Reply to Reviewer's Comments on "Training a gaming agent on brainwaves"

We are grateful to all the reviewers for their tremendously helpful feedback.
In the following, we discuss how we dealt with each raised issue.

. . . . . .

REVIEWER #1 TRANSCRIPT:

This paper proposes the use of a gaming agent to be trained using Reinforcement Learning (RL) as well as the feedback obtained from EEG brainwaves of human critic observers. The idea of using brain patterns for boosting ML is highly interesting and well suited for the journal. Although it is not a new concept, it is nicely re-purposed for online use-case.

However, I am not sure if there is something new in the paper compared to the 2010 proof-of-concept (Iturrate et al.: Robot Reinforcement Learning using EEG-based reward signals). The presentation style seems a bit confusing and needs serious improvements. Background is very limited (some are incorporated in the introduction) and substantial work is required. There has been a lot of work in BCIs over the past decade and some in the area of games.

From the scientific part, the results are not so convincing. In particular, the proposed idea does not seem to be too practical method. The classification does not provide something significant as it stands at the moment. The sample used is also quite small and the complexity of the game is limited.

Nevertheless, the paper could be improved in many ways. An obvious point would be to improve the classifiers. Apart from this, I would suggest that authors would look at ways of manipulating the stimulus, stimulus presentation or giving out incentives to participants.

REVIEWER #2 TRANSCRIPT:

The paper is interesting and relevant. However in my opinion it needs revision for purposes of clarification and to improve the contribution. Please see attached document for my comments.

Training a Gaming Agent on Brainwaves There are too many grammatical errors to list individually. //Does the header refer to a general template? JOURNAL OF LATEX CLASS FILES, VOL. 14, NO. 8, AUGUST 2015 //The abstract read more like an introduction, rather than an indication of the work undertaken for the paper. 25 "Results show that there is an effective transfer of information and that the agent learns successfully to solve the game efficiently." //This is too vague. What are the outcomes of the study? 39 This information is used to make a gaming agent improves its operational performance using electroencephalography (EEG) signals as feedback of the performed task, obtained from an observational human critic. //There is a grammar issue. I don't understand this sentence. 54 RL should be defined in main body of the paper. Also the term 'agent' needs to be defined/clarified. How does it relate to the Game Manager from Figure 1? 17 col 2 Recently, this technique has seen a come-back. //this is not scientific language 58 col 3 The precision of 25.125 is inappropriate P2, 46 col 2 What is "state information"? P3, 2 red normally indicates an error P3, 11 I assume that the "observational human critic" is the player/participant/subject/human observer. This role should be clarified and terminology used consistently. P3, 41 define MNE P3, 56 Thus, each epoch is composed of a matrix 500 x 8. // add channels P4, 32 Hence, following the iterative procedure based on Equation 1, the Q-Table is updated in each iteration. After the algorithm finishes iterating through all the training episodes, the Q-Table is stored to test the performance of the agent. // Will the game always terminate? How long does the game take? Does smooth progression toward finish affect the err potential? P4 Fig 3. Is the chance score = 0.5?

P4 the labels referred to in Fig 5 should correlate with the test here, i.e., A= subject 1 etc. P4 referring to Fig 6; a comparison ROC curve with subject 1 would be more interesting

P5, 34 what is meant by “experiences”? P5, 29 col 2 – remove Average steps per Q-Table legend. P5 50, col 2 The collected data show that ErrP signals can in fact be classified and used to train an agent effectively. // how has effectiveness been determined here? Page 6 Fig 10 the text is not legible, it should be improved. Page 6, 40 “However, even though this implies that the agent misses frequently that an action taken is wrong, this is not hindering the overall performance and the agent is still learning.” // Your results show that this is subject dependent P6, 56 Results show that training a classifier with data of one subject, but using it to classify the events of experiences of another subject does not lead to an improvement on the performance of the agent. // could a pre-trained generic classifier provide a better initial state, subsequently trained with observer data to converge more quickly? Are there differential err potential for up, down, left, right?

REVIEWER #3 TRANSCRIPT:

Comments to the Author The following work shows the usage of ErrPs for training a gaming agent using reinforcement learning. Authors attempt different conventional classification approaches, as well as intersubject classification.

It is unclear to me the benefit of evaluating single subject to single subject offline classification accuracy. Please elaborate on this. Since different subjects attained different performance, all combinations of subjects used for training and testing should be inspected for the 1:1 evaluation setup. Maybe classification accuracy could be reported at different steps (depending on the amount of training data – gradual increasing the number of subjects) What is the ratio of hit/no hit segments of action (epochs)? Page 3, column 2, lines 4-9. It is unclear at this stage what action from the starting point would generate an ErrP. moves-¿ move Page 3, column 2, lines 11-13. Has the MinMax Scaler been applied in any other BCI related study or elsewhere? I think including a reference would be useful. Page 4. Figure 3. I would suggest adding another set of bars for the grand average classification scores. Page 4, Figure 4. This figure is redundant and could be removed. Page 5, Figure 5. The titles of the subplots would be more descriptive when replacing the letters with the corresponding subjects number. Page 5, column 1, line 52, Not -¿no Page 5, column1, line 55, accumulative-¿cumulative Page 6, Figure 8. It is unclear to me what is the benefit of showing both A and B. Since the behavior is similar in both cases. Page 6, Figure 10. This figure is, in my opinion, incomplete. What is the reason for showing this subset of subjects? I suggest showing the confusion matrices of all subjects and/or their average. Page 6, column 2, line 40. show -¿ shows

Reviewer 1 General Comments

This paper proposes the use of a gaming agent to be trained using Reinforcement Learning (RL) as well as the feedback obtained from EEG brainwaves of human critic observers. The idea of using brain patterns for boosting ML is highly interesting and well suited for the journal. Although it is not a new concept, it is nicely re-purposed for online use-case.

However, I am not sure if there is something new in the paper compared to the 2010 proof-of-concept (Iturrate et al.: Robot Reinforcement Learning using EEG-based reward signals). The presentation style seems a bit confusing and needs serious improvements. Background is very limited (some are incorporated in the introduction) and substantial work is required. There has been a lot of work in BCIs over the past decade and some in the area of games.

We conducted a new BCI and games literature review and updated the Introduction accordingly, adding more updated background. We highlight now what we think is our contribution, which is a simple game that can be used to trigger the ErrP response in a simple scenario that can be used to train an agent. We were unable to find a similar work from that same perspective.

From the scientific part, the results are not so convincing. In particular, the proposed idea does not seem to be too practical method. The classification does not provide something significant as it stands at the moment. The sample used is also quite small and the complexity of the game is limited.

Issues raised by the Reviewer are accurate and timely described. We aimed to find a suitable game that at the same time can be trained with a Reinforcement Learning technique and that triggered the ErrP response as well. We modified the manuscript, emphasized three results of our experiments. First we show that the basic Q-Learning algorithm is robust to noisy signals. Additionally, we show the futility of transfer learning approaches using the brainwave signals, and finally that the cumulative contribution of the rewards obtained from different subjects enhances the performance of an agent.

Nevertheless, the paper could be improved in many ways. An obvious point would be to improve the classifiers. Apart from this, I would suggest that authors would look at ways of manipulating the stimulus, stimulus presentation or giving out incentives to participants.

We added a new classifier and performed two more experiments. First we included kNeighbours as a new classification algorithm. This can be seen in Section C and in Figure 3. In that same Figure, we also added results obtained while performing signal averaging to enhance the SNR of the ErrP component. Moreover, we performed a complete transfer learning experiment, verifying if it was possible to improve the efficiency of the gaming agent and verified that the best results for the improvement (i.e. reduction) on the average number of steps, required to reach the target in 200 plays, is achieved when a classifier is trained with ErrP signals from the same subject that is used to identify the ErrP. This is added in Section IV and in Figure 7. On the other hand, the structure of the game scenario allows to easily extend the experiments to verify what is the outcome of manipulating the stimulus, their presentation or the influence on the results if incentives are given to participants. We added this compelling ideas to the Section IV.

Reviewer 2 General Comments

The paper is interesting and relevant. However in my opinion it needs revision for purposes of clarification and to improve the contribution. Please see attached document for my comments.

Training a Gaming Agent on Brainwaves There are too many grammatical errors to list individually.

We truly appreciate this level of detail in all the comments and the requests for information.

//Does the header refer to a general template? JOURNAL OF LATEX CLASS FILES, VOL. 14, NO. 8, AUGUST 2015

This error has been fixed. Thanks for pointing it out.

//The abstract read more like an introduction, rather than an indication of the work undertaken for the paper. 25 “Results show that there is an effective transfer of information and that the agent learns successfully to solve the game efficiently.” //This is too vague. What are the outcomes of the study?

We performed several changes in the manuscript. We modified the Introduction describing those changes and emphasizing what we think may be our main contributions. We rewrote the Abstract completely.

39 This information is used to make a gaming agent improves its operational performance using electroencephalography (EEG) signals as feedback of the performed task, obtained from an observational human critic. //There is a grammar issue. I don't understand this sentence.

We modified the sentence. We thanks the reviewer for pointing out this issue.

54 RL should be defined in main body of the paper. Also the term 'agent' needs to be defined/clarified. How does it relate to the Game Manager from Figure 1?

The acronym was revised and we also verified all the other acronyms used in the text. The term "agent" was defined and presented in the Introduction, and their relation with the Game Manager in Figure 1 was established.

17 col 2 Recently, this technique has seen a come-back. // this is not scientific language

This mistake has been corrected. We apologize for the bad choice of words.

58 col 3 The precision of 25.125 is inappropriate P2, 46 col 2 What is "state information"?

The precision was adjusted to two decimals for the age mean and the standard deviation. We removed "state information" and replaced the wording of the phrase in order to clarify more clearly the information that we wanted to convey: we refer to state information to the sequence of movements that the gaming agent performs on the game.

P3, 2 red normally indicates an error

We understand the Reviewer's point, and we agree with her/him that red is not the right color to represent the completion of a task in the game, and in this case it may have cognitive implications regarding the generalized concept that red indicates an error. It is known that the ErrP response may be affected by colors and shapes (Eimer 1997, *An event-related potential (ERP) study of transient and sustained visual attention to color and form*). The experiments that we performed were implemented with the board showed on Figure 2 of the manuscript. We added this very important issue at the conclusions in relation with possible future works.

P3, 11 I assume that the "observational human critic" is the player/participant/subject/human observer. This role should be clarified and terminology used consistently.

Absolutely. We preferred to use OHC "observational human critic" to emphasize what the subject is actually doing while using the system. We clarified this notation at the Introduction and along the manuscript.

P3, 41 define MNE

Historically MNE standed for Minimum Error Estimate, a software package developed in Martinos Center of Harvard University. Now MNE is the name of a entire software platform to perform analysis of several type of brain signals like Magnetoencephalograpy and Electroencephalography. We added a clarification about this on Section II.C.

P3, 56 Thus, each epoch is composed of a matrix 500 x 8. // add channels

Excellent. We added that information.

P4, 32 Hence, following the iterative procedure based on Equation 1, the Q-Table is updated in each iteration. After the algorithm finishes iterating throughout all the training episodes, the Q-Table is stored to test the performance of the agent. // Will the game always terminate? How long does the game take? Does smooth progression toward finish affect the err potential?

When the gaming agent moves randomly on the board of Figure 2, it takes on average 100 steps to arrive to the final location on the grid. Each step, the movement direction is performed once every 2 seconds, so on average it takes around 200 seconds to finish the game. After training during *run sessions* (check new Section II.F), the agent starts to use the Q-Table to decide the next movement. However, the Q-Table could potentially end up producing endless loops or deadlocks so if the steps count arrives to 200 the game is interrupted and it starts all over again. So far we didn't test if the Error Potential is affected by the smooth progression toward the end, which is indeed something interesting to verify in this simple scenario. We added this very interesting extension in Section IV.

P4 Fig 3. Is the chance score = 0.5?

Yes, it is. We added that information in the Figure's caption.

P4 the labels referred to in Fig 5 should correlate with the test here, i.e., A= subject 1 etc.

The figure labels on Figure 4(now) were modified to reflect which subfigure references which subject. Additionally, we use the same notation on the new Figure 5.

P4 referring to Fig 6; a comparison ROC curve with subject 1 would be more interesting

We added on Figure 5(now) all the ROC curves that we obtained for all the different subjects, including the one for Subject 1.

P5, 34 what is meant by “experiences”?

We added a new Section II.F to clarify this part of the experiment which was not explained at all. First we rewrote how we separated the two parts of the experiment, the ”Cognitive Game Procedure” and the ”Gaming Agent Learning Procedure”. We replaced ”experiences” with *run session*. This all is explained in the above-mentioned section.

P5, 29 col 2 – remove Average steps per Q-Table legend.

This was modified in all the Figures, adjusting fonts and labels. We appreciate a lot the time to point out this issue.

P5 50, col 2 The collected data show that ErrP signals can in fact be classified and used to train an agent effectively. // how has effectiveness been determined here?

Each time the gaming agent plays this simple game, it takes on average around 100 steps to reach the target spot. We asked an observational human critic to watch the gaming agent play the game, while we recorded their brainwaves. We trained first a classifier to be able to recognize Error Potentials from these brainwaves. We started the game again, and the gaming agent started to move around the board, but this time once an Error Potential was identified from the brainwaves, it was used as a negative reward to train a new Q-Table for the agent. The game was iteratively repeated, but the gaming agent used the Q-Table that was updated in the previous run session. We verified that by doing this experiment, the agent required less and less steps on average to reach the goal until it arrives to the optimal number of around 10 steps. We found that although the effectiveness of this improvement depended on the accuracy of the classifier, even with very low values (just above chance level), the agent learns and the number of average number of steps is reduced. We verified that if trained the agent with sham signals, completely uncorrelated with the reward, the agent learned nothing, and the average number of steps was not reduced at all. Finally, we also verified that if we train a classifier with Error Potentials from one subject and used that classifier to provide the rewards for the experiment, the number of steps is not reduced, and that do not depend on the subjects, is an intrinsic result which is produced when intermixing the OHCs. Only the usage of a trained classifier from the same OHC produces this improvement. Finally, we verified that rewards obtained from different OHCs, can be used to improve the performance of the gaming agent collaboratively.

Page 6 Fig 10 the text is not legible, it should be improved.

This is now Figure 9, we increased the fontsize, added a colormap and included the confusion matrices for all the subjects.

Page 6, 40 “However, even though this implies that the agent misses frequently that an action taken is wrong, this is not hindering the overall performance and the agent is still learning.” // Your results show that this is subject dependent

Yes, the Reviewer is correct. We modified this sentence to state more clearly the message that we wanted to convey, which is, that results are completely subject dependent. But that even when many action should be tagged as ”wrong” (triggering an ErrP) only with a few of them right, it is enough to train the agent towards the optimal solution.

P6, 56 Results show that training a classifier with data of one subject, but using it to classify the events of experiences of another subject does not lead to an improvement on the performance of the agent.
// could a pre-trained generic classifier provide a better initial state, subsequently trained with observer data to converge more quickly? Are there differential err potential for up, down, left, right?

The statement is absolutely right. We extended this experience per Reviewer's suggestion and we verified that effectively the only combination of training/testing in the recognition of ErrP that produces an improvement on the performance of the agent is the one that uses information from the same subject (check new Figure 7). We think a generic-classifier to identify ErrP potentials for any subject (i.e. without calibration) is a research goal of the BCI community because this will allow to have more robust, easier to use, with shorter setups BCI devices (Suller 2012,*Online use of error-related potentials in healthy users and people with severe motor impairment increases performance of a P300-BCI*). If such a generic pre-trained classifier could be conceived, this simple game can be used to verify if the number of training sessions required is reduced and this will confirm that the pre-trained classifier is helping to do it more quickly. We didn't verify if the error potential for up, down, left, right per se produces a different error potential, though we verified that in general getting farther from the target (in Manhattan distance) produces a signal which is slightly different from the one that is produced when the gaming agent moves closer to the target. We added these very interesting lines of further exploration in Section IV.

. . . . . .

Reviewer 3 General Comments

Comments to the Author The following work shows the usage of ErrPs for training a gaming agent using reinforcement learning. Authors attempt different conventional classification approaches, as well as intersubject classification.

We appreciate a lot the time dedicated to review our manuscript.

It is unclear to me the benefit of evaluating single subject to single subject offline classification accuracy. Please elaborate on this.

We tested many variants of simple games and we found that only the one proposed here produced a distinctive ErrP response. The single subject accuracy that is presented on Figure 5(now) shows the level of ErrP single trial identification that we achieved for different subjects (which is low). We wanted to emphasize the finding that even with such low levels of identification, it was possible for the agent to learn an efficient strategy.

Since different subjects attained different performance, all combinations of subjects used for training and testing should be inspected for the 1:1 evaluation setup.

This is an excellent idea that we now explored with a new experiment. We showed the results on Figure 7, where a classifier is trained to recognize ErrP signals using the information obtained from one Trainer subject and used to generate rewards for a gaming agent on a different Tester subject. We added information about this on Section IV.

Maybe classification accuracy could be reported at different steps (depending on the amount of training data – gradual increasing the number of subjects)

The classification accuracy obtained for all subject is now being shown on Figure 5 with ROC curves. What is shown progressively is the experiment itself, where after each gaming agent training match, a new Q-Table is generated based on the rewards tagged from the brainwave session. The agent improves their efficiency by reaching the target in a progressively less number of steps on average for each run session.

What is the ratio of hit/no hit segments of action (epochs)?

The ratio hit/no-hit segments is 50/50, so for each brainwave session match, as the average number of steps is around 100, it will be 50 segments of hit (moving further from the target) and 50 of no-hit (moving closer).

Page 3, column 2, lines 4-9. It is unclear at this stage what action from the starting point would generate an ErrP. moves-¿ move

The Reviewer is absolutely right in that there isn't any possible action that the gaming agent can perform at the beginning of the experience (the game agent is located at the upper-left corner) that may trigger an ErrP response. We modified the phrase to emphasize the message that we wanted to convey. Additionally, there is a very important caveat of the experiment. The gaming agent always has a 5% chance of wandering and selecting an action randomly, regardless of their Q-Train. This is precisely to avoid loops or deadlocks situations.

On the other hand, we fixed the pointed out grammatical error as well.

Page 3, column 2, lines 11-13. Has the MinMax Scaler been applied in any other BCI related study or elsewhere? I think including a reference would be useful.

We added a reference where the MinMaxScaler is used in a BCI application (new page 3, second column).

Page 4. Figure 3. I would suggest adding another set of bars for the grand average classification scores.

We added a new bar on the Figure 3 where we included the obtained classification accuracy when 5 segments are ensemble averaged to produce an averaged signal (with an improved SNR in terms of the ErrP). These scores are reported in the Figure 4 as "GrandAvg". There is a balance to be considered because adding more segments to the averaged signal reduces the number of available samples used for training the classifier.

Page 4, Figure 4. This figure is redundant and could be removed.

The figure was removed, thank you very much for your comment.

Page 5, Figure 5. The titles of the subplots would be more descriptive when replacing the letters with the corresponding subjects number.

We agree with Reviewer's comment. We removed the letters and replace them with subject's numbers in Figures 4 and 5.

Page 5, column 1, line 52, Not -¿no

Fixed. Thank you very much !

Page 5, column1, line 55, accumulative-¿cumulative

Fixed.

Page 6, Figure 8. It is unclear to me what is the benefit of showing both A and B. Since the behavior is similar in both cases.

We included now all the ROC curves for all the subjects.

Page 6, Figure 10. This figure is, in my opinion, incomplete. What is the reason for showing this subset of subjects? I suggest showing the confusion matrices of all subjects and/or their average.

Absolutely right. We included all the confusion matrices for all the subjects. The reason for this is that we wanted to convey the message that our hypothesis is that as false positives

are low, those ErrP that are triggered are indeed good ”rewards” that are useful for the training algorithm and the gaming agent is using that information, tough scarce, in a useful way to learn and improve it performance.

Page 6, column 2, line 40. show -¿ shows

Fixed.

. . . . . .