Responses to Reviewers

Contents

Reviewer	Page No.	Total No. of Pages
Reviewer 1	2	2
Reviewer 2	4	1
Reviewer 3, Round 1	5	6
Reviewer 3, Round 2*	11	2

*Miscellaneous confidential communication between Corresponding Author (SD) and Energies editors and MDPI board have been withheld.

Reviewer-1

"Dear Authors, This paper provided an in-depth review of the applications of reinforcement learning (RL) for home energy management systems (HEMS). However, there is some scope for improvement."

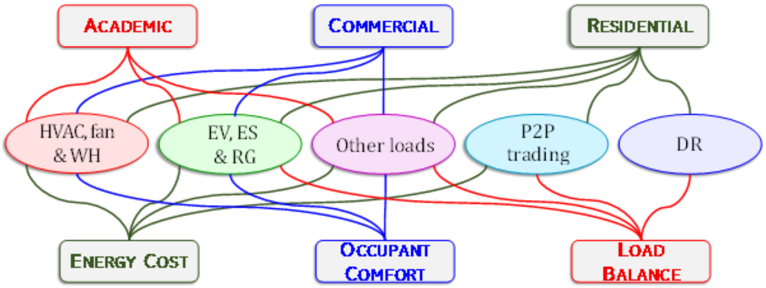
Thank you. We have tried to address all the concerns in the modified manuscript.

"Please mention the full form of the abbreviation DNN on the first page."

We would like to thank the referee for pointing this out, which is also mentioned in the journal's guidelines. Accordingly, we have included the full form of the acronym DNN in the abstract as well as in the list of keywords.

"Kindly provide a block diagram describing the home energy management system."

We have provided a block diagram that connects HEMS applications with building types as well as RL objectives in Figure 13 "Building Types and Objectives" in the revised manuscript. Instead of this diagram, if the esteemed reviewer had a pictorial representation in mind, we may add that the communication protocol is the only factor that Figure 13 omits. Figure 13 is shown below:



"On page 3, second sentence, the word "win" should be replaced with "wins"."

We have rearranged the relevant paragraphs and reworded that sentence for better English. The word "wins" is now replaced with "advantageous" or "preferred".

"Table 2 has not been referred to in the text."

We apologize for this omission. The revised manuscript contains a paragraph:

> In a similar manner, Table 2 considers references that used DQN. Most algorithms in the survey used DQN. However, DDQN was also popular in the HEMS research community. The survey found that dueling-DQN was applied in only one article.

"Kindly include the conclusion and future work at the end."

We apologize for this rather egregious omission. Section 8 (Conclusion) has now been added to the revised manuscript.

"Presenting a short commentary on what type of RL technique to select for a given HEMS application will be a good contribution. This can be included after section 7."

Although the specific approach would be very case-specific, the revised manuscript contains recommendations that would help the reader decide on the specific approach (or not to use RL at all). This appears in the Conclusion section:

There is a diverse array of algorithms in RL. Since tabular methods require discrete states and actions, and that the spaces be of low cardinalities, are discrete, they may not be much use for most HEMS applications. Not surprisingly, this survey shows that tabular methods have been less frequently than DNN methods. In future, as the HEMS community investigates increasingly complex HEMS domains, tabular methods would become even less likely to be used. Consequently, the choice of algorithm would usually be confined to DNN methods.

Out of the DNN methods, it must be noted that DQN and its derivatives can only be used in applications only when the action space is finite and small, such as in controlling OFF-ON switches. The survey reveals that actor-critic methods, which include Q-learning and policy learning, are the most popular in HEMS applications. Another deciding factor is whether to use policy-free or policy-based RL. On-policy learning may be used is applications where abandoning the policy in the initial stages may occasionally very negatively impact the environment. Thus, they may be used if the environment does not require too much exploration. On the other hand, off-policy RL can discover more novel policies.

Reviewer-1

Reviewer-2

"This is by far one of the highest quality papers in the energy journals I have reviewed. I have a few comments."

We thank the esteemed reviewer for such encouraging remark. We have tried to address each comment in the revised manuscript.

"1. In Section 2.1, the authors focus on Zigbee. I have researched zigbee 10 years ago, and I am curious if this technology is not yet obsolete? Almost all smart homes are now using wifi for communication."

This is an excellent comment. Wi-Fi is used for relatively smaller sets of appliances. We have inserted the following passage in the revised manuscript:

> The choice of communication protocol for home automation is an open question. To a large extent, it depends on the user's personal requirements. If it is desired to automate a smaller set of home appliances with ease of installation, and operability in a plug-and-play manner, Wi-Fi is the appropriate one to use. However, with more ex-tensive automation requirements, involving tens through hundreds of smart devices, Wi-Fi is no longer the optimal choice. There are issues relating to scalability and signal interference in Wi-Fi. More importantly, due to its relatively high energy consumption, Wi-Fi is not appropriate for battery powered devices.

A new reference has been inserted in the very next paragraph.

"2. In Section 2.3, don't HEMS control algorithms have methods based on MPC and its variants? Some applications of MPC and its variants are described in this reference 'Model Controlled Prediction: A Reciprocal Alternative of Model Predictive Control'. I suggest the authors cite this paper in this section and add some description related to MPC."

The revised manuscript contains a paragraph that cites this reference [172]:

> The only truly viable alternative to RL is to use nonlinear control, more specifically model predictive control (MPC) [171]. MPC is widely used in various engineering applications [cf. 172]. The benefit of MPC is in the explicit manner by which it handles physical constraints. At each iteration, MPC considers a receding time horizon in-to the future and applies a constrained optimization algorithm to determine the best control actions. However, in most cases MPC uses linear or quadratic objective functions. This is a basic limitation that must be taken into account before applying MPC to large-scale problems and is in sharp contrast to RL that does not place any restriction of the reward signal. Moreover MPC is a model based approach, whereas an over-whelming majority of references in this survey used model-free RL methods ([138] being the sole exception).

"3. In the second page, the authors give many examples of rl, and I add a class of applications where there are many intelligent transportation algorithms based on rl, such as vehicle dispatching. I suggest the authors cite some relevant papers, such as 'Deep dispatching: A deep reinforcement learning approach for vehicle dispatching on online ride-hailing platform'."

We have added the sentence, and the suggested reference [36]:

> In intelligent transportation systems, RL is used in a range of applications such as vehicle dispatching in online ride-hailing platforms [36].

Reviewer-3

Round 1

Reviewer-3, Round-1

“The paper will be ready for publication after major revision. ”

We thank the esteemed reviewer for such encouraging remark. We have tried to address each comment in the revised manuscript as best as we can.

“The literature should be supported by more publications from Energies. ”

All papers from Energies that we could retrieve in reinforcement learning that were specific to HEMS applications were included in the original manuscript. Several new references have been added in the revised manuscript, including those from from other MDPI journals.

“What does HEMS stand for? It should be mentioned at least in abstract.”

We would like to thank the referee for pointing this out, which is also mentioned in the journal’s guidelines. Accordingly, we have included the full form of the acronym HEMS in the abstract. The extended form of HEMS has also been included as a keyword.

“The authors need to interpret the meanings of the variables. Please highlight your contributions in introduction.”

We have added several sentences/phrases at various places in the introduction of the revised manuscript to highlight the contributions.

“More statistical measures used to assess the model accuracy should be mentioned Modeling of solar energy systems using artificial neural network: A comprehensive review”

Unlike in case of supervised learning, validating a reinforcement learning is nearly impossible. This is because one cannot define accuracy, and there is no “desired output” available for each action. Ergo, mean squared error, average absolute error, correlation coefficients, linear regression coefficient, etc. that are so commonly used in supervised learning, cannot be used in RL. We make this issue clear in the following explanatory figure. However, we have added a few sentences as well as many new references to address this issue in the Conclusion section:

> Unlike in the unsupervised and supervised learning where simple performance metrics are readily available, performance evaluation in RL is an open problem [JCC+20]. The steadily increasing reward with iteration is the best means for any real application. The authors suggest that the following four criteria should be considered.
> (*i*) Maximum at Saturation: The value of the reward must be relatively high at saturation.
> (*ii*) Variance: The reward must not have a large amount variance at saturation.
> (*iii*) Rate: The number of iterations before the reward starts to saturate should not be large.
> (*iv*) Initial Minimum: At the initial stages the minimum reward received due to more exploration, must not be so low that the environment is adversely affected.

$$\text{Mean Squared Error} = \frac{1}{N}\sum_{n}\left(y(n) - \boxed{t(n)}\right)^2$$

$t(n)$ is the desired (or ‘target’) output of sample n
Supervised learning: $t(n)$ is available in the training data.
Reinforcement learning: $t(n)$ is not available. Must use indirect methods [e.g. Eqn. (6)].

Reviewer-3, Round-1

“The paper is well-written, I have to thank you to your effort. ”

We would like to thank the esteemed reviewer for appreciating our work.

“Figure 1. The ANN structure. Need a reference. ”

We have added the following in the Fig 1. legend:

> “Although the agent is depicted as a neural network (cf. [REF]), it may be in the form of a tabular structure.”

“What are the main features in Figure 2?”

Since Section 3.2 describes the various ways to classify RL methods, we have added the following in the Fig 2. legend:

> “Section 3.2 provides a description of each class.”

“Actor-critic methods are hybrid approaches that borrow features from value based as well as policy-based RL. Add a reference.”

We have added a reference [54] in that sentence. It now reads:

> “Actor-critic methods are hybrid approaches that borrow features from value-based as well as policy-based RL [54].”

“The introduction should be supported by: A new optimized artificial neural network model to predict thermal efficiency and water yield of tubular solar still; Modeling ultrasonic welding of polymers using an optimized artificial intelligence model using a gradient-based optimizer; Productivity forecasting of solar distiller integrated with evacuated tubes and external condenser using artificial intelligence model and moth-flame optimizer; Fine-tuned artificial intelligence model using pigeon optimizer for prediction of residual stresses during turning of Inconel 718.”

Please see response:
”**Why Reinforcement Learning is not an Optimization Metaheuristic**”.

“These publications present the integration between ANN and metaheuristic optimizers which may support the introduction section. ”

Please see response:
”**Why Reinforcement Learning is not an Optimization Metaheuristic**”.

“Discuss the following in the introduction: A new fine-tuned random vector functional link model using Hunger games search optimizer for modeling friction stir welding process of

polymeric materials; A new optimized predictive model based on political optimizer for eco-friendly MQL-turning of AISI 4340 alloy with nano-lubricants."

Please see response:
"**Why Reinforcement Learning is not an Optimization Metaheuristic**".

"The abstract should be rewritten to reflect the significance of the proposed work. The current abstract shows a lot of background information."

Thank you for this very useful observation. We have completely rewritten the abstract.

"Conclusion: What are the advantages and disadvantages of this study compared to the existing studies in this area?"

The revised manuscript contains a Conclusion section. This section provides a brief outline of the attractive features as well as the limitations of each RL method. As such, RL is incomparable with any other existing study as it solves a different problem.

"The inspiration of your work must further be highlighted."

The manuscript contains the following:

> In contrast to the previous reviews, the scope of our review is broad enough to cover all areas of HEMS, including HEMS interfacing with the energy grid. More importantly, it provides a comprehensive overview of all major RL methods, providing a sufficient level of explanation for readers' understanding. Therefore, this article would be of benefit for researchers and practitioner in other areas of the energy systems, and beyond, to acquire a theoretical level understanding of basic RL techniques.

Which may have been overlooked. We have added a sentence to the abstract in the revised manuscript:

> The steep rise in reinforcement learning (RL) in various applications in energy, as well as the penetration of home automation in recent years are the motivation for this article.

Lastly, we have inserted words and phrases at other places to further address this issue.

"Some suggested recent literatures should be added."

We have covered in the survey all papers that were published in leading HEMS journals within the past five years. But in light of this concern (among other issues) several new references have been cited in the revised manuscript.

"From a machine learning standpoint, stochastic policies help explore and assess the effects of the entire repertoire of actions available in . Such exploration is critical during the initial stages of the learning algorithm. ; what do you mean by exploration?"

This is an excellent suggestion. We have added the explanation to bring out the difference between exploration and exploitation.

> Exploration is a applied to stochastically search and evaluate the available repertoire of actions at each state, before converging towards the optimal ones. It is an essential component of value-based RL. Since exploitation is the strategy of picking the best actions in Π, it should not be applied until the algorithm has all actions in a sufficient manner. However, endowing the learning algorithm with too much exploration slows down the learning. Identifying the right tradeoff between exploration and exploitation is a widely studied problem in machine learning [REF]. It is for this reason that the parameters ϵ in Eqn. (9), and τ in Eqn. (10) are steadily lowered as learning progresses.

Reviewer-3, Round-1

Reviewer-3, Round-1

The above paragraph includes a new reference.

“Add the publication year as a column in Table 1-5.”

We understand that it is frequent practice in review articles. Unfortunately, we discovered that adding a new column was impossible without reducing the font sizes, a practice that is counter to MDPI guidelines. Since the survey spans a relatively short period of five years, the inclusion of such a column might not provide as much new information to the reader. However, in the list of references, we have arranged all references used in the tables alphabetically, and then using publication year.

“Copyrights should be obtained for any figure copied from another published study.”

All figures are our own, where we sometimes used smaller images that were placed in public domain. Thus, there are no copyright issues.

“Some recent references in ANN applications may be included probably from MDPI.”

Please see response:
”**Why Reinforcement Learning is not an Optimization Metaheuristic**”.

“What are factors have been considered to evaluate the model accuracy?”

RL relies solely on reinforcement feedback signals. In many situations this signal is zero, or just “good” or “bad”. For instance, when training an agent to play chess, how can every move evaluated for accuracy? The only evaluation is at the very end – the RL is “accurate” if it can defeat a human, otherwise it “inaccurate”. A similar situation holds for an RL agent that is trained to drive a vehicle. All agents that don’t cause traffic accidents are “accurate”. Hence, model accuracy is extremely difficult/impossible to formulate in RL.

We have added a brief explanation addressing this issue. This appears in the last section (Conclusion) of the revised manuscript. We have included a new reference.

We have also proposed a set of four metrics for performance evaluation. These are

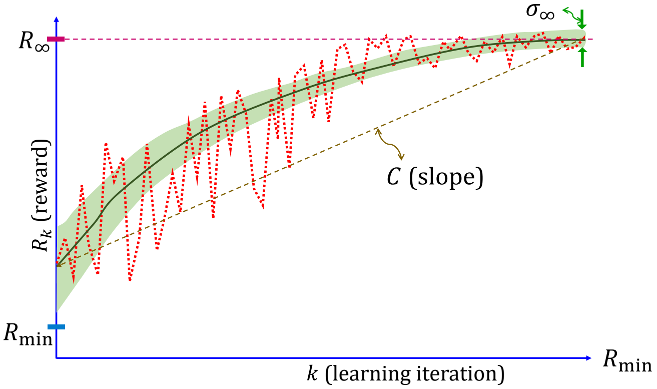
$$R_\infty \approx \frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}} R^i_{K_i^{\max}}\,;$$

$$\sigma_\infty \approx \frac{1}{|\mathcal{J}|-1}\sum_{i\in\mathcal{J}}\left(R^i_{K_i^{\max}} - \hat{R}_\infty\right)^2;$$

$$R_{\min} \approx \min_{i\in\mathcal{J}}\min_{k} R^i_k \;\left(\text{or}, R_{\min} \approx \min_{i\in\mathcal{J}} R^i_1\right);$$

$$C \approx \frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}} \frac{R^i_0 - R^i_{K_i^{\max}}}{K_i^{\max}}.$$

Detailed explanation of each metric is provided in the Conclusion section, which also includes a new figure (Figure 17), which is also shown below:



Reviewer-3, Round-1

“The space between value and units may be eliminated.”

Done.

“Looking and wishes for the revised version.”

Thank you.

Why Reinforcement Learning is not an Optimization Metaheuristic

Reinforcement Learning vs. Optimization Metaheuristics

Criterion	Reinforcement Learning	Optimization Metaheuristic
Continuous Time Analogy	**Optimal Control.**	**None exists.**
Use of Heuristics	**None.** There are formal convergence proofs for value-based and policy-based RL.	**Heuristic-based.** Convergence may be statistically proven. A proof may not even exist for new methods (many exist).
Search Method	**Dynamic Programming** (value-based) **Gradient** (policy-based).	**Various** (e.g. Darwinian evolution, foraging strategy of swarms), based on some biological or social phenomenon.
Stochastic/ Deterministic	**Deterministic.** Tabular learning are 100% deterministic. Often the agent’s policy itself is stochastic, but not the RL algorithm itself. Stochasticity is sometimes useful for exploration.	**Stochastic.** E.g. in a genetic algorithm mutation & crossover *must* be randomized.
Number of Agents	**One agent** (or two for actor-critic RL as it combines value-based and policy-based RL).	**Many agents** (**Population/Swarm**). It is necessary for the algorithm to maintain a “population” or “swarm” of candidate solutions.
Performance Evaluation	**Very difficult/impossible** to formulate. However, methods can be devised to *indirectly* evaluate the performance of a fully trained RL agent, e.g. with an actual human being. Moreover, performance evaluation is **not required** by the algorithm as it only uses reinforcement signals as feedback.	**Very easy** to formulate performance measures. Standard metrics such as RMS or absolute error can be readily used. A performance measure is also **required** by the algorithm to operate. It is generally called as the candidate solution’s “fitness”.
Use of Standard Backpropagation	**Cannot be used** (unless/except in conjunction with an Experience Replay Buffer).	**Should be used**. These metaheuristics run on top of standard backpropagation. (In theory, backpropagation may be discarded, but doing so is always a bad idea!)
Definition of “Error”	The “error” (*actual output – desired output*) in back-propagation cannot be used. In RL, back-propagation is replaced with another formula based on the reward. Even when using an Experience Replay Buffer, the “error” is defined as, ***current true output – previous true output***	The error is defined in the usual manner: ***actual output – desired output.***

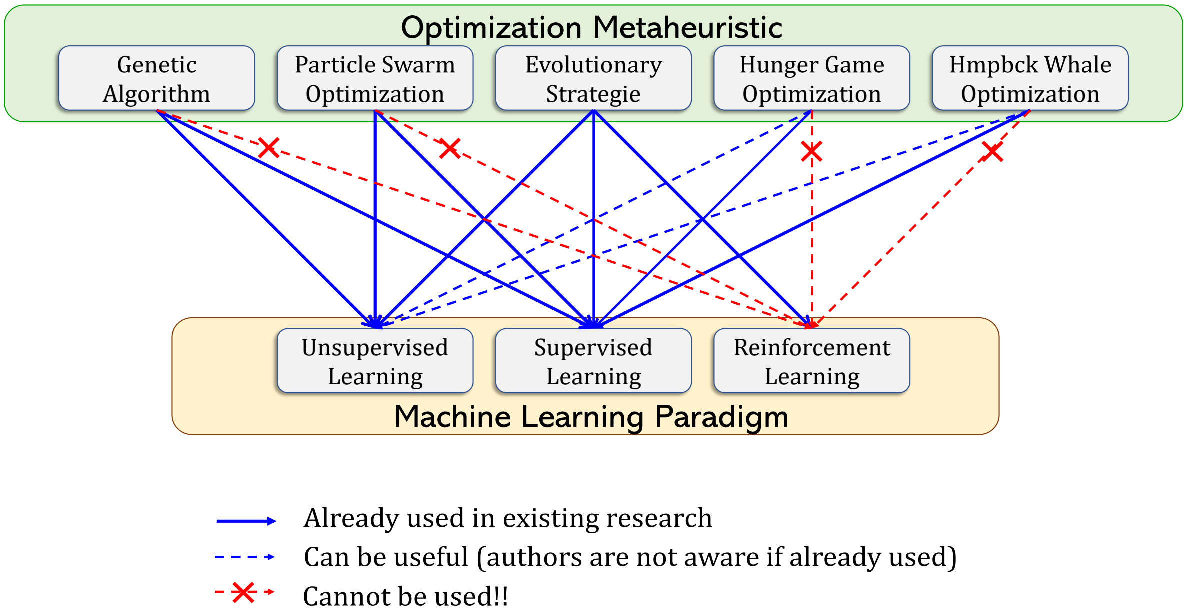
Reinforcement learning is one of the three fundamental machine learning paradigms. These three machine learning paradigms are:

Unsupervised learning. Examples: clustering, dimension reduction, manifold learning,
Metric learning.

Supervised learning. Examples: classification, regression, forecasting, prediction.

Reinforcement learning. Examples: Learning to play a strategic game such as Chess or Go,
Learning how to navigate a passenger car.

In our view, reinforcement learning is significantly the more difficult than the other two learning paradigms. For instance one can imagine why training a learning agent to play a game of chess with an opponent is much more difficult than training it to forecast the energy demand of a distribution system for every hour. Reinforcement learning is the ultimate goal of AI (Artificial Intelligence) because it allows a learning agent to learn and behave like a human being. Metaheuristics (evolutionary algorithms, swarm intelligence, etc.) can be applied to any of these machine learning paradigms.


Optimization Metaheuristic
Genetic Algorithm
Particle Swarm Optimization
Evolutionary Strategie
Hunger Game Optimization
Hmpbck Whale Optimization
Unsupervised Learning
Supervised Learning
Reinforcement Learning
Machine Learning Paradigm
Already used in existing research
Can be useful (authors are not aware if already used)
Cannot be used!!

Reviewer-3, Round-1

Reviewer-3

Round 2

My email query sent to agency responsible for copyright issues:

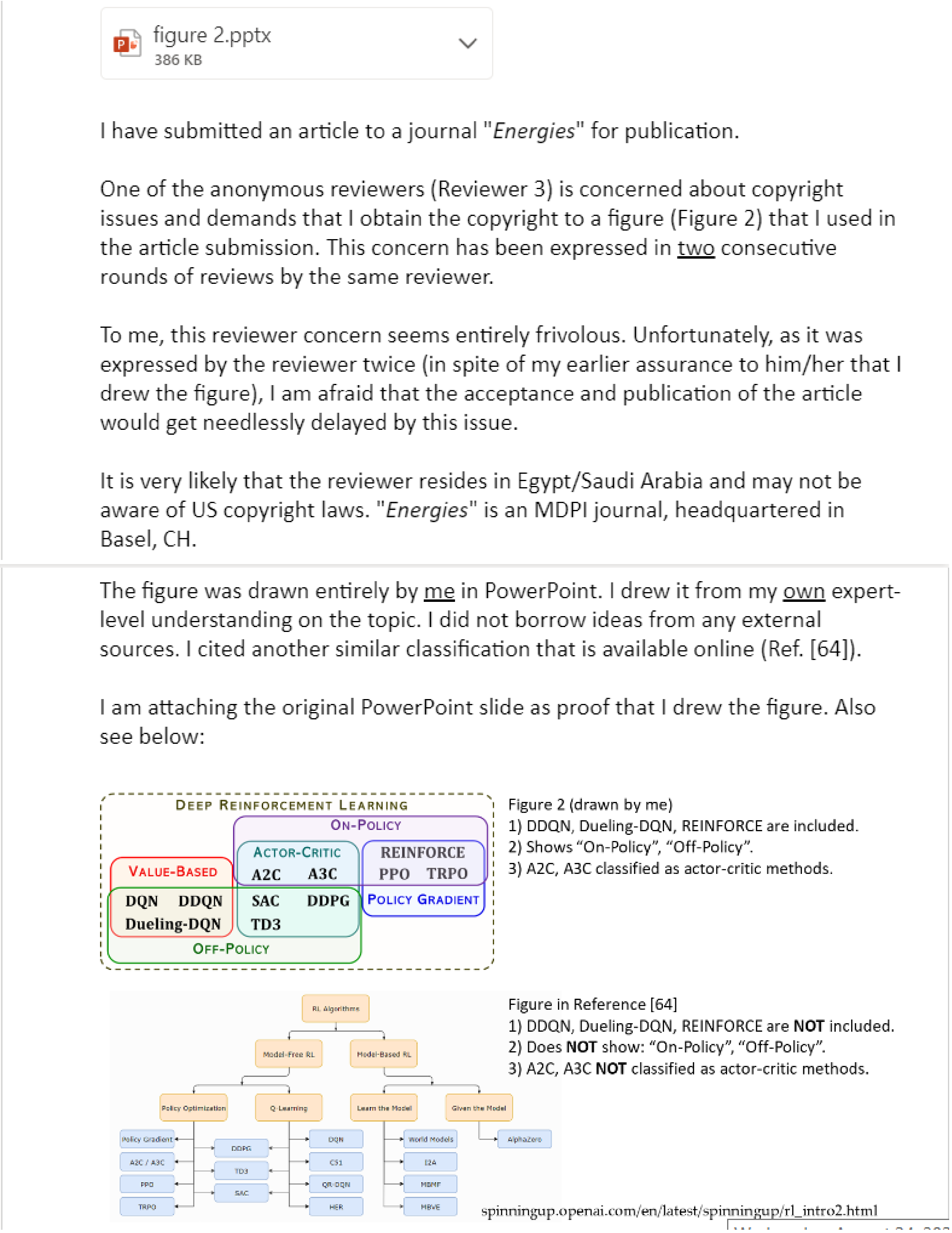


figure 2.pptx
386 KB

I have submitted an article to a journal "*Energies*" for publication.

One of the anonymous reviewers (Reviewer 3) is concerned about copyright issues and demands that I obtain the copyright to a figure (Figure 2) that I used in the article submission. This concern has been expressed in <u>two</u> consecutive rounds of reviews by the same reviewer.

To me, this reviewer concern seems entirely frivolous. Unfortunately, as it was expressed by the reviewer twice (in spite of my earlier assurance to him/her that I drew the figure), I am afraid that the acceptance and publication of the article would get needlessly delayed by this issue.

It is very likely that the reviewer resides in Egypt/Saudi Arabia and may not be aware of US copyright laws. "*Energies*" is an MDPI journal, headquartered in Basel, CH.

The figure was drawn entirely by <u>me</u> in PowerPoint. I drew it from my <u>own</u> expert-level understanding on the topic. I did not borrow ideas from any external sources. I cited another similar classification that is available online (Ref. [64]).

I am attaching the original PowerPoint slide as proof that I drew the figure. Also see below:

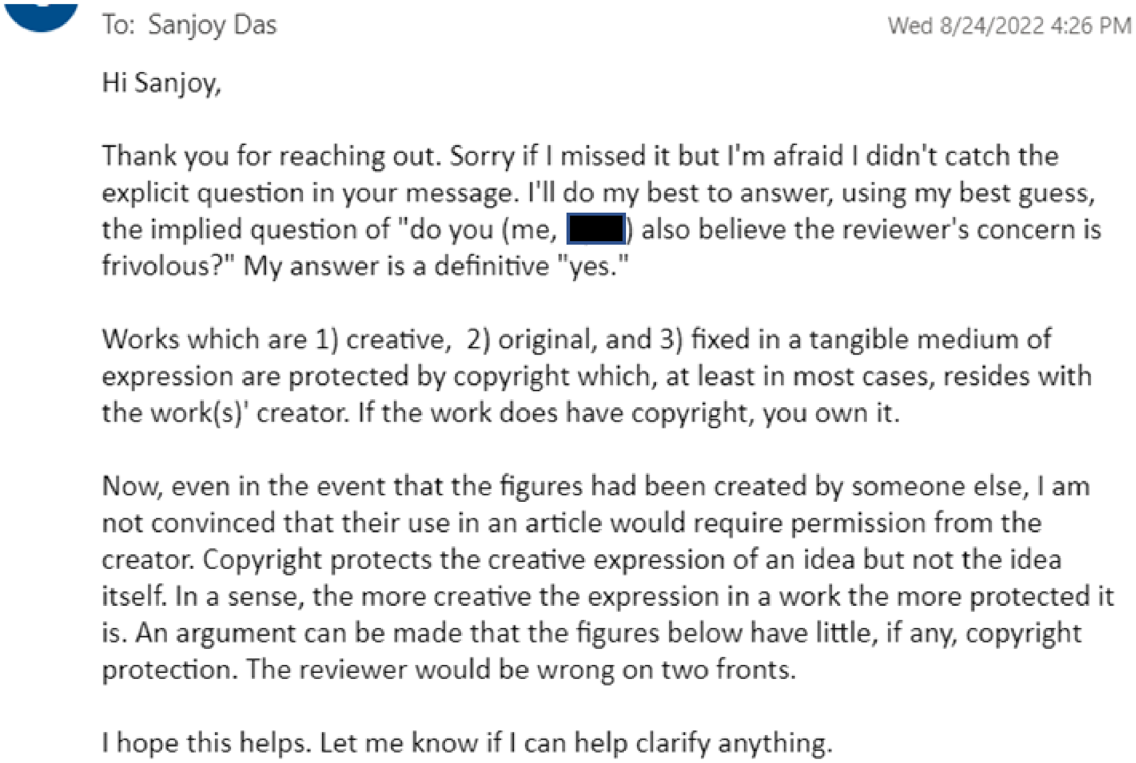
Figure 2 (drawn by me)
1) DDQN, Dueling-DQN, REINFORCE are included.
2) Shows "On-Policy", "Off-Policy".
3) A2C, A3C classified as actor-critic methods.

Figure in Reference [64]
1) DDQN, Dueling-DQN, REINFORCE are **NOT** included.
2) Does **NOT** show: "On-Policy", "Off-Policy".
3) A2C, A3C **NOT** classified as actor-critic methods.

spinningup.openai.com/en/latest/spinningup/rl_intro2.html

Reviewer-3, Round-2

Response from CADS to my query:



To: Sanjoy Das

Wed 8/24/2022 4:26 PM

Hi Sanjoy,

Thank you for reaching out. Sorry if I missed it but I'm afraid I didn't catch the explicit question in your message. I'll do my best to answer, using my best guess, the implied question of "do you (me, [redacted]) also believe the reviewer's concern is frivolous?" My answer is a definitive "yes."

Works which are 1) creative, 2) original, and 3) fixed in a tangible medium of expression are protected by copyright which, at least in most cases, resides with the work(s)' creator. If the work does have copyright, you own it.

Now, even in the event that the figures had been created by someone else, I am not convinced that their use in an article would require permission from the creator. Copyright protects the creative expression of an idea but not the idea itself. In a sense, the more creative the expression in a work the more protected it is. An argument can be made that the figures below have little, if any, copyright protection. The reviewer would be wrong on two fronts.

I hope this helps. Let me know if I can help clarify anything.

Reviewer-3, Round-2

Note: Miscellaneous confidential communication between corresponding author (SD) and Energies editors as well as MDPI board have been withheld.