monthly salary）、WH_m [hours/month]は月の勤務時間数（monthly working hours）である。重み関数をこのように定義すれば、Eq. 15 の各段の第 1 項と第 2 項は同じ床面積あたりのコスト（USD/m^2）という単位に揃えられる。これは重要な点で、試行錯誤による恣意的な値ではなく、このような物理的に意味のある概念にすることができれば、他の研究との比較可能性が高まる。

快適性のコスト関数は次式で定義する。

$$J_C(PMV) = f_{PD}(PMV) \qquad \text{Eq. 18}$$

$$J_{ALV}(PMV) = f_{PD}(PMV^2) \qquad \text{Eq. 19}$$

ここで $f_{PD}(\cdot)$は PMV の値に応じた生産性の減少関数（Performance decrement function）である。ここでは Lan らの研究報告（Lan et al. 2011）を応用して次式で定義した。

$$f_{PD}(PMV) = 0.00135 + 0.000351PMV^3 + 0.005294PMV^2 + 0.00215PMV \qquad \text{Eq. 20}$$

Eq. 15 の上段は快適性のコスト関数が登場しているが、これは快適性を考慮することが目的ではなく、中段の式との連続性を保証することが目的である。同様に、下段のエネルギー性能のコスト関数は、エネルギーによるペナルティを課すことが目的ではなく、中断の領域よりも常に報酬が小さくなることを保証することが目的である。しかし acceptable range の境界で関数を滑らかに連続させることはできていないという限界はある。このような不連続点は、勾配ベースの最適化手法を用いる一部の強化学習アルゴリズムにおいて、学習の不安定性を引き起こす可能性がある。また、Eq. 19 右辺では全体を 2 乗せず、PMV のみを 2 乗している点は重要である。この領域では PMV は 1.0 を必ず上回るため、Eq. 15 の下段の

式で PMV の増加に対するコストの増加が加速することが保証される。もっとも、本例に限ればそもそも f_{PD} は PMV の増加に対して非線形に増えるため、この工夫は必須ではない。

以上、典型的な報酬関数の型を示したが、これは既往研究の多くの報酬関数が示したパターンの意図を読み取り、それらを論理的に連結させたものである。この報酬関数の型が真に有効であるかどうかは実際の問題を通して検証されるべきであり、将来の研究が必要であろう。

5. Conclusions

本研究では、空調制御における強化学習の応用において、報酬関数の設計が性能に与える影響の重要性に着目した。特に、快適性とエネルギー性能という複数の指標のトレードオフをどのようにバランスさせて報酬関数に統合しているかという観点から、2020 年以降に発表された 78 件の学術論文を詳細にレビューした。また、それらの報酬関数の比較を容易にするために、変数記号の共通化、誤差関数や標準化関数の導入といった抽象化・標準化の手法を提案した。

レビューを通じて、報酬関数の定義には極めて大きな多様性が存在し、これが研究成果の相互比較を著しく困難にしている現状が明らかになった。エネルギー性能と快適性の統合は、大半の研究で理論的根拠の乏しい重み係数による線形和に依存しており、両概念を統合するための洗練された手法は未だ確立されていない。快適性評価に用いられる状態量としては乾球温度が最も多く、PMV や PPD といった統合指標の利用は限定的であり、その活用には観測可能性との兼ね合いが課題となる。また、快適性の式形状に見られる執務者の在不在情報（OC）、快適限界（CL）、誤差の指数化（P）、許容限界（AL）といった典型的な工夫は、それぞれ異なる目的を持つものの、その適用方法によっては報酬設計の原則に反したり、エネルギーと快適性の統合をさらに困難にしたりする危険もあった。

レビューで得られた知見に基づき、快適範囲、許容範囲、許容範囲外の 3 つの領域で異なる評価を行う区分的な報酬関数の型を提示した 。これは、エネルギー効率の追求、快適性とのバランス、そして逸脱状態からの迅速な回復という、状況に応じた異なる要求を報酬関数に体系的に組み込む試みであり、今後の報酬関数設計における一つの指針となり得る。

本分野における今後の重要な課題は、エネルギーと快適性という異なる尺度をより理論的かつ定量的に統合する手法を確立することである。この点に関連して、個々の執務者の快適性を単純に合計したり平均化したりして全体の快適性指標とすることが、果たして適切なのかという根源的な問いも存在する。異なる個人の快適性要求をどのように集約し、あるいは個別に扱うべきかという問題は、空調制御における公平性や個人の尊重といった、より哲学的、倫理的な側面からの議論を必要とするかもしれない。そして情報分野ではこのような「公平性」をどのように強化学習に組み込むのかについて、既に検討が始まっている（S. Zhang et al 2024）。

この問題に取り組むためにはそもそも快適性に関わる個人の嗜好を知る必要があり、本文で例示したように、いくつかの研究では執務者ごとの温冷感を機械学習で特定するという試みが始まっている。執務者は入退室し、また、それぞれの温冷感は確率的だから、報酬関数は時間的に大きく変動する非常にダイナミックな特徴を持つだろう。このような場面でも実時間に遅れずに学習を成功させるためには Adaptive reward shaping（Chahoud et al. 2025）が技術的な解決策になるかもしれない。

一方で、上記のようなエネルギー性能と快適性の統合は短期的には達成できまい。そうであれば当面は Multi-objective optimization の技術を進化させるという選択も有効だろう。この手法でパレートフロントを明らかにすることができれば、将来的にエネルギー性能と快適性の統合式を組み合わせることで運転点を唯一に特定できるようになる。少なくとも現状のように根拠の不明瞭な重み係数を使うことでパレートフロントの中の 1 点のみを最適点の候補として指し示すよりは価値があるだろう。しかし、本文でも指摘したように、パレートフロントの探索はそれ自体では具体的な運転点の特定には至らないということは正確に認識しなくてはならない。やはり我々が最も必要としているものは両性能指標の統合である。

また、研究間の比較可能性を高めるため、報酬関数の設計パターンに関する共通認識の形成や、ベンチマークとなるような標準的な報酬関数の枠組みを整備していく必要があろう。さらに、制御の安定性や実システムへの適用性といった、本レビューでは限定的な言及に留まった要素も、今後は報酬関数設計においてより深く考慮されるべきである。

実システムへの適用に関しては、本研究では十分に扱うことができなかった重要な課題も残されている。例えば、今回収集された研究のほぼすべてが各タイムステップで与えられる即時報酬を前提としていたが、実環境では目標達成時にのみ与えられるような疎な遅延報酬しか得られないケースも少なくない。エネルギー性能では、建物の熱的遅れの影響や蓄熱・蓄電システムのようにエネルギーがシフトする設備がある場合が該当する。また、快適性についても、本文で例示したようなウェアラブルデバイスによる即時の生理量把握が可能になるまでにはかなりの時間が必要なはずで、それまでは離散的なアンケート調査が精々だろう。このような状況下での効果的な報酬関数設計は、今後の研究が待たれる領域である。

さらに、本研究では報酬関数の構造自体に焦点を当てたが、本来、最適な報酬関数はエージェントが取り得る行動の選択肢（行動空間）や、観測可能な情報（観察空間）の設計と密接に関連し、相互に影響し合う。これらの関係性まで踏み込んだ包括的な設計論の構築は、今後の大きな挑戦と言える。これらの課題の存在は、報酬関数に関する議論が依然として発展途上であり、今後も継続的かつ多角的に深められなければならないことを強く示唆している。

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]
$f_{PD}(\cdot)$	: performance-decrement function	[-]
GTS	: Group thermal sensation	[-]
$J_{ALV}(\cdot)$	: Cost function for comfort when outside acceptable limits (Acceptable Limit Violation)	[-]
$J_C(\cdot)$	: Cost function for comfort within acceptable limits	[-]
$N(\cdot)$	: 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	[-]
$\epsilon(\cdot)$	: Function determining energy consumption	[-]
λ_E	: Energy purchase price	[USD/GJ]
$\lambda_{sell,E}$	: Energy selling price	[USD/GJ]
ϕ	: Relative humidity	[%]
$\chi(\cdot)$	: Function determining the comfort indicator	[-]

Subscript:

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

[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.

References:

1）Ahn, K. U., & Park, C. S. (2020). Application of deep Q-networks for model-free optimal control balancing between different HVAC systems. Science and Technology for the Built Environment, 26(1), 61-74. https://doi.org/10.1080/23744731.2019.1680234

2）Ajifowowe, I., Chang, H., Lee, C. S., & Chang, S. (2024). Prospects and challenges of reinforcement learning-based HVAC control. Journal of Building Engineering, 98, 111080. https://doi.org/10.1016/j.jobe.2024.111080

3）Al Mindeel, T., Spentzou, E., & Eftekhari, M. (2024). Energy, thermal comfort, and indoor air quality: Multi-objective optimization review. Renewable and Sustainable Energy Reviews, 202, 114682. https://doi.org/10.1016/j.rser.2024.114682

4）Al Sayed, K., Boodi, A., Sadeghian Broujeny, R., & Beddiar, K. (2024). Reinforcement learning for HVAC control in intelligent buildings: A technical and conceptual review. Journal of Building Engineering, 95, 110085. https://doi.org/10.1016/j.jobe.2024.110085

5）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

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

7）American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE). (2017). *ANSI/ASHRAE Standard 55-2017: Thermal environmental conditions for human occupancy*.

8）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

9）Andrés, E., Cuéllar, M. P., & Navarro, G. (2022). On the Use of Quantum Reinforcement Learning in Energy-Efficiency Scenarios. Energies, 15(16), 6034.

10）Azimi, A., & Akbari, O. (2024). 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, Article 100700. https://doi.org/10.1016/j.prime.2024.100700

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

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

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

14） 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

15）Brandi, S., Fiorentini, M., & Capozzoli, A. (2022). Comparison of online and offline deep reinforcement learning with model predictive control for thermal energy management. Automation in Construction, 135, 104128. https://doi.org/10.1016/j.autcon.2022.104128

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

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

18）Campoy-Nieves, A., Manjavacas, A., Jiménez-Raboso, J., Molina-Solana, M., & Gómez-Romero, J. (2025). Sinergym – A virtual testbed for building energy optimization with Reinforcement Learning. Energy and Buildings, 327, 115075. https://doi.org/10.1016/j.enbuild.2024.115075

19）Chahoud, M., Sami, H., Mizouni, R., Bentahar, J., Mourad, A., Otrok, H., & Talhi, C. (2025). Reward shaping in DRL: A novel framework for adaptive resource management in dynamic environments. Information Sciences, 715, 122238. https://doi.org/10.1016/j.ins.2025.122238

20）Chatterjee, A., & Khovalyg, D. (2023). Dynamic indoor thermal environment using Reinforcement Learning-based controls: Opportunities and challenges. Building and Environment, 244, 110766. https://doi.org/10.1016/j.buildenv.2023.110766

21）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.

22）Chen, L., Meng, F., & Zhang, Y. (2023). Fast Human-in-The-Loop Control for HVAC Systems via Meta-Learning and Model-Based Offline Reinforcement Learning. IEEE Transactions on Sustainable Computing, 8(3), 504-521. https://doi.org/10.1109/TSUSC.2023.3251302

23）Chen, Z., Yu, L., Zhang, S., Hu, S., & Shen, C. (2024). Multiagent Hierarchical Deep Reinforcement Learning for Operation Optimization of Grid-Interactive Efficient Commercial Buildings. IEEE Transactions on Artificial Intelligence, 5(8), 4280-4292. https://doi.org/10.1109/TAI.2024.3366869

24）Coraci, D., Brandi, S., Piscitelli, M. S., & Capozzoli, A. (2021). Online implementation of a soft actor-critic agent to enhance indoor temperature control and energy efficiency in buildings. Energies, 14(4), Article 997. https://doi.org/10.3390/en14040997

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

26）Dawood, S. M., Hatami, A., & Homod, R. Z. (2022). Trade-off decisions in a novel deep reinforcement learning for energy savings in HVAC systems. Journal of Building Performance Simulation, 15(6), 809-831. https://doi.org/10.1080/19401493.2022.2099465

27）Deng, X., Zhang, Y., & Qi, H. (2022). Toward smart multizone HVAC control by combining context-aware system and deep reinforcement learning. IEEE Internet of Things Journal, 9(21), 21010-21024.

https://doi.org/10.1109/JIOT.2022.3175728

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

29）Devasenan, M., & Madhavan, S. (2024). Thermal intelligence: exploring AI's role in optimizing thermal systems – a review. Interactions, 245(1), 282. https://doi.org/10.1007/s10751-024-02122-6

30）Devidze, R. (2025). Reward Design for Reinforcement Learning Agents. ArXiv, abs/2503.21949.

31）Ding, X., Cerpa, A., & Du, W. (2024a). Exploring Deep Reinforcement Learning for Holistic Smart Building Control. ACM Transactions on Sensor Networks, 20(3), Article 70. https://doi.org/10.1145/3656043

32）Ding, X., Cerpa, A., & Du, W. (2024b). Multi-Zone HVAC Control With Model-Based Deep Reinforcement Learning. IEEE Transactions on Automation Science and Engineering, 1-19. https://doi.org/10.1109/TASE.2024.3410951

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

34）Ding, Z.-K., Fu, Q.-M., Chen, J.-P., Wu, H.-J., Lu, Y., & Hu, F.-Y. (2022). Energy-efficient control of thermal comfort in multi-zone residential HVAC via reinforcement learning. Connection Science, 34(1), 2364-2394. https://doi.org/10.1080/09540091.2022.2120598

35）Dinh, H. T., & Kim, D. (2022). MILP-Based Imitation Learning for HVAC Control. IEEE Internet of Things Journal, 9(8), 6107-6120. https://doi.org/10.1109/JIOT.2021.3111454

36）Dmitrewski, A., Molina-Solana, M., & Arcucci, R. (2022). CNTRLDA: A building energy management control system with real-time adjustments. Application to indoor temperature. Building and Environment, 215, Article 108938. https://doi.org/10.1016/j.buildenv.2022.108938

37）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

38）Esrafilian-Najafabadi, M., & Haghighat, F. (2022). Towards self-learning control of HVAC systems with the consideration of dynamic occupancy patterns: Application of model-free deep reinforcement learning. Building and Environment, 226, Article 109747. https://doi.org/10.1016/j.buildenv.2022.109747

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

40）Fan, Y., Fu, Q., Chen, J., Wang, Y., Lu, Y., & Liu, K. (2025). A deep reinforcement learning control method for multi-zone precooling in commercial buildings. Applied Thermal Engineering, 260, 124987. https://doi.org/10.1016/j.applthermaleng.2024.124987

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

42）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

43）Fisk, W. J. (2000). 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.

44）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

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

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

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

48）Gao, G., Li, J., & Wen, Y. (2020). DeepComfort: Energy-Efficient Thermal Comfort Control in Buildings Via Reinforcement Learning. IEEE Internet of Things Journal, 7(9), 8472-8484. https://doi.org/10.1109/JIOT.2020.2992117

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

50）Gunnarsen, L., & Ole Fanger, P. (1992). Adaptation to indoor air pollution. Environment International, 18(1), 43-54. https://doi.org/10.1016/0160-4120(92)90209-M

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

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

53）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

54）Han, M., Zhao, J., Zhang, X., Shen, J., & Li, Y. (2021). The reinforcement learning method for occupant behavior in building control: A review. Energy and Built Environment, 2(2), 137-148. https://doi.org/10.1016/j.enbenv.2020.08.005

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

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

57）Heidari, A., & Khovalyg, D. (2023). DeepValve: Development and experimental testing of a Reinforcement Learning control framework for occupant-centric heating in offices. Engineering Applications of Artificial Intelligence, 123, Article 106310. https://doi.org/10.1016/j.engappai.2023.106310

58）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

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

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

61） Kannari, L., Kantorovitch, J., Piira, K., & Piippo, J. (2023). Energy Cost Driven Heating Control with Reinforcement Learning. Buildings, 13(2), Article 427. https://doi.org/10.3390/buildings13020427

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

63）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

64）Kwon, K. B., Park, J. Y., Hong, S. M., Heo, J. H., & Jung, H. (2024). Development of machine learning-based energy management agent to control fine dust concentration in railway stations. Journal of Electrical Engineering and Technology, 19(4), 2757-2766. https://doi.org/10.1007/s42835-023-01730-6

65）Lan, H, Huiying, H, & Zhonghua, G., & Gou, Z. (2024). User-centric approach to optimizing thermal comfort in university classrooms: Utilizing computer vision and Q-XGBoost reinforcement learning. Energy and Buildings, 323, Article 114808. https://doi.org/10.1016/j.enbuild.2024.114808

66）Lan, L., Wargocki, P., & Lian, Z. (2011). Quantitative measurement of productivity loss due to thermal discomfort. Energy and Buildings, 43(5), 1057-1062. https://doi.org/10.1016/j.enbuild.2010.09.001

67）Lei, Y., Zhan, S., Ono, E., Peng, Y., Zhang, Z., Hasama, T., & Chong, A. (2022). A practical deep reinforcement learning framework for multivariate occupant-centric control in buildings. Applied Energy, 324, Article 119742. https://doi.org/10.1016/j.apenergy.2022.119742

68）Li, R., & Zou, Z. (2025). How far back shall we peer? Optimal air handling unit control leveraging extensive past observations. Building and Environment, 269, Article 112347.

https://doi.org/10.1016/j.buildenv.2024.112347

69）Li, W., Wu, H., Zhao, Y., Jiang, C., & Zhang, J. (2024). Study on indoor temperature optimal control of air-conditioning based on Twin Delayed Deep Deterministic policy gradient algorithm. Energy and Buildings, 317, Article 114420. https://doi.org/10.1016/j.enbuild.2024.114420

70）Li, W., Zhao, Y., Zhang, J., Jiang, C., Chen, S., Lin, L., & Wang, Y. (2023). Indoor temperature preference setting control method for thermal comfort and energy saving based on reinforcement learning. Journal of Building Engineering, 73, Article 106805. https://doi.org/10.1016/j.jobe.2023.106805

71）Li, Z., Sun, Z., Meng, Q., Wang, Y., & Li, Y. (2022). Reinforcement learning of room temperature set-point of thermal storage air-conditioning system with demand response. Energy and Buildings, 259, Article 111903. https://doi.org/10.1016/j.enbuild.2022.111903

72）Lim, S. H., Kim, T. G., Yeom, D. J., & Yoon, S. G. (2024). Robust deep reinforcement learning for personalized HVAC system. Energy and Buildings, 319, Article 114551. https://doi.org/10.1016/j.enbuild.2024.114551

73）Lin, X., Yuan, D., & Li, X. (2023). Reinforcement Learning with Dual Safety Policies for Energy Savings in Building Energy Systems. Buildings, 13(3), Article 580. https://doi.org/10.3390/buildings13030580

74）Liu, B., Akcakaya, M., & McDermott, T. E. (2021). Automated Control of Transactive HVACs in Energy Distribution Systems. IEEE Transactions on Smart Grid, 12(3), 2462-2471. https://doi.org/10.1109/TSG.2020.3042498

75）Liu, X., & Gou, Z. (2024a). Occupant-centric HVAC and window control: A reinforcement learning model for enhancing indoor thermal comfort and energy efficiency. Building and Environment, 250, Article 111197. https://doi.org/10.1016/j.buildenv.2024.111197

76）Liu, X., Wu, Y., & Wu, H. (2024b). Enhancing HVAC energy management through multi-zone occupant-centric approach: A multi-agent deep reinforcement learning solution. Energy and Buildings, 303, Article 113770. https://doi.org/10.1016/j.enbuild.2023.113770

77）Liu, X., Ren, M., Yang, Z., Yan, G., Guo, Y., Cheng, L., & Wu, C. (2022). A multi-step predictive deep reinforcement learning algorithm for HVAC control systems in smart buildings. Energy, 259, Article 124857. https://doi.org/10.1016/j.energy.2022.124857

78）Manjavacas, A., Campoy-Nieves, A., Jiménez-Raboso, J., Molina-Solana, M., & Gómez-Romero, J. (2024). An experimental evaluation of deep reinforcement learning algorithms for HVAC control. Artificial Intelligence Review, 57(7), Article 173. https://doi.org/10.1007/s10462-024-10819-x

79）Marzullo, T., Dey, S., Long, N., Leiva Vilaplana, J., & Henze, G. (2022). A high-fidelity building performance simulation test bed for the development and evaluation of advanced controls. Journal of Building Performance Simulation, 15(3), 379-397. https://doi.org/10.1080/19401493.2022.2058091

80）Masdoua, Y., Boukhnifer, M., & Adjallah, K. H. (2024). Active fault-tolerant control based on DDQN architecture applied to HVAC system. Transactions of the Institute of Measurement and Control, 01423312241273767. https://doi.org/10.1177/01423312241273767

81）Miao, C., Cui, Y., Li, H., & Wu, X. (2024). Efficient multi-agent reinforcement learning HVAC power consumption optimization. Energy Reports, 12, 5420-5431. https://doi.org/10.1016/j.egyr.2024.11.011

82）Naug, A., Quinones-Grueiro, M., & Biswas, G. (2022). Deep reinforcement learning control for non-stationary

building energy management. Energy and Buildings, 277, Article 112584. https://doi.org/10.1016/j.enbuild.2022.112584

83）Ng, A., Harada, D., & Russell, S. J. (1999). Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 278-287

84）Nguyen, A. T., Pham, D. H., Oo, B. L., Santamouris, M., Ahn, Y., & Lim, B. T. H. (2024). Modelling building HVAC control strategies using a deep reinforcement learning approach. Energy and Buildings, 310, Article 114065. https://doi.org/10.1016/j.enbuild.2024.114065

85）Paoluccio, J. P. (1978). Dead band controls guide (CR 79.002).

86）Park, H.-A., Byeon, G., Son, W., Kim, J., & Kim, S. (2023). Data-Driven Modeling of HVAC Systems for Operation of Virtual Power Plants Using a Digital Twin. Energies, 16(20), 7032.

87）Qin, H., Yu, Z., Li, T., Liu, X., & Li, L. (2023). Energy-efficient heating control for nearly zero energy residential buildings with deep reinforcement learning. Energy, 264, Article 126209. https://doi.org/10.1016/j.energy.2022.126209

88）Quang, T. V., & Phuong, N. L. (2024). Using Deep Learning to Optimize HVAC Systems in Residential Buildings. Journal of Green Building, 19(1), 29-50. https://doi.org/10.3992/jgb.19.1.29

89）Razzano, G., Brandi, S., Piscitelli, M. S., & Capozzoli, A. (2025). 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. Applied Energy, 381, Article 125046. https://doi.org/10.1016/j.apenergy.2024.125046

90）Scarcello, L., Cicirelli, F., Guerrieri, A., Mastroianni, C., Spezzano, G., & Vinci, A. (2023). Pursuing Energy Saving and Thermal Comfort With a Human-Driven DRL Approach. IEEE Transactions on Human-Machine Systems, 53(4), 707-719. https://doi.org/10.1109/THMS.2022.3216365

91）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,

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

93）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

94）Shin, M., Kim, S., Kim, Y., Song, A., & Kim, H. Y. (2024). Development of an HVAC system control method using weather forecasting data with deep reinforcement learning algorithms. Building and Environment, 248, Article 111069. https://doi.org/10.1016/j.buildenv.2023.111069

95）Sierla, S., Ihasalo, H., & Vyatkin, V. (2022). A review of reinforcement learning applications to control of heating, ventilation and air conditioning systems. Energies, 15(10), 3526.

96）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

97）Su, Y., Zou, X., Tan, M., Peng, H., & Chen, J. (2024). Integrating few-shot personalized thermal comfort model and reinforcement learning for HVAC demand response optimization. Journal of Building Engineering, 91, Article 109509. https://doi.org/10.1016/j.jobe.2024.109509

98）Sun, L., Hu, Z., Mae, M., & Imaizumi, T. (2024). Individual room air-conditioning control in high-insulation residential building during winter: A deep reinforcement learning-based control model for reducing energy consumption. Energy and Buildings, 323, Article 114799. https://doi.org/10.1016/j.enbuild.2024.114799

99）Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.

100） Togashi, E., Miyata, M., & Yamamoto, Y. (2020). The first world championship in cybernetic building optimization. Journal of Building Performance Simulation, 13(3), 391-408. https://doi.org/10.1080/19401493.2020.1741685

101）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

102）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

103）Wang, H., Chen, X., Vital, N., Duffy, E., & Razi, A. (2024). Energy optimization for HVAC systems in multi-VAV open offices: A deep reinforcement learning approach. Applied Energy, 356, Article 122354. https://doi.org/10.1016/j.apenergy.2023.122354

104）Wang, M., & Lin, B. (2023). MF^2: Model-free reinforcement learning for modeling-free building HVAC control with data-driven environment construction in a residential building. Building and Environment, 244, Article 110816. https://doi.org/10.1016/j.buildenv.2023.110816

105）Wang, X., Mahdavi, N., Sethuvenkatraman, S., & West, S. (2025). An environment-adaptive SAC-based HVAC control of single-zone residential and office buildings. Data-Centric Engineering, 6, Article e3. https://doi.org/10.1017/dce.2024.57

106）Wang, Z., & Hong, T. (2020). Reinforcement learning for building controls: The opportunities and challenges. Applied Energy, 269, 115036. https://doi.org/10.1016/j.apenergy.2020.115036

107）Wei, T., Ren, S., & Zhu, Q. (2021). Deep Reinforcement Learning for Joint Datacenter and HVAC Load Control in Distributed Mixed-Use Buildings. IEEE Transactions on Sustainable Computing, 6(3), 370-384. https://doi.org/10.1109/TSUSC.2019.2910533

108）Xia, M., Chen, F., Chen, Q., Liu, S., Song, Y., & Wang, T. (2023). Optimal Scheduling of Residential Heating, Ventilation and Air Conditioning Based on Deep Reinforcement Learning. Journal of Modern Power Systems and Clean Energy, 11(5), 1596-1605. https://doi.org/10.35833/MPCE.2022.000249

109）Xia, Y., Wang, X., Yin, X., Bo, W., Wang, L., Li, S., & Li, K. (2024). Federated Accelerated Deep Reinforcement Learning for Multi-Zone HVAC Control in Commercial Buildings. IEEE Transactions on Smart Grid. https://doi.org/10.1109/TSG.2024.3524756

110）Xin, X., Zhang, Z., Zhou, Y., Liu, Y., Wang, D., & Nan, S. (2024). A comprehensive review of predictive control strategies in heating, ventilation, and air-conditioning (HVAC): Model-free vs. model. Journal of Building Engineering, 94, 110013. https://doi.org/10.1016/j.jobe.2024.110013

111）Xu, D. (2022). Learning Efficient Dynamic Controller for HVAC System. Mobile Information Systems, 2022, Article 4157511. https://doi.org/10.1155/2022/4157511

112）Xue, W., Jia, N., & Zhao, M. (2025). Multi-agent deep reinforcement learning based HVAC control for multi-zone buildings considering zone-energy-allocation optimization. Energy and Buildings, 329, Article 115241. https://doi.org/10.1016/j.enbuild.2024.115241

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

114）Yang, J., Yu, J., & Wang, S. (2024). Heating ventilation air-conditioner system for multi-regional commercial buildings based on deep reinforcement learning. Advanced Control for Applications, 6(4), e190. https://doi.org/10.1002/adc2.190

115）Yu, H., Tam, V. W. Y., & Xu, X. (2024). A systematic review of reinforcement learning application in building energy-related occupant behavior simulation. Energy and Buildings, 312, 114189. https://doi.org/10.1016/j.enbuild.2024.114189

116）Yu, L., Sun, Y., Xu, Z., Shen, C., Yue, D., Jiang, T., & Guan, X. (2021). Multi-Agent Deep Reinforcement Learning for HVAC Control in Commercial Buildings. IEEE Transactions on Smart Grid, 12(1), 407-419. https://doi.org/10.1109/TSG.2020.3011739

117）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

118）Yu, L., Xu, Z., Zhang, T., Guan, X., & Yue, D. (2022). Energy-efficient personalized thermal comfort control in office buildings based on multi-agent deep reinforcement learning. Building and Environment, 223, Article 109458. https://doi.org/10.1016/j.buildenv.2022.109458

119）Yuan, X., Pan, Y., Yang, J., Wang, W., & Huang, Z. (2021). Study on the application of reinforcement learning in the operation optimization of HVAC system. Building Simulation, 14(1), 75-87. https://doi.org/10.1007/s12273-020-0602-9

120）Zenginis, I., Vardakas, J., Koltsaklis, N. E., & Verikoukis, C. (2022). Smart Home's Energy Management Through a Clustering-Based Reinforcement Learning Approach. IEEE Internet of Things Journal, 9(17), 16363-16371. https://doi.org/10.1109/JIOT.2022.3152586

121）Zhang, B., Hu, W., Ghias, A. M. Y. M., Xu, X., & Chen, Z. (2022). Multi-agent deep reinforcement learning-based coordination control for grid-aware multi-buildings. Applied Energy, 328, Article 120215. https://doi.org/10.1016/j.apenergy.2022.120215

122）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.

123）Zhao, H., Zhao, J., Shu, T., & Pan, Z. (2021). Hybrid-Model-Based Deep Reinforcement Learning for Heating, Ventilation, and Air-Conditioning Control. Frontiers in Energy Research, 8, Article 610518. https://doi.org/10.3389/fenrg.2020.610518

124）Zhong, X., Zhang, Z., Zhang, R., & Zhang, C. (2022). End-to-End Deep Reinforcement Learning Control for HVAC Systems in Office Buildings. Designs, 6(3), Article 52. https://doi.org/10.3390/designs6030052

125）Zhou, S. L., Shah, A. A., Leung, P. K., Zhu, X., & Liao, Q. (2023). A comprehensive review of the applications of machine learning for HVAC. DeCarbon, 2, 100023. https://doi.org/10.1016/j.decarb.2023.100023

126）Zhuang, D., Gan, V. J. L., Duygu Tekler, Z., Chong, A., Tian, S., & Shi, X. (2023). Data-driven predictive control for smart HVAC system in IoT-integrated buildings with time-series forecasting and reinforcement learning. Applied Energy, 338, Article 120936. https://doi.org/10.1016/j.apenergy.2023.120936

127）Zou, Z., Yu, X., & Ergan, S. (2020). Towards optimal control of air handling units using deep reinforcement learning and recurrent neural network. Building and Environment, 168, Article 106535. https://doi.org/10.1016/j.buildenv.2019.106535