

# Scalable multi-agent reinforcement learning for distributed control of residential energy flexibility

Flora Charbonnier<sup>a,\*</sup>, Thomas Morstyn<sup>b</sup>, Malcolm D. McCulloch<sup>a</sup>

<sup>a</sup>*Department of Engineering Science, University of Oxford, UK*

<sup>b</sup>*School of Engineering, University of Edinburgh, UK*

---

## Abstract

This paper proposes a novel scalable type of multi-agent reinforcement learning-based coordination for distributed residential energy. Cooperating agents learn to control the flexibility offered by electric vehicles, space heating and flexible loads in a partially observable stochastic environment. In the standard independent Q-learning approach, the coordination performance of agents under partial observability drops at scale in stochastic environments. Here, the novel combination of learning from off-line convex optimisations on historical data and isolating marginal contributions to total rewards in reward signals increases stability and performance at scale. Using fixed-size Q-tables, prosumers are able to assess their marginal impact on total system objectives without sharing personal data either with each other or with a central coordinator. Case studies are used to assess the fitness of different combinations of exploration sources, reward definitions, and multi-agent learning frameworks. It is demonstrated that the proposed strategies create value at individual and system levels thanks to reductions in the costs of energy imports, losses, distribution network congestion, battery depreciation and greenhouse gas emissions.

**Keywords:** Energy management system, multi-agent reinforcement learning, demand-side response, peer-to-peer, prosumer, smart grid.

---

## Highlights

- Privacy-preserving multi-agent reinforcement learning is used to coordinate residential energy
- Learning from optimisations improves coordination scalability in stochastic environments
- Marginal reward signals further enhance cooperation relative to previous approaches
- The curse of dimensionality is mitigated by the use of fixed-size Q-tables
- Case studies with large real-life datasets yield 33.7% local and global cost reductions

## 1. Introduction

This paper addresses the scalability issue of distributed domestic energy flexibility coordination in a cost-efficient and privacy-preserving manner. A novel class of coordination strategies using optimisation-based multi-agent reinforcement learning (MARL<sup>1</sup>) with fixed Q-table size is proposed for household-level decision-making, tackling the challenge of scalability for simultaneously learning independent agents under partial observability in a stochastic environment [1]. Multiple versions of the novel strategy are assessed to maximise the statistical expectation of system-wide benefits, including local battery costs, grid costs and greenhouse gas emissions.

Widespread electrification of primary energy provision and decarbonisation of the power sector are two vital prerequisites for limiting anthropogenic global warming to 1.5°C above pre-industrial levels. To reduce

risks of climate-related impacts on health, livelihood, security and economic growth, intermittent renewable power supplies could be required to supply 70% to 85% of electricity by 2050 [2]. However, this poses the challenges of the intermittency and limited controllability of resources [3]. Therefore, a robust, decarbonised power system will rely on two structural features: decentralisation and demand response (DR) [4]. The coordination of distributed flexible energy resources can help reduce costs for transmission, storage, peaking plants and capacity reserves, improve grid stability, align demand with decarbonised energy provision, promote energy independence and security, and lower household energy bills [5, 6].

Residential sites constitute a significant share of potential DR, representing for example 38.5% of the 2019 UK electricity demand, and 56.4% of energy consumption if including transport and heat, which are both

---

\*Corresponding author

Email address: [flora.charbonnier@eng.ox.ac.uk](mailto:flora.charbonnier@eng.ox.ac.uk) (Flora Charbonnier)

<sup>1</sup>A full nomenclature is available in Appendix A

undergoing electrification [7]. Increasing ownership of EVs and PV panels has been facilitated by regulatory changes, with many countries committing to internal combustion car phase-outs in the near future, and by plummeting costs, with an 82% and 87% levelised cost drop between 2010 and 2019 for EVs and PV panels [8, 9]. This potential is so far underexploited, as DR primarily focuses on larger well-known industrial and commercial actors that require less coordination and data management [10], with most customers still limited to trade with utility companies [11]. The primary hurdles to unlocking residential flexibility are the high capital cost of communication and control infrastructure as the domestic potential is highly fragmented [4], concerns about privacy and hindrance of activities [12, 6], and computational challenges for real-time control at scale [13].

Traditionally, convex optimisation would be used to maximise global coordination objectives in convex problems with variables known ahead of time. Techniques such as least-squares and linear programming have been well-studied for over a century [14]. However, residential energy coordination presents challenges to its application. Firstly, optimisations that are centralised are hindered by privacy, acceptance, and communication constraints, and present exponential time complexity at the scale of millions of homes [15]. Secondly, standard optimisation methods cannot be used without full knowledge of the system’s inputs and dynamics [16]. In residential energy, agents only have partial observability of the system due to both the stochasticity and uncertainty of environment variables such as individual residential consumption and generation profiles, and to the privacy and infrastructure cost constraints that hinder communication between agents during implementation [17]. Not relying on shared information may also improve the robustness of the solutions to failure of other agents, communication delays, and unreliable information, and improve adaptability to changing environments [18]. Finally, the real-life complex electricity grid environment may not be amenable to a convex model representation. Due to the heterogeneity of users and behaviours needing different parameters and models, the large-scale use of model-based controllers is cumbersome [19]. A model-free approach instead avoids modelling non-trivial interactions of parameters, including private information [15].

Given these challenges to residential energy flexibility coordination, and the specific constraints of the problem at play which renders traditional approaches unsuitable, we seek to develop a novel coordination mechanism which satisfies the following criteria, as tested in real-life scenarios:

- Computational scalability: minimal and constant computation burden during implementation as the system size increases;
- Performance scalability: no drop in coordination

performance as the system size increases, measured in savings obtained per hour and per agent;

- Acceptability: local control of appliances, no communication of personal data, thermal discomfort, or hindrance/delay of activities.

The rest of this paper is organised as follows. In Section 2 we motivate the novel MARL approach with a literature review and a gap analysis. In Section 3, a system model is presented that includes household-level modelling of EVs, space heating, flexible loads and PV generation. Section 4 lays out the MARL methodology, with various methodological options for independent agents to learn to cooperate. In Section 5, the input data used to populate the model is presented. In Section 6, the performance of different MARL strategies is compared to lower and upper bounds in case studies. Finally, we conclude in Section 7.

## 2. MARL-based energy coordination: literature review and gap analysis

Reinforcement learning (RL) can overcome the constraints faced by centralised convex optimisation for residential energy coordination, by allowing for decentralised and model-free decision-making based on partial knowledge. RL is an artificial intelligence (AI) framework for goal-oriented agents<sup>2</sup> to learn sequential decision-making by interacting with an uncertain environment [22]. As an increasing wealth of data is collected in local electricity systems, RL is of growing interest for the real-time coordination of distributed energy resources (DERs) [23, 5]. Instead of optimising based on inherently uncertain data, RL more realistically searches for statistically optimal sequential decisions given partial observation and uncertainty, with no *a priori* knowledge [16]. Approximate learning methods may be more computationally scalable, more efficient in exploring high-dimensional state spaces and therefore more scalable than exact global optimisation with exponential time complexity [24, 15].

As classified in [25], numerous RL-based coordination methods have been proposed in the literature for residential energy coordination, though with remaining limitations in terms of scalability and privacy protection. On the one hand, in RL-based direct control strategies, a central controller directly controls individual units, and households directly forfeit their data and control to a central RL-based scheduler [26]. While most existing AI-based DR research thus assumes fully observable tasks [23], direct controllability of resources from different owners with different objectives and resources and subject to privacy, comfort and security concerns is challenging [27]. Moreover, centralised policies do not scale due to the curse of dimensionality as the state and action spaces grow exponentially with the system size [28]. On the other hand, RL-based indirect control strategies consider decision-making at the

---

<sup>2</sup>Here agents are independent computer systems acting on behalf of prosumers [20]. Prosumers are proactive consumers with distributed energy resources actively managing their consumption, production and storage of energy [21].

prosumer level, entering the realm of MARL. This can be achieved using different communication structures, with either centralised, bilateral, or no sharing of personal information, as presented below.

Firstly, agents may share information with a central entity, which in turn broadcasts signals based on a complete picture of the coordination problem. For example, the central entity may send unidirectional price signals to customers based on information such as prosumers' costs, constraints and day-ahead forecasts. RL can inform both the dynamic price signal [29, 30], and the prosumer response to price signals [30, 31]. The central entity may also collect competitive bids and set trades and match prosumers centrally, where RL algorithms are used to refine individual bidding strategies [32, 33, 34, 35, 36] or to dictate the auction market clearing [11, 37]. Units may also use RL to cooperate towards common objectives with the mediation of a central entity that redistributes centralised personal information [38, 39, 40, 41]. However, information centralisation also raises costs, security, privacy and scalability of computation issues. Biased information may lead to inefficient or even infeasible decisions [42].

Secondly, RL-based coordination has been proposed where prosumers only communicate information bilaterally without a central authority. For example, in [43] agents use transfer learning with distributed W-learning to achieve local and system objectives. Bilateral peer-to-peer communication offers autonomy and expression of individual preferences, though with remaining risks around privacy and bounded rationality [44]. There is greater robustness to communication failures compared situations with a single point of failure. However, as the system size increases, the number of communication iterations until algorithmic convergence increases, requiring adequate computational resources and limited communication network latency for feasibility [45]. The safe way of implementing distributed transactions to ensure data protection is an ongoing subject of research [25].

Finally, in RL-based implicit coordination strategies, prosumers rely solely on local information to make decisions. For example, in [46, 47], competitive agents in isolation maximise their profits in RL-based energy arbitrage, though they do not consider the impacts of individual actions on the rest of the system, with potential negative impacts for the grid. For example, a concern is that all loads receive the same incentive, the natural diversity on which the grid relies may be diminished [48], and the peak potentially merely displaced, with overloads on upstream transformers. Implicit cooperation, which keeps personal information at the local level while encouraging cooperation towards global objectives, has been thus far under-researched beyond frequency control. In [49], agents learn the optimal way of acting and interacting with the environment to restore frequency using local information only. This is a promising approach for decentralised control. However, the applicability in more complex scenarios with residential electric vehicles and smart heating load scheduling problems has not been consid-

ered. Moreover, the convergence slows down for increasing number of agents, and scalability beyond 8 agents has not been investigated. Indeed, fundamental challenges to the coordination of simultaneously learning independent agents at scale under partial observability in a stochastic environment have been identified when using traditional RL algorithms [1]: independent learners may reach individual policy equilibria that are incompatible with a global Pareto optimal, the non-stationarity of the environment due to other concurrently learning agents affects convergence, and the stochasticity of the environment prevents agents from discriminating between their own contribution to global rewards and noise from other agents or the environment. Novel methods are therefore needed to develop this approach.

We seek to bridge this gap, using implicit coordination to unlock the so-far largely untapped value from residential energy flexibility to provide both individual and system benefits. We propose a new class of MARL-based implicit cooperation strategies for residential DR, to make the best use of the flexibility offered by increasingly accessible assets such as photovoltaic (PV) panels, electric vehicle (EV) batteries, smart heating and flexible loads. Agents learn RL policies using a data-based, model-free statistical approach by exploring a shared environment and interacting with decentralised partially observable Markov decision processes (Dec-POMDPs), either through random exploration or learning from convex optimisation results. In the first rehearsal phase [50] with full understanding of the system, they learn to cooperate to reach system-wide benefits by assessing the global impact of their individual actions, searching for trade-offs between local, grid and social objectives. The pre-learned policies are then used to make decisions under uncertainty given limited local information only.

This approach satisfies the computational scalability, coordination scalability and acceptance criteria set out in this paper.

Firstly, the real-time control method is computationally scalable thanks to fixed-size Q-tables which avoid the curse of dimensionality, and there is only minimal, constant local computation required to implement the pre-learned policies during implementation. No further communication is required for implementation. This increases robustness to communication issues and data inaccuracy relative to when relying on centralised and bilateral communication, and cuts the costs of household computation and two-way communication infrastructure.

Secondly, we address the outstanding MARL coordination performance scalability issue for agents with partial observability in a stochastic environment seeking to maximise rewards which also depend on other concurrently learning agents [51, 1]. The case studies in this paper show that allowing agents to learn from omniscient, stable, and consistent optimisation solutions can successfully act as an equilibrium-selection mechanism, while the use of marginal rewards improves learn-

ability<sup>3</sup> by isolating individual contributions to global rewards. This novel methodological combination offers significant improvements on MARL scalability and convergence issues, with high coordination performance maintained as the number of agents increases, where that of standard MARL drops at scale.

Finally, this method tackles acceptability issues, with no interference in personal comfort nor communication of personal data.

The specific novel contributions of this paper are (a) a novel class of decentralised flexibility coordination strategies, MARL-based implicit cooperation, with no communication and fixed-size Q-tables to mitigate the curse of dimensionality; (b) a novel MARL exploration strategy for agents under partial observability to learn from omniscient, convex optimisations prior to implementation for convergence to robust cooperation at scale; and (c) the design and testing with large banks of real-world data of combinations of reward definitions, exploration strategies and multi-agent learning frameworks for assessing individual impacts on global energy, grid and storage costs. Methodologies are identified which outperform a baseline with increasing numbers of agents despite uncertainty.

### 3. Local system description

In this section, the variables, objective function and constraints of the problem are described. This sets the frame for the application of the RL algorithms presented in Section 4.

#### 3.1. Variables

We consider a set of time steps  $t \in \mathcal{T} = \{t_0, \dots, t_{\text{end}}\}$  and a set of prosumers  $i \in \mathcal{P} = \{1, \dots, n\}$ . Decision variables are *italicised* and input data are written in roman. Energy units are used unless specified otherwise. Participants have an EV, a PV panel, electric space heating and generic flexible loads.

The EV at-home availability  $\mu_i^t$  (1 if available, 0 otherwise), EV demand for required trips  $d_{\text{EV},i}^t$ , household electric demand  $d_i^t$ , PV production  $p_{\text{PV},i}^t$ , external temperature  $T_e^t$  and solar heat flow rate  $\phi^t$  are specified as inputs for  $t \in \mathcal{T}$  and  $i \in \mathcal{P}$ .

The local decisions by prosumers are the energy flows in and out of the battery  $b_{\text{in},i}^t$  and  $b_{\text{out},i}^t$ , the electric heating consumption  $h_i^t$  and the prosumer consumption  $c_i^t$ . These have both local and system impacts (Figure 1). Local impacts include battery energy levels  $E_i^t$ , losses  $\epsilon_{\text{ch},i}^t$  and  $\epsilon_{\text{dis},i}^t$ , prosumer import  $p_i^t$ , building mass temperature  $T_{\text{m},i}^t$  and indoor air temperature  $T_{\text{air},i}^t$ . System impacts arise through the costs of total grid import  $g^t$  and distribution network trading. Distribution network losses and reactive power flows are not included.

#### 3.2. Objective function

Prosumers cooperate to minimise system costs consisting of grid ( $c_g^t$ ), distribution ( $c_d^t$ ) and storage ( $c_s^t$ ) costs. This objective function will be maximised both in convex optimisations off-line – to provide an upper bound for the achievable objective function, and in some cases to provide information to the learners during the simulated learning phase – and in the learning of MARL policies for decentralised online implementation.

$$\max F = \sum_{\forall t \in \mathcal{T}} \hat{F}_t = \sum_{\forall t \in \mathcal{T}} -(c_g^t + c_d^t + c_s^t) \quad (1)$$

$$c_g^t = C_g^t (g^t + \epsilon_g) \quad (2)$$

Where losses incurred by imports and exports from and to the main grid are approximated as

$$\epsilon_g = \frac{R}{V^2} (g^t)^2 \quad (3)$$

The grid cost coefficient  $C_g^t$  is the sum of the grid electricity price and the product of the carbon intensity of the generation mix at time  $t$  and the Social Cost of Carbon which reflects the long-term societal cost of emitting greenhouse gases [52]. The impacts of local decisions on upstream energy prices are neglected. Grid losses are approximated using the nominal root mean square grid voltage  $V$  and the average resistance between the main grid and the distribution network  $R$  [53], based on the assumption of small network voltage drops and relatively low reactive power flows [54]. The second-order dependency disincentivises large power imports and exports, which helps ensure interactions of transmission and distribution networks do not reduce system stability.

$$c_d^t = C_d \sum_{i \in \mathcal{P}} \max(-p_i^t, 0) \quad (4)$$

Distribution costs  $c_d^t$  are proportional to the distribution charge  $C_d$  on exports. The resulting price spread between individual imports and exports decreases risks of network constraints violation by incentivising the use of local flexibility first [55]. Distribution network losses due to power flows between prosumers are neglected so there is no second-order dependency.

$$c_s^t = C_s \sum_{i \in \mathcal{P}} (b_{\text{in},i}^t + b_{\text{out},i}^t) \quad (5)$$

Storage battery depreciation costs  $c_s^t$  are assumed to be proportional to throughput using the depreciation coefficient  $C_s$ , assuming a uniform energy throughput degradation rate [56].

<sup>3</sup>“the sensitivity of an agent’s utility to its own actions as opposed to actions of others, which is often low in fully cooperative Markov games” [1]

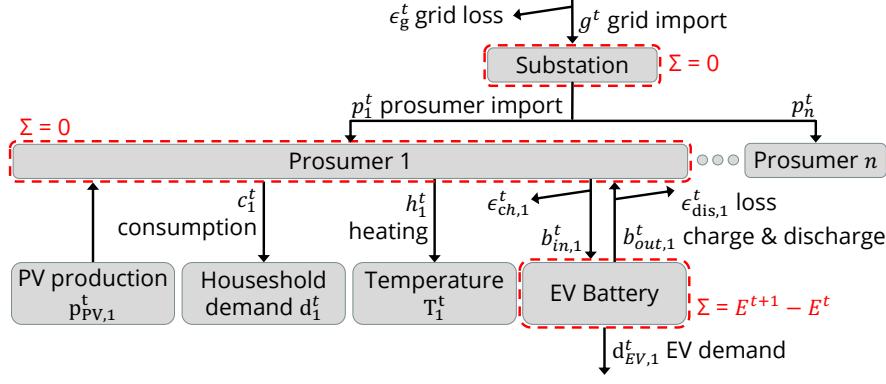


Figure 1: Local system model. Red dotted lines denote energy balances.

### 3.3. Constraints

Let  $E_0$ ,  $\underline{E}$  and  $\bar{E}$  be the initial, minimum and maximum battery energy levels,  $\eta_{ch}$  and  $\eta_{dis}$  the charge and discharge efficiencies, and  $\bar{b}_{in}$  the maximum charge per time step. Demand  $d_{i,k}^{t_D}$  is met by the sum of loads consumed  $\hat{c}_{i,k,t_C,t_D}$  at time  $t_C$  by prosumer  $i$  for load of type  $k$  (fixed or flexible) demanded at  $t_D$ . The flexibility boolean  $f_{i,k,t_C,t_D}$  indicates if time  $t_C$  lies within the acceptable range to meet  $d_{i,k}^{t_D}$ . A Crank-Nicholson scheme [57] is employed to model heating, with  $\kappa$  a  $2 \times 5$  matrix of temperature coefficients, and  $\underline{T}_i^t$  and  $\bar{T}_i^t$  lower and upper temperature bounds. System constraints for steps  $\forall t \in \mathcal{T}$  and prosumers  $\forall i \in \mathcal{P}$  are:

- Prosumer and substation energy balance (see Figure 1)

$$p_i^t = c_i^t + h_i^t + \frac{b_{in,i}^t}{\eta_{ch}} - \eta_{dis} b_{out,i}^t - p_{PV,i}^t \quad (6)$$

$$\sum_{i \in \mathcal{P}} p_i^t = g^t \quad (7)$$

- Battery energy balance

$$E_i^{t+1} = E_i^t + b_{in,i}^t - b_{out,i}^t - d_{EV,i}^t \quad (8)$$

- Battery charge and discharge constraints

$$E_0 = E_i^{t_0} = E_i^{t_{end}} + b_{in,i}^{t_{end}} - b_{out,i}^{t_{end}} - d_{EV,i}^{t_{end}} \quad (9)$$

$$\mu_i^t \underline{E}_i \leq E_i^t \leq \bar{E}_i \quad (10)$$

$$b_{in,i}^t \leq \mu_i^t \bar{b}_{in} \quad (11)$$

$$b_{out,i}^t \leq \mu_i^t \bar{E}_i \quad (12)$$

- Consumption flexibility — the demand of type  $k$  at time  $t_D$  by prosumer  $i$  must be met by the sum of partial consumptions  $\hat{c}_{i,k,t_C,t_D}$  at times  $t_C \dots t_C + n_{flex}$  within the time frame  $n_{flex}$  specified by the flexibility of each type of demand in matrix  $f_{i,k,t_C,t_D}$

$$\sum_{t_C \in \mathcal{T}} \hat{c}_{i,k,t_C,t_D} f_{i,k,t_C,t_D} = d_{i,k}^{t_D} \quad (13)$$

- Consumption — the total consumption at time  $t_C$  is the sum of all partial consumptions  $\hat{c}_{i,k,t_C,t_D}$  meeting parts of demands from current and previous time steps  $t_D$ :

$$\sum_{t_D \in \mathcal{N}} \hat{c}_{i,k,t_C,t_D} = c_{i,k}^{t_C} \quad (14)$$

- Heating — the workings to obtain this equation are included in Appendix C:

$$\begin{bmatrix} T_{m,i}^{t+1} \\ T_{air,i}^{t+1} \end{bmatrix} = \kappa [1, T_{m,i}^t, T_e^t, \Phi^t, h_i^t]^\top \quad (15)$$

$$\underline{T}_i^t \leq T_{air,i}^t \leq \bar{T}_i^t \quad (16)$$

- Non-negativity constraints

$$c_i^t, h_i^t, E_i^t, b_{in,i}^t, b_{out,i}^t, \hat{c}_{i,l,t_C,t_D} \geq 0 \quad (17)$$

While the proposed framework could accommodate the use of idiosyncratic satisfaction functions to perform trade-offs between flexibility use and users' comfort, no such trade-offs are considered in this paper, with comfort requirements for temperature and EV usage always being met. Field evaluations have shown that programmes that do not maintain thermal comfort are consistently overridden, increasing overall energy use and costs [58], while interference in consumption patterns and temperature set-points cause dissatisfaction [5]. Meeting fixed domestic loads, ensuring sufficient charge for EV trips, and maintaining comfortable temperatures are therefore set constraints.

### 4. Reinforcement learning methodology

The MARL approach is now presented in which independent prosumers learn to make individual decisions which together maximise the statistical expectation of the objective function in Section 3.

At time step  $t \in \mathcal{T}$ , each agent is in a state  $s_i^t \in \mathcal{S}$  corresponding to accessible observations (here the time-varying grid cost), and selects an action  $a_i^t \in \mathcal{A}$  as defined in Section 4.3. This action dictates the decision

variables in Section 3.1  $b_{\text{in},i}^t$ ,  $b_{\text{out},i}^t$ ,  $h_i^t$  and  $c_i^t$ . The environment then produces a reward  $r^t \in \mathcal{R}$  which corresponds to the share  $\hat{F}_t$  of the system objective function presented in Section 3.2 and agents transition to a state  $s_i^{t+1}$ . Agents learn individual policies  $\pi_i$  by interacting with the environment using individual, decentralised fixed-size Q-tables.

We first introduce the Q-learning methodology. Then, the mapping between the RL agent action and the decision variables in Section 3.1 is presented. Finally, we propose variations on the learning method, with different experience sources, multi-agent structures and reward definitions.

#### 4.1. Q-Learning

While any reinforcement learning methodology could be used with the framework proposed in this paper, here we focus on Q-learning, a model-free, off-policy RL methodology. Its simplicity and proof of convergence make it suited to developing novel learning methodologies in newly defined environments [5]. State-actions values  $Q(s, a)$  represent the expected value of all future rewards  $r_t \forall t \in \mathcal{T}$  when taking action  $a$  in state  $s$  according to policy  $\pi$ :

$$Q(s, a) \triangleq E^\pi[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} \dots | s_t = s, a_t = a] \quad (18)$$

where  $\gamma$  is the discount factor setting the relative importance of future rewards. Estimates are refined incrementally as

$$\hat{Q}(s, a) \leftarrow \hat{Q}(s, a) + \alpha \delta \quad (19)$$

where  $\delta$  is the temporal-difference error,

$$\delta = (r_t + \gamma \hat{V}(s^{\text{next}}) - \hat{Q}(s, a)) \quad (20)$$

$\hat{V}$  is the state-value function estimate,

$$\hat{V}(s) = \max_{a^* \in \mathcal{A}(s)} \hat{Q}(s, a^*) \quad (21)$$

and  $\alpha$  is the learning rate. In this work we use hysteric learners, i.e. chiefly optimistic learners that use an increase rate superior to the decrease rate in order to reduce oscillations in the learned policy due to actions chosen by other agents [1, 59]. For  $\beta < 1$ :

$$\alpha = \begin{cases} \alpha_0 & \text{if } \delta > 0 \\ \alpha_0 \beta & \text{otherwise} \end{cases} \quad (22)$$

Agents follow an  $\epsilon$ -greedy policy to balance exploration of different state-action pairs and knowledge exploitation. The greedy action with highest estimated rewards is selected with probability  $1 - \epsilon$  and random actions otherwise.

$$a^* = \begin{cases} \arg \max_{a^* \in \mathcal{A}} \hat{Q}(s, a^*) & \text{if } x \sim U(0, 1) > \epsilon \\ a \sim p(a) = \frac{1}{|\mathcal{A}|} \forall a \in \mathcal{A} & \text{otherwise} \end{cases} \quad (23)$$

Henceforth, we refer to the estimates  $\hat{Q}$  and  $\hat{V}$  as  $Q$  and  $V$  to reduce the amount of notation.

#### 4.2. Agent state

The agent state is defined by the time-dependent grid cost coefficient  $C_g^t$ , i.e. the sum of the grid electricity price and the product of the carbon intensity of the generation mix at time  $t$  and the social cost of carbon.

To convert the RL policy action into local decisions, the agent also requires information on their current PV generation, battery level, flexible loads and indoor air temperature, as described below in Section 4.3.

#### 4.3. Agent action

Large action spaces compound the curse of dimensionality in Q-learning and waste exploration resources [28]. At each time step, the decision variables in Section 3 controlling the flows in and out of the battery  $b_{\text{in},i}^t$  and  $b_{\text{out},i}^t$ , the electric heating consumption  $h_i^t$  and the prosumer consumption  $c_i^t$  for household  $i$  are therefore synthesised into a single variable  $\psi \in [0, 1]$  controlling the use of available local flexibility. Figure 2 shows how consumption (for domestic loads and heat), imports and storage change with  $\psi$ .

At each step, the fixed requirements for loads, heat and upcoming EV trips are first met. The  $\psi$  decision then applies to the remaining flexibility. In conditions deemed optimal for energy exports  $\psi = 0$ , all initial storage and residual PV generation is exported and flexible loads are delayed. On the other end, a *passive* agent does not utilise its flexibility and uses the *default* action  $\psi = 1$ , maximising imports with EVs charged when plugged in and no flexible loads delayed. Intermediate imports trade-offs are mapped on Figure 2:

1. From exporting all to none of the initial storage  $E_i^t$
2. From meeting fixed loads  $d_{i,\text{fixed}}^t$  with the energy stored to importing the required amount
3. From no to maximum flexible consumption  $d_{i,\text{tot}}^t$
4. From exporting to storing PV energy  $p_{\text{PV},i}^t$  remaining after meeting loads
5. From importing no additional energy to filling up the battery to capacity  $\bar{E}_i$

Costlier actions incurring battery depreciation, losses and export costs are towards either  $\psi$  extreme, only used in highly beneficial situations (convex local costs function in the lower plot of Figure 2). Ranking actions consistently ensures agents do not waste resources trialling sub-optimal combinations of decisions. For example, it is more cost-efficient to first absorb energy imports by consuming flexible loads, and only use the battery (incurring costs) if imports are large.

Note that although this action space is continuous, it can be discretised into intervals for implementation in Q-learning.

#### 4.4. Variations of the learning method

Different experience sources, reward definitions and MARL structures are proposed within the MARL approach. The performance of these combinations of algorithmic possibilities will be assessed in Section 6 to inform effective model design.

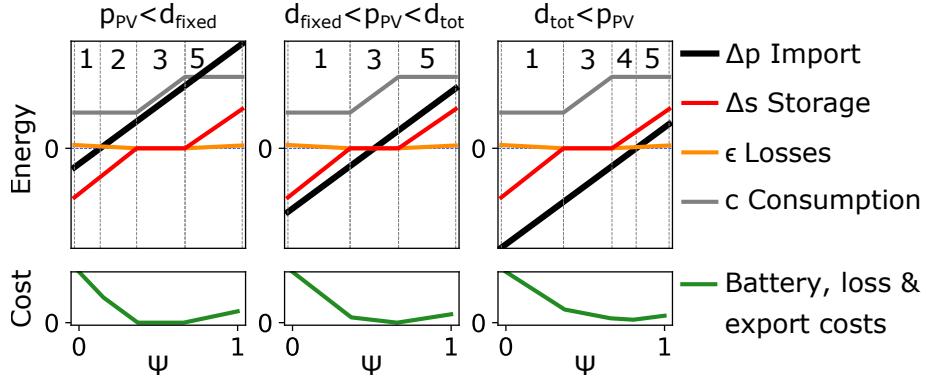


Figure 2: Decision variable  $\psi$ . Sections 1–5 denote the trade-off regimes described in Section 4.3. At each step, the fixed requirements for loads, heat and upcoming EV trips are first met. The  $\psi$  decision then applies to the remaining flexibility, from maximal energy exports (full use of flexibility) at  $\psi = 0$ , to maximal energy imports (no use of flexibility) at  $\psi = 1$ .  $d_{tot}$  and  $d_{fixed}$  are the sum of household and heating loads with and without their flexible component. If fixed loads cannot be fully met by PV energy, the residual is met by storage and imports (2). If there is additional PV energy after meeting all loads, it can be stored or exported (4).

#### 4.4.1. Experience sources

In data-driven strategies, the learning is determined by the collected experience.

- **Environment exploration.** Traditionally, agents collect experience by interacting with an environment [22].
- **Optimisations.** A novel approach collects experience from optimisations. Learning from entities with more knowledge or using knowledge more effectively than randomly exploring agents has previously been proposed, as with agents “mimicking” humans playing video games [60]. Similarly, agents learn from convex “omniscient” optimisations on historical data with perfect knowledge of current and future variables. This experience is then used under partial observability and control for stable coordination between prosumers at scale. Note in this case that, although the MARL learning and implementation are model-free, a model of the system is used to run the convex optimisation and produce experience to learn from. A standard convex optimiser uses the same data that would be used to populate the environment explorations but solves over the whole day-horizon with perfect knowledge of all variables using the problem description in Section 3. Then, at each time step, the system variables are translated into equivalent RL  $\{s_t, a_t, r_t, s_{t+1}\}$  tuples for each agent, which are used to update the policies in the same way as for standard Q-learning as presented below.

#### 4.4.2. MARL structures

Both the centralised and decentralised structures proposed use fixed-size  $|\mathcal{S}| \times |\mathcal{A}|$  Q-tables corresponding to individual state-action pairs. The size of a global Q-table referencing all possible combinations of states and actions would grow exponentially with the number of agents. This would limit scalability due to memory

limitations and exploration time requirements. Moreover, as strategies proposed in this paper are privacy-preserving, only local state-action pairs are used for individual action selection, wasting the level of detail of a global Q-table.

- **Distributed learning.** Each agent  $i$  learns its  $Q_i$  table with its own experience. No information is shared between agents.
- **Centralised learning.** A single table  $Q_c$  uses experience from all agents during pre-learning. All agents use the centrally learned policy for decentralised implementation.

#### 4.4.3. Reward definitions

The reward definition is central to learning as its maximisation forms the basis for incrementally altering the policy [22]. Assessing the impact of individual actions on global rewards accurately is key to the effective coordination of a large number of prosumers. In the following, the Q-tables  $Q^0$ ,  $Q^{\text{diff}}$ ,  $Q^A$  and  $Q^{\text{count}}$  may be either agent-specific  $Q_i$  or centralised  $Q_c$  based on the MARL structure. We proposed four variations of the Q-table update rule for each experience step tuple collected  $(s_i^t, a_i^t, r^t, s_i^{t+1})$ .

$$Q(s_i^t, a_i^t) \leftarrow Q(s_i^t, a_i^t) + \alpha \delta \quad (24)$$

- **Total reward.** The instantaneous total system reward  $r^t = \hat{F}_t$  is used to update the Q-table  $Q^0$ .

$$\delta = r^t + \gamma V^0(s_i^{t+1}) - Q^0(s_i^t, a_i^t) \quad (25)$$

- **Marginal reward.** The difference in total instant rewards  $r^t$  between that if agent  $i$  selects the greedy action and that if it selects the default action is used to update  $Q^{\text{diff}}$  [61]. The default action  $a_{\text{default}}$  corresponds to  $\psi = 1$ , where no flexibility is used. The default reward  $r_{a_i=a_{\text{default}}}^t$ , where all agents perform their greedy action apart from agent  $i$  which performs the default action, is

obtained by an additional simulation.

$$\delta = (r^t - r_{a_i=a_{\text{default}}}^t) + \gamma V^{\text{diff}}(s_i^{t+1}) - Q^{\text{diff}}(s_i^t, a_i^t) \quad (26)$$

- **Advantage reward.** The post difference between  $Q^0$  values when  $i$  performs the greedy and the default action is used. This corresponds to the estimated increase in rewards not just instantaneously but over all future states, analogously to in [62]. No additional simulations are required as the Q-table values are refined over the normal course of explorations.

$$\delta = (Q^0(s_i^t, a_i^t) - Q^0(s_i^t, a_{a_i=a_{\text{default}}})) - Q^A(s_i^t, a_i^t) \quad (27)$$

- **Count.** The Q-table stores the number of times each state-action pair is selected by the optimiser.

$$\alpha\delta = 1 \quad (28)$$

## 5. Input Data

This section presents the data that is fed into the model presented in Section 3. Interaction with this data will shape the policies learned through RL [22] and should reflect resource intermittency and uncertainty to maximise the expectation of rewards in a robust way without over-fitting. EV demand  $d_{EV,i}^t$  and availability  $\mu_i^t$ , PV production  $p_{PV,i}^t$  and electricity consumption  $d_i^t$  are drawn from large representative datasets.

### 5.1. Data selection and pre-processing

Load and PV generation profiles are obtained from the Customer Led Network Revolution (CLNR), a UK-based smart grid demonstration project [63, 64], and mobility data from the English National Travel Survey (NTS) [65]. The NTS does not focus on EVs only and offers a less biased view into the general population’s travel pattern than small-scale EV trials data, both due to the smaller volume of data available compared to for generic cars and because the self-selected EV early trial participants may not be representative of patterns once EVs become widely adopted. It is implicitly assumed that electrification will not affect transport patterns [66].

NTS data from 82,455 households from 2002 to 2017 results in 1,272,834 full days of travel profiles. Load and PV data from 11,907 customers between 2011 and 2014 yields 620,702 and 22,670 full days of data, respectively. Profiles are converted to hourly resolution and single missing points replaced with the figure from the same time the day or week before or after which has the lowest sum of squares of differences between the previous and subsequent point. Tested with available data, this yields absolute errors with mean 0.13 and 0.08 kWh and 99th percentile 1.09 and 0.81 kWh for PV and load data. PV sources have nominal capacities between 1.35 and 2.02 kWp.

The at home-availability of the vehicles is inferred from the recorded journeys’ origin and destination. EV

energy consumption profiles are obtained using representative consumption factors from a tank-to-wheel model proposed in [66], dependent on travel speed and type (rural, urban, motorway).

### 5.2. Markov chain

During learning, agents continuously receive experience to learn from. However, numerous subsequent days of data are not available for single agents. We design a Markov chain mechanism to feed consistent profiles for successive days, using both consistent scaling factors and behaviour clusters.

Daily profiles for load and travel are normalised such that  $\sum_{t=0..24} x^t = 1$ , and clustered using K-means, minimising the within-cluster sum-of-squares [67] in four clusters for both weekday and weekend data (with one for no travel). The features used for load profiles clustering are normalised peak magnitude and time and normalised values over critical time windows, and those for travel are normalised values between 6 am and 10 pm. PV profiles were grouped per month.

Probabilistic Markov chain transition rules are shown in Table 1. Transition probabilities for clusters  $k$  and scaling factors  $\lambda$  are obtained from available transitions between subsequent days in the datasets for each week day type  $w$  (week day or weekend day). Figure 3 shows that subsequent PV and load scaling factors follow strong linear correlation, with the residuals of the perfect correlation following gamma distributions with zero mean, whereas EV load scaling factors follow more complex patterns, so transitions probabilities are computed between 50 discrete intervals.

## 6. Case study results and discussion

This section compares the performance of the residential flexibility coordination strategies presented in Section 4 to baseline and upper bound scenarios for increasing numbers of prosumers. The performance of traditionally used MARL strategies drops at scale, while that of the novel optimisation-based methodology using marginal rewards is maintained.

### 6.1. Set-up

The MARL algorithm is trained in off-line simulations using historical data prior to online implementation. This means agents do not trial unsuccessful actions with real-life impacts during learning. Moreover, the computation burden is taken prior to implementation, while prosumers only apply pre-learned policies, avoiding the computational challenges of large-scale real-time control.

The learning occurs over 50 epochs consisting of an exploration, an update and an evaluation phase. First, the environment is explored over two training episodes of duration  $|\mathcal{T}| = 24$  hours. Learning in batches of multiple episodes helps stabilise learning in the stochastic environment. Then, Q-tables are updated based on the rules presented in Section 4.4. Finally, an evaluation is performed using a deterministic greedy policy on new evaluation data. Ten repetitions are performed such

	Normalised profile	Scaling factor
PV	Randomly selected from current month bank $b_{t+1} = (m)$	Computed as $\lambda_{t+1} = \lambda_t + x$ , where $x \sim \Gamma(\alpha(b_t, b_{t+1}), \beta(b_t, b_{t+1}))$
Load	Cluster selected based on transition probability $p(k_{t+1} k_t, w_t, w_{t+1})$	
EV	Normalised profile randomly selected from bank $b_{t+1} = (k_{t+1}, w_{t+1})$	Random variable from discrete distribution $p(\lambda_{t+1} \lambda_t, b_t, b_{t+1})$

Table 1: Markov chain mechanism for selecting behaviour clusters, profiles and scaling factors for input data in subsequent days

that the learning may be assessed over different trajectories.

The Social Cost of Carbon is set at 70 £/tCO<sub>2</sub>, consistent with the UK 2030 target [68]. Weather [69], electricity time-of-use prices [70] and grid carbon intensity [71] are from January 2020, where relevant specified for London, UK. The low solar heat gains in January are neglected [72]. Other relevant parameters for the case studies are listed in Appendix B.

As performed on a Intel(R) Core(TM) i7-9800X CPU @ 3.80GHz, computation time for a learning trajectory is 2'45" for one agent and 97'5" for 30 agents, including evaluation points. The policy can then be directly applied at the household level during operation.

Case study results using different experience sources, reward definitions and MARL structures are presented in Figure 4. Acronyms for each strategy are tabulated in the legend. Positive values denote savings relative to a baseline scenario where all agents are passive, i.e. not using their flexibility with EVs charged immediately and no flexible loads delayed. As the Q-learning policies are first initialised with zero values, in the first epoch of learning completely random action values are chosen, which provides rewards far below the baseline. As agents collect experience and update their policies at each epoch, improved policies are learned, some of which are able to outperform the baseline. An upper bound is provided by results from “omniscient” convex optimisations, which are however not achiev-

able in practice for three main reasons. Firstly, they use perfect knowledge of all the environment variables in the present and future, despite uncertainty in renewable generation, mix of the grid, and customer behaviour. Optimisation with inaccurate data would lead to suboptimal results. Secondly, prosumers may not be willing to yield their data and direct control to an external entity. Finally, central optimisations become computationally expensive for real-time control of large numbers of prosumers.

## 6.2. Results

Results presented in Figure 4 show that only the algorithms learning from optimisations maintained stable coordination performance at scale, while the performance of traditionally used MARL algorithms would drop in this context of stochasticity and partial observation. The optimisation-based algorithm which uses marginal rewards (MO) performed best. We further elaborate on the results in the subsections below.

### 6.2.1. Environment exploration-based learning

The centralised MARL structure is favoured for environment exploration-based learning (continuous lines in Figure 4). A single policy uses experience collected by all agents, rather than each agent learning from their own experience only.

Figure 4 shows that environment exploration-based MARL using total rewards (TE, orange), the baseline

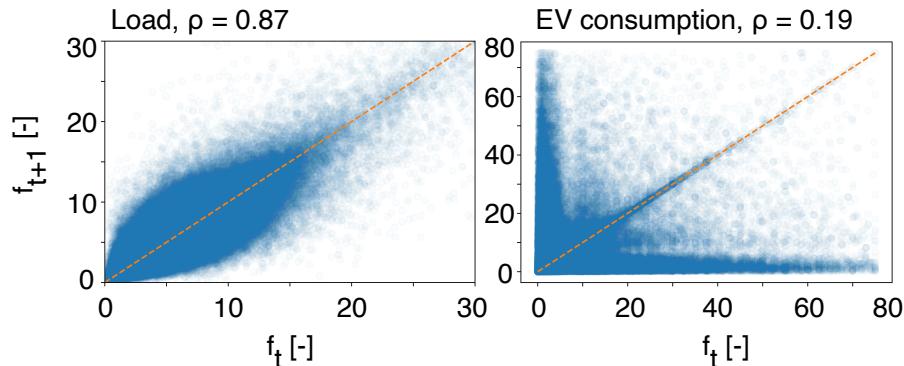


Figure 3: Scaling factors for normalised profiles (i.e. total daily loads in kWh) in subsequent days. Linear correlation can be observed for the load profiles, while more complex patterns are exhibited for EV consumption.  $\rho$  is the Pearson correlation coefficient.

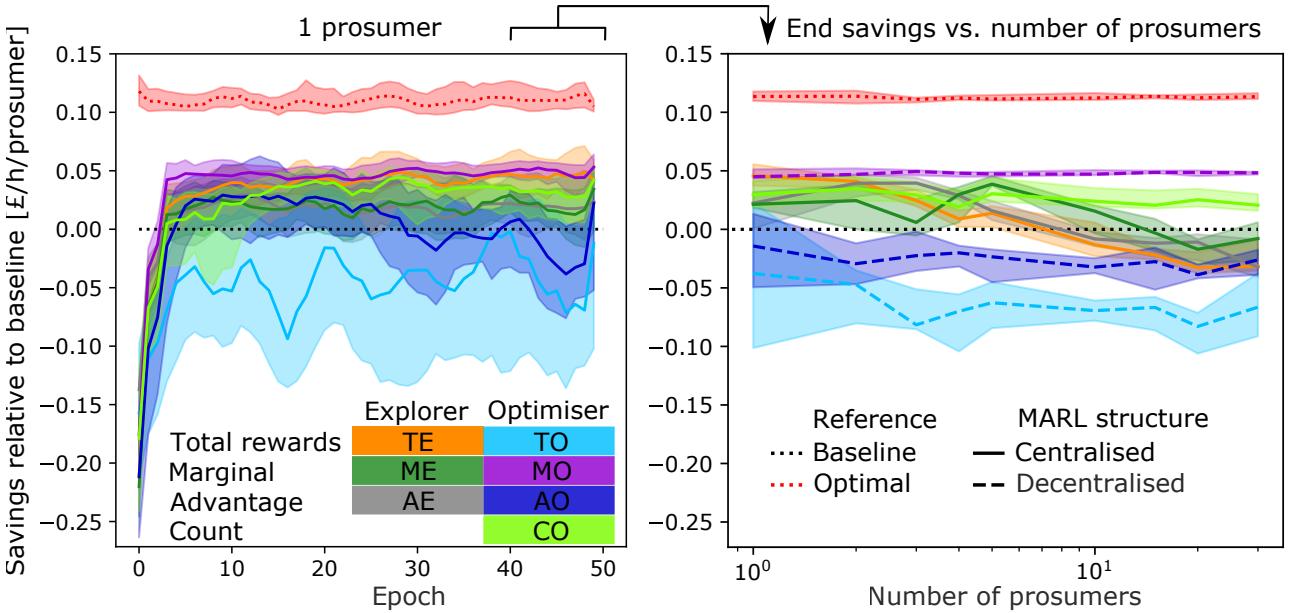


Figure 4: The left-hand side plot shows the five-epoch moving average of evaluation rewards relative to baseline rewards for a single prosumer. The right-hand side plot shows the mean of the final 10 evaluations against the number of prosumers. Lines show median values and shaded areas the 25th and 75th percentiles over the 10 repetitions. The best-performing MARL structure is displayed for each exploration source and reward definition pair. The performance of the baseline MARL algorithm (TE, orange) drops as the number of concurrently learning agents in the stochastic environment increases; the best-performing alternative algorithm proposed (MO, purple) maintains high performance at scale.

MARL framework, exhibits a high performance for a single agent. However, savings drop as the number of cooperating agents increases, down to around zero from ten agents. Coordination challenges arise for independent learners to isolate the contribution of their actions to total rewards from the stochasticity of the environment, compounded by other simultaneously learning agents’ random explorations, and the non-stationarity of their on-policy behaviour [1].

Using advantage rewards (AE, grey), based on estimates of the long-term value of actions relative to that of the baseline action, yields superior results beyond two agents. However, as AE uses the total reward  $Q^0$ -table as an intermediary step, results similarly drops for increasing numbers of agents.

Using marginal rewards (ME, dark green), the value of each agent’s action relative to the baseline action is singled out immediately by an additional simulation and used as a reward at each time step. This improves the performance relative to TE and AE for five agents and more, though still with declining performance as the number of agents increases.

### 6.2.2. Optimisation-based learning

Optimisation-based learning generally favours the distributed MARL structure, with agents able to converge to distinct compatible policies (dashed lines in Figure 4).

Comparing trajectories in Figure 4, learning from the total rewards obtained by an optimiser (TO, light blue) yields lower savings than when using environment explorations (TE). The learned policies yield negative savings, i.e. would provide worse outcomes than inflexible agents. The omniscient optimiser takes precise,

extreme decisions thanks to its perfect knowledge of all current and future system variables, importing at very high  $\psi$  values when it is optimal to do so. RL algorithms on the other hand are used under partial observability, aiming for actions that statistically perform well under uncertainty. Agents independently picking TO-based decisive actions in a stochastic environment do not yield optimal outcomes. Assessing the long-term advantage of actions from optimisations (AO, dark blue) follows a similar trend, whilst providing marginally superior savings relative to TO.

Optimisation-based learning using marginal rewards (MO, purple) offers the highest savings as the additional baseline simulations are best able to isolate the contribution of individual actions from variations caused by both the environment and other agents. When increasing the number of agents, the strategy is able to learn from optimal, stable, consistently behaving agents. Savings of 6.18p per agent per hour, or £45.11 per agent per month are obtained on average for 30 agents, corresponding to a 33.7% reduction from baseline costs. 65.9% of savings stem from reduced battery depreciation, 20.32% from distribution grid congestion, 11.1% from grid energy, and 2.7% from greenhouse gas emissions.

The count-based strategy learning from optimisations (CO, light green) seeks to reproduce the state-action patterns of the omniscient optimiser with perfect knowledge of system variables and perfect control of agents for local decision-making under partial observability. It provides results lower than the high performances of MO, though with a stable performance at scale. Savings of £21.09 per agent per month on average for 30 agents are obtained. The battery and dis-

tribution grid costs increase by an equivalent of 6.0% and 7.7% of total savings respectively, while grid energy and greenhouse gas emissions costs reductions represent 59.7% and 54.0% of total savings.

Both the MO and CO strategies exhibit stable performance at scale, though converging to different types of policy. The MO policy saves more by smoothing out the charging and distribution grid utilisation profiles despite smaller savings in imports and emissions costs, while CO derives a larger advantage from the grid price differentials in grid imports, though with higher battery and distribution grid costs. The weight applied to each of those competing objectives in the objective function directly impacts the policies that are learned. Examples of how the individual home energy management system decision variables (heating, energy consumption, battery charging) vary based on the controller are illustrated in Appendix D.

Overall, the new class of optimisation-based learning performs significantly better across different numbers of prosumers, with higher savings and lower inter-quartile range than environment-based learning at scale. This superior performance requires computations to run optimisations on historical data, and to perform baseline simulations to compute marginal rewards, though computational time for pre-learning is not strictly a limiting factor as it is performed off-line ahead of implementation.

A fundamental challenge in MARL has been the trade-off between fully centralised value functions, which are impractical for more than a handful of agents, or, in a more straightforward approach, independent learning of individual action-value functions by each agent in independent Q-learning (IQL) [73]. However, an ongoing issue with this approach has been that of convergence at scale, as agents do not have explicit representations of interactions between agents, and each agent's learning is confounded by the learning and exploration of others [74]. As shown in Figure 4, the Pareto selection, non-stationarity and stochasticity issues presented in Section 2 have prevented environment exploration-based learners from achieving successful MARL cooperation at scale for agents under partial observability in a stochastic environment. This case study of coordinated residential energy management shows that the novel combination of marginal rewards, which help agents isolate their marginal contribution to total rewards, and the learning from results of convex optimisations, where agents learn successful policy equilibria from omniscient, stable, and consistent solutions, offer significant improvements on these scalability and convergence issues.

## 7. Conclusion

In this paper, a novel class of strategies has addressed the scalability issue of residential energy flexibility coordination in a cost-efficient and privacy-preserving manner. The combination of off-line optimisations with multi-agent reinforcement learning provides high, stable coordination performance at scale.

We identified in the literature that the concept of RL-based implicit energy coordination, where energy prosumers cooperate towards global objectives based on local information only, had been under-researched beyond frequency droop control with limited number of agents. The scalability of such methods was identified as a key gap that we have sought to bridge. The novel coordination mechanism proposed in this paper thus satisfies the criteria for successful residential energy coordination set out in the introduction, as tested with large banks of real data in the case studies:

- Computational scalability: The scalability of traditional learning algorithms is significantly improved thanks to fixed-size Q-tables to avoid the curse of dimensionality, so that policies can be learned for larger number of agents. The proposed method does not require expensive communication and control appliances at the prosumer level, as pre-learned policies are directly applied with no further communication and no exponential time real-time optimisations needed. This is a crucial benefit for applications with physical limitations in hardware availability and processing time.
- Performance scalability: The coordination performance remains high for increasing numbers of prosumers despite the challenges of partial observability, environment stochasticity and concurrently learning of agents, thanks to learning from the results of global omniscient optimisations on historical data, and to rewards signals that isolate individual contributions to global rewards. Significant value of £45.11 per agent per month was obtained in the presented case study for 30 agents, thanks to savings in energy, prosumer storage and societal greenhouse gas emissions-related costs. Those savings do not drop with increasing number of agents, as opposed to with standard MARL approaches.
- Acceptability: The approach does not rely on sharing of personal data, thermal discomfort, or hindrance/delay of activities, and the appliances are controlled locally. This cost-efficient and privacy-preserving implicit coordination approach could help integrate distributed energy resources such as residential energy, otherwise excluded from energy systems' flexibility management.

Important future work is a more detailed assessment of the impacts of the coordination strategies on power flows, as well as an evaluation of the generalisation and adaptability potential of policies when used by other households or if household characteristics change over time. Moreover, while all agents readily reduce individual costs through participation in the framework, further game-theoretic tools could be used to design a post-operation reward scheme.

## Acknowledgement

This work was supported by the Saven European Scholarship and by the UK Research and Innovation and the Engineering and Physical Sciences Research Council (award references EP/S000887/1, EP/S031901/1, and EP/T028564/1).

## References

- [1] L. Matignon, G. Laurent, N. Le Fort-Piat, Independent reinforcement learners in cooperative Markov games: A survey regarding coordination problems, *Knowledge Engineering Review* 27 (1) (2012) 1–31. doi:10.1017/S0269888912000057.
- [2] Masson-Delmotte, V., Global Warming of 1.5C. An IPCC Special Report on the impacts of global warming of 1.5C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change (2018).
- [3] S. Bose, S. Low, Some Emerging Challenges in Electricity Markets, in: Smart Grid Control, power elec Edition, 2019, pp. 29–45. doi:10.1007/978-3-319-98310-3\_2.
- [4] T.-O. Léautier, Imperfect Markets and Imperfect Regulation: An Introduction to the Microeconomics and Political Economy of Power Markets, MIT Press, 2019.
- [5] J. Vázquez-Canteli, Z. Nagy, Reinforcement learning for demand response: A review of algorithms and modeling techniques, *Applied Energy* 235 (Oct 2018) (2019) 1072–1089. doi:10.1016/j.apenergy.2018.11.002.
- [6] K. Pumphrey, S. Walker, M. Andoni, V. Robu, Green hope or red herring? Examining consumer perceptions of peer-to-peer energy trading in the United Kingdom, *Energy Research and Social Science* 68 (September 2019) (2020) 101603. doi:10.1016/j.erss.2020.101603.
- [7] Department for Business Energy and Industrial Strategy, Energy consumption in the UK (2021).
- [8] I. R. E. Agency, Renewable Power Generation Costs in 2018, 2018. arXiv:arXiv:1011.1669v3, doi:10.1007/SpringerReference\_7300.
- [9] BloombergNEF, 2019 Battery Price Survey (2019).
- [10] Charles River Associates, An assessment of the economic value of demand-side participation in the Balancing Mechanism and an evaluation of options to improve access (2017).
- [11] T. Chen, W. Su, Indirect Customer-to-Customer Energy Trading with Reinforcement Learning, *IEEE Transactions on Smart Grid* 10 (4) (2019) 4338–4348. doi:10.1109/TSG.2018.2857449.
- [12] D. Bugden, R. Stedman, A synthetic view of acceptance and engagement with smart meters in the United States, *Energy Research and Social Science* 47 (January 2018) (2019) 137–145. doi:10.1016/j.erss.2018.08.025.
- [13] F. Moret, P. Pinson, Energy Collectives: A Community and Fairness Based Approach to Future Electricity Markets, *IEEE Transactions on Power Systems* 34 (5) (2019) 3994–4004. doi:10.1109/TPWRS.2018.2808961.
- [14] S. Boyd, Convex optimization theory, Vol. 25, 2009.  
URL <http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.214.7707>
- [15] S. Dasgupta, Computer Science: A Very Short Introduction, Oxford University Press, 2016. doi:10.1093/actrade/9780198733461.001.0001.
- [16] B. Recht, A tour of reinforcement learning: The view from continuous control, arXiv (2018). arXiv:1806.09460, doi:10.1146/annurev-control-053018-023825.
- [17] V. François Lavet, Contributions to deep reinforcement learning and its applications in smartgrids (2017).
- [18] S. Sen, M. Sekaran, J. Hale, Learning to coordinate without sharing information, *Proceedings of the National Conference on Artificial Intelligence* 1 (1994) 426–431.
- [19] F. Ruelens, Residential Demand Response of Thermostatically Controlled Loads Using Batch Reinforcement Learning, *IEEE Transactions on Smart Grid* 8 (5) (2017) 2149–2159. doi:10.1109/TSG.2016.2517211.
- [20] M. Wooldridge, Intelligent Agents: The Key Concepts, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002.
- [21] T. Morstyn, N. Farrell, S. Darby, M. McCulloch, Using peer-to-peer energy-trading platforms to incentivize prosumers to form federated power plants, *Nat Energy* 3 (2) (2018) 94–101.
- [22] R. S. Sutton, A. G. Barto, Reinforcement learning : an introduction [electronic resource], Adaptive computation and machine learning, MIT Press, Cambridge, Mass., 1998.
- [23] I. Antonopoulos, Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review, *Renewable and Sustainable Energy Reviews* 130 (April) (2020) 109899. doi:10.1016/j.rser.2020.109899.
- [24] C. Schellenberg, J. Lohan, L. Dimache, Comparison of metaheuristic optimisation methods for grid-edge technology that leverages heat pumps and

- thermal energy storage, Renewable and Sustainable Energy Reviews 131 (June) (2020) 109966. doi:10.1016/j.rser.2020.109966. URL <https://doi.org/10.1016/j.rser.2020.109966>
- [25] F. Charbonnier, T. Morstyn, M. McCulloch, Coordination of resources at the edge of the electricity grid: systematic review and taxonomy (2022). arXiv:2202.03786.
- [26] D. O'Neill, M. Levorato, A. Goldsmith, U. Mitra, Residential Demand Response Using Reinforcement Learning, 2010 First IEEE International Conference on Smart Grid Communications (2010) 409–414doi:10.1109/smartgrid.2010.5622078.
- [27] S. J. Darby, Demand response and smart technology in theory and practice: Customer experiences and system actors, Energy Policy 143 (April) (2020) 111573. doi:10.1016/j.enpol.2020.111573. URL <https://doi.org/10.1016/j.enpol.2020.111573>
- [28] W. Powell, Approximate dynamic programming: solving the curses of dimensionality, 2nd Edition, Wiley series in probability and statistics, J. Wiley & Sons, Hoboken, N.J., 2011.
- [29] R. Lu, S. H. Hong, Incentive-based demand response for smart grid with reinforcement learning and deep neural network, Applied Energy 236 (December 2018) (2019) 937–949. doi:10.1016/j.apenergy.2018.12.061.
- [30] B. Kim, Y. Zhang, M. Van Der Schaar, J. Lee, Dynamic Pricing and Energy Consumption Scheduling With Reinforcement Learning, IEEE Transactions on Smart Grid 7 (5) (2016) 2187–2198.
- [31] M. Babar, P. H. Nguyen, V. Cuk, I. G. Kamphuis, M. Bongaerts, Z. Hanelka, The evaluation of agile demand response: An applied methodology, IEEE Transactions on Smart Grid 9 (6) (2018) 6118–6127. doi:10.1109/TSG.2017.2703643.
- [32] M. G. Vayá, L. B. Roselló, G. Andersson, Optimal bidding of plug-in electric vehicles in a market-based control setup, Proceedings - 2014 Power Systems Computation Conference, PSCC 2014 (2014). doi:10.1109/PSCC.2014.7038108.
- [33] Y. Ye, D. Qiu, M. Sun, D. Papadaskalopoulos, G. Strbac, Deep Reinforcement Learning for Strategic Bidding in Electricity Markets, IEEE Transactions on Smart Grid 11 (2) (2020) 1343–1355. doi:10.1109/TSG.2019.2936142.
- [34] D. Dauer, C. M. Flath, P. Ströhle, C. Weinhardt, Market-based EV charging coordination, Proceedings - 2013 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2013 2 (2013) 102–107. doi:10.1109/WI-IAT.2013.97.
- [35] Y. Sun, A. Somani, T. Carroll, Learning based bidding strategy for HVAC systems in double auction retail energy markets, Proceedings of the American Control Conference 2015-July (2015) 2912–2917. doi:10.1109/ACC.2015.7171177.
- [36] J. G. Kim, B. Lee, Automatic P2P energy trading model based on reinforcement learning using long short-term delayed reward, Energies 13 (20) (2020). doi:10.3390/en13205359.
- [37] B. J. Claessens, S. Vandael, F. Ruelens, K. De Craemer, B. Beusen, Peak shaving of a heterogeneous cluster of residential flexibility carriers using reinforcement learning, 2013 4th IEEE/PES Innovative Smart Grid Technologies Europe, ISGT Europe 2013 (2013) 1–5doi:10.1109/ISGTEurope.2013.6695254.
- [38] X. Zhang, T. Bao, T. Yu, B. Yang, C. Han, Deep transfer Q-learning with virtual leader-follower for supply-demand Stackelberg game of smart grid, Energy 133 (2017) 348–365. doi:10.1016/j.energy.2017.05.114.
- [39] I. Dusparic, Maximizing renewable energy use with decentralized residential demand response, 2015 IEEE 1st International Smart Cities Conference, ISC2 2015 (2015). doi:10.1109/ISC2.2015.7366212.
- [40] I. Dusparic, Multi-agent residential demand response based on load forecasting, 2013 1st IEEE Conference on Technologies for Sustainability, SusTech 2013 (2013) 90–96doi:10.1109/SusTech.2013.6617303.
- [41] L. A. Hurtado, E. Mocanu, P. H. Nguyen, M. Gibescu, R. I. Kamphuis, Enabling Cooperative Behavior for Building Demand Response Based on Extended Joint Action Learning, IEEE Transactions on Industrial Informatics 14 (1) (2018) 127–136. doi:10.1109/TII.2017.2753408.
- [42] T. Morstyn, M. McCulloch, Peer-to-Peer Energy Trading, Analytics for the Sharing Economy: Mathematics, Engineering and Business Perspectives (March) (2020). doi:10.1007/978-3-030-35032-1.
- [43] A. Taylor, Accelerating Learning in multi-objective systems through Transfer Learning, Proceedings of the International Joint Conference on Neural Networks (2014) 2298–2305doi:10.1109/IJCNN.2014.6889438.
- [44] S. Herbert, Models of bounded rationality, MIT Press, Cambridge, Mass. ; London, 1982.
- [45] J. Guerrero, D. Gebbran, S. Mhanna, A. C. Chapman, G. Verbić, Towards a transactive energy system for integration of distributed energy resources: Home energy management, distributed optimal power flow, and peer-to-peer energy trading, Renewable & sustainable energy reviews 132 (2020).

- [46] J. Cao, Deep Reinforcement Learning Based Energy Storage Arbitrage With Accurate Lithium-ion Battery Degradation Model, *IEEE Transactions on Smart Grid* 14 (8) (2019) 1–9.
- [47] Y. Yang, J. Hao, Y. Zheng, C. Yu, Large-Scale Home Energy Management Using Entropy-Based Collective Multiagent Deep Reinforcement Learning Framework (2019) 630–636.
- [48] C. Crozier, D. Apostolopoulou, M. McCulloch, Mitigating the impact of personal vehicle electrification: A power generation perspective, *Energy Policy* 118 (2013) (2018) 474–481. doi:10.1016/j.enpol.2018.03.056.
- [49] S. Rozada, D. Apostolopoulou, E. Alonso, Load frequency control: A deep multi-agent reinforcement learning approach, *IEEE Power and Energy Society General Meeting 2020-August* (2020) 0–4. doi:10.1109/PESGM41954.2020.9281614.
- [50] L. Kraemer, B. Banerjee, Multi-agent reinforcement learning as a rehearsal for decentralized planning, *Neurocomputing* 190 (2016) 82–94. doi:10.1016/j.neucom.2016.01.031.
- [51] L. Buşoniu, R. Babuška, B. De Schutter, A comprehensive survey of multiagent reinforcement learning, *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews* 38 (2) (2008) 156–172. doi:10.1109/TSMCC.2007.913919.
- [52] M. Parry, Climate change 2007: impacts, adaptation and vulnerability, Published for the Intergovernmental Panel on Climate Change [by] Cambridge University Press, Cambridge, 2007.
- [53] T. Morstyn, M. McCulloch, Multiclass Energy Management for Peer-to-Peer Energy Trading Driven by Prosumer Preferences, *IEEE Transactions on Power Systems* 34 (5) (2019) 4005–4014. doi:10.1109/TPWRS.2018.2834472.
- [54] C. Coffrin, P. Van Hentenryck, R. Bent, Approximating line losses and apparent power in AC power flow linearizations, *IEEE Power and Energy Society General Meeting* (2012) 1–8doi:10.1109/PESGM.2012.6345342.
- [55] T. Morstyn, A. Teytelboym, C. Hepburn, M. McCulloch, Integrating P2P Energy Trading with Probabilistic Distribution Locational Marginal Pricing, *IEEE Transactions on Smart Grid* 11 (4) (2020) 3095–3106. doi:10.1109/TSG.2019.2963238.
- [56] R. Dufo-López, J. M. Lujano-Rojas, J. L. Bernal-Agustín, Comparison of different lead-acid battery lifetime prediction models for use in simulation of stand-alone photovoltaic systems, *Applied energy* 115 (2014) 242–253.
- [57] ISO, Calculation of Energy Use for Space Heating and Cooling ISO/FDIS 13790:2007(E) (2007).
- [58] O. Sachs, Field Evaluation of Programmable Thermostats (2012).
- [59] L. Matignon, G. J. Laurent, N. Le Fort-piat, Hysteretic Q-Learning : an algorithm for Decentralized Reinforcement Learning in Cooperative Multi-Agent Teams ., in: *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2007, pp. 64–69.
- [60] O. Vinyals, Grandmaster level in StarCraft II using multi-agent reinforcement learning, *Nature* 575 (November) (2019). doi:10.1038/s41586-019-1724-z.
- [61] D. Wolpert, K. Tumer, Optimal payoff functions for members of collectives, *Advances in Complex Systems* 04 (03 2002). doi:10.1142/S0219525901000188.
- [62] J. N. Foerster, G. Farquhar, T. Afouras, N. Nardelli, S. Whiteson, Counterfactual multi-agent policy gradients, *32nd AAAI Conference on Artificial Intelligence, AAAI 2018* (2018) 2974–2982arXiv:1705.08926.
- [63] R. Wardle, Dataset (TC1a): Basic Profiling of Domestic Smart Meter Customers (2014).
- [64] R. Wardle, Dataset (TC5): Enhanced Profiling of Domestic Customers with Solar Photovoltaics (PV) (2014).
- [65] Department for Transport, National Travel Survey 2002-2017 (2019). doi:<http://doi.org/10.5255/UKDA-SN-5340-10>.
- [66] C. Crozier, D. Apostolopoulou, M. McCulloch, Numerical analysis of national travel data to assess the impact of UK fleet electrification, *20th Power Systems Computation Conference, PSCC 2018* (2018) 1–7arXiv:1711.01440, doi:10.23919/PSCC.2018.8450584.
- [67] S. Lloyd, Least squares quantization in PCM, *IEEE Transactions on Information Theory* 28 (2) (1982) 129–137. doi:10.1109/TIT.1982.1056489.
- [68] D. Hirst, Commons Briefing Paper SNO5927: Carbon Price Floor (CPF) and the price support mechanism (2018).
- [69] Weather Wunderground, London City Airport weather history (2020).
- [70] Octopus Energy, Octopus Energy API (2019).
- [71] National Grid ESO, Environmental Defense Fund Europe, University of Oxford Department of Computer Science, WWF, Carbon Intensity API (2020).
- [72] J. Brown, J. Chambers, A. Rogers, SMITE : Using Smart Meters to Infer the Thermal Efficiency of Residential Homes, in: *The 7th ACM International*

- Conference on Systems for Energy- Efficient Buildings, Cities, and Transportation (BuildSys '20), 2020.
- [73] M. Tan, Multi-Agent Reinforcement Learning : Independent vs . Cooperative Agents (1993).
  - [74] T. Rashid, G. Farquhar, B. Peng, S. Whiteson, Weighted QMIX: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning, Advances in Neural Information Processing Systems 2020-December (2020). [arXiv:2006.10800](https://arxiv.org/abs/2006.10800).
  - [75] HOMER Energy, HOMER Pro 3.14 User Manual (2020).
  - [76] W. Schram, Empirical evaluation of V2G round-trip efficiency, SEST 2020 - 3rd International Conference on Smart Energy Systems and Technologies (October) (2020). doi:10.1109/SEST48500.2020.9203459.
  - [77] V. Becker, W. Kleiminger, V. Coroamă, F. Mattern, Estimating the savings potential of occupancy-based heating strategies, Energy Informatics 1 (S1) (2018). doi:10.1186/s42162-018-0022-6.
  - [78] BRE, SAP 2012 9.92 The Government's Standard Assessment Procedure for Energy Rating of Dwellings (2014). [arXiv:9809069v1](https://arxiv.org/abs/9809069v1).
  - [79] British Standards, Heating systems in buildings. Method for calculation of the design heat load, Ics 91.140.10 (January) (2009) 1–89.

## Appendix A Nomenclature

Acronyms	
AE	MARL with advantage rewards and exploration-based learning
AO	MARL using advantage rewards and optimisation-based learning
AI	artificial intelligence
CLNR	customer-led network revolution
CO	MARL using count rewards and optimisation-based learning
ME	MARL using marginal rewards and exploration-based learning
MO	MARL using marginal rewards and optimisation-based learning
Dec-POMDP	decentralised partially observable Markov decision process
DER	distributed energy resource
DR	demand response
EV	electric vehicle
MARL	multi-agent reinforcement learning
NTS	national travel survey
PV	photovoltaic
RL	reinforcement learning
TE	MARL using total rewards and exploration-based learning
TO	MARL using total rewards and optimisation-based learning
UK	United Kingdom

Variables	
$b_{in}$	charge into the battery [kWh]
$\bar{b}_{in}$	maximum charge into the battery [kWh]
$b_{out}$	discharge out of the battery [kWh]
$c$	household consumption [kWh]
$\hat{c}$	partial consumption for load type and time demanded [kWh]
$c_d$	distribution cost [£]
$C_d$	distribution charge [£/kWh]
$c_g$	grid cost [£]
$C_g$	grid cost coefficient [£/kWh]
$c_s$	storage cost [£]
$C_s$	battery depreciation coefficient [£/kWh]
$d$	household demand [kWh]
$d_{EV}$	electric vehicle demand [kWh]
$d_{fixed}$	sum of non-flexible household and heating loads [kWh]
$d_{tot}$	sum of household all heating loads [kWh]
$E$	battery energy level [kWh]
$E_0$	initial battery energy level [kWh]
$\underline{E}$	minimum battery energy level [kWh]
$\bar{E}$	maximum battery energy level [kWh]
$f$	flexibility boolean
$F$	objective function [£]
$\hat{F}$	share of objective function for given time step
$g$	total grid import to the group of prosumers [kWh]
$h$	heating energy consumption [kWh]
$k$	behaviour cluster for transport or household consumption profile
$p$	prosumer import [kWh]
$P_{PV}$	PV generation [kWh]
$Q$	Q value [£]
$\hat{Q}$	Q value estimate [£]
$r$	global reward [£]
$R$	average resistance between the main grid and the prosumers [ $\Omega$ ]
$T_e$	external temperature [ $^{\circ}\text{C}$ ]
$T_m$	building mass temperature [ $^{\circ}\text{C}$ ]
$T_{air}$	indoor air temperature [ $^{\circ}\text{C}$ ]
$\underline{T}$	minimum indoor air temperature [ $^{\circ}\text{C}$ ]
$\bar{T}$	maximum indoor air temperature [ $^{\circ}\text{C}$ ]
$U$	uniform distribution function

$V$	nominal root mean square grid voltage [V]
$\hat{V}$	state-value estimate [ $\mathcal{L}$ ]
Greek letters	
$\alpha_0$	base learning rate [-]
$\alpha$	learning rate [-]
$\beta$	hysteretic learning rate reduction factor [-]
$\gamma$	discount factor [-]
$\delta$	loss [ $\mathcal{L}$ ]
$\epsilon_{ch}$	battery charging losses [kWh]
$\epsilon_{dis}$	battery discharging losses [kWh]
$\epsilon$	share of random action selection during exploration [-]
$\eta_{ch}$	battery charging efficiency [-]
$\eta_{dis}$	battery discharging efficiency [-]
$\kappa$	matrix of heating model coefficients
$\lambda$	scaling factor for transport or household consumption profile [kWh]
$\mu$	electric vehicle availability boolean
$\pi$	policy
$\phi$	solar heat flow rate [ $J \cdot s^{-1}$ ]
$\psi$	local flexibility use decision variable [-]
Indexes	
$a$	action
$i$	prosumer
$s$	state
$t$	time step
$t_C$	consumption time step
$t_D$	demand time step
$w$	day type (week day or weekend day)
Sets	
$\mathcal{A}$	set of actions
$\mathcal{T}$	set of time steps
$\mathcal{P}$	set of prosumers
$\mathcal{S}$	set of states

Table 2: Nomenclature

## Appendix B Case study input data

- Learning parameters: The depreciation, learning and exploration rates are  $\gamma = 0.99$ ,  $\alpha_0 = 0.01$  and  $\epsilon = 0.5$ . The hysteretic learning rate reduction parameter for negative errors is  $\beta = 0.5$ . The states are defined by three uniform grid cost intervals for each day. The action space is discretised in 10 equal  $\psi$  intervals.
- Battery:  $\eta_{\text{ch}} = \eta_{\text{dis}} = \sqrt{\eta_{\text{round trip}}}$  [75], where  $\eta_{\text{round trip}} = 0.87$  [76], capacity  $\bar{E} = 75$  kWh, max. charging rate  $\bar{b}_{\text{in}} = 22$  kW, depreciation  $C_s = 20$  USD/MWh-throughput [53], initial and min. charge  $E_0 = 0.5\bar{E}$  and  $\underline{E} = 0.1\bar{E}$ .
- Grid: nominal voltage  $V = 415$  [V], average resistance to prosumers  $R = 0.084$  [ $\Omega$ ] [53].
- Flexible loads: 10% deferrable for up to  $n_{\text{flex}} = 5$  hours.
- Heating: housing of  $76 \text{ m}^2$ , 2.4m height. Comfort temperature  $20^\circ\text{C}$  between 7-10am and 5-10pm, setback  $16^\circ\text{C}$ . Variations of  $3^\circ\text{C}$  acceptable. U-values from [77], other heating inputs from [57, 78, 79]. Pre-heating up to five hours in advance. Coefficients after re-arranging:

$$\kappa = \begin{bmatrix} 6.84e-2, 9.08e-1, 9.15e-2, 2.62e-4, 2.52e-1 \\ 2.40e-1, 8.80e-1, 1.20e-1, 3.46e-4, 1.46 \end{bmatrix} \quad (29)$$

- EV consumption factors [kWh/10km]: 2.25 for motorway, 1.62 for urban and 1.36 for rural travel [66].
- Distribution network export charge: 0.01 £/kWh

## Appendix C Heating model

We use the simple hourly method heating model laid out in [57].

The input data used in the heating model in this paper is tabulated below. Note that these heating model and input data are meant as a generic building example that can be used to test the relative performances of the MARL coordination algorithms in the case study in Section 6. Models and parameters used for the detailed study of a specific building should be validated with experimental data.

Symbol	Definition	Value	Unit	Reference
$A_d$	door area	$1.4 \times 2$	$\text{m}^2$	
$A_f$	floor area	76	$\text{m}^2$	
$A_{wd}$	window area	$1.4 \times 1.4$	$\text{m}^2$	
$e$	shielding coefficient	0.03	-	[79]
$h$	height of rooms	2.4	m	
$h_{is}$	heat transfer coefficient between the air node $\theta_{\text{air}}$ and the surface node $\theta_s$	3.45	$\text{W} \cdot \text{m}^{-2} \cdot \text{K}^{-1}$	[57]
$h_{ms}$	heat transfer coefficient between nodes m and s	9.1	$\text{W} \cdot \text{m}^{-2} \cdot \text{K}^{-1}$	[57]
$k_{\text{party}}$	fraction of floor space that is party floor rather than on ground, for one-storey building	0.5	[ $\cdot$ ]	
$n_{\min}$	minimum external air exchange rate per hour for a habitable room	0.5	$\text{h}^{-1}$	[79]
$n_{50}$	Air exchange rate resulting from a pressure difference of 50 Pa between the inside and the outside of the building, including the effects of air inlets, medium construction family dwelling	6	$\text{h}^{-1}$	[79]
$U_g$	U-value for ground (old build)	1.0	$\text{W} \cdot \text{m}^{-2} \cdot \text{K}^{-1}$	[77]
$U_r$	U-value for roof (old build)	1.0	$\text{W} \cdot \text{m}^{-2} \cdot \text{K}^{-1}$	[77]
$U_w$	U-value for walls, ceiling against outside (old build)	1.5	$\text{W} \cdot \text{m}^{-2} \cdot \text{K}^{-1}$	[77]
$U_{wd}$	U-value for windows (old build)	4.3	$\text{W} \cdot \text{m}^{-2} \cdot \text{K}^{-1}$	[77]
$\epsilon$	height correction factor	1	-	[79]
$\Lambda_{\text{at}}$	dimensionless ratio between the internal surfaces area and the floor area	4.5	[ $\cdot$ ]	[57]
$\tau$	time step	3600	s	

Table 3: Input parameters to the heating model

We obtain the following intermediate parameter values:

- Effective mass area  $A_m$  [ $\text{m}^2$ ] for a medium-class building [57]:

$$A_m = 2.5 A_f \quad (30)$$

- Internal heat capacity of the building zone for medium-class building  $\text{J} \cdot \text{K}^{-1}$  [57]:

$$C_m = 165,000 A_f \quad (31)$$

- Area of all surfaces facing the building zone  $A_{\text{tot}}$  [ $\text{m}^2$ ] [57]:

$$A_{\text{tot}} = \Lambda_{\text{at}} A_f \quad (32)$$

- The coupling conductance [ $\text{W} \cdot \text{K}^{-1}$ ] [57]:

$$H_{\text{tr},is} = h_{is} A_{\text{tot}} \quad (33)$$

- The coupling conductance between nodes m and s [ $\text{W} \cdot \text{K}^{-1}$ ] [57]:

$$H_{\text{tr},ms} = h_{ms} A_m \quad (34)$$

- Wall area (excluding windows and doors)

$$A_w = 4\sqrt{A_f}h - 8A_{wd} - A_d \quad (35)$$

- The thermal transmission coefficient of walls [W.K<sup>-1</sup>] [57]:

$$H_{tr,w} = A_w U_w \quad (36)$$

- The thermal transmission coefficient of the roof [W.K<sup>-1</sup>] [57]:

$$H_{tr,r} = A_f U_r \quad (37)$$

- The thermal transmission coefficient of the floor [W.K<sup>-1</sup>] [57]:

$$H_{tr,f} = A_f(1 - k_{party})U_g \quad (38)$$

- The heat transfer coefficient for opaque elements  $H_{tr,op}$  [W.K<sup>-1</sup>] [57]:

$$H_{tr,op} = H_{tr,w} + H_{tr,r} + H_{tr,f} \quad (39)$$

- The opaque heat transfer coefficient is split between conductance transfer and  $H_{tr,em}$  [57]:

$$H_{tr,em} = \frac{1}{\frac{1}{H_{tr,op}} - \frac{1}{H_{tr,ms}}} \quad (40)$$

- The thermal transmission coefficient of windows [W.K<sup>-1</sup>] [57]:

$$H_{tr,wd} = (A_{wd} + A_d)U_{wd,eff} \quad (41)$$

- The conditioned air volume [m<sup>3</sup>]

$$V_r = A_f h \quad (42)$$

- The hygiene minimum air flow rate of a heated space  $V_{min}$  [m<sup>3</sup>.h<sup>-1</sup>] [79]:

$$V_{min} = n_{min} V_r \quad (43)$$

- The infiltration through building envelope  $V_{inf}$  [m<sup>3</sup>.h<sup>-1</sup>] [79]:

$$V_{inf} = 2V_r n_{50} e \epsilon \quad (44)$$

- The air flow rate of heated space [m<sup>3</sup>.h<sup>-1</sup>] [79]:

$$V = \max(V_{min}, V_{inf}) \quad (45)$$

- The heat transfer by ventilation  $H_{ve}$  [W.K<sup>-1</sup>] [79]:

$$H_{ve} = 0,34V \quad (46)$$

- The effective window U-value, corrected for the assumed use of curtains [W.m<sup>-2</sup>.K<sup>-1</sup>] [57]:

$$U_{wd,eff} = \frac{1}{\frac{1}{U_{wd}} + 0.04} \quad (47)$$

- Three helper transmission coefficients [57]:

$$H_{tr,1} = \frac{1}{\frac{1}{H_{ve}} + \frac{1}{H_{tr,is}}} \quad (48)$$

$$H_{tr,2} = H_{tr,1} + H_{tr,w} \quad (49)$$

$$H_{tr,3} = \frac{1}{\frac{1}{H_{tr,2}} + \frac{1}{H_{tr,ms}}} \quad (50)$$

- The heat flow rate from internal heat sources  $\Phi_{\text{int}}$  [W] is taken as the sum of the average heat flow rate from appliances  $\Phi_{\text{int},A}$  and occupants  $\Phi_{\text{int},OC}$  [57]:

$$\Phi_{\text{int}} = \Phi_{\text{int},A} + \Phi_{\text{int},OC} = 2A_f + 1.5A_f = 3.5A_f \quad (51)$$

- the part of the heat flow rate from internal heat sources going to the air node  $\theta_{\text{int}}$   $\Phi_{\text{ia}}$  [W] [57]

$$\Phi_{\text{ia}} = \frac{1}{2}\Phi_{\text{int}} \quad (52)$$

Given these input parameters, the Crank-Nicholson scheme is defined in [57] is applied. We seek to find the temperature of the internal air node  $\theta_{\text{air}}$  [ $^{\circ}\text{C}$ ] and of the building mass  $\theta_{m,t}$  at each time step given the heating or cooling power  $\Phi_{\text{HC}}$  (positive for heating and negative for cooling), the external air temperature  $\theta_e$  [ $^{\circ}\text{C}$ ] and the heat flow rates from solar heat sources  $\Phi_{\text{sol}}$ .

The air node temperature  $\theta_{\text{air}}$  is given as

$$\theta_{\text{air}} = \frac{H_{\text{tr,is}}\theta_s + H_{\text{ve}}\theta_{\text{sup}} + \Phi_{\text{ia}} + \Phi_{\text{HC}}}{H_{\text{tr,is}} + H_{\text{ve}}} \quad (53)$$

Where the surface node temperature  $\theta_s$  is defined as:

$$\theta_s = \frac{H_{\text{tr,ms}}\theta_m + \Phi_{\text{st}} + H_{\text{tr,w}}\theta_e + H_{\text{tr,1}}\left(\theta_{\text{sup}} + \frac{\Phi_{\text{ia}} + \Phi_{\text{HC}}}{H_{\text{ve}}}\right)}{H_{\text{tr,ms}} + H_{\text{tr,w}} + H_{\text{tr,1}}} \quad (54)$$

The average temperature over the hour of the building mass  $\theta_m$ :

$$\theta_m = \frac{1}{2}(\theta_{m,t-1} + \theta_{m,t}) \quad (55)$$

$$\theta_{m,t} = \frac{\theta_{m,t-1}\left(\frac{C_m}{\tau} + \frac{1}{2}(H_3 + H_{\text{tr,em}})\right) + \Phi_{\text{mtot}}}{\frac{C_m}{\tau} + \frac{1}{2}(H_{\text{tr,3}} + H_{\text{tr,em}})} \quad (56)$$

$$\Phi_{\text{mtot}} = \Phi_m + H_{\text{tr,em}}\theta_e + H_{\text{tr,3}}\frac{\Phi_{\text{st}} + H_{\text{tr,w}}\theta_e + H_{\text{tr,1}}\left(\frac{\Phi_{\text{ia}} + \Phi_{\text{HC}}}{H_{\text{ve}}} + \theta_{\text{sup}}\right)}{H_{\text{tr,2}}} \quad (57)$$

The part of heat flow rates from internal and solar heat sources going to the internal nodes  $\theta_s$

$$\Phi_{\text{st}} = \left(1 - \frac{A_m}{A_t} - \frac{H_{\text{tr,w}}}{9.1A_t}\right)\left(\frac{1}{2}\Phi_{\text{int}} + \Phi_{\text{sol}}\right) \quad (58)$$

The part of heat flow rates from internal and solar heat sources going to the internal nodes  $\theta_m$

$$\Phi_m = \frac{A_m}{A_t}\left(\frac{1}{2}\Phi_{\text{int}} + \Phi_{\text{sol}}\right) \quad (59)$$

$$\theta_{\text{sup}} = \theta_e \quad (60)$$

We rearrange the equations of this model in order to obtain a linear recursive formulation. We first define some helper variables:

$$A = \frac{C_m}{\tau} + \frac{1}{2}(H_{\text{tr,3}} + H_{\text{tr,em}}) \quad (61)$$

$$B = 1 - \frac{A_m}{A_t} - \frac{H_{\text{tr,w}}}{9.1A_t} \quad (62)$$

$$C = \frac{B\Phi_{\text{int}}}{2} \quad (63)$$

$$D = \frac{A_m\Phi_{\text{int}}}{2A_t} + \frac{H_{\text{tr,3}}}{H_{\text{tr,2}}}\left(C + \frac{H_{\text{tr,1}}\Phi_{\text{ia}}}{H_{\text{ve}}}\right) \quad (64)$$

$$E = H_{\text{tr,em}} + \frac{H_{\text{tr,3}}}{H_{\text{tr,2}}}(H_{\text{tr,w}} + H_{\text{tr,1}}) \quad (65)$$

$$H_{\text{tr,ms}} + H_{\text{tr,w}} + H_{\text{tr,1}} \quad (66)$$

$$G = \frac{1}{F} \left( \frac{H_{\text{tr,ms}} a_{\text{T}}}{2} + C + \frac{H_{\text{tr},1} \Phi_{\text{ia}}}{H_{\text{ve}}} \right) \quad (67)$$

$$H = \frac{H_{\text{tr,ms}}}{2F} (1 + b_{\text{T}}) \quad (68)$$

$$I = \frac{1}{F} \left( \frac{H_{\text{tr,ms}} c_{\text{T}}}{2} + H_{\text{tr,w}} + H_{\text{tr},1} \right) \quad (69)$$

$$J = \frac{1}{F} \left( \frac{H_{\text{tr,ms}} d_{\text{T}}}{2} + B \right) \quad (70)$$

$$K = \frac{1}{F} \left( \frac{H_{\text{tr,ms}} e_{\text{T}}}{2} + \frac{H_{\text{tr},1}}{H_{\text{ve}}} \right) \quad (71)$$

$$a_{\text{T}} = \frac{D}{A} \quad (72)$$

$$b_{\text{T}} = \frac{\left( \frac{C_m}{\tau} + 0.5(H_3 + H_{\text{tr,em}}) \right)}{A} \quad (73)$$

$$c_{\text{T}} = \frac{E}{A} \quad (74)$$

$$d_{\text{T}} = \frac{\frac{A_m}{A_t} + \frac{H_{\text{tr},3}B}{H_{\text{tr},2}}}{A} \quad (75)$$

$$e_{\text{T}} = \frac{H_{\text{tr},3}H_{\text{tr},1}}{H_{\text{tr},2}H_{\text{ve}}A} \quad (76)$$

$$a_{\text{air}} = \frac{H_{\text{tr,is}}G + \Phi_{\text{ia}}}{H_{\text{tr,is}} + H_{\text{ve}}} \quad (77)$$

$$b_{\text{air}} = \frac{H_{\text{tr,is}}H}{H_{\text{tr,is}} + H_{\text{ve}}} \quad (78)$$

$$c_{\text{air}} = \frac{H_{\text{tr,is}}I + H_{\text{ve}}}{H_{\text{tr,is}} + H_{\text{ve}}} \quad (79)$$

$$d_{\text{air}} = \frac{H_{\text{tr,is}}J}{H_{\text{tr,is}} + H_{\text{ve}}} \quad (80)$$

$$e_{\text{air}} = \frac{H_{\text{tr,is}}K + 1}{H_{\text{tr,is}} + H_{\text{ve}}} \quad (81)$$

$$\kappa = \begin{bmatrix} a_{\text{T}} & b_{\text{T}} & c_{\text{T}} & d_{\text{T}} & e_{\text{T}} \\ a_{\text{air}} & b_{\text{air}} & c_{\text{air}} & d_{\text{air}} & e_{\text{air}} \end{bmatrix} \quad (82)$$

Rearranging eqs. (58), (62) and (63):

$$\Phi_{\text{st}} = C + B\Phi_{\text{sol}} \quad (83)$$

Rearranging eqs. (57), (59), (60) and (83):

$$\begin{aligned} \Phi_{\text{mtot}} = & \left( \frac{A_m}{2A_t} \Phi_{\text{int}} + \frac{A_m}{A_t} \Phi_{\text{sol}} \right) + H_{\text{tr,em}} \theta_e + \frac{H_{\text{tr},3}}{H_{\text{tr},2}} (C + B\Phi_{\text{sol}}) + \\ & \frac{H_{\text{tr},3}}{H_{\text{tr},2}} H_{\text{tr,w}} \theta_e + \frac{H_{\text{tr},3}}{H_{\text{tr},2}} H_{\text{tr},1} \left( \frac{\Phi_{\text{ia}} + \Phi_{\text{HC}}}{H_{\text{ve}}} + (\theta_e) \right) \end{aligned} \quad (84)$$

$$\begin{aligned} \Phi_{\text{mtot}} = & \frac{A_m \Phi_{\text{int}}}{2A_t} + \frac{A_m}{A_t} \Phi_{\text{sol}} + H_{\text{tr,em}} \theta_e + \frac{H_{\text{tr},3}C}{H_{\text{tr},2}} + \frac{H_{\text{tr},3}B}{H_{\text{tr},2}} \Phi_{\text{sol}} + \frac{H_{\text{tr},3}H_{\text{tr,w}}}{H_{\text{tr},2}} \theta_e + \\ & \frac{H_{\text{tr},3}H_{\text{tr},1}\Phi_{\text{ia}}}{H_{\text{tr},2}H_{\text{ve}}} + \frac{H_{\text{tr},3}H_{\text{tr},1}}{H_{\text{tr},2}H_{\text{ve}}} \Phi_{\text{HC}} + \frac{H_{\text{tr},3}H_{\text{tr},1}}{H_{\text{tr},2}} \theta_e \end{aligned} \quad (85)$$

$$\begin{aligned}\Phi_{\text{mtot}} = & \left( \frac{A_m \Phi_{\text{int}}}{2A_t} + \frac{H_{\text{tr},3}}{H_{\text{tr},2}} \left( C + \frac{H_{\text{tr},1} \Phi_{\text{ia}}}{H_{\text{ve}}} \right) \right) + \left( \frac{A_m}{A_t} + \frac{H_{\text{tr},3} B}{H_{\text{tr},2}} \right) \Phi_{\text{sol}} \\ & + \left( H_{\text{tr},\text{em}} + \frac{H_{\text{tr},3}}{H_{\text{tr},2}} (H_{\text{tr},w} + H_{\text{tr},1}) \right) \theta_e + \frac{H_{\text{tr},3} H_{\text{tr},1}}{H_{\text{tr},2} H_{\text{ve}}} \Phi_{\text{HC}}\end{aligned}\quad (86)$$

Rearranging eqs. (64), (65) and (86):

$$\Phi_{\text{mtot}} = D + \left( \frac{A_m}{A_t} + \frac{H_{\text{tr},3} B}{H_{\text{tr},2}} \right) \Phi_{\text{sol}} + E \theta_e + \frac{H_{\text{tr},3} H_{\text{tr},1}}{H_{\text{tr},2} H_{\text{ve}}} \Phi_{\text{HC}} \quad (87)$$

From eqs. (56) and (61)

$$\theta_{m,t} = \frac{\left( \frac{C_m}{\tau} + 0.5(H_3 + H_{\text{tr},\text{em}}) \right)}{A} \theta_{m,t-1} + \frac{\Phi_{\text{mtot}}}{A} \quad (88)$$

From eqs. (87) and (88)

$$\begin{aligned}\theta_{m,t} = & \frac{D}{A} + \frac{\left( \frac{C_m}{\tau} + 0.5(H_3 + H_{\text{tr},\text{em}}) \right)}{A} \theta_{m,t-1} + \frac{E}{A} \theta_e + \frac{\frac{A_m}{A_t} + \frac{H_{\text{tr},3} B}{H_{\text{tr},2}}}{A} \Phi_{\text{sol}} + \\ & \frac{H_{\text{tr},3} H_{\text{tr},1}}{H_{\text{tr},2} H_{\text{ve}} A} \Phi_{\text{HC}}\end{aligned}\quad (89)$$

From eqs. (72) to (76) and (89):

$$\theta_{m,t} = a_T + b_T \theta_{m,t-1} + c_T \theta_e + d_T \Phi_{\text{sol}} + e_T \Phi_{\text{HC}} \quad (90)$$

From eqs. (55) and (90):

$$\theta_m = \frac{1}{2} (\theta_{m,t-1} + a_T + b_T \theta_{m,t-1} + c_T \theta_e + d_T \Phi_{\text{sol}} + e_T \Phi_{\text{HC}}) \quad (91)$$

$$\theta_m = \frac{a_T}{2} + \frac{1+b_T}{2} \theta_{m,t-1} + \frac{c_T}{2} \theta_e + \frac{d_T}{2} \Phi_{\text{sol}} + \frac{e_T}{2} \Phi_{\text{HC}} \quad (92)$$

From eqs. (54), (60), (66), (83) and (92)

$$\begin{aligned}\theta_s = & \frac{H_{\text{tr},\text{ms}} a_T}{2F} + \frac{H_{\text{tr},\text{ms}}}{F} \frac{1+b_T}{2} \theta_{m,t-1} + \frac{H_{\text{tr},\text{ms}} c_T}{2F} \theta_e + \frac{H_{\text{tr},\text{ms}} d_T}{2F} \Phi_{\text{sol}} + \\ & \frac{H_{\text{tr},\text{ms}} e_T}{2F} \Phi_{\text{HC}} + \frac{C}{F} + \frac{B}{F} \Phi_{\text{sol}} + \frac{H_{\text{tr},w}}{F} \theta_e + \frac{H_{\text{tr},1}}{F} \theta_e + \frac{H_{\text{tr},1} \Phi_{\text{ia}}}{FH_{\text{ve}}} + \frac{H_{\text{tr},1}}{FH_{\text{ve}}} \Phi_{\text{HC}}\end{aligned}\quad (93)$$

$$\begin{aligned}\theta_s = & \frac{1}{F} \left( \frac{H_{\text{tr},\text{ms}} a_T}{2} + C + \frac{H_{\text{tr},1} \Phi_{\text{ia}}}{H_{\text{ve}}} \right) + \frac{H_{\text{tr},\text{ms}}}{2F} (1+b_T) \theta_{m,t-1} + \\ & \frac{1}{F} \left( \frac{H_{\text{tr},\text{ms}} c_T}{2} + H_{\text{tr},w} + H_{\text{tr},1} \right) \theta_e + \\ & \frac{1}{F} \left( \frac{H_{\text{tr},\text{ms}} d_T}{2} + B \right) \Phi_{\text{sol}} + \frac{1}{F} \left( \frac{H_{\text{tr},\text{ms}} e_T}{2} + \frac{H_{\text{tr},1}}{H_{\text{ve}}} \right) \Phi_{\text{HC}}\end{aligned}\quad (94)$$

From eqs. (67) to (71) and (94):

$$\theta_s = G + H \theta_{m,t-1} + I \theta_e + J \Phi_{\text{sol}} + K \Phi_{\text{HC}} \quad (95)$$

From eqs. (53), (60) and (95):

$$\begin{aligned}\theta_{\text{air}} = & \frac{H_{\text{tr},\text{is}} G}{H_{\text{tr},\text{is}} + H_{\text{ve}}} + \frac{H_{\text{tr},\text{is}} H}{H_{\text{tr},\text{is}} + H_{\text{ve}}} \theta_{m,t-1} + \frac{H_{\text{tr},\text{is}} I}{H_{\text{tr},\text{is}} + H_{\text{ve}}} \theta_e + \frac{H_{\text{tr},\text{is}} J}{H_{\text{tr},\text{is}} + H_{\text{ve}}} \Phi_{\text{sol}} + \\ & \frac{H_{\text{tr},\text{is}} K}{H_{\text{tr},\text{is}} + H_{\text{ve}}} \Phi_{\text{HC}} + \frac{H_{\text{ve}}}{H_{\text{tr},\text{is}} + H_{\text{ve}}} \theta_e + \frac{\Phi_{\text{ia}}}{H_{\text{tr},\text{is}} + H_{\text{ve}}} + \frac{1}{H_{\text{tr},\text{is}} + H_{\text{ve}}} \Phi_{\text{HC}}\end{aligned}\quad (96)$$

$$\begin{aligned}\theta_{\text{air}} = & \frac{H_{\text{tr},\text{is}} G + \Phi_{\text{ia}}}{H_{\text{tr},\text{is}} + H_{\text{ve}}} + \frac{H_{\text{tr},\text{is}} H}{H_{\text{tr},\text{is}} + H_{\text{ve}}} \theta_{m,t-1} + \frac{H_{\text{tr},\text{is}} I + H_{\text{ve}}}{H_{\text{tr},\text{is}} + H_{\text{ve}}} \theta_e + \frac{H_{\text{tr},\text{is}} J}{H_{\text{tr},\text{is}} + H_{\text{ve}}} \Phi_{\text{sol}} + \\ & \frac{H_{\text{tr},\text{is}} K + 1}{H_{\text{tr},\text{is}} + H_{\text{ve}}} \Phi_{\text{HC}}\end{aligned}\quad (97)$$

From eqs. (77) to (81) and (97):

$$\theta_{\text{air}} = a_{\text{air}} + b_{\text{air}}\theta_{m,t-1} + c_{\text{air}}\theta_e + d_{\text{air}}\Phi_{\text{sol}} + e_{\text{air}}\Phi_{\text{HC}} \quad (98)$$

Note that the notation from [57] was used in this appendix. In this paper,  $T_{\text{air},i}^{t+1} \leftarrow \theta_{\text{air}}$ ,  $T_{m,i}^{t+1} \leftarrow \theta_{m,t}$ ,  $T_{m,i}^t \leftarrow \theta_{m,t-1}$ ,  $T_e^t \leftarrow \theta_e$ ,  $\Phi^t \leftarrow \Phi_{\text{sol}}$ ,  $h_i^t \leftarrow \Phi_{\text{HC}}$ , such that from eqs. (82), (90) and (98):

$$\begin{bmatrix} T_{m,i}^{t+1} \\ T_{\text{air},i}^{t+1} \end{bmatrix} = \kappa [1, T_{m,i}^t, T_e^t, \Phi^t, h_i^t]^\top \quad (99)$$

This is equivalent to eq. (15).

## Appendix D Residential energy management: commented illustrative day

Here we look in detail at the actions selected by an agent which learned to coordinate using different MARL strategies. This is meant to illustrate how example RL  $\psi$  actions translate into local energy management system behaviour. Note however that the MARL algorithms aim to generate statistically favourable outcomes when averaged over longer durations and over larger number of agents. As such, while the average outcomes are predictable, as described in Section 6, this individual case is not meant to be generally representative but rather simply an example day in a stochastic environment.

Figure 5 shows an example of an evaluation day during which the final policies learned is used deterministically on a day-long batch of data. Four different policies are compared:

- Baseline: no flexibility used
- Optimal: the actions selected by a central optimiser with perfect knowledge and control of all current and future variables
- CO: a policy seeking to replicate action patterns by the optimiser by counting the number of actions taken for each grid coefficient level (state) during pre-learning (see Section 4)
- MO: learning from optimisations and using marginal rewards (see Section 4)

The baseline and optimal act as reference points while the two latter policies have been identified in Section 6 as scalable policies when the number of agents increases. While both policies reduce costs relative to the baseline on average at scale, MO was shown to be the best-performing policy at scale.

Subplot A shows the wholesale prices and the grid carbon intensity for the example day, as well as the resulting grid cost coefficient  $C_g$  given a social cost of carbon of 70 £/tCO<sub>2</sub>. This coefficient informs the choice of action  $\psi$ .

Subplot B shows the action  $\psi$  selected by each policy over time. The CO policy seeks to imitate patterns by the optimiser – though with more limited information than is available to the optimiser – and takes more extreme actions, for example with maximum delaying of consumption ( $\psi = 0$ ) at time intervals with high network cost coefficient  $C_g^t$ , whereas the MO policy selects intermediate  $\psi$  values.

Subplot C shows the total energy consumption over time. This includes both household loads and heating consumption. MO takes intermediate  $\psi$  actions and so follows more closely the baseline, non-flexible consumption profile than CO which delays more loads when taking lower  $\psi$  values. Both strategies are seen to shave consumption peaks and/or displace them to lower-price time intervals. As the total household electric demand is fixed, displacing consumption does not increase total consumption. However, variation in heating loads within the acceptable temperature bounds may increase overall consumption. Thus, in this example day the MO strategy consumes more 26.1% more energy than the baseline, though overall incurring lower costs and greenhouse gas emissions.

Subplot D shows cumulative rewards over time each of the policies. While the CO strategy was seen to take advantage more closely of grid price differentials, overall the costs incurred are higher than with the MO strategy. With the MO strategy, savings of £1.20 are obtained compared to the baseline over the example day, corresponding to a 57.5% reduction from baseline costs. 86.3% of savings stem from reduced battery depreciation, 22.3% from reduced distribution grid congestion, while grid energy costs increased by 8.6%. In this example day the CO strategy achieved savings of £0.76, with 66.1% stemming from reduced grid energy costs, 30.0% from reduced battery costs, and 3.9% from reduced grid congestion.

An interplay is thus illustrated by the two policies between the costs of battery depreciation and distribution network congestion on the one hand, and the opportunity for energy arbitrage to save on grid energy and emissions costs on the other. Both the MO and CO strategies exhibit stable performance at scale, though converging to different types of policy. The MO policy saves more by smoothing out the charging and distribution grid utilisation profiles despite smaller savings in imports and emissions costs, while CO derives a larger advantage from the grid price differentials in grid imports, though with higher battery and distribution grid costs. The weight applied on each of those competing objectives in the objective function will have a direct impact on the policies that are learned.

Subplot E shows the heating energy profile, resulting in the temperatures in subplot F. The baseline profile maintains the median desired temperature, whereas the flexible policies can go above or below that median, within the desired temperature bounds. Both policies are more likely to absorb energy imports through heating (no marginal costs) rather than storage (battery depreciation costs) relative to the baseline policy. Consumption peaks are shaved or displaced to lower-cost time intervals with both policies.

Subplot G shows the EV at-home availability and consumption. In this example day, the electric vehicle (EV) leaves home at 9 am, consumes 2.2 kWh on an outbound trip, remains parked at its destination for 1 hour, and consumes 2.2 kWh on the inbound trip back to the home at 12 pm. The car can therefore not be charged during this time interval, and enough charge has to be available beforehand for these travelling loads.

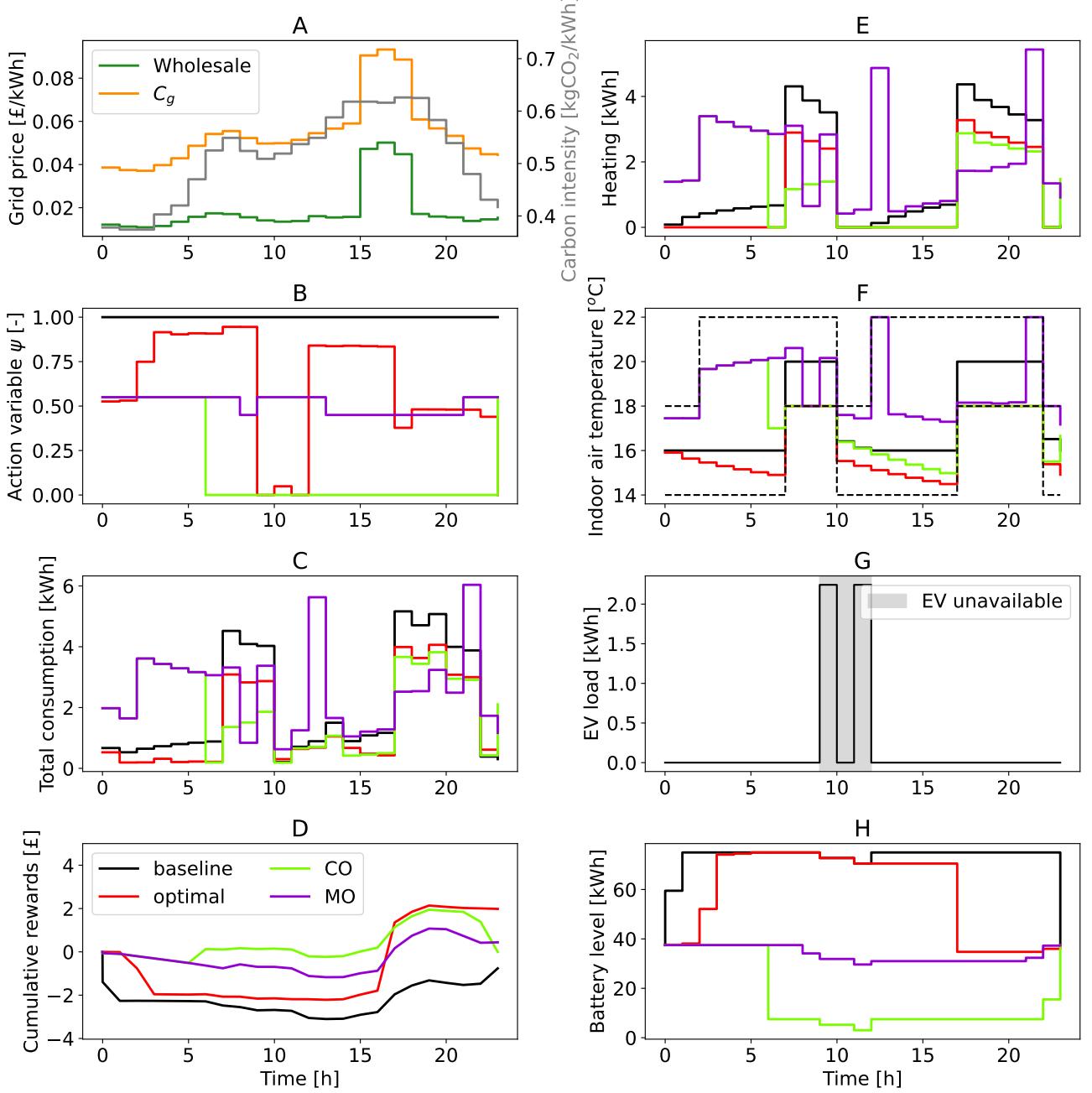


Figure 5: Example of local home energy system variables for one agent using different policies learned ahead of implementation (optimisation-based learning using marginal rewards (MO) and optimiser state-action pairs selection counts (CO), compared with the inflexible baseline and omniscient centralised optimiser control actions

Subplot H shows the battery level profiles. In the baseline, the EV is charged as soon as it is plugged in, given battery capacity and charging rate constraints. The CO policy sells energy from the battery when prices increase to take advantage of the price differentials, whereas the MO policy flattens out the charging profile.