Prospects for Reinforcement Learning

A guide for energy sector applications

Christopher Lee, Dr. Charlotte Avery, Samuel Young
Digital Team

April 2023

Content

DISCLAIMER

This document has been prepared by Energy Systems Catapult Limited. For full copyright, legal information and defined terms, please refer to the "Licence / Disclaimer" section at the back of this document.

All information is given in good faith based upon the latest information available to Energy Systems Catapult Limited. No warranty or representation is given concerning such information, which must not be taken as establishing any contractual or other commitment binding upon the Energy Systems Catapult Limited or any of its subsidiary or associated companies.

Glossary

Acronyms

Shorthand	Expression
RL	Reinforcement learning
ML	Machine Learning
AI	Artificial Intelligence
MPC	Model Predictive Control
EV	Electric Vehicle
DSR	Demand Side Response
DHW	Domestic Hot Water
HVAC	Heating, Ventilation and Air Conditioning

Definitions

Term	Definition
AI	A form of automation where computers can perform tasks which would otherwise require human intelligence. Encompasses fields such as ML, optimisation, making predictions/forecasting, analysis and much more.
ML	Methods by which computers can derive information without an explicit set of instructions.
Digital Twin	A digital model which integrates two-way data flow between the model and physical object or system. Making a change to one can change the other.
Transfer Learning	Transfer learning is a field of research in machine learning that centres on retaining the knowledge acquired from solving one problem and utilizing it to address a different but connected problem. As an illustration, the knowledge acquired from identifying cars could be transferred and utilised to identify trucks.
Online learning	Learning while interacting with the real physical environment.
Offline learning	Learning prior to interacting with the real physical environment. For the purpose of this paper, that could be learning from historical data logs, using a Digital Twin, or from a source environment and using it for transfer learning.
Stochasticity	A probability distribution or pattern that is subject to statistical analysis but cannot be accurately predicted, due to its random nature.
DSR	Demand-side response: energy consumers altering their energy consumption behaviour in response to large-scale demand on the grid.
Model Predictive Control	Iterative optimal control technique which minimises a cost function which evaluates future states within a horizon, whilst manipulating variable inputs which define behaviours across a given time interval.

1. Executive summary

Reinforcement learning (RL) is a branch of machine learning (ML) that focuses on optimal decision making through trial and error— analogous to a child exploring and learning in their surrounding world. We are starting to see promising applications of RL for control systems within the energy sector, including Deepmind's cooling systems [1], and Carbon Re's cement production [2]. In addition, recent advances in 'Reinforcement Learning from Human Feedback' (e.g. in refining the ChatGPT core model to respond more appropriately) have the potential to support embedding of expert knowledge and consumer preferences into existing energy sector models. It is an exciting time for RL within the energy sector, but it is still early days and much of the knowledge exists in highly technical academic literature or is seldom shared by industry. Therefore, this paper aims to bridge a gap between industry and academia by presenting the most salient advantages and challenges of RL, providing a framework to guide the reader on how to scope an RL approach to a given problem, and highlighting the landscape of near-term and future applications through a series of industry and academic use cases. We hope to encourage innovators to see the potential for RL within the energy sector.

Here we summarise key takeaways from this paper:

1) RL is primarily useful for optimal decision making and control within complex systems which have underlying uncertainty or randomness within the system. It is an attractive approach owing to several advantages it holds over alternative methods:

Key advantages of RL

Uncertainty

RL naturally handles uncertainty, which is often more restrictive in alternative methods.

Complex systems

RL can handle complex systems that are difficult to model.

Human feedback

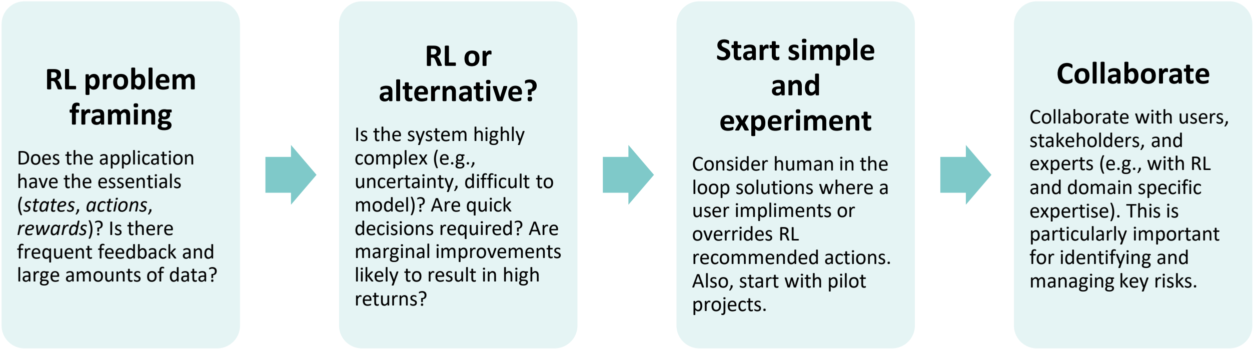
Human feedback and preferences can be incorporated, which is difficult to capture in alternative methods.

Quick decision making

Computations can occur prior to deployment when the algorithm is learning. Once deployed, quick decisions are made in real time.

2) Drawing from a review of the literature, use cases, and expert consultation, we provide a general framework below for using RL for a particular application:

Framework for using RL



RL problem framing

Does the application have the essentials (*states*, *actions*, *rewards*)? Is there frequent feedback and large amounts of data?

RL or alternative?

Is the system highly complex (e.g., uncertainty, difficult to model)? Are quick decisions required? Are marginal improvements likely to result in high returns?

Start simple and experiment

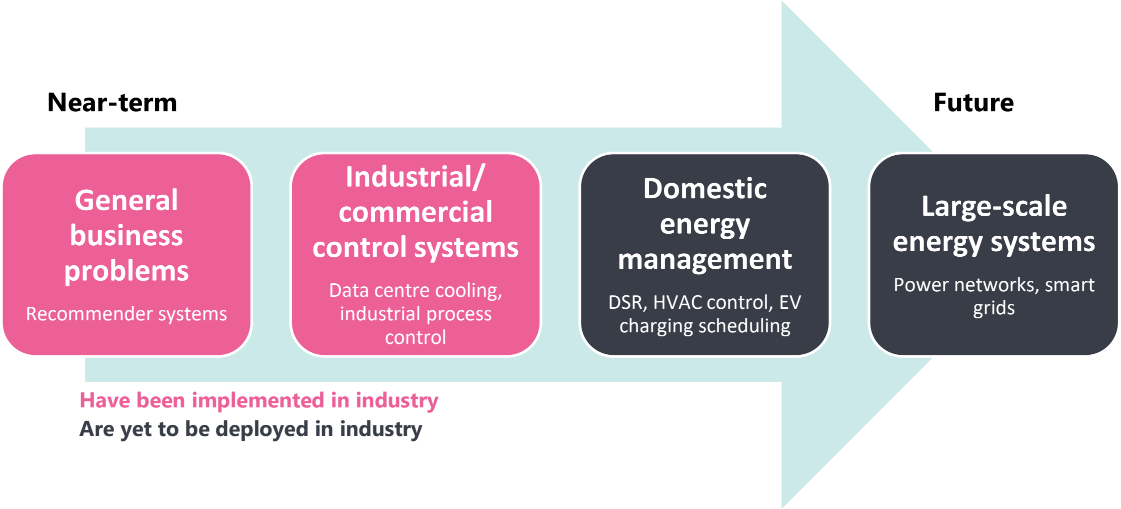
Consider human in the loop solutions where a user impliments or overrides RL recommended actions. Also, start with pilot projects.

Collaborate

Collaborate with users, stakeholders, and experts (e.g., with RL and domain specific expertise). This is particularly important for identifying and managing key risks.

3) RL has potential to help create a more flexible, low-carbon approach to using energy. In Sections 6 and 7, we present a range of use cases drawing from both the industry examples and use cases demonstrated in academic literature. Below we show an overview of the potential near-term and future applications of RL in the energy sector. RL has been used in general business problems, which can indirectly benefit the energy sector (e.g. improving customer journeys using recommender systems). It has also had direct impact on the energy sector through industrial/commercial control systems. Domestic and large-scale applications tend to remain in the research phase, as there are stricter safety, regulatory and security hurdles to overcome.

Landscape for RL applications



Near-term

Future

General business problems

Recommender systems

Industrial/ commercial control systems

Data centre cooling, industrial process control

Domestic energy management

DSR, HVAC control, EV charging scheduling

Large-scale energy systems

Power networks, smart grids

Have been implemented in industry

Are yet to be deployed in industry

4) It's important to recognise that RL is still a developing field and faces many challenges centred around safety, practical/technical limitations, and the need for confidence and trust.

Key challenges of RL

Learning with limited data

Real-world systems are limited in data generation, unlike simulating millions of games like AlphaGo.

Real-world consequences and safety constraints

Real-world systems present safety risks to equipment and users that are not present in virtual environments.

Offline learning

Learning from historical data logs can reduce safety risks by reducing interactions with the physical environment but is often limited in how much can be learnt.

Interpretability and reliability

Users often require interpretable solutions for reassurance, but RL is often a black box. RL can also result in occasional extreme behavior in an unpredictable way.

Who is this paper for?

This paper is designed as a useful guide for business leaders within the energy sector considering experimenting with RL or RL practitioners outside of energy who might be interested in making impact in the energy space. We conclude with two calls to action:

- **Industry**: As much as possible, share challenges and successes of real-world implementations of RL to inform future applications — see Deepmind's work [3] as an example. Consider sharing datasets that can be used to develop RL solutions.
- **Academia**: Conduct fair comparisons between RL and alternatives to better understand where RL might excel in the future and understand how they can potentially complement each other.

2. Introduction

Reinforcement Learning (RL) is a type of Machine learning (ML) and Artificial Intelligence (AI). AI and ML have gained vast amounts of attention in the last decade due to their ability to transform various sectors and create business value. Examples include the facial recognition algorithms used by Apple and language translation used by Google. More recently, ML has been transforming the energy sector [4] with applications in solar photovoltaic forecasting [5] and in segmenting customers for custom energy tariffs utilising smart meter data [6].

Many of these applications are only achievable due to the breakthroughs in *deep neural networks* in 2012 [7] and increases in computational power. Huge progress was further made upon the invention of the *transformer* [8] in 2017 leading to breakthroughs in human-like large language models like OpenAI's GPT-3, which can generate poems, stories and even code. Transformers also enabled the development of Deepmind's Alphafold [9], solving the decades-long challenge of protein folding, which can help tackle diseases and discover new medicines faster.

The examples described above fall under branches of ML known as *supervised* and *unsupervised learning*. RL examples are less common but, nonetheless, RL offers the potential to provide powerful solutions to industry problems. RL has been most notably recognised in the application of game playing with Google Deepmind's AlphaGo. AlphaGo involved training RL on thousands of amateur Go games (and later millions more playing against itself) until it could consistently beat the top human Go players. It discovered effective Go strategies not previously used by human players – strategies human players are now starting to adopt.

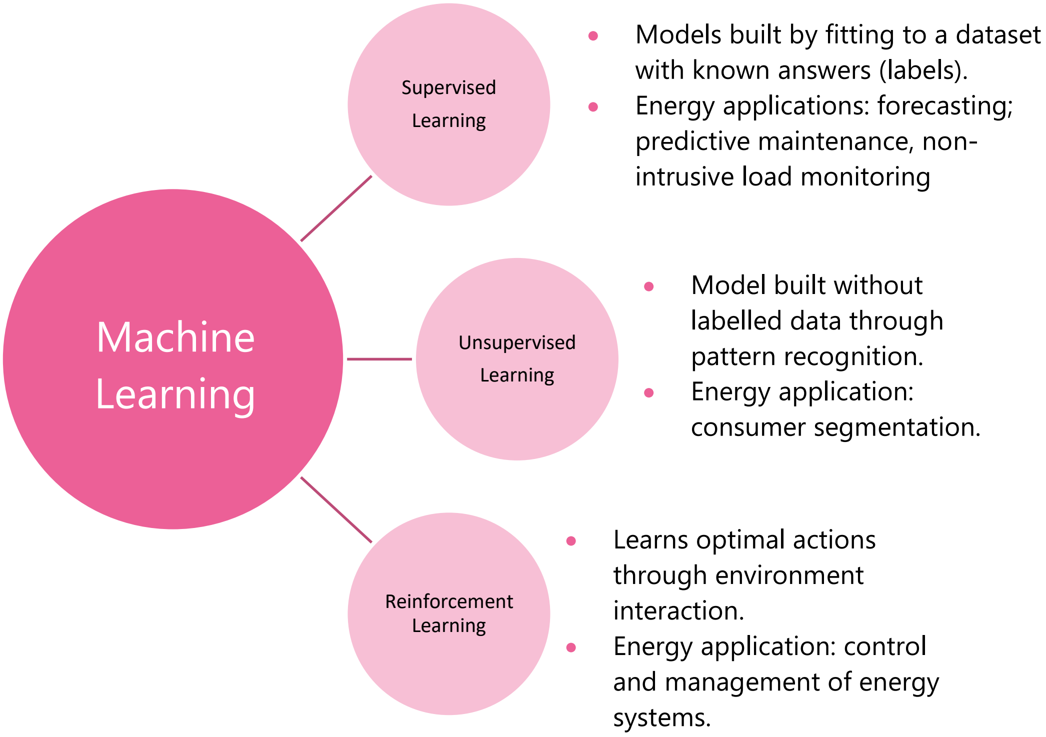
More recently, the use of RL to refine text generation outputs has helped propel language generation models like ChatGPT into the mainstream. The Large Language Model that underpins ChatGPT was primarily developed using other forms of ML but in its raw form it often struggled to align its outputs with human intent and values. One of ChatGPT's critical breakthroughs was to use RL to incorporate human feedback into that base model. Whilst generative models like ChatGPT themselves have many potential uses in the energy sector (which are largely outside the scope of this report), this approach of incorporating human feedback into models using RL could also offer significant benefits to applications in the sector.

With the increasing success of ML applications in the energy sector and the progress in RL applications, now is an appropriate time to review the current state of RL applications within the energy sector. At the same time, Digital Twin technologies have been emerging which can greatly compliment and enable RL. This paper presents an overview of RL (Section 3), its associated challenges (Section 3.1)/potential solutions (Section 3.2), and alternative methodologies (Section 3.3). We provide guidance for innovators and organizations to start thinking about how RL can transform the sector in the near and long-term future (Sections 4 and 5). Sections 6 and 7 highlight promising use cases in industry and discuss where there is expected to be the most potential over the next several years as algorithms and computational methods become increasingly advanced. These risks and ethical considerations involved in deploying RL are considered in Section 8, along with an illustrative example in the energy sector. Lastly, we conclude in Section 9 with a call to action for industry and academia.

Our purpose is to bridge the gap between industry and academia since academic literature reviews are technical, making it difficult for a non-specialist to easily draw insight from them, and often do not reference industry examples. Conversely, industry material is often marketing focused and thus shies away from discussing the nuances/challenges.

3. Reinforcement Learning theory

Reinforcement Learning is a type of Machine learning and Artificial Intelligence that focuses on sequential decision-making tasks to maximise a numeric goal. Figure 1 illustrates the key differences between other ML methods (e.g. supervised and unsupervised learning) and RL. The former methods learn by identifying trends or patterns in a dataset to build a generalised model to make predictions from new, unseen data. RL, on the other hand, focuses on control and learns by interacting with its external environment through a trial-and-error based approach.



- Models built by fitting to a dataset with known answers (labels).
- Energy applications: forecasting; predictive maintenance, non-intrusive load monitoring

- Model built without labelled data through pattern recognition.
- Energy application: consumer segmentation.

- Learns optimal actions through environment interaction.
- Energy application: control and management of energy systems.

Figure 1: Branches of ML and their example uses cases in the field of energy.

In comparison to RL, supervised learning applications, such as predicting solar power, have a clear correct answer that the algorithm will try to predict—the actual power output. RL problems, on the other hand, do not have an obvious unique answer. Using chess as an example, it is not clear what the correct move is, but if you have a long-term goal (winning) then you can learn through trial and error which moves are likely to result in a win—similar to how humans learn.

More formally, an RL problem consists of an *agent* interacting with an *environment* that learns to maximise some objective. It consists of three main components:

- **States**: this provides a picture of what the agent sees in its surrounding environment (e.g. location of pieces on chess board).
- **Actions**: physical changes that the agent can take (e.g. possible pieces to move).
- **Rewards**: feedback signals processed by the agent. The agent acts to maximise the expected future cumulative reward signal.

The agent observes the **state** of its environment, takes certain **actions** within its environments, and receives a **reward** (good or bad) based on its actions; it then tries to maximise its long-term rewards as it learns through trial and error (Figure 2). The final output is known as a *policy*, which is an instruction for the agent to follow by looking at the current state of its environment and deciding which action to take next.

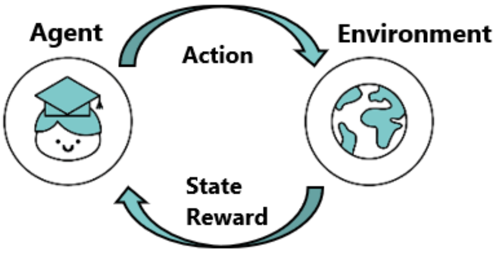


Figure 2: Reinforcement learning schematic

RL algorithms can be classified as either model-free or model-based. Model-free RL does not contain an internal model of the environment that is used to predict changes in state that will result from actions – instead it relies on associations between actions and rewards that have been learned by trial and error. Note that an external 'model' of the environment (e.g. a Digital Twin) can be used to repeatedly simulate taking actions and the resulting reward in order to train the model-free RL, but this remains external to the RL agent; the model-free RL agent is agnostic as to whether the environment is real or a Digital Twin.

Conversely, model-based RL contains an internal model of how likely different actions are to cause transitions into different states. It uses this to do forward-looking short-term predictions of the rewards from various actions and factor these into its decision-making alongside the learned rewards from historical actions. This model of action-state transition probabilities and associated rewards can be defined based on expert knowledge, or trained using supervised learning based on historical transition data.

Model-free RL offers some advantages due to its high flexibility, as it does not require any prior knowledge of the problem (i.e. a model). On the other hand, model-based RL, although less common, is gaining popularity since often it requires less data and can provide greater robustness and interpretability.

3.1. CHALLENGES IN RL

Whilst RL shows great promise, it does not come without its challenges. The implementation of RL in real-world applications faces challenges that are not typically present in virtual applications like game playing (e.g. AlphaGo). These challenges have been documented in several sources [10]–[13] and we have summarised a non-exhaustive list of the main challenges faced when applying RL in the energy sector. Note, many of these challenges are not exclusive to RL; data-driven alternative methods are often subject to similar challenges. Throughout Sections 6 and 7, we describe how these challenges impact specific use cases of RL in the energy sector.

Challenges:

- Learning with limited data
- Real-world consequences and safety constraints
- Offline learning
- Interpretability and reliability

Learning with limited data

Unlike virtual applications like game playing (e.g. AlphaGo), real systems are constrained by the sampling frequency of a physical decision-making process. When the agent is interacting with the physical environment, it cannot speed up time and collect vast amounts of data in the same way as simulating millions of games. Therefore, sample efficient algorithms are incredibly important; that is, algorithms that can learn faster with less data. It is usually not practical to wait several years for an RL agent to learn to perform optimally.

Real-world consequences and safety constraints

Real-world problems have safety, security, and economic consequences that are not present with virtual problems like AlphaGo. Therefore, safety constraints need to be imposed, or a human needs to interpret the recommended actions and act accordingly. However, the need for exploration is essential for RL to find the true optimal solution; this often requires suboptimal actions to be taken, which can result in these unacceptable consequences. Imposed constraints can limit exploration and can result in a suboptimal solution. Additionally, complex safety constraints can be difficult to quantify and can lead to unintended consequences [14], and so cannot guarantee against violations. Furthermore, in a complex environment it becomes difficult for a human to define a solid reward function without loopholes that the agent may slip through – as demonstrated by this list of examples [15]. Before the RL solution is deployed on large-scale in the real world, any potentially dangerous actions the agent may take whilst going through these loopholes need to have been identified.

Safety and ethics should also be a consideration when developing an RL solution. This could come from the study of the behaviour of humans in response to RL. For example, if an RL agent is given a goal, the agent does not care whether it is exploiting human psychological vulnerabilities to maximise that goal, so developers have a responsibility to ensure negative human responses are not invoked because of the RL [16]. See section 8 for a further discussion on this.

Offline learning

For the purpose of this paper, *online* learning or an online deployment refers to the controller being utilised on the live physical system. In the case of online learning, the RL agent interacts with the physical system and collects real-time data to learn from. In contrast to this, RL can also learn *offline* without directly interacting with the physical environment. For example, learning offline from historical logs of data has proven to be an effective way to get the RL agent up to speed, such that when it is deployed online, it is performing relatively well and safely. However, the historical controller (usually based on existing rule-based methods) is typically conservative and suboptimal, and therefore learning is typically restricted.

Interpretability and reliability

As with most machine learning solutions, a RL solution is typically a black-box and the process taken to obtain an output lacks transparency. In other words, the developer cannot clearly explain the steps and reasoning taken by the agent. Especially when an action/recommendation is dramatically different to what an operator may expect, they would typically need some reassurance of the reasoning behind the decision. This is difficult with RL models today. However, some research has focused on using interpretable methods (e.g. decision trees) to map to the learnt black-box policies (e.g. neural networks), which can help with interpretability in the final solution [17]. In terms of reliability, whilst RL solutions can typically provide optimal recommendations/ actions on average, the variability in responses can sometimes lead to unacceptable actions that lead to poor outcomes (e.g. a high penalty or negative reward). These instances can be difficult to explain due to the black-box and stochastic nature of RL.

3.2. HOW TRANSFER LEARNING AND DIGITAL TWINS CAN HELP ADDRESS CHALLENGES

Many of the challenges highlighted centre around safety concerns, limited data, and the need for safe exploration. Two potential ways of addressing these challenges are utilising *transfer learning or Digital Twins*.

Transfer learning is an area of ML that focuses on transferring the knowledge from a *source* environment to help accelerate the learning in a similar *task* environment. This means that an RL agent can be pre-trained with a large amount of data in a safe and controlled source environment. Once deployed in the target environment, it will not be as random and won't require as much data to achieve high performance. This is the same approach used in training large language models like ChatGPT; pretraining occurs in a source environment where the model learns a language by training on large amounts of data scraped from the internet. This understanding of the language can then be transferred to specific tasks and can be finetuned on a small dataset (e.g. customer service documents), allowing it to perform at a high level with minimal data and be optimised for a specific task.

Transfer learning can be valuable when applied to RL in energy. For example, in the application of building controls, an RL agent could be trained in a source environment learning to reduce energy consumption for a particular occupant’s behaviour. This knowledge could then be transferred to a new target building with different occupant behaviour and accelerate learning when compared to starting from scratch.

On the other hand, a Digital Twin can be defined as: ’A digital model which integrates **two-way data** flow between the model and physical object or system, where making a change to one can change the other. For example, a control centre network map which displays real time system status and enables engineers to control assets to mitigate issues ‘[18].

In the context of RL, the RL agent can bridge this two-way data exchange between digital model and physical system. The main benefit of utilising a Digital Twin is the ability to safely explore and optimise a policy in a digital environment, whilst simulating thousands of iterations that would not be possible in the real world. The pairing of Digital Twins and RL is increasingly common in manufacturing and robotics [19], and there are also some use cases in industrial and commercial control systems which are discussed in Section 6.

A possible Digital Twin and RL configuration is shown in Figure 3. The agent can train offline with the digital model, whilst simulating thousands of scenarios considering disturbances in the system and anomalies for the agent to learn to deal with. Once an optimal policy is found it can be pushed to the physical system to execute actions in the real world. Data is fed back to the RL agent in both cases, to learn and train on, but a higher importance weighting can be given to data generated from the physical system to account for any inherent discrepancies with the digital model. Essentially, this is a form of transfer learning as there are inherent differences between the simulated source environment (Digital Twin) and the target environment (real-world); this is also referred to as Sim2Real in the literature [20], which is an open area of research trying to address the challenge of closing the gap between simulation and reality.

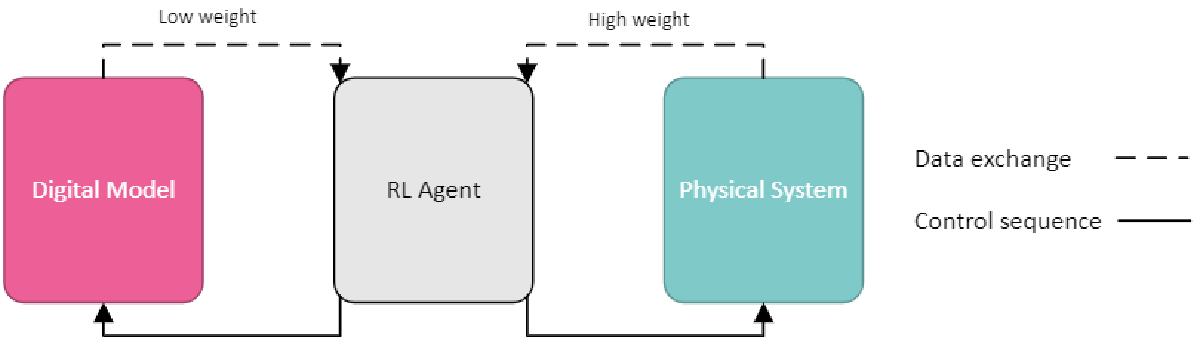


Figure 3: Using Digital Twins alongside RL agents: an example setup.

3.3. HOW DOES RL COMPARE TO ALTERNATIVE METHODS?

Throughout this paper we discuss where RL can potentially outperform alternative approaches, but it is also important to know where it does not (at least at present). We highlight two alternatives commonly used in the use cases presented in the later sections: rule-based control and model predictive control (MPC)/optimisation.

Rule-based

Rule-based controls adopt the form of: *'if x is true, do y'*. For a relatively simple system which does not change unpredictably with time, a rule-based approach may be the best solution given that they are relatively simple to implement at low cost. They are also relatively easy to understand and maintain, and as a result are fairly widely used in industry. These are typically based on domain knowledge, such as a static rule increasing hot water temperature setpoints when it is colder outside to compensate for higher heat losses.

Rule-based controls are therefore a useful benchmark when developing more complex algorithms such as RL. However, energy systems have the potential to be too complex and stochastic for this approach to achieve optimal performance. This is because rule-based approaches are typically not fully adaptive to the environment (e.g. complex weather patterns, building loads) and are therefore often not optimal.

Model Predictive Control/optimisation

MPC is a state-of-the-art method, like RL, which utilises a model to select optimal actions. MPC is a framework that typically consists of a model of a system which forecasts the future state of the system and then uses an optimisation-based solution method (e.g. linear programming, mixed integer programming, evolutionary algorithms) to determine the optimal control parameters that will maximise/minimise the desired objective function. Once the optimal control actions are pushed to the physical system, actual measurements are fed back to update the model, and the process repeats into the next timestep. Figure 4 illustrates this process.

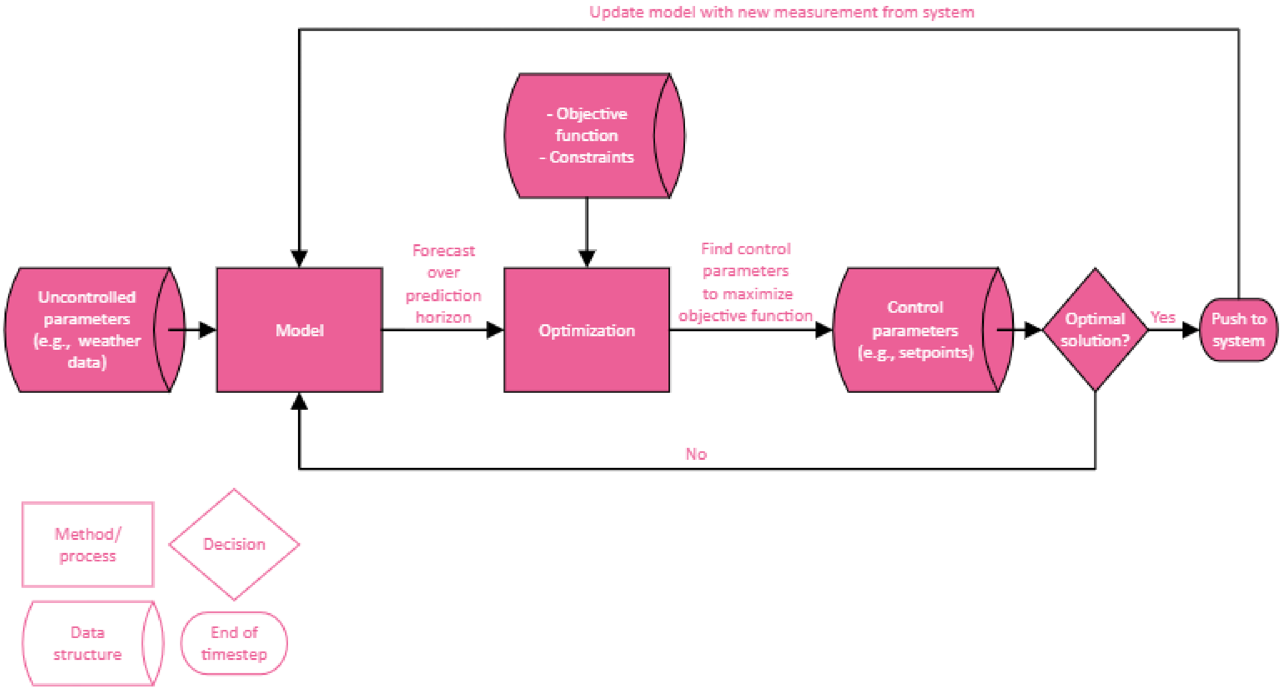

Update model with new measurement from system
- Objective function
- Constraints
Uncontrolled parameters (e.g., weather data)
Model
Forecast over prediction horizon
Optimization
Find control parameters to maximize objective function
Control parameters (e.g., setpoints)
Optimal solution?
Yes
Push to system
No
Method/ process
Decision
Data structure
End of timestep

Figure 4: MPC framework schematic

Note that optimisation techniques can also be used without a model in some applications, such as in the unit commitment problem described in Section 7.6. We loosely use MPC and optimisation interchangeably in this paper to keep categories simple and since they share the commonality of the optimisation techniques described above.

Comparison

Many control systems today use rule-based methods, rather than MPC or RL. These have the significant advantage of being transparent and comparatively easy to understand and refine. Where the complexity and coupling between systems are low enough for system behaviours to be relatively stable and predictable, rule-based methods are probably to be preferred.

However, for advanced control problems where the system behaviour is less straightforward, MPC and RL are potentially more suitable. Comparisons of the two are rare, and if done, they are often biased towards one of the communities, using the other as a 'strawman'. The debate is quite contentious and nuanced, but we try to distil the differences here and highlight where one has an advantage over the other. The reader is referred to [21] for a more technical comparison.

One of the key differences between MPC and (model-free) RL is how tightly coupled the model, objective, and optimisation are[1].

For MPC, the optimisation is tightly coupled with the model and the objective function. This means that changes to the mathematical formulation of the model or objective function (for example introducing non-linearities) can dramatically affect the design of the optimisation algorithm and how well the optimisation performs.

What this means in practice is that the computational time/cost of obtaining the solution is often the dominant constraint, and the design choices around the model and objective function end up

[1] Note, RL tends to use different vocabulary than MPC, but for the purpose of this comparison, objective function and reward function, and optimisation and learning/training are analogous.

being made with the aim of keeping the problem tractable using the available optimisation algorithms.

In contrast, with model-free RL the optimisation (RL agent) and the objective function are only very weakly coupled, with design of the RL agent not fundamentally depending on the mathematical formulation of the objective function. The model of the environment and the RL agent are completely decoupled – the RL agent considers there to be an external environment and is completely agnostic to whether this is the real world or a Digital Twin/model.

This results in several differences in practice, detailed in *Table 1*.

Table 1: Comparison of alternatives

Rule-based	MPC	Model-free RL
Requirements for a model		
• Does not require a model of the system • Rules may or may not be transferrable between similar systems	• Requires an accurate model of each system (often physics-based, but can also be data-driven) • Complexity of model is limited by tractability of optimisation • Developing and maintaining these models (physics-based) requires domain expertise and is often time-consuming and expensive • Models may not transfer well between similar systems	• Does not require a model of the system • Can benefit from an accurate model of the system (e.g. Digital Twin) [23] for offline training (and complexity of the model is not limited) • Transfer learning may enable RL agents to transfer easily between similar systems
Online vs offline computational cost		
• All computational cost is online (i.e. at decision time) • Computational cost/time is very low	• All computational cost is online (i.e. at decision time) • Online computational cost/time is significant and often a limiting factor	• Most of the computational cost can be shifted offline with pre-training • Online computational cost/time can therefore be low • Offline computational (i.e. training) cost/time can be very high, depending on complexity of state space
Flexibility in definition of objective/reward function		
• Objectives are narrow (individual rules/constraints)	• Objectives can be general / high-level • Mathematical formulation of objective function often limited to certain analytical forms by the requirements of optimisation algorithms	• Very high flexibility in reward function • Very few limitations on formulation or scope • Easy to incorporate human feedback directly into reward
Interpretability and human understanding of the system		
• Very easily interpretable • Iterating on rules naturally improves human understanding of the system	• Largely interpretable • Easy for people to refine model based on domain expertise	• Less interpretable • Refinements and tuning are less directly linked to domain expertise

(mental model) over time – particularly around edge cases	• Refinements to model naturally improve human understanding of the system (mental model) – particularly around system dynamics, but also edge cases	• Refinements do not naturally link to and improve human understanding of the system (mental model) due to black-box nature
Inclusion of uncertainty		
• Difficult to factor in uncertainty/randomness	• Typically, methods are deterministic and do not account for uncertainty. • Uncertainty/randomness can be included (e.g. via stochastic MPC), but must be explicitly modelled (which comes with computational cost)[24]	• Uncertainty/randomness implicitly included within learning
Optimality measurement and constraint guarantees		
• Typically not optimal • Cannot measure how far from optimal a solution is • Can guarantee constraints (i.e. rules) are satisfied	• Can guarantee an optimal solution (to a strict mathematical formulation of the problem) • Can measure how far from optimal a solution is • Can guarantee constraints are satisfied	• Solutions are often close to optimal • Cannot measure how far from optimal a solution is • Cannot guarantee constraints are satisfied
Prediction horizon		
• No prediction	• Finite in length due to computational cost limitations. This can potentially lead to suboptimal solutions if rewards are sparse, which creates a short-sighted view.	• Effectively considers an infinite horizon as future rewards are discounted (e.g. similar to the time value of money in finance)

Though RL has the potential to be better suited in certain applications, it is less mature and often underperforms when compared to MPC/optimisation methods (as some use cases will show), and it can be difficult to justify the additional costs and complexity over existing rule-based methods. In fact, combinations of two or more of these approaches can often exploit individual strengths and outperform any single approach [22]. For example, in practice RL is often augmented with rule-based constraints to prevent completely unacceptable behaviours – as can be seen in some of the case studies in Section 6.

Similarly, model-based RL (which is not included in the table above) starts to blur the boundaries between model-free RL and MPC. It includes a model of how likely different actions are to cause transitions into different states, and uses this to do forward-looking short term predictions of the rewards from various actions (like MPC), and factor these into its decision-making alongside the learned rewards from historical actions (using RL). This leads to higher online computational costs when compared to model-free RL (whilst still being lower than MPC), but maintains the flexibility in reward function and ability to learn long-term policies from model-free RL.[2]

RL is, however, a very active area of research and so there may be key breakthroughs in coming years that result in significant improvements in performance. As shown in Figure 5, which is taken from reference [22], whilst MPC research has grown linearly in recent years, RL research shows signs of growing exponentially, suggesting that RL could continue to advance at a faster pace.

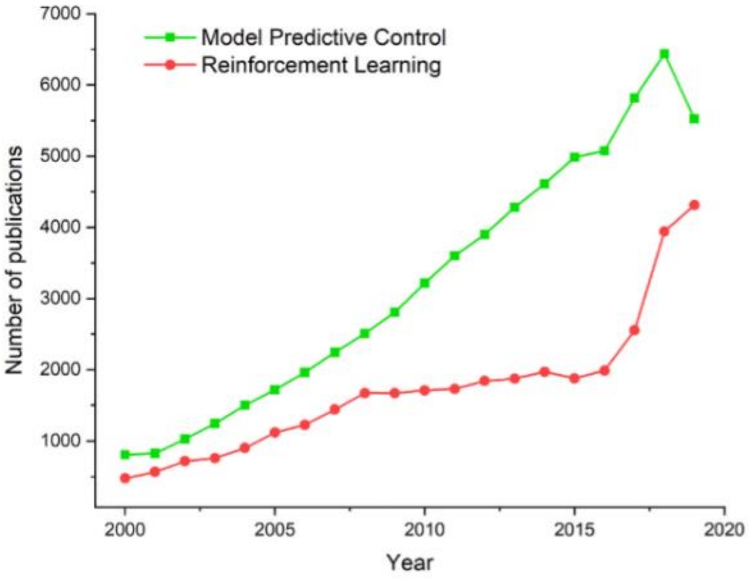


Figure 5: Number of academic publications for MPC and RL over time, figure by Perera and Kamalaruban (2011) [22]

[2] Note that there is some ambiguity in terminology here – sometimes model-based RL effectively refers to training model-free RL using a simulated environment (e.g. a Digital Twin), in which case it shares the same advantages and disadvantages as model-free RL.

4. Framework for using RL

Here we provide a general guide for the reader to follow to determine whether RL is a potential solution to a business problem.

(1) Determine if the problem can be framed as an RL problem.

Must haves for a problem to be solved with RL

Sequential decision-making process	Mathematically defined reward	Enough data to adequately define the environment state and reward at each timestep	Definable action space
The problem can be framed as a series of decisions about actions that are made based on observations of the environment.	The goals that you are trying to achieve can be quantified (i.e. scored numerically).	Data exists that would allow you to measure the impact of actions on the environment.	This may be a discrete (on/off) or continuous (charging between empty and full) action space. This affects the algorithm you design.

Additional considerations for RL implementations

Offline training in a simulated environment	Frequent feedback	Large amounts of data	Less worried about interpretability
Not always safe to implement a model-free algorithm in the real-world. You may need a safe space for the agent to learn and make mistakes.	More frequent feedback will allow the agent to learn faster.	If there is uncertainty in the input data or stochasticity in the process, it is important that you have lots of training data to avoid random fluctuations having a big influence.	The motivation behind the actions of RL agents is typically not identifiable, especially if the algorithm involves neural networks.

(2) Determine if RL is an appropriate choice over other methods.

RL is less mature than rule based or MPC approaches, and so when RL is likely to work well is less understood. A potentially effective way of identifying promising RL applications may therefore be to consider where alternative approaches struggle and RL, in theory, would not. At a high level, the following are probably a good initial filter for RL applications:

Marginal improvements/ high return

RL is likely to be costly (e.g., training, Digital Twin) and therefore applications where marginal improvements result in high returns are more likely to justify the costs.

Complex system

The system is complex, non-linear, influenced by many variables, and/or evolving, so alternative model-based approaches struggle to adequately capture the system dynamics.

Quick decision making

If the application requires fast decision making, then RL can be preferable over alternative advanced strategies like MPC.

Numerous heterogenous systems

If there are numerous heterogenous systems and it is not economical to develop individual models/Digital Twins, then model-free RL and transfer learning can potentially be a scalable solution.

Uncertainty

If the problem has a high degree of uncertainty coming from many different sources, RL may be able to incorporate this more effectively than alternatives.

Human feedback

If the reward would benefit from incorporating human feedback, then RL can be more flexible at handling this compared to alternatives.

(3) Start simple and experiment first.

To mitigate safety concerns and understand the performance of the RL agent, consider starting with a human in the loop implementation where the RL agent recommends actions, and a human user ultimately makes the final decision. This is particularly useful it can be harder to identify potential edge/failure cases for RL in advance, and real-world testing may help identify these. This implementation works well with supervisory level control, where the action or recommendation is a setpoint (e.g. temperature setpoint with HVAC). This keeps things simple as the RL controller does not need to directly control the modulation of a heat pump for example; this is left to a simpler PID (proportional integral derivative) controller. A general schematic of various levels of autonomy in RL enabled systems is illustrated in Figure 6.

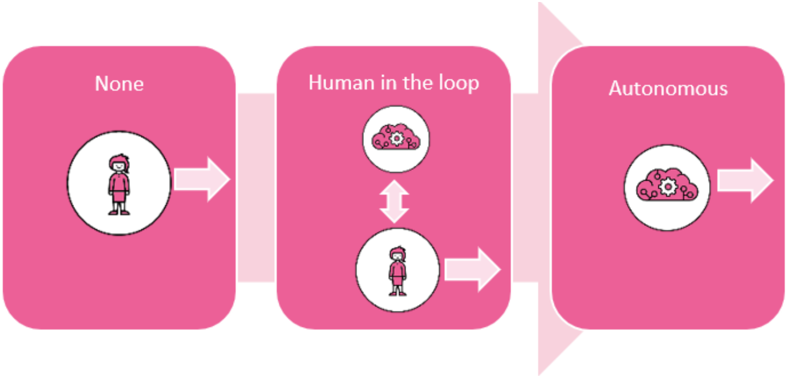


Figure 6: Levels of autonomy for RL enabled systems. (Left) human operated, rule-based control or another human-interpretable optimisation control system should be considered first. (Middle) 'Human in the loop' still utilises human intervention to assess recommended actions from the RL agent before implementing. (Right) Fully autonomous system, where errors in the system are still monitored by humans but, in general, RL is doing a good job.

(4) Collaborate with users, stakeholder, and experts:

Applying RL to solve energy related problems requires collaboration amongst domain experts and ML specialists. This was apparent in many of the use cases. Additionally, it is important to engage with the end users (sometimes the domain experts) and stakeholders as they are the ones who will need buy in for the risk concerns and business case. Domain experts are particularly important to engage with when considering the reward function and thinking about edge cases, as there may be pitfalls that are obvious to them but not to an ML specialist.

5. How RL can benefit the energy sector

Reducing wasteful energy consumption and carbon emissions is talked about extensively as a necessity to reach Net Zero targets, and to reduce energy costs for consumers, which have reached a record high in Great Britain. Optimising energy consumption is a difficult problem to address for several reasons: the increasing penetration of intermittent (usually weather dependent) renewable energy sources makes generation more difficult to predict; consumers' actions are stochastic in nature making demand difficult to predict; and demand is evolving with an increased uptake of new electricity-intensive technologies, e.g. electric vehicles (EVs) and heat pumps.

Steps towards finding a solution to this problem require a control strategy which considers the factors that influence both electricity demand and generation. This optimisation can happen at the small-scale: increasing energy efficiency of individual buildings; medium-scale: optimising the energy consumption of small energy systems which connect multiple consumers and generation facilities; and at the large-scale: improving the interactions between different energy systems, or managing a large scale network. RL is considered an exciting potential solution given that it is useful for complex systems where the interplay between the many influential factors is too difficult for a human, or simple algorithm, to optimise. For example, in a game of chess there is not one solution to the problem, instead the solution depends on how the game evolves. Optimising energy consumption is similar since the moves taken by the algorithm (such as drawing energy from the grid) depend on what the environment looks like at a given time, like the state of a chess board during a game. The optimisation problem will need to consider the complex interplay between demand, generation, and demand side response simultaneously, and potentially their forecast predictions. One important factor that must be considered in any optimisation strategy is the fact that consumers have different preferences/requirements in the way they use energy which need to be addressed. RL algorithms can be used to learn and/or directly consider consumer preferences.

Whilst the aforementioned applications of RL in the energy sector are exciting, many are likely years or even decades away. However, we are starting to see some promising use cases of RL within industry today. Figure 7 illustrates the landscape of potential near-term and future applications of RL within the energy sector and in the next section we present specific industry and academic examples to showcase each use case.

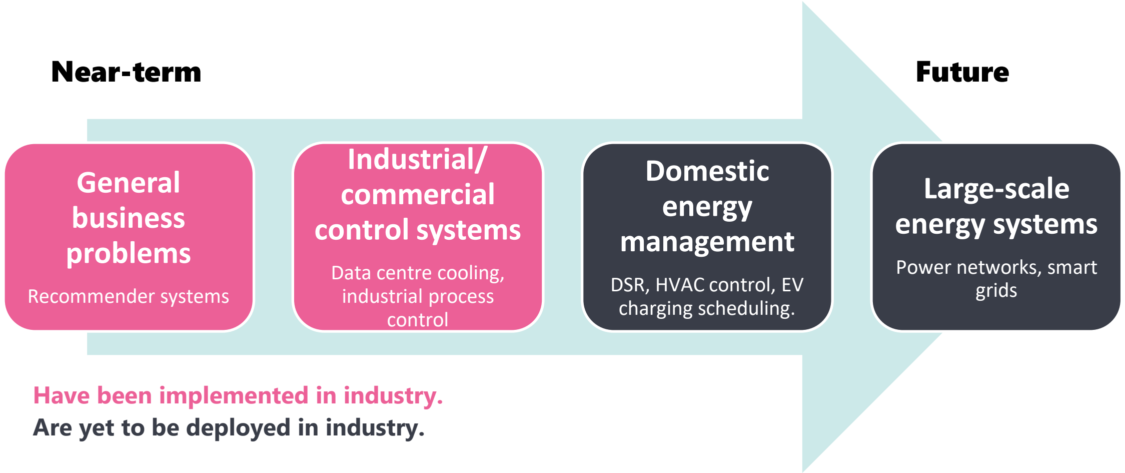

Near-term
Future
General business problems
Recommender systems
Industrial/ commercial control systems
Data centre cooling, industrial process control
Domestic energy management
DSR, HVAC control, EV charging scheduling.
Large-scale energy systems
Power networks, smart grids
Have been implemented in industry.
Are yet to be deployed in industry.

Figure 7: Landscape of near-term and future applications of RL within the energy sector.

Despite being an immature technology that is frequently not yet ready for production, we see RL penetrating other industries, such as the tech sector, faster than in energy. This would be consistent with the broader trend of ML in general, since the energy sector has been a slow adopter of ML. Ultimately, this could indirectly benefit the energy sector as RL can address general business problems that are not industry specific, such as utilising recommender systems or language models to support customers of an energy supplier.

In the energy sector specifically, we are starting to see industry applying RL to industrial and commercial control systems. These applications generally have higher margins and face lower barriers compared to domestic applications, which are still primarily in the research phase and subject to stricter safety regulations. Additionally, operating power networks and smart grids necessitates a high level of security, as any issues could impact many customers. Therefore, implementing RL in these areas is further from becoming a reality.

6. Successful use cases of RL in industry

6.1. GOOGLE DEEPMIND

DeepMind is an AI research lab that have recently focused on applying their expertise towards real-world applications, such as understanding protein folding, identifying eye disease faster, and energy reduction in data centre cooling. They reported on the use of RL in reducing data centre cooling energy consumption by up to 40%; the solution was later productionised in multiple Google data centres with greater autonomy given to the RL controller [1], [23]. Building on this, they extended this work by conducting experiments in collaboration with a building management system provider, Trane, on two commercial building cooling systems [3].

Problem: A chiller plant typically provides the cooling to commercial buildings (similar to air conditioning in a home). The optimisation of a chiller plant is challenging as there is a trade-off between energy consuming components in the system. The chiller energy consumption decreases as the condenser water temperature decreases, but this comes at the cost of consuming more energy by the cooling tower fans—thus a balance must be met for optimal operation. On top of this, external and internal environmental conditions, and a large number of control parameters (e.g. setpoints and number of equipment) make the problem increasingly difficult to optimise.

Alternative approaches: HVAC controls are typically rule-based. Setpoints are either fixed based on an operator's experience or they can be adjusted on a reset schedule, where the setpoint can vary based on outdoor temperatures. MPC has also been explored in research and in limited industry applications. This often requires developing and maintaining expensive physics-based models for each building, which may not be scalable since buildings are very heterogenous.

How reinforcement learning was applied:

- **States:** Sensors (e.g. temperature, equipment statuses, etc.)
- **Actions:** Setpoints (e.g. chilled water temperature) and on/off control of equipment
- **Rewards:** Minimise chiller plant energy consumption

First, the RL controller was trained offline on a log of historical data produced by the rule-based controller; this is typical in real-world examples as it is safer since no interaction with the real world is required. The RL controller was then deployed, allowing for exploration and improvement of the control strategy as it interacted with the physical system. The actions of the RL agent were simply automated setpoint recommendations to the building management system, which could be automatically overwritten with the rule-based controller if safety constraints were violated or if the operator manually decided to intervene.

Key benefits: A 3-month experiment was conducted to compare the energy consumption between the RL and rule-based controller for 2 commercial properties. Two different experiments were conducted using different methodologies, resulting in approximately 9% and 13% energy savings. The RL controller was able to learn the optimal trade-off between chiller and cooling tower energy consumption. Interestingly the RL controller was also able to adapt to sensor miscalibration.

Challenges and future thoughts: Several challenges were highlighted, such as learning from limited data, complex constraints, non-stationarity (e.g. equipment degradation and building occupancy changing over time), real-time decision making, and different operating modes. Success was dependent on extensive domain knowledge and numerous constraints. In the future, the authors suggest using detailed simulation environments to learn offline and transfer knowledge (*transfer learning*) to overcome some of the limitations in training offline from limited historical logs. Incorporating direct human feedback in teaching the agent could be an area to explore.

6.2. CARBON RE

Carbon Re is a startup that has developed a software product (Delta Zero) that utilises RL to optimise cement production by reducing energy consumption and carbon emissions [2].

Problem: Cement production is a complex process with constantly changing inputs (fuels, raw materials), conditions (state of equipment, shift changes), and competing priorities (throughput, control limits). A UK case study [24] identified a 7% variation in day-to-day energy consumption due to operational variations, implying potential energy saving opportunities if the plant could continuously operate like the 'best days'. Key drivers to operational improvements consisted of the carbon footprint of the fuel, excess oxygen, and combustion properties of the fuel; optimising these drivers indicated a potential energy reduction of 8.5%—thus motivating the inception of Carbon Re.

Alternative approaches: Traditionally, rule-based 'Expert Systems' are used for cement production. These systems are difficult to update with new rules, do not handle uncertainty well, and are limited in the number of factors considered (e.g. energy, emissions). Predictive process control and automation systems (or model predictive control) are also used extensively but require high-cost upgrades for automated controls, have limitations in their models, are expensive to upgrade and incorporate new features (e.g. exploiting alternative fuels, reduce emissions and material costs), and specialised training is required for operators.

How reinforcement learning was applied:

- **States**: Fuel mix (e.g. coal, waste, biomass), kiln (where thermodynamics and physical processing occurs) feed, etc.
- **Actions**: Recommended process setpoints (e.g. kiln feed rate, fan speed, etc.)
- **Rewards**: Reduce carbon costs whilst maintaining production and safety constraints.

Carbon Re utilises a Digital Twin which is developed based on historical sensor data and uses this to train the RL agent by simulating 100 years of training in a single day. The final product is a 'human-in-the-loop' solution where the RL agent makes process setpoint recommendations to a human operator to act on.

Key benefits: The company reports potential energy savings of 10% and carbon emissions reduction by up to 20%.

Challenges and future thoughts: A highly accurate Digital Twin played a key role in enabling a well-performing RL solution. Highly instrumented, closed processes like manufacturing lend themselves to accurate Digital Twins and therefore potentially RL solutions. The company also aims to utilise the Digital Twin to better understand and address maintenance issues like kiln blockage, which can result in costly downtime, as well as potentially evaluate capital investments.

6.3. E.ON

E.ON is one of the UK's largest energy suppliers, with a global reach. E.ON is also a major windfarm player in many countries, and have piloted a dynamic yield optimisation project using RL [25]

Problem: During the pre-construction phase of a wind farm, layout optimisation of wind turbines is utilised to minimise negative aerodynamic interactions (wake effects) between turbines, which reduce power generation. However, residual wake effects still occur during operation which can result in power losses of a few percentage points.

Alternative approaches: Typically, all turbines operate to maximise their individual power generation. The control strategy does not consider the trade-offs between reducing an individual turbine's power generation to increase downstream turbine power generation via reduced residual wake effects. Computational fluid dynamics is a method that can model and simulate the aerodynamic phenomena described above. Conceivably, it could be used in an MPC framework, but can potentially be too computationally expensive to use for online control. Additionally, the model would be site specific and inflexible to changes in the wind farm environment.

How reinforcement learning was applied: E.ON developed a simulation environment of the wind farm to test their RL algorithm. After a successful demonstration they piloted it in the field at the Champion wind park in Texas USA. The trial consisted of 3 rows of 3 turbines each, where the algorithm was able to learn within 20 iterations. The RL problem was framed as follows:

- **States**: Wind direction and speed, turbulence intensity, curtailment level (blade pitch angle settings).
- **Actions**: Curtailment level (blade pitch angle settings).
- **Rewards**: Gain/loss in power with respect to nominal operation (no curtailment) for the same wind conditions.

Key benefits: The average power increase of the 3 rows of turbines was 0.8% for the pilot project. Though the energy gains were small, when scaled to multiple wind farms this could be significant.

Challenges and future thoughts: The current control only allowed for the changing of blade pitch angle. Future work can focus on adjusting the orientation of the rotor as well, which has shown to improve power output in research.

6.4. MICROSOFT

As a large, multinational corporation, Microsoft's offices consume significant amounts of energy. To achieve their carbon targets they need to reduce this considerably.

Problem: Microsoft looked to reduce energy consumption in their commercial buildings by changing the way the HVAC systems were operated [26].

How reinforcement learning was applied: Microsoft used their platform called Project Bonsai that enables engineers to build industrial control systems powered by AI [27]. It implements a technique called *machine teaching* [28], which utilises domain expert knowledge to accelerate the learning speed of the RL agent. HVAC experts worked with AI experts to design the objectives and constraints for the RL agent.

- **States**: Outside air humidity, outside air temperature, etc.
- **Actions**: Temperature set point, the speed of the pump delivering chilled water to the condenser unit, differential pressure between supply & return water flow pipe.
- **Reward**: Reduce energy consumption while maintaining desired temperature.

Key benefits: The company reported an expected (modelled) energy reduction of around 15%, with the RL agent learning within 2 weeks. The RL agent also uncovered counterintuitive recommendations to what a human would assume to reduce energy consumption such as increasing setpoints in the cooling tower and chiller and increasing pump speeds.

Challenges and future thoughts: This is another example of a human-in-the-loop application of RL – the algorithm recommends new set points for building managers to implement. The company sees further potential in optimising similar areas due to the success in changing only a few setpoints.

6.5. EXAMPLES OUTSIDE THE ENERGY SECTOR

Although the focus of this paper is the application of RL within the energy sector, there has been significant success in other areas which should be looked to in order to gain inspiration about how RL can be applied in the energy sector. Other industry applications include self-driving cars, robotics, healthcare [29], game playing, etc. [12]. We list some examples in Table 2.

It's important to recognise that companies within the energy sector can face similar business challenges and opportunities to any other sector. Customer facing energy companies can leverage recent advancements in RL applications in recommender systems and language models. In addition to optimising control systems, RL can also serve as an optimisation layer on top of traditional supervised learning models used in recommender systems and language models. For example, ChatGPT used RL with human feedback in the reward model to optimise a supervised learning language model for more human-like responses and to reduce toxic content. Similarly, Spotify's recommender system incorporates human feedback by learning which tracks are relevant through user song skipping. Traditional supervised learning-based recommender system can also be optimised with RL to balance a trade-off between exploiting existing user satisfaction and exploring diverse music content. Energy companies can adopt these AI solutions into their existing practices or experiment with them further using in-house data and RL techniques, thus building trust in the RL systems and explore additional use cases in the future.

Table 2: Industry examples of applying RL. Note, we are unable to verify if these were used in production applications.

Company	Application	Sector	States	Actions	Rewards
Royal Bank of Canada [30]	Trade execution platform for multiple strategies	Financial services	200-plus market-related data inputs	Sell, buy, hold stocks	To trade as close as possible to VWAP, a common price metric
IBM [31]	Financing trading	Financial services	Stock prices	Sell, buy, hold stocks	Profit/loss generated by each trade
Spotify [32]	Diverse content recommendations	Entertainment	Previously recommended songs	Recommended track	Maximise diversity and relevance of recommended track
Salesforce [33]	Text summarisation	Technology	Original long document	Summary	Used human labelled reference summary to guide learning
DiDi [34]	Order dispatching	Ride hailing	Number of idle vehicles, number of orders, location, destination	Match driver to passenger	Minimise pickup time and maximise revenue
Deepmind [35]	Game playing (AlphaGO/ AlphaZero)	Technology	Game board and position of pieces	Next move of player	Win game
OpenAI [36]	Large language model (ChatGPT)	Technology	Text prompt (e.g. asking questions)	Text response (e.g. answers to questions)	Fine tune language model with human feedback (e.g. remove toxicity, follow instructions better)

7. Promising use cases for innovators to be aware of

Researchers are continuously working on delivering improved RL algorithms applicable to a wide range of use cases, making RL an exciting topic of conversation when it comes to thinking about innovation in energy management and control.

Where Section 6 provides an overview of specific case studies where RL applications have already been shown to be successful in industry, this section considers future applications of RL within the energy sector. Key areas where RL applications could be most beneficial in the energy sector are discussed, based on the current state of academic research. Most of these relate to control of different systems, but it is worth noting that the use of RL with human feedback (RLHF) [36] also proposes significant opportunities—as highlighted in Section 4 and Section 6.5. RLHF looks likely to drive a paradigm shift in the way we think of RL. Whilst Deepmind's AlphaGo focused on removing human influence to achieve superhuman performance in gameplay, OpenAI's ChatGPT did the opposite by incorporating human feedback to improve language understanding. Both approaches are likely to have a role in future RL research, but RLHF has already demonstrated its potential in various energy-related applications. For example, learning from a occupants' subjective thermal comfort in buildings through an app [21], and improving driver comfort in autonomous vehicles [37].

A summary of the methods described in this section of how RL can add value to energy solutions is shown in Table 3: promising use cases for RL in the energy sector, and what the focus of the optimisation (i.e. the reward function) could look like.

Table 3: promising use cases for RL in the energy sector, and what the focus of the optimisation (i.e. the reward function) could look like.

	Application	Optimisation goal focus		
		Cost	**Emissions**	**Revenue (e.g., from flexibility)**
Asset Management	EV charging	✓	✓	✓
	Fleet EV charging	✓	✓	✓
	Battery charging/discharging	✓	✓	✓
	Smart appliance scheduling	✓	✓	✓
	Heating and cooling system control	✓	✓	✓
System Management	Whole home/building energy management	✓	✓	✓
	Wind farm design and operation	✓		✓
	Energy trading within local energy markets or microgrids	✓	✓	✓
	Generation dispatch and grid balancing	✓	✓	
	Energy pricing/bidding strategies			✓

7.1. ENERGY MANAGEMENT FOR DEMAND FLEXIBILITY

Consumers can save money by drawing energy from the grid at times away from when there is high-demand, or when the electricity price is low and energy from renewables is readily available [38]. When consumers shift their electricity consumption in this way, this is known as demand-side response (DSR) and it is used to prevent the grid from being overloaded (leading to power cuts); and to prevent the startup of emergency power supplies which run on greenhouse gases. DSR will become increasingly important as more consumers invest in electric vehicles (EVs) and electric heating solutions (e.g. heat pumps) intensifying the load on the grid.

One way to maximise energy savings when participating in DSR would be to use a tool which automatically controls intelligent assets to shift and/or reduce their energy usage, or to automatically advise humans on what actions to take to optimise their energy use. Such algorithms must consider consumer comfort and consumer preferences (e.g. reaching the temperature setpoints determined by the consumer) whilst simultaneously optimising the demand for grid management. RL is invoked as a promising solution to this challenge given that it can learn human preferences which change over time and is further capable of adapting to specific environments which vary from building to building. The following sections consider RL applications to optimise assets in intelligent ways with the goal of either maximising their energy efficiency and/or to allow flexibility to initiate demand response.

7.2. HVAC AND DHW SYSTEMS

Heating, ventilation, and air conditioning (HVAC) systems and domestic hot water (DHW) systems are huge contributors to building energy consumption and, in most cases, these systems are managed by sub-optimal control systems which do not consider grid response mechanisms (e.g. simple rule-based control, or set by humans based on 'a feeling') and can result in significant energy wastage.

RL solutions for optimising the energy consumption of the heating control of DHW systems [39], to optimise the control HVAC systems in office spaces [40]–[42], and to manage commercial cooling centres [13] have all been tested in real, functioning buildings. The fact that, in each of these cases, the deployment of the RL solution led to a significant reduction in energy consumption (where the latter example is described in detail in Section 6.1) showcases the exciting potential for RL-based control. Promisingly, these successes are not limited to simple environments (e.g. a single isolated room) but include relatively complex buildings with multiple zones and floors [43]. However, this doesn't mean the implementation of such RL algorithms into the real-world is trivial, nor does it mean that RL is always the optimal method for control: there is progress to be made towards developing RL-based solutions which have the potential to be productised so that any consumer can install these solutions in their homes or commercial buildings. A lot of this progress is being made by researchers, where many studies of HVAC and DHW control in simulated building environments (as opposed to real buildings) have been conducted e.g. [44]. In general, these solutions have been shown to be successful in saving 10 - 20% of building energy consumption.

Large-scale deployment of RL solutions is a challenge due to variations in the physical properties of HVAC and DHW systems, variations in the buildings themselves (where even similar types of homes will differ in their capacity to retain thermal energy), and variation in occupant behaviour from building to building. Because of this, to achieve optimal control, agents monitoring different buildings may need to behave very differently from one another. When productionising an RL solution, large amounts of training data will be needed to allow for exploration into many different states for the agent to learn the best actions for each scenario. Another important point to note is that many RL algorithms may need to be combined with rule-based methods to ensure that agents do not take actions which will be unsafe or cause occupant discomfort, which may be non-trivial to implement (see e.g. [45]).

MPC has also made good progress in HVAC controls within buildings. Whilst fair comparisons between RL and MPC are rare, a study [21] was conducted in a simulated environment, which ultimately showed MPC outperformed RL in terms of energy costs and discomfort. However, a combination of the two methods showed comparable results to MPC. This potentially opens opportunities to learn perceived discomfort through human feedback, account for uncertainties (which was not incorporated in the simulator) and optimise over longer periods; these benefits could compliment the performance and interpretability advantages of MPC. In contrast to this, a similar study showed RL can potentially outperform MPC by up to 6% [46]. Thus, more comparison studies like these are encouraged and needed to better understand the methods.

7.3. SMART APPLIANCES

To initiate DSR, household occupiers may be incentivised to use their energy consuming appliances (e.g. washing machines, cooking appliances, lighting) at optimal times to reduce their electricity usage during peak times. RL can address the problem of optimising the DSR of individual homes by performing a scheduling assistant for smart devices. The scheduling of appliances to optimise energy consumption needs to be in line with the preferences of the consumer; this can be accounted for by determining whether the *switching on* of appliances is deferrable or non-deferrable [47] and/or by incorporating consumer feedback directly into the algorithm [48]. RL provides a promising solution to such optimisation problems due to its ability to cope with consumer preferences and learn in a stochastic environment. RL has shown potential to outperform other algorithms [49], including MPC, however the training time for RL algorithms can be very long [50]. It is important to note that RL solutions tend to be developed assuming a single domestic consumer with multiple appliances, however there is often more than one consumer in a household which needs to be investigated. Furthermore, the consumer feedback incorporated in training during these studies is usually not real data but simulated data so before deployment it may be appropriate to use more realistic consumer feedback in training. Such feedback could be gained from testbeds such as the Living Lab [51].

7.4. ELECTRIC VEHICLES

Optimal approaches can be defined when it comes the charging of EVs. Since EVs consume substantial amounts of electricity, a consumer may want to participate in DSR by optimising the plug-in times for charging their EV at their property. Furthermore, during a journey, a driver may want to optimise how they charge their vehicles at public charging stations to simultaneously minimise the time required for travel and the charging cost. These optimisation challenges are difficult to solve given the variability in the departure and arrival times of the EV and the requirement for the EV to be charged fully before departure. Further complications come with the variation in traffic conditions and waiting times for EV chargers. RL is shown to be an effective solution given its ability to adapt to consumer preferences and reach a solution even without having prior knowledge about the randomness [52].

RL can be further used to optimise the charging a fleet of EVs where the charging properties of the vehicles which vary from vehicle to vehicle (e.g. battery size) are unknown [53]. Day ahead charging plan solutions can also be addressed with RL. In [54], a simulated environment is used to show that their RL solution gives results comparable to an optimisation method called stochastic programming in which the details about the charging of cars is needed to be known, unlike in the RL case.

A further use for RL is to optimise hybrid vehicles [16]. The challenge here is to increase fuel efficiency whilst driving even when the driving conditions vary from journey to journey, and so can the style of the driver. RL can be used to optimise without giving the algorithm information about the route to be driven [51] [55]. RL solutions can improve on the standard performance on HEVs, but it is important to note RL is only one method to get an optimal solution here [56], and optimal solutions reached with RL are close to those reached with dynamic programming [57].

7.5. ENERGY TRADING

Local Energy Markets (LEMs), where electricity is generated, bought, traded, stored, and consumed in a decentralised manner, are expected to play a significant role in helping us reach a Net Zero energy system [58]. The future electricity grid will need to be able to cope with the increased uptake of EVs and electric heating technologies which come with the transition to Net Zero and reach the demand from volatile renewable sources. LEMs provide a solution by locally establishing a balance between generation and consumption which would otherwise be a struggle on a centralised national distribution grid. RL comes into play to facilitate an intelligent market by having agents which are responsible for determining an optimal pricing strategy for household bidding, and agents which are responsible for trading [59]. In [59], the RL algorithm is rewarded for achieving low energy prices during times of high renewable energy generation and higher energy prices when the consumption is high (to encourage DSR) in a simulated environment. RL is also seen as a promising solution for trading between these microgrids to reduce energy prices and maximise profits [60]. Drawbacks of the solutions in the literature are described in the relevant papers and overall further research needs to be done before it can be well established that RL solutions will indeed provide optimal solutions in a real-life LEM scenario.

In general, the RL techniques applied to the financial sector for trading (see Table 2) are naturally extendable to energy trading. RL algorithms can learn a trading strategy to determine when to buy, sell, or hold assets from its experience of market behaviours.

7.6. LARGE-SCALE MANGEMENT OF ENERGY SYSTEMS

A lot of work has focused on the implementation of single RL agents aiding the management of individual buildings and their associated assets, or a small complex of similar buildings. However, what about if you want to consider the optimisation of a smart grid system where lots of houses are connected? This is a more difficult challenge given the increased complexity of the system where the individual domestic and industrial properties require very different treatments relative to each other and will have different reward functions to optimise. To increase the penetration of renewables into the energy system through low carbon tech and flexible demand maintained by DSR solutions, the future system may comprise local energy markets which are interconnected and individual consumers are able to trade energy generated by, for example, their solar panels, back to the grid. This enhances the complexity compared to a large-scale centralised system, where the optimisation of the distribution may be best addressed with RL. Furthermore, as technological advances are made which enable long-term storage of electricity, there is a question as to how to best manage the charging and discharging of batteries to maximise energy generation through renewables.

Progress is being made in this area by the set-up of OpenAI Gym Environments, such as CityLearn [61]. These environments allow anyone to build algorithms to optimise the control of energy distribution in the environment, mimicking a simplified version of the real-world energy systems. Challenges posed to the public have led to the development of novel algorithms progressing the field further [62]. Promising results have been shown for RL to reduce peak demand on smart grids using energy storage and avoid a 'rebound effect', where the lack of coordination leads to shifting a peak in time rather than reducing it [63]. Additionally, transfer learning methods have shown promise in addressing the challenge of limited samples of data, as you can transfer the learnings of one building to another and get up to speed quicker [64] . However, RL still was outperformed by optimised rule-based controllers [10] and adaptive optimisation methods [65] indicating further research is required. More recently, a competition was conducted to encourage creative and collaborative approaches to the CityLearn problem [66]. Many of the top solutions employed a combination of RL, MPC, and rule-based approaches, further supporting the convergence of a hybrid method capitalising on the benefits of each method.

Perhaps the ultimate application of RL to the energy sector is in dynamically balancing the whole electricity network. Demand and supply need to be perfectly balanced for the grid to remain stable, and this involves manging multiple electrical parameters like power and frequency simultaneously, over a range of timescales from days and hours down to seconds. This is becoming increasingly difficult as renewable and embedded generation becomes more widespread and transport and heating increasingly electrify. As a result, the cost of doing so is increasing, with system balancing and constraint costs in Great Britain rocketing in recent years [67] to £4.2bn in 2022. This poses a significant opportunity for RL.

Production applications of RL to this problem are probably many years in the future, but early research in this area is already occurring. The Learn to Run a Power Network competition [68] focused on learning to find the optimal network topology to relieve network congestion at minimum costs whilst preventing cascading failures leading to blackouts. Additionally, a similar application is choosing the optimal power generator dispatching schedule for a network (known as the unit commitment problem). RL was shown to outperform existing optimisation methods for this problem [69]. However, when conducting a more rigours comparison (including stochastic optimisation), RL underperformed [70]. Nonetheless, RL demonstrated clear superiority in terms of online computational time, which is likely key for managing large systems in real time. Once again, a hybrid approach was shown to outperform both methods individually.

Due to the safety critical nature of the grid and the extremely high consequences of failure, progress in this area is likely to be very cautious, and the human-in-the-loop approach is likely to dominate. It does, however, present an exciting opportunity for RL as the technology matures.

8. Risks and ethical considerations

Use of RL shares many of the risks and ethical considerations that arise in the use of AI in general, which are covered at length in other reports e.g., [71]–[74]. This section focuses those that are particularly important for RL (and applications of RL in energy).

The greatest risk with RL is that it successfully optimises for something that we do not actually want. In other words, people specify the reward function in a way that does not adequately capture everything we value. This is known as the alignment problem [75]. The consequences of misalignment can be minor – an amusing loophole the AI finds that is easily fixed – or critical – a system actively causing harm to people.

Achieving alignment is often very difficult technically. It is hard to robustly specify objectives in a reward function that does not leave loopholes for the RL agent to exploit, and there are many examples of specification gaming [15] and reward hacking in RL systems. It is, however, even more difficult from an ethical and philosophical perspective. There is no universally accepted ethical framework, and disagreements on what is right - and optimal for society - are widespread. Developers of AI systems that are optimising for 'something' therefore have a responsibility to consider the ethical implications of what they are optimising for.

Some of the ethical considerations that are particularly relevant for RL are:

- Autonomy and control – to what extent should individuals have control taken away from them?
- Bias and fairness – are different groups treated fairly by the system? Does the algorithm perform equally well for different groups?

This may sound very abstract and academic, so the example below considers a typical use case in energy. This example also draws out two further risks:

1. An RL agent may learn to operate well on the data and situations it is trained on, but when it encounters situations that are dramatically different from those it has encountered before, it may produce completely unacceptable solutions and/or take a long time to learn how to handle those situations.
2. Multiple RL agents that have different objectives but are controlling overlapping systems may demonstrate unexpected or undesirable behaviour as they feed into each other's feedback loops. This is particularly important in the energy sector, where RL could conceivably be employed at multiple levels simultaneously, from individual devices to whole home energy management, to energy trading and grid balancing. How RL agents behave when interacting with other agents in the wider system is important to consider and test.

To illustrate all of these, let's consider a typical use case in energy: an RL agent designed to optimise heating in a home. We'll work through a hypothetical deployment of RL in this use case.

At a high level, the RL agent is given the objective of minimising energy costs whilst meeting the temperature set points provided by the occupants. This seems straightforward, but then the question arises of how precisely the set points must be met. Is a tolerance of +/-1 degree ok? The wider the tolerance, the more flexibility the RL agent has to reduce energy costs. But different people may have different acceptable tolerances.

Fortunately, RL is well suited to adapt to this – it can include human feedback in the reward function. The obvious thing to do is allow quite wide tolerances but add a penalty for every time that an occupant adjusts or overrides the set points (as that is an indicator they are uncomfortable). However, there are multiple occupants in the home and they have different heating preferences (for example, men may prefer a lower temperature to women [76]). Not only that, but they have different access to (and comfort with adjusting) the digital controls for the thermostat (with men being more likely to dominate the control [77]). As a result, the RL agent learns to optimise for preferences of the occupant who more frequently adjusts the controls, resulting in an unfair outcome.

Realising this, the reward function is adjusted to place less weight on thermostat adjustments (although this won't mitigate the issue fully). Since the manual adjustments often increase energy costs (which it is highly penalised for), and the adjustments are only weakly penalised, the RL agent learns that if it ignores the adjustments for several weeks the occupants will eventually realise their feedback is having no impact and give up trying to make adjustments – which means the RL agent no longer suffers a penalty for them. The RL designer eventually works out a way to prevent that happening, and the RL agent is rolled out to heating controllers across the country.

At this point two things occur. Firstly, the RL agent encounters a heat pump for the first time (having been trained on homes with gas boilers). Not only does the heat pump have a peak efficiency when used in a different pattern compared to a boiler, but it also contains its own internal RL agent that aims to not just maximise efficiency, but also minimise stress on the components of the heat pump to improve reliability. The goals of the two RL agents (home heating controller and heat pump) aren't fully compatible and they proceed to have a tug of war over control which leads to high energy costs and more rapid deterioration of the heat pump.

Secondly, the RL agent is deployed in the home of someone experiencing financial hardship. They are very tightly constrained by finances, and are often making trade-offs between heating and food. Having control is very important to them, and they make very frequent adjustments to the thermostat based on current finances, rather than comfort. The RL agent is designed to minimise cost whilst meeting the target temperature set points, not to minimise discomfort whilst meeting a fixed budget. It also can't cope with so many changes to the thermostat settings that aren't obviously related to comfort, so it produces a poor, suboptimal solution that costs the consumer significantly more than manual control would.

As this example illustrates, employing RL in a robust and ethical way takes considerable thought and care. Innovators developing RL solutions should make sure to invest sufficient time in this, and identify people with appropriate expertise to support that – for example experts in AI risks and ethics. It is also important to consider and engage a broad range of stakeholders from an early stage – including a diverse set of end users and consumers.

9. Conclusions

Reinforcement Learning is a powerful tool for optimising control methods throughout the energy sector. Its ability to learn through a trial-and-error approach allows it to be implemented without the developer having a detailed knowledge of the physics underlying a system, or relationships between different variables within a system. This is particularly useful when dealing with complex systems with lots of variables which are difficult to model.

For innovators, there are some key use cases throughout the energy system which could potentially benefit from RL solutions. These include drawing inspiration from RL applications within general business problems, industrial and commercial control systems, scheduling of household appliances and EV charging, control of HVAC and DHW systems, and enabling the systems to participate in demand-response initiatives to reduce load on the grid. Although the maturity of many RL implementations is still at an early stage, the increasing complexity involved in balancing distributed renewable generation and flexible demand means that the benefits of smart, autonomous, reinforcement-learning solutions are likely to steadily increase.

Whilst RL offers much promise for the energy sector and decarbonisation, much work remains before it can be fully realised. In particular, our collective understanding of where RL is the best solution and how to get the most out of it remains relatively limited. Therefore our call to action is to contribute to developing this understanding by:

- **Industry**: As much as possible, share challenges and successes of real-world implementations of RL — see Deepmind's work [13] as an example. Consider sharing datasets that can be used to develop RL solutions.
- **Academia**: Conduct fair comparisons between RL and alternatives (e.g. MPC/optimisation) to better understand where RL might excel in the future and understand how they can potentially complement each other.

10. Acknowledgements

We express our gratitude to the individuals who contributed their time and expertise towards this paper. Their consultations helped us gain a deeper understanding of the challenges and opportunities of RL in the energy sector and ensure the technical details were accurate. We want to clarify that these experts were not asked to endorse the conclusions or recommendations of the report and provided their insights in a personal, non-representative capacity. We are thankful for their contributions, and Energy Systems Catapult acknowledges their valuable input.

Name	Organisation
Dr. Patrick de Mars	University College London (Until September 2022) UK Power Networks (Since October 2022)
Dr. Zoltan Nagy	University of Texas at Austin
Dr. Aidan O’Sullivan	Carbon Re, University College London, Alan Turing Institute
Dr. Stephen Haben	Energy Systems Catapult, University of Oxford

11. References

[1] "DeepMind AI Reduces Google Data Centre Cooling Bill by 40%." https://www.deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-by-40 (accessed Feb. 20, 2023).

[2] "How AI is helping cement plant operators reduce energy consumption and carbon emissions — Carbon Re." https://carbonre.com/resources/cementwhitepaper (accessed Feb. 20, 2023).

[3] "Controlling Commercial Cooling Systems Using Reinforcement Learning." https://www.deepmind.com/publications/controlling-commercial-cooling-systems-using-reinforcement-learning (accessed Feb. 20, 2023).

[4] "AI for Energy report outlines opportunities for applying AI in the energy sector." https://www.techuk.org/resource/ai-for-energy-report-outlines-opportunities-for-applying-ai-in-the-energy-sector.html (accessed Feb. 20, 2023).

[5] "Forecasting | Open Climate Fix." https://openclimatefix.org/projects/forecasting (accessed Mar. 28, 2023).

[6] M. Jang, H. C. Jeong, T. Kim, and S. K. Joo, "Load Profile-Based Residential Customer Segmentation for Analyzing Customer Preferred Time-of-Use (TOU) Tariffs," *Energies 2021, Vol. 14, Page 6130*, vol. 14, no. 19, p. 6130, Sep. 2021, doi: 10.3390/EN14196130.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Adv Neural Inf Process Syst*, vol. 25, 2012, Accessed: Feb. 21, 2023. [Online]. Available: http://code.google.com/p/cuda-convnet/

[8] A. Vaswani *et al.*, "Attention Is All You Need," *Adv Neural Inf Process Syst*, vol. 2017-December, pp. 5999–6009, Jun. 2017, doi: 10.48550/arxiv.1706.03762.

[9] "AlphaFold." https://www.deepmind.com/research/highlighted-research/alphafold (accessed Feb. 21, 2023).

[10] K. Nweye, B. Liu, P. Stone, and Z. Nagy, "Real-world challenges for multi-agent reinforcement learning in grid-interactive buildings," *Energy and AI*, vol. 10, p. 100202, Nov. 2022, doi: 10.1016/J.EGYAI.2022.100202.

[11] G. Dulac-Arnold, D. Mankowitz, and T. Hester, "Challenges of Real-World Reinforcement Learning," Apr. 2019, doi: 10.48550/arxiv.1904.12901.

[12] P. Osborne, K. Singh, and M. E. Taylor, *Applying Reinforcement Learning on Real-World Data with Practical Examples in Python*, 1st ed. Springer International Publishing, 2022. doi: https://doi.org/10.1007/978-3-031-79167-3.

[13] J. Luo *et al.*, "Controlling Commercial Cooling Systems Using Reinforcement Learning," pp. 2022–2034, 2022.

[14] C. Perrow, *Normal Accidents*. Basic Books, 1984. Accessed: Mar. 28, 2023. [Online]. Available: https://www.theisrm.org/public-library/Charles%20Perrow%20-%20Normal%20Accidents.pdf

[15] "Specification gaming examples in AI - master list - Google Drive." https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml (accessed Mar. 28, 2023).

[16] B. Suh, "5 Rules to Manage AI's Unintended Consequences," May 2021. https://hbr.org/2021/05/5-rules-to-manage-ais-unintended-consequences (accessed Mar. 07, 2023).

[17] Z. Ding, P. Hernandez-Leal, G. W. Ding, C. Li, and R. Huang, "CDT: Cascading Decision Trees for Explainable Reinforcement Learning", Accessed: Mar. 24, 2023. [Online]. Available: https://github.com/openai/gym/blob/master/gym/envs/box2d/lunar

[18] "BEIS Energy System Digital Twin Demonstrator - Energy Systems Catapult." https://es.catapult.org.uk/report/beis-energy-system-digital-twin-demonstrator/ (accessed Mar. 24, 2023).

[19] K. Xia *et al.*, "A digital twin to train deep reinforcement learning agent for smart manufacturing plants: Environment, interfaces and intelligence," *J Manuf Syst*, vol. 58, pp. 210–230, Jan. 2021, doi: 10.1016/J.JMSY.2020.06.012.

[20] "3rd Workshop on Closing the Reality Gap in Sim2Real Transfer for Robotics | Full Day Workshop at R:SS 2022 (New York City, New York, USA), June 27th." https://sim2real.github.io/ (accessed Mar. 24, 2023).

[21] J. Arroyo, C. Manna, F. Spiessens, and L. Helsen, "Reinforced model predictive control (RL-MPC) for building energy management," *Appl Energy*, vol. 309, p. 118346, Mar. 2022, doi: 10.1016/J.APENERGY.2021.118346.

[22] A. T. D. Perera and P. Kamalaruban, "Applications of reinforcement learning in energy systems," *Renewable and Sustainable Energy Reviews*, vol. 137, p. 110618, Mar. 2021, doi: 10.1016/J.RSER.2020.110618.

[23] "Safety-first AI for autonomous data centre cooling and industrial control." https://www.deepmind.com/blog/safety-first-ai-for-autonomous-data-centre-cooling-and-industrial-control (accessed Feb. 20, 2023).

[24] D. L. Summerbell, C. Y. Barlow, and J. M. Cullen, "Potential reduction of carbon emissions by performance improvement: A cement industry case study," *J Clean Prod*, vol. 135, pp. 1327–1339, Nov. 2016, doi: 10.1016/J.JCLEPRO.2016.06.155.

[25] J. B. Moreno, M. Timms, and K. Wildberger, "How Artificial Intelligence is accelerating the Energy Transition An overview of AI activities at E.ON".

[26] "Microsoft Customer Story-Getting Microsoft to carbon negative with the help of cutting edge AI." https://customers.microsoft.com/EN-US/story/845665-microsoft-autonomoussystems-projectbonsai (accessed Feb. 20, 2023).

[27] S. Thongsawang, "Bringing Autonomy to Industrial Control Systems," 2020.

[28] P. Y. Simard *et al.*, "Machine Teaching: A New Paradigm for Building Machine Learning Systems," Jul. 2017, doi: 10.48550/arxiv.1707.06742.

[29] "Transforming Healthcare with Reinforcement Learning." https://www.mobiquity.com/insights/transforming-healthcare-with-reinforcement-learning (accessed Feb. 20, 2023).

[30] "Aiden: Reinforcement Learning for Electronic Trading - Borealis AI." https://www.borealisai.com/product/aiden/ (accessed Feb. 20, 2023).

[31] "Reinforcement Learning: The Business Use Case, Part 2 | by Aishwarya Srinivasan | IBM Data Science in Practice | Medium." https://medium.com/ibm-data-ai/reinforcement-learning-the-business-use-case-part-2-c175740999 (accessed Mar. 08, 2023).

[32] C. Hansen, R. Mehrotra, C. Hansen, B. Brost, L. Maystre, and M. Lalmas, "Shifting Consumption towards Diverse Content on Music Streaming Platforms," *WSDM 2021 - Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 238–246, Aug. 2021, doi: 10.1145/3437963.3441775.

[33] R. Paulus, C. Xiong, and R. Socher, "A Deep Reinforced Model for Abstractive Summarization," *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, May 2017, doi: 10.48550/arxiv.1705.04304.

[34] Z. Qin *et al.*, "Ride-Hailing Order Dispatching at DiDi via Reinforcement Learning," *https://doi.org/10.1287/inte.2020.1047*, vol. 50, no. 5, pp. 272–286, Sep. 2020, doi: 10.1287/INTE.2020.1047.

[35] D. Silver *et al.*, "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm," Dec. 2017, doi: 10.48550/arxiv.1712.01815.

[36] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," Mar. 2022, doi: 10.48550/arxiv.2203.02155.

[37] J. Xiang and L. Guo, "Comfort Improvement for Autonomous Vehicles Using Reinforcement Learning with In-Situ Human Feedback," *SAE Technical Papers*, Jan. 2022, Accessed: Mar. 31, 2023. [Online]. Available: https://par.nsf.gov/servlets/purl/10335895

[38] "Demand side response (DSR) | National Grid ESO." https://www.nationalgrideso.com/industry-information/balancing-services/demand-side-response-dsr (accessed Mar. 08, 2023).

[39] H. Kazmi, F. Mehmood, S. Lodeweyckx, and J. Driesen, "Gigawatt-hour scale savings on a budget of zero: Deep reinforcement learning based optimal control of hot water systems," *Energy*, vol. 144, pp. 159–168, Feb. 2018, doi: 10.1016/J.ENERGY.2017.12.019.

[40] Y. Lei *et al.*, "A practical deep reinforcement learning framework for multivariate occupant-centric control in buildings," *Appl Energy*, vol. 324, p. 119742, Oct. 2022, doi: 10.1016/J.APENERGY.2022.119742.

[41] Z. Zhang, A. Chong, Y. Pan, C. Zhang, and K. P. Lam, "Whole building energy model for HVAC optimal control: A practical framework based on deep reinforcement learning," *Energy Build*, vol. 199, pp. 472–490, Sep. 2019, doi: 10.1016/J.ENBUILD.2019.07.029.

[42] B. Chen, Z. Cai, and M. Bergés, "Gnu-RL: A Precocial Reinforcement Learning Solution for Building HVAC Control Using a Differentiable MPC Policy KEYWORDS Deep Reinforcement Learning, HVAC Control," 2019, doi: 10.1145/3360322.3360849.

[43] H. Y. Liu, B. Balaji, S. Gao, R. Gupta, and D. Hong, "Safe HVAC Control via Batch Reinforcement Learning," *Proceedings - 13th ACM/IEEE International Conference on Cyber-Physical Systems, ICCPS 2022*, pp. 181–192, 2022, doi: 10.1109/ICCPS54341.2022.00023.

[44] D. Azuatalam, W. L. Lee, F. de Nijs, and A. Liebman, "Reinforcement learning for whole-building HVAC control and demand response," *Energy and AI*, vol. 2, p. 100020, Nov. 2020, doi: 10.1016/J.EGYAI.2020.100020.

[45] J. Luo *et al.*, "Controlling Commercial Cooling Systems Using Reinforcement Learning," pp. 2022–2034, 2022.

[46] Y. Fu, S. Xu, Q. Zhu, Z. O'Neill, and V. Adetola, "How good are learning-based control v.s. model-based control for load shifting? Investigations on a single zone building energy system," *Energy*, vol. 273, p. 127073, Jun. 2023, doi: 10.1016/J.ENERGY.2023.127073.

[47] M. Khan, J. Seo, and D. Kim, "Real-Time Scheduling of Operational Time for Smart Home Appliances Based on Reinforcement Learning," *IEEE Access*, vol. 8, pp. 116520–116534, 2020, doi: 10.1109/ACCESS.2020.3004151.

[48] F. Alfaverh, M. Denai, and Y. Sun, "Demand Response Strategy Based on Reinforcement Learning and Fuzzy Reasoning for Home Energy Management," *IEEE Access*, vol. 8, pp. 39310–39321, 2020, doi: 10.1109/ACCESS.2020.2974286.

[49] M. Khan, J. Seo, and D. Kim, "Real-Time Scheduling of Operational Time for Smart Home Appliances Based on Reinforcement Learning," *IEEE Access*, vol. 8, pp. 116520–116534, 2020, doi: 10.1109/ACCESS.2020.3004151.

[50] H. Li, S. Member, Z. Wan, and H. He, "Real-Time Residential Demand Response", doi: 10.1109/TSG.2020.2978061.

[51] "Living Lab | Design, Test & Launch Energy Innovations." https://es.catapult.org.uk/tools-and-labs/living-lab/ (accessed Mar. 10, 2023).

[52] Z. Wan, H. Li, H. He, and D. Prokhorov, "Model-Free Real-Time EV Charging Scheduling Based on Deep Reinforcement Learning," *IEEE Trans Smart Grid*, vol. 10, no. 5, pp. 5246–5257, Sep. 2018, doi: 10.1109/TSG.2018.2879572.

[53] S. Vandael, B. Claessens, D. Ernst, T. Holvoet, and G. Deconinck, "Reinforcement Learning of Heuristic EV Fleet Charging in a Day-Ahead Electricity Market," *IEEE Trans Smart Grid*, vol. 6, no. 4, pp. 1795–1805, Jul. 2015, doi: 10.1109/TSG.2015.2393059.

[54] S. Vandael, B. Claessens, D. Ernst, T. Holvoet, and G. Deconinck, "Reinforcement Learning of Heuristic EV Fleet Charging in a Day-Ahead Electricity Market," *IEEE Trans Smart Grid*, vol. 6, no. 4, pp. 1795–1805, Jul. 2015, doi: 10.1109/TSG.2015.2393059.

[55] X. Lin, Y. Wang, P. Bogdan, N. Chang, and M. Pedram, "Reinforcement learning based power management for hybrid electric vehicles," *IEEE/ACM International Conference on Computer-Aided Design, Digest of Technical Papers, ICCAD*, vol. 2015-January, no. January, pp. 32–38, Jan. 2015, doi: 10.1109/ICCAD.2014.7001326.

[56] S. Goel, R. Sharma, and A. K. Rathore, "A review on barrier and challenges of electric vehicle in India and vehicle to grid optimisation," *Transportation Engineering*, vol. 4, p. 100057, Jun. 2021, doi: 10.1016/J.TRENG.2021.100057.

[57] H. Lee and S. W. Cha, "Reinforcement Learning Based on Equivalent Consumption Minimization Strategy for Optimal Control of Hybrid Electric Vehicles," *IEEE Access*, vol. 9, pp. 860–871, 2021, doi: 10.1109/ACCESS.2020.3047497.

[58] "Smart Local Energy Systems - Energy Systems Catapult." https://es.catapult.org.uk/tools-and-labs/our-place-based-net-zero-toolkit/smart-local-energy-systems/ (accessed Mar. 28, 2023).

[59] S. Bose, E. Kremers, E. M. Mengelkamp, J. Eberbach, and C. Weinhardt, "Reinforcement learning in local energy markets," *Energy Informatics*, vol. 4, no. 1, pp. 1–21, Dec. 2021, doi: 10.1186/S42162-021-00141-Z/FIGURES/5.

[60] D. J. B. Harrold, J. Cao, and Z. Fan, "Renewable energy integration and microgrid energy trading using multi-agent deep reinforcement learning," *Appl Energy*, vol. 318, Jul. 2022, doi: 10.1016/J.APENERGY.2022.119151.

[61] "GitHub - intelligent-environments-lab/CityLearn: Official reinforcement learning environment for demand response and load shaping." https://github.com/intelligent-environments-lab/CityLearn (accessed Mar. 24, 2023).

[62] K. Nweye, S. Sankaranarayanan, and Z. Nagy, "MERLIN: Multi-agent offline and transfer learning for occupant-centric energy flexible operation of grid-interactive communities using smart meter data and CityLearn".

[63] J. R. Vazquez-Canteli, G. Henze, and Z. Nagy, "MARLISA: Multi-Agent Reinforcement Learning with Iterative Sequential Action Selection for Load Shaping of Grid-Interactive Connected Buildings," *BuildSys 2020 - Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pp. 170–179, Nov. 2020, doi: 10.1145/3408308.3427604.

[64] K. Nweye, S. Sankaranarayanan, and Z. Nagy, "MERLIN: Multi-agent offline and transfer learning for occupant-centric energy flexible operation of grid-interactive communities using smart meter data and CityLearn".

[65] V. Khattar and M. Jin, "Winning the CityLearn Challenge: Adaptive Optimization with Evolutionary Search under Trajectory-based Guidance," 2023, Accessed: Mar. 24, 2023. [Online]. Available: www.aaai.org

[66] "CityLearn Challenge 2022 - Workshop at NeurIPS 22 - YouTube." https://www.youtube.com/watch?v=Yel5zybmvwg (accessed Mar. 24, 2023).

[67] "Cost of balancing Britain's power grid shatters record - Nuclear Industry Association." https://www.niauk.org/cost-of-balancing-britains-power-grid-shatters-record/ (accessed Mar. 28, 2023).

[68] A. R. Fuxjäger, K. Kozak, M. Dorfer, P. M. Blies, and M. Wasserer, "Reinforcement Learning Based Power Grid Day-Ahead Planning and AI-Assisted Control", Accessed: Feb. 20, 2023. [Online]. Available: https://eur-lex.

[69] P. de Mars and A. O'Sullivan, "Applying reinforcement learning and tree search to the unit commitment problem," *Appl Energy*, vol. 302, p. 117519, Nov. 2021, doi: 10.1016/J.APENERGY.2021.117519.

[70] C. O'Malley, P. de Mars, L. Badesa, and G. Strbac, "Reinforcement Learning and Mixed-Integer Programming for Power Plant Scheduling in Low Carbon Systems: Comparison and Hybridisation," Dec. 2022, [Online]. Available: http://arxiv.org/abs/2212.04824

[71] L. Floridi *et al.*, "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds Mach (Dordr)*, vol. 28, no. 4, pp. 689–707, Dec. 2018, doi: 10.1007/S11023-018-9482-5/FIGURES/2.

[72] R. Eitel-Porter, "Beyond the promise: implementing ethical AI," *AI and Ethics 2020 1:1*, vol. 1, no. 1, pp. 73–80, Oct. 2020, doi: 10.1007/S43681-020-00011-6.

[73] M. Nadimpalli, "Artificial Intelligence Risks and Benefits Big Data & Analytics View project Artificial Intelligence View project Artificial Intelligence Risks and Benefits," *International Journal of Innovative Research in Science, Engineering and Technology (An ISO*, vol. 3297, 2007, Accessed: Mar. 31, 2023. [Online]. Available: https://www.researchgate.net/publication/319321806

[74] B. Cheatham, K. Javanmardian, and H. Samandari, "Confronting the risks of artificial intelligence With great power comes great responsibility. Organizations can mitigate the risks of applying artificial intelligence and advanced analytics by embracing three principles," 2019.

[75] I. Gabriel, "Artificial Intelligence, Values, and Alignment," vol. 30, pp. 411–437, 2020, doi: 10.1007/s11023-020-09539-2.

[76] S. Karjalainen, "Gender differences in thermal comfort and use of thermostats in everyday thermal environments," *Build Environ*, vol. 42, no. 4, pp. 1594–1603, Apr. 2007, doi: 10.1016/J.BUILDENV.2006.01.009.

[77] N. D. Sintov, L. V. White, and H. Walpole, "Thermostat wars? The roles of gender and thermal comfort negotiations in household energy use behavior," *PLoS One*, vol. 14, no. 11, p. e0224198, Nov. 2019, doi: 10.1371/JOURNAL.PONE.0224198.

LICENCE/DISCLAIMER

CATAPULT
Energy Systems

OUR MISSION

TO UNLEASH INNOVATION AND OPEN NEW MARKETS TO CAPTURE THE CLEAN GROWTH OPPORTUNITY.

ENERGY SYSTEMS CATAPULT
7TH FLOOR, CANNON HOUSE,
18 PRIORY QUEENSWAY,
BIRMINGHAM, B4 6BS.

ES.CATAPULT.ORG.UK
@ENERGYSYSCAT