**Reply to Reviewers**

In this reply, Reviewer and Editor comments are in black and our replies in blue. *Italic font style* is used wherever text from the revised manuscript is quoted.

Dear Editor of PLoS Biology,

We would like to thank you for allowing us to submit a revised manuscript despite, admittedly, a lack of focus and clarity, as well as exaggerated claims in the original manuscript.

Following your recommendations and those of the reviewers, we have proceeded with a substantial revision and provide a point-by-point reply to all the reviewer comments. Our revision briefly entails:

* An explicit and clear statement of our hypothesis that mutual learning is a crucial element for translational, non-invasive motor imagery (MI) brain-computer interaction (BCI). Related to this, we clarify that we use the term “mutual learning” equivalently to co-adaptation, but prefer the former as “co-adaptation” has been, in our opinion, too strongly linked to adaptive classifier approaches in the field of MI BCI.
* A complete shift of the article’s focus from our team’s victory in the Cybathlon BCI race and the demonstration of “translational impact of non-invasive MI BCI” to what is actually the scientifically significant finding in our study: The substantiation of subject learning effects, which are further facilitated by the application design, next to and correlated with machine learning. All claims made on the grounds of our team’s victory have been toned down.
* An explicit statement of the significance of our study, namely, the promotion of a much needed paradigm shift in BCI user training, from the current machine-learning-oriented protocols towards training strategies that take into account all three learning pillars: the subject, the machine and the application.
* A complete rewriting of the Introduction and Discussion sections to reflect the aforementioned changes.
* Additional important references and discussion of those, as pointed out by the reviewers, as well as a better and far more elaborate discussion of our findings with respect to the relevant literature.
* Additional results to corroborate the role of the application in BCI training and the merits of our user training strategy in comparison to that of our Cybathlon competitors.
* The provision of further information and references that have now become available on the methods adopted and the specifics of the Cybathlon competition, as requested by the reviewers.

We are confident that our much revised manuscript is devoid of the original one’s identified deficiencies. Furthermore, we believe that it reports on significant findings which should have a strong impact on the field of translational BCI. Therefore, we truly think that our manuscript is a useful addition to the literature and a good fit to the PLoS Biology journal.

We hope for a positive evaluation of our revised manuscript.

Sincerely yours,

The authors

**Reviewer 1**

“Authors present a manuscript on MI-driven BCI, using data based on their participation on the Cybathlon BCI race 2016. They present information on the longitudinal training of two participants with chronic spinal cord injury. Their key finding is that one of their participants won the BCI race, while the other established the record time. It is argued that translational MI-BCI should be considered a mutual learning skill. Moreover, the achieved performances are considered evidence of the maturity of translational MI-BCI technology. The authors and participants are to be congratulated for winning the race. Reading the manuscript as a report on the Cybathlon was informative. However, several issues need to be addressed from a scientific perspective”

Comment 1.1: “What was the hypothesis of this study? How was it operationalized as research questions? How are they addressed in the manuscript? Which evidence is provided?”

We apologize that our original manuscript has been indeed obscure and confusing with respect to this essential point, and we thank the reviewer for giving us the opportunity to clarify it.

**Main hypothesis**

In the revised manuscript, our study’s hypothesis is explicitly stated right away in the Introduction: *“In this study, we investigated the hypothesis that mutual learning is a critical underlying factor for the success of motor imagery (MI) BCI in translational applications. Contrary to a popular trend of focusing almost exclusively on the machine learning aspects of MI training, our hypothesis propounds that a holistic mutual learning training approach grounded symmetrically on all three learning pillars (at the machine, subject and application level) would be the optimal training apparatus for preparing two end-user participants for the Cybathlon BCI race, the first international BCI competition”*. Of note, we have completely removed the, admittedly, far-fetched argumentation regarding the demonstration of “substantial translational impact thanks to our findings” and generally toned down all our claims.

**Research questions**

Derived from our hypothesis, the specific research questions we have set out to answer were the following:

1. Was our “mutual learning” end-user training approach able to bring about a) application performance improvement, b) machine learning and c) subject learning effects, involving end-users suffering from Spinal Cord Injury (SCI) and in harsh training and operation conditions (home-training, stressful tournament)?
2. Were such effects relevant, i.e. can a mechanism be established where the accomplishment of subject and machine learning objectives explains the improvement in application performances?
3. How does our mutual learning methodology compare to competitive approaches (which we believe are in general strongly focused towards machine learning)?
4. Can critical factors of a good mutual learning approach be identified?

**Data for addressing the research questions**

We address the first two research questions by analyzing the data (EEG imaging, application performances) acquired with two end-user participants throughout the training period leading up to (and including) the Cybathlon 2016 BCI race event. All variables of interest (primary and secondary outcomes, see below) of the proposed “mutual learning” approach are monitored over time (we define the “run” as the time-point of evaluation) and the same protocol was applied to both participants. For the remaining two research questions, we further make use of the competition outcomes and information collected from our competitors regarding their user training approach.

**Evidences provided**

Specific evidence is provided to answer (to the extent possible) all the identified research questions. Of note, competition outcomes are, as explained below, supporting evidence but not our key finding.

1. The training effects of our method under the aforementioned conditions are shown with **strong and significant correlations between the variables of interest (at the machine, subject and application levels) and the defined chronological point indices (runs)**. We also regularly provide differences and their statistical significance between the naive performances (first sessions/runs) and the learned outcome (last sessions/runs). In other words, we show that all application, machine and subject-related variables improve significantly with the amount of training: Such comprehensive information is usually not shown in the literature. These variables are, first, the race completion time, which we consider to be the primary outcome as the main goal of any BCI training is to achieve effective and efficient application control. Additionally, the secondary outcomes of BCI command accuracy and “time on pad” (which simultaneously evaluates command delivery speed) are employed as the metrics to assess the machine learning, and the discriminancy of sensorimotor rhythm (SMR) features for the subject learning merits of our protocol. **Importantly, we consider such detailed demonstration of subject learning to be a major finding of our study and a very useful addition to the relevant literature** (see our reply to Comment 2.3 by Reviewer 2).
2. The impact of subject learning and machine learning on application performance improvement (i.e. how improved SMR modulation allows a well calibrated BCI to better decode the subjects’ intentions) is shown with **strong, significant correlations between all these variables, what establishes the mechanism of our mutual learning paradigm**. In other words, we show that when one variable improves, the other follows. Although correlation does not necessarily imply causation, causes and effects are in this case pretty clear thanks to the employment of a standard pattern recognition framework: increasing discriminancy of brain features allows properly calibrated decoders to also increase BCI command accuracy and delivery speed, which in turn allows for decrease of the race completion time. **Of note, such assessment of impact of subject learning is regularly missing in the relevant literature.** Interestingly, our results also **suggest a possible role of the application design not only in the corresponding performances, but also on subject learning, what we think is a novel finding.** The relevance of the achieved learning outcomes is also showcased with comparisons to, first, the performance of a perfect input on the same task and, second, by presenting also the performances of our competitors in the Cybathlon.
3. Although our longitudinal study was conceived as an uncontrolled, observational study, information collected from most other competitors via a questionnaire (see new supplementary Table S2), allows us to have a fair control group as they have essentially adopted a training methodology mainly based on machine learning, while we followed a more holistic mutual learning methodology. Despite the limitations of this comparison, we feel that the **Cybathlon BCI race performances** (the race times achieved, rather than the fact that our team won and the overall final ranking), even if representing rather circumstantial evidence, **seem to support our claim that a mutual learning approach based on all three learning pillars will be more effective for translational applications than machine learning-oriented approaches that are dominant in recent years.**
4. Lastly, based on the same evidence (questionnaires, comparisons with competition and previous work) we have been able to isolate distinctive differences of the followed training approaches, so as to at least reasonably **speculate on the elements that must be crucial for a mutual learning approach**. Specifically, we explain, how a) longitudinal (but, not necessarily year-long) training with adequate intensity, b) strong motivational contexts including training with a real application, c) incorporation of educational principles like incremental difficulty and, most likely, d) parsimonious machine adaptation that coordinates the machine and subject learning objectives are key factors for the emergence of mutual learning effects.

We believe that all the above points are now clearly presented in the Introduction and Discussion sections of our revised manuscript.

Comment 1.2: “It is argued that the performance in the Cybathlon is evidence for the technological maturity of non-invasive BCI. How does this match with the statement on p.13 the "main shortcoming of our BCI system is unsatisfactory robustness" and "lack of robustness is a well-known issue”.”

We agree with the reviewer’s assessment that this claim has been exaggerated and we have completely removed it from the revised manuscript. We have completely shifted the focus from the demonstration of “a translational impact of non-invasive MI BCI” to the substantiation of the merits of a comprehensive mutual learning approach. Yet, we comment on the issue of unsatisfactory robustness in a subsection *“Limitations”* in the Discussion.

.

Comment 1.3: “Authors argue that they offer "unique proof" and are the "very first" (p.2, Â§3) to show evidence that MI-BCI is a skill that can be learned. This is not the case. Authors acknowledge the existence of earlier evidence for learning in the same paragraph. There is also evidence for MI-BCI learning in other studies and groups they do not cite. Authors might consider revising their claims and including earlier work in more detail.”

Please, see also the related reply to Comment 2.3 by Reviewer 2. We apologize for exaggerated and misunderstandable claims in the original submission. In the revised manuscript, we have abstained from any claim of uniqueness and priority. Please, note also that it was by no means our intention to argue, already in the original manuscript, that we are the first and only to hypothesize and provide evidence on the fact that BCI is “as skill to be learned”. As also pointed out by the reviewer, we had already cited literature bringing such evidence. The claim we have really wished to make, and by which we still stand, is that our results **bring stronger and more complete evidence, in particular at the neurophysiological level, towards the possibility and magnitude of subject learning effects in non-invasive MI BCI with respect to previous literature**, and that such evidence is much needed to reverse the current attitude in BCI training of focusing almost exclusively on machine learning.

The completely re-written Discussion section of the revised manuscript includes now a much extended literature review of subject learning and discusses elaborately the place of our work therein (section “*Subject learning in MI BCI*”). For more details, see our reply to Comment 2.3 by Reviewer 2.

Comment 1.4: “Presenting limited data on two subjects cannot be considered "beyond reasonable doubt at all levels".”

We again apologize for exaggerated claims in the original manuscript. This and similar phrases have been now removed in the revised manuscript. However, it must be noted that the limitation of including only two individuals had been acknowledged already in the original manuscript.

Furthermore, please consider that the phrase “beyond reasonable doubt at all levels” was not meant to suggest that our findings should necessarily extend to all end-users. Regarding this point, both the original and the revised manuscripts simply report “cautious optimism”, but absolutely no such certainty: *“Still, the fact that both participants exhibited the same training effects and comparable performances, makes us confident that our conclusions should generalize, at least to populations with similar clinical profile.”*

Instead, this phrase was meant to convey that our work is one of very few to report learning effects at all three identified levels –neuroimaging, machine and application– and provides strong evidence on the potential impact subject learning might have within the mutual learning scheme.

Comment 1.5: “Authors argue that it is unclear whether MI-BCI is a skill learning or a neural decoding problem, and that their "work contributes to enlighten this dilemma". Lotte (2013), cited by the authors, do not make such claims. They might consider whether they really wanted to imply a dilemma, i.e. a paradoxical situation requiring a decision, or just "two domains offering space to explore for improving MI-BCI".”

The reviewer is right: This unfortunate expression has essentially twisted the message we want to convey. It has been removed from the revised manuscript. In fact, we do believe that subject and machine learning are, as the reviewer puts it, "two domains offering space to explore for improving MI-BCI". It is for this reason that we advocate for “mutual learning” protocols. In the revised manuscript, it is now clear that we do not regard the term “mutual learning” as any different from the more common one, “co-adaptation”, while the merits of machine learning in our findings are better highlighted in the new Discussion section.

The message we actually want to pass forward is that subject learning is of vital importance and must be promoted in a mutual learning framework. We consider this message to be important as, in our opinion (but also quite evidently in the literature), machine-learning-oriented approaches are vastly dominating non-invasive MI BCI, including many paradigms that claim to be “co-adaptive”. On the contrary, we find that subject learning has been very much neglected in this field in the past 10 years. The goal of our manuscript is to push towards the restoration of subject learning as equally important to machine learning by showing its great potential, and not to downgrade or invalidate machine learning (which, of course, our training framework is also making use of).

Consequently, we think that the work of Lotte et al. (2013) and similar publications that have followed since, are aligned with our position, because the ideas proposed there for improving BCI training protocols (incorporation of educational principles like explanatory feedback, incremental learning, etc) are mostly focused on what could be improved on the subject learning side.

Comment 1.6: “Please provide convincing evidence to answer this important research question (whether MI-BCI is mutual learning vs. pure skill learning vs. pure neural decoding), e.g. "infrequent calibration of the BCI" (p12), i.e. "twice per subject" (p.11, table 2) and the contrast between different control paradigms (e.g. figure 7 or 1C) seems not enough to substantiate these claims.”

We apologize that our original manuscript has been confusing regarding this essential message. In the revised manuscript, we make it clear and provide evidence that BCI should be treated as a mutual learning problem. Specifically, we show and discuss how thanks to a mutual learning protocol based on all three learning pillars (subject, machine and application) a) SMR modulation improved for both end-users b) BCI accuracy increased and reached high accuracy levels, and c) as a results of these, application performances improved.

In this respect, and as further clarified in the revised manuscript (Discussion, sub-section *“Evidence of mutual learning during training for the Cybathlon”*), although the decoders were re-calibrated only twice per subject, *“possible new classifiers were trained after every session with the new recorded data and the simulated performances were evaluated. In such an iterative process, most of classifiers were discarded during the training period due to similar performances. One might argue that such an infrequent BCI re-calibration contradicts the mutual learning hypothesis. However, this approach is substantiated by the fact that BCI decoding achieved high-level accuracy (Fig. 3A) for both users after the initial re-calibrations. Thus, we had assumed that the machine learning model was sufficiently optimized.”*

Comparison to other participating teams via the questionnaire seems to indicate that we were the only to adopt a comprehensive mutual learning approach, and competition outcomes suggest that this was the differential factor. This is clarified in the revised version of the manuscript (Discussion, sub-section *“Limitations”*): *“we can rely on our competitors as a fair control group because they have essentially adopted a training methodology mainly based on machine learning, as per the results of the questionnaire (Table S2), while we followed a more holistic mutual learning methodology. Indeed, their approach involved frequent classifier recalibration and feature re-selection, as well as training protocols that were either relatively short and/or not particularly intense. Of note, the differences in machine learning methods of all participating teams were too subtle to explain the competition outcomes according to the organizers [8].”*

Finally, new Figure 6D shows how the control paradigm facilitates subject learning (Results, subsection “*Effects of the application in BCI control and learning”*): *“Interestingly, we can show that the involvement of the application in the learning process also creates bi-directional interactions. Specifically, Fig. 6D shows the feature discriminancy of the first and last ten runs of the training periods with the three control paradigms. Interestingly, the discriminancy significantly increased only in the case of control paradigm 4 (0.27±0.07 to 0.34±0.05, p=.045, two-sided Wilcoxon ranksum test), while no difference (or even a reduction) is reported for the other two paradigms. Results suggest that the refinement of the control paradigm might have had a critical role in facilitating subject learning.”*

Comment 1.7: “Figure 1C shows that control paradigms were not evenly and randomly sampled across the whole training duration. The claim that ~10 s performance increase can be attributed to the control paradigm change based on a simple Wilcoxon rank sum test (Figure 7C) cannot be considered "beyond reasonable doubt at all levels"”

Thank you for this observation. In accordance with this comment, in the revised manuscript, we have avoided to speculate about the exact magnitude of the improvement brought forward by the experimentation with different control paradigms, because it is difficult to separate the effects of subject learning and those of the control paradigm on race time. Especially since new results presented in Fig. 6D (former Fig. 7 in the original manuscript) show that subject learning was only evident during the application of paradigm 4 (and not paradigm 3), which means that indeed the reported test must be confounded with the subject learning.

Nevertheless, when including in the “effects of the control paradigm” their ability to facilitate subject learning, what we believe is one of our important findings, then the existing statistical test fairly depicts what was gained thanks to the control paradigm. For this reason, we have preserved this result in the revised manuscript.

In addition to this, motivated by this comment, in the revised manuscript we provide evidence to suggest that there was also some immediate effect of the control paradigm (beyond the combined one confounded with the induced subject learning). Specifically, we report a significant reduction of race time between the last 10 races with paradigm 3 and the first 10 races with paradigm 4, where subject learning could not have had substantial impact (paradigm 3: 130.1 ± 17.2 s paradigm 3: 112.4 ± 15.1 s, p=0.03) given the fact that these 20 samples are adequately interleaved (Fig. 1A).

Lastly, as already mentioned, the exaggerated claim “beyond reasonable doubt” has been removed, while it is better communicated that “at all levels” is meant to stress that learning/training effects could be derived at the subject (Fig. 4), machine (Fig. 2/3) and application (Fig. 1) levels of the interface.

Comment 1.8: “Please tone down the claims.”

Following similar comments by all reviewers and recommendation of the editor, we have toned down all claims made in the revised manuscript and apologize for exaggerations and overstatements in our original submission.

Comment 1.9: “Figures are not interpretable when printed in gray/bw. Authors might consider revising them and using different markers instead of colors when appropriate.”

In the revised manuscript we have proceeded with modifications to make the figures more readable in gray/bw, like the use of striping and the increase of line width where needed. We have opted to not completely remove information in color, since there was already information provided through symbols (Fig. 1) and we think that four additional ones would make the figures more difficult to interpret. We prefer to use colors for the sake of readability, at least during the review process.

**Reviewer 2**

Comment 2.1: “This lengthy manuscript describes the sensorimotor rhythm(SMR)-based BCI training and performance of two subjects with high-level spinal cord injuries. In addition to the BCI data from two subjects who are very disabled, its distinctive feature is that these subjects ultimately performed, and triumphed, in a public BCI competition. This is interesting and impressive. The strong emphasis on the essential role of subject learning in BCI performance is commendable; this principle is unfortunately still often ignored in the BCI literature and in BCI study design.”

We would like to thank the reviewer for acknowledging the merits of our manuscript and for sharing the opinion that subject learning is most often ignored in the literature and BCI design. Indeed, assisting in reversing this attitude has been the main motivation behind this work.

We would like to also point out that the main finding of our study is not our team’s victory in the Cybathlon competition but, rather, the demonstration of learning/training effects at the subject, machine and application level during training of our two end-users with a comprehensive mutual learning protocol. As elaborated in the remaining of this reply, we take full responsibility of failing to communicate this and other important messages in the original manuscript.

Lastly, we apologize for the length of our original and revised submissions. Although we have tried to keep the article as concise as possible, we are afraid that, in order to comprehensively respond to all reviewer comments, the revised manuscript is in fact a little longer than the original one (especially the Discussion section).

Comment 2.2: “At the same time, the paper makes unduly exaggerated statements as to the new insights it provides and the unprecedented excellence of its results. These seem to derive principally from inadequate attention to the developments in BCI understanding, methods, and performance over the past 25 years.”

We apologize for exaggerated claims in our original manuscript, a fact pointed out by all reviewers. All claims are toned down in the revised manuscript and we avoid any statement of priority and uniqueness.

However, we do believe there is considerable contribution of our work to the state of the art, which, admittedly, has not been communicated properly. We argue that, as explained below, this failure is related to a lack of focus and clarity in our original submission rather than insufficient knowledge or understanding of the relevant literature on our part. We are putting every possible effort in this reply and in the revised manuscript to persuade this reviewer that this is really the case. In particular, in the new Discussion of the revised manuscript (sub-section *“Subject learning in MI BCIs”*), we have extended our reference list with important works we have originally missed, elaborately discussed the findings of previous studies and, especially, explicitly identified their limitations, so as to better place our own work in this literature and highlight its contribution.

Comment 2.3: “The fact that effective BCI performance is a skill that the subject acquires through practice has been evident since the earliest SMR studies in the early 1990s. Fifteen years ago, this principle was emphasized and discussed in detail (under the heading “BCI Use is a Skill”) in the most widely cited BCI review article (Clin Neurophysiol 113:767-791, 2002); and it has been discussed in many other primary papers, review articles, and chapters over the past 25 years. The occurrence and importance of improvement with training has also been consistently apparent in the reports from the major single-neuron-based BCI groups (e.g., Schwartz, Andersen, Carmena). In the case of the single neurons in these animal studies, their relationships to actual movement served as logical starting points for BCI control, neuronal behaviors changed with training; and, as a result, performance improved. In similar fashion, with SMR-based BCIs in humans, motor imagery is useful mainly to provide starting points for the subsequent adaptations (by the subject and the BCI) that lead to better SMR-based BCI performance. Indeed, subjects who achieve the most complex forms of SMR-based control achieved to-date – graded control of multidimensional movements – typically report that they do not use imagery, rather they perform BCI control in an automatic fashion similar to normal muscle-based control (though not as yet as rapid or reliable). Furthermore, the claim of the manuscript that it gives the first detailed documentation of the acquisition of control by the subject with training is not correct; Fig. 2 in the paper first demonstrating 3D SMR-based control (J. Neural Eng. 7 (2010) 036007) shows the gradual acquisition of control by the subjects. Note that the algorithm that derived the control parameters from the subjects’ past data did not change over this period; thus, if the subjects’ SMR control had not improved, their performances would not have improved.”

Please, refer also to our reply to Comment 1.3 by Reviewer 1. Indeed, the well-known review of Wolpaw et al. (2002) elaborates on the need for co-adaptive interfaces, what has probably been a critical factor in sparking the interest in the design and experimentation with such interfaces since then. We cite this paper in the revised manuscript. Furthermore, it is also true that the issue of co-adaptation has been experimentally very well established in invasive BCI, where a lot of evidence has been provided at the level of neural correlates. Our original submission was citing two of these works (Ganguly et al., 2009; Orsborn et al., 2014) and the revised manuscript explicitly acknowledges this fact and cites more similar ones. Our original manuscript also cited many “co-adaptive” BCI works that claim learning (despite some important omissions, see below). We hope the reviewer trusts that we have neither ignored nor neglected this literature, which we believe is well known to every BCI researcher. We are also well aware of how much subject learning is claimed in general BCI papers, even ones unrelated to user training. However, we stand by our opinion that subject learning in the framework of co-adaptation, specifically in non-invasive MI BCI, has been far more claimed and discussed than actually shown to occur.

We admit that the actual message we wanted to convey regarding the impact of subject learning within a mutual learning framework has not been properly communicated in the original manuscript, and we again apologize for that. We believe that the Introduction and Discussion sections of the revised manuscript, which have been completely rewritten and refocused, make our argumentation clearer. While we now avoid claims of novelty and priority, we still think our work brings some important new information to the field.

In the revised manuscript (Discussion, sub-section *“Subject learning in MI BCIs”*), we discuss what we believe are the limitations of the literature on subject learning in this particular field. Briefly, first, while we agree with the reviewer’s assessment that “The fact that effective BCI performance is a skill that the subject acquires through practice has been evident since the earliest SMR studies in the early 1990s”, we believe that, unfortunately, the possible extent and impact of subject learning has not been well established. We hope that more and better evidences of subject learning, in particular on end-user populations, could establish balanced and comprehensive mutual learning methodologies as the way to go.

In our opinion, subject learning was well substantiated in early works like Birbaumer et al. (1999) --where the BCI implementation was very close to neurofeedback systems and not a pattern recognition system employing multivariate brain activity-- as well as in a very limited number of SMR studies we refer to below. Hence, we postulate that most of the claims for the existence of subject learning are extrapolations from neurofeedback theories and invasive BCI studies, and not actual demonstrations of it. Like all extrapolations, they are not guaranteed to generalize to non-invasive MI BCI, and indeed, there are very few MI BCI studies that show evidences of subject learning at the level of neural correlates. A great body of literature claims learning based on accuracy and application performance increase. We provide a number of reasons why those do not necessarily imply subject learning. Moreover, when evolution of SMR patterns is provided, it is not longitudinal enough to establish subject learning because the observed changes might be only indicative of transitory effects. Also, most such studies are on able-bodied populations and thus lack translational implications.

In the revised manuscript, by conducting a more comprehensive literature review, we have been able to isolate only three studies (notably, 8-18 years old, which is typical of the aforementioned current trend on machine learning) that are longitudinal, involve an end-user, and (two of them) show some evolution of neuroimaging over time: *Pfurtscheller et al. (2000)*, *Neuper et al. (2003)* and *McFarland et al. (2010)*. We are grateful to this reviewer for pointing out the latter publication to us: Although we were aware of this line of research (citing Wolpaw et al., 2004), we had missed this particular work. However, even these studies suffer some limitations. *Neuper et al. (2003)* show how performance boost came as a result of the adoption of more suitable mental strategies. We do not consider this to be an instance of subject learning, as in this case spontaneous and not “learned” SMRs where seemingly still employed throughout training. In *Pfurtscheller et al. (2000)*, only 4 ERD/ERS maps are given over time, where it is shown that beta band activity during “resting” eventually disappeared, what allowed greater separability between resting and an active mental task to emerge. The depth of ERD/ERS of the active task is not reported to increase. It is unclear to what extent this was the result of feedback training, given the limited time resolution of SMR evidence. Lastly, the work of *McFarland et al. (2010)* is indeed the most complete and multi-faceted one, as, beyond task-dependent performances, it also shows SMR topographies of an able-bodied user improving over time. Even in this case, however, the time resolution of the neuroimaging is limited (begin/middle/end). Also, as the authors state “Across the four users, performance did not correlate with amount of 3D training.” In this comment, the reviewer argues that, in the reported study, subject learning is key for performance improvement. Along this line, our work also provides direct evidences of this subject learning process and its impact on BCI accuracy and race completion time. In fact, none of the works report positive correlations between indices of subject learning and BCI/application performances.

In the same subsection of the Discussion in the revised manuscript we also refer to the issue of “automaticity” in BCI control reported by expert users. We now mention that our pilot P2 reported the same (but not P1, who insisted he has always employed MI). Nevertheless, we argue how for this type of learning to drive BCI performance, increased discriminancy of SMR activity is still a prerequisite and, in our opinion, worth showing.

The present work provides all the above elements on top of what has been already shown in the literature. As stated in the Discussion of our revised manuscript: “*The main contribution of this work is the provision of quantitative evidences regarding the possible extent of operant subject learning in longitudinal MI training, how it can drive both BCI and task performance, and how it can be facilitated by the refinement of the application control paradigm.*”. Of note, while we are confident we now cite the ensemble of relevant literature, even potential further omissions providing evidence along the lines we describe as necessary, would not substantially affect the conclusion that more and better direct subject learning evidences need to be presented in order to enhance the case of subject learning within co-adaptive schemes and convince the field for a paradigm shift in user training. Our work is, we hope, a step towards this goal.

Comment 2.4: “Second, while the victory in competition is impressive, it appears that the actual level of SMR control is not markedly different from, nor superior to, that reported in many previous SMR-based BCI studies. The high inter-session variability in performance evident in Figures 1C and 4B of the manuscript is characteristic of the SMR control reported to date; and the accuracies in Figure 3A are not unusually impressive, given that they are for a two-class discrimination.

We agree that we are not the first to report BCI command accuracy that reaches almost 100% for end-users in a 2-class BCI. Nevertheless, the important finding in our study is that both users learned to use the BCI so as to achieve this very good level of control: they did not possess good enough spontaneous modulation of brain rhythms to spontaneously exhibit this level of performance at training onset. We believe our results clearly demonstrate this fact.

Regarding stability, both the original and revised manuscripts acknowledge that our mutual learning method did not seem to have a considerable effect, at least not for P1 who also had a complete performance loss in the final. For pilot P2, it is interesting to note that there is a notable stabilization between the beginning and end of training, what warrants further investigations.

Comment 2.5: “The inclusion of the topographies is commendable; it would also be worthwhile to include frequency spectra to indicate how well focused in frequency the control was.”

We provide the additional figure in the revised manuscript (new Supplementary Fig. S3), as requested by the reviewer. Please, note that, as shown in this new figure, all discriminancy is concentrated in the beta band for both users, which is why we have originally only provided topographies for this frequency band.

Comment 2.6: “The fact that the two subjects have tetraplegia is important, since as the paper notes, most previous studies (though not all (e.g., PNAS 101:17849-54, 2004)) have been confined to subjects without disabilities.”

We thank the reviewer for acknowledging this important point. Indeed, it is also our position that moving BCI towards translational applications can only happen if experimentation is done with actual end-users, as conclusions of able-bodied user studies can hardly generalize to end-user populations.

Comment 2.7: “The design of the BCI and the training procedure is interesting, however at present their description in the paper is wholly inadequate. This reviewer for one is left with many questions as to exactly how the application operated. The authors should develop a figure and accompanying detailed legend that clearly and completely describes the application and its operation. They should have this evaluated by naïve individuals to ensure that it is completely clear.”

We have proceeded with several changes in the revised manuscript to address this lack of information regarding the application and its control through the BCI. Specifically:

1. More details on the tournament format and the race are provided in the revised Introduction. Please, note that in the interest of conciseness, only the most essential details needed to comprehend the results are provided there, while for more details we refer the reader to the Materials and Methods section, where elaborate information is provided.
2. There is an explanatory supplementary video (S1 Movie, already available in the original submission) that illustrates how the BCI drives the avatar during the race.
3. The revised manuscript cites, early on, the recent publication of the Cybathlon organizers on the BCI race (Novak et al., 2018) which provides all necessary information on the game interface and the tournament, as well as the official results. We further cite our related publication in IEEE Spectrum Magazine, where the task and the control paradigm is explained in “layman” terms.

Comment 2.8: “In summary, this manuscript is significant primarily because of the competitive venue, the fact that the subjects have tetraplegia, and the emphasis on the importance of training both the subject and the BCI (i.e., emphasis on the principle that BCI use is a skill to be acquired). It needs revisions to place it and its findings appropriately in the context of previous BCI studies, to adequately describe the BCI application, and to fully present the results (e.g. frequency spectra to complement the topographies).”

We are grateful to this reviewer for recognizing the significance of our work. We believe the revised version alleviates the identified deficiencies of the original one, including defining its place with respect to the literature and the provision of additional results and methods.

**Reviewer 3**

Comment 3.1: “The paper describes an EEG BCI competition with paralyzed persons controlling avatars in a race. One of the players of the authors' team won the gold medal. The authors attribute the success of their BCI system to "mutual learning" of human and BCI. I was not convinced that the authors can indeed attribute the success to mutual learning. The paper lacks important details about the BCIs employed by other groups, and does not make any direct comparison with the other teams. Only the bottom-line performance in the race is compared. We therefore lack details about how other systems were trained. It is possible that other systems also employed "mutual learning" - we simply lack the details. It is therefore impossible to infer from the manuscript what is the true reason for the BCI success and whether indeed the mutual learning is the reason.”

Please, refer also to Comment 1.1 and Comment 1.6 by Reviewer 1, as well as to Comment 4.2 by Reviewer 4. We apologize that our original manuscript has been vague regarding the hypothesis, the research questions answered and the evidence provided.

We think that showing mutual learning effects (at the subject, machine and application level) taking place with two end-users (within-subject results) in non-invasive MI BCI is already important evidence towards the merits of this approach. However, the reviewer is right that our study lacked a proper control in its effort to show that a mutual learning approach should be the ideal strategy for user training. This effort was hampered by the fact that, initially, given the competitive nature of the Cybathlon event, precise details about the BCIs employed by other teams were not disclosed, so that our study was designed as an uncontrolled (observational), longitudinal, prospective, two-case study.

To mitigate to some extent the absence of control, in the revised manuscript, information has been collected from most other competitors via a questionnaire (see new supplementary Table S2), which allow us to have a fair control group as it is shown that they have essentially adopted a training methodology mainly based on machine learning, while we followed a more holistic mutual learning methodology. Furthermore, a recent publication from the Cybathlon organizers (Novak et al., 2018) highlighted the fact that all teams used similar state-of-the-art signal processing/machine learning methods, so that they conclude that this factor cannot explain the tournament outcomes. Please, also note that all teams but one, recruited pilots with equally or less severe disability than our team. Therefore, we postulate that the only other factor that seems likely to have determined the Cybathlon performances should be the training methodology followed.

Hence, although we acknowledge this comparison is based on rather circumstantial evidence, we consider that the Cybathlon BCI race performances support our claim that a mutual learning approach based on all three learning pillars will be more effective for translational applications than machine learning-oriented approaches that are dominant in recent years. The revised Introduction and Discussion sections highlight this line of argumentation.

Comment 3.2: “ Whereas one of their participants achieved the fastest performance in the race, the other was the last in the Final A race. This fact is obscured in the cumbersome Table 1 (see below). This casts doubt about the authors' conclusion that their method is superior to that of the other team. The authors convey that their other participant (last in the Final A) "establishes the record time", but this was not in any of the finals. It is also not explained how and why the excellent result in the Qualifier race (90s) turns to almost the worst (190s) in all the races described (rank: 17th out of 18 races!).”

In the revised manuscript, we have clarified the structure of the competition (Qualifier and Final A/B races) and modified Table 1 accordingly.

The bad performance of pilot P1 in the final is pointed out (Fig. 5 in the revised version) and its causes deliberated in the Results and Discussion of both the original and the revised manuscript, where we also refer to the known problem of robustness in MI BCI. This is a common and open issue which, we acknowledge, our mutual learning approach did not seem to resolve. Still, it is interesting to note that for pilot P2 there is a notable stabilization between the beginning and end of training. Please, see also our reply to Comment 2.4 by Reviewer 2.

Importantly, we believe that the effectiveness of our methodology is not considerably challenged by the result of P1 in the final, as it is evident by this pilot’s performances throughout the training period (Fig. 1A) that it was an exception and not the rule (apparent outlier). Other than this, our team’s pilots achieved the three best race times over all the tournament (90.1 s, 122.5 s, 125.3 s) followed by 132 s, 135 s, 136 s and 146 s reached by the closest competitors throughout the tournament (qualifiers and finals). Despite not representing definite proof, we posit these results can hardly be coincidental. A secondary reason for the P1’s poor race time in the final is that, as in any other kind of sport competition, motivation plays a fundamental role and, as P1 later reported, once he failed to deliver correct mental commands and lagged behind the other pilots, he lost concentration and motivation to continue the race.

Comment 3.3: “According to Table1: On the Qualifier race, P1 was far faster than P2, consistent with the double training P1 received (35 vs. 16 sessions according to the Methods).However, these results reversed in Final A. This apparently seems to contradict the main conclusion of the paper - superiority of the mutual learning approach.”

Please, refer to our reply to Comment 3.2 regarding the issue of stability. Since, with respect to the performances of our own team’s pilots, we are in possession of a large dataset collected throughout their training period, we find it reasonable to study the effect of our training approach on performance on the grounds of this dataset (182 races for P1, 57 races for P2), including, and not isolating, the competition performances. Therefore, in our opinion, a positive effect of our mutual training approach on application performances is convincingly depicted in Fig. 1A. In addition to that, 3 out of 4 performances achieved during the Cybathlon event seem to replicate well the average race time towards which our pilots converged at the end of training.

Comment 3.4: “There are no details about the training that other participants in the competition had. If other participants had by far less training, the reasons for the gold medal may not be algorithmic at all.”

Please, refer to our replies to Comment 1.1 by Reviewer 1 and Comment 3.1 by Reviewer 3. In summary, as shown in new supplementary Table S1, the amount of training of our pilots was in the range of other participants. In addition, by comparing the questionnaire (S1 Table) and the competition results (Table 1), the amount of training (in month) seems to not have a direct effect on the final outcomes of the race.

Comment 3.5: “Fig. 1A is the only figure of the race, but is unclear. Moreover: No explanation is provided of this picture at all. Is there a video of the competition?”

Please, refer to our reply to Comment 2.7 by Reviewer 2. Fig. 1A has been better explained in the main text and in the corresponding caption in the revised manuscript. Furthermore a video of the training procedure, explaining the BCI setup, control paradigm and the game interface is reported in Movie S1 of the supplementary material. Furthermore, a video of the event’s media coverage by the Swiss public TV is available at the following link:

https://www.srf.ch/kultur/wissen/srf-menschmaschine-das-volle-tv-programm-hier-nochmal-erleben-2?ns\_source=app?ns\_source=app?ns\_source=app?ns\_source=app

Comment 3.6: “The abstract should have defined the criteria for winning the "gold medal" (e.g., speed, completing the race first, etc.). It also unclear in the abstract how "gold medal" differs from the "time record".”

Given the journal’s instructions for authors regarding the abstract (“The Abstract succinctly introduces the manuscript. It should mention the techniques used without going into methodological detail and mention the most important results.”), we have preferred to not include this information (tournament format). Furthermore, please note that, in the revised manuscript, we have clarified that the competition outcomes are not the key finding of our study. Please, see our reply to Comment 1.1 by Reviewer 1.

Furthermore, please note that more details on the tournament and race formats are provided in the revised Introduction. Additionally, in the interest of conciseness, only the most essential details needed to comprehend the results are provided there, while for more details we refer the reader to the Materials and Methods section, where elaborate information is provided. Please, see also our reply to Comment 2.7 by Reviewer 2.

Comment 3.7: “The paper could have been written more clearly. From the abstract and introduction, the reader does not understand how many participant attended the competition.”

In the light of a refocus of our paper towards investigating the hypothesis that mutual learning is a crucial element for translational, non-invasive MI BCI, the Abstract, Introduction and Discussion sections of the revised manuscript have been completely rewritten. We apologize for a lack of clarity in the original manuscript and hope that in the revised version this issue is addressed and all necessary information is provided.

Comment 3.8: “Table 1 is intermixed between 3 different races, which obscures the real results and make it difficult to follow the coarse of the competition.”

Please, refer to our reply to Comment 3.2 by Reviewer 3. Table 1 has been rearranged in the revised manuscript so as to exactly follow the tournament format.

Comment 3.9: “Finally, the term "pilot" may not be appropriate here, as the race does not seem related to aviation at all. This is not a common term in the field. "driver", "controller" or simply "user" is far more appropriate.”

We appreciate the reviewer’s recommendation towards making our manuscript clearer. However, we have finally opted to keep the term “pilot”, as this is the terminology adopted by the organizers and communicated to all popular and scientific audiences. Please, see also the recent publications by Novak et al. (2018) and [Statthaler](http://paperpile.com/b/dIT9gv/qwi2) et al (2017). Nevertheless, if required by the reviewer or by the editor we can proceed with this modification for a final version of our manuscript.

**Reviewer 4**

Comment 4.1: “The paper reports on a competition and two participants who won the competition. The report details measures of performance and changes over time due to training. The report differs from other papers in that it describes trainig for a competition and then competing in a challenging environment, rather than using metrics of offline or carefully controlled online BCI performance that have been used elsewhere.”

We would like to thank the reviewer for recognizing the importance of the training and operation conditions in our study and their translational implications.

Comment 4.2: “Although i appreciate the way the competition was carried out with quantifiable metrics and performance was measured , I have a hard time positioning this manuscript in PLoS Biology. My main concern about the manuscript is that there is no testing of hypotheses which is a hallmark of papers in neuroscience journals like Plos Biology. Understandably one has to determine parmeters for a longer learning study but there is no comparison to anything that allows readers to gauge the level of performance (other than within-subject training effects). Statements that mutual learning and user-centered design contributed to the results are not quite based on data presented (there was understandably no comparison to other training schemes and decoders), but on other literature. Thus these parameters are chosen beforehand as described in the Discussion, and are not evaluated in a quantitative manner. This is no problem in itself, but it is not quite suitable for this journal. I can imagine that this manuscript in this format appeals more to engineering and gaming fields. For this journal the manuscript would need to be written more critical and with much more focus on the science and the improvements compared to previous MI BCI studies than the competition.”

Please, refer also to our replies to Comment 1.1 by Reviewer 1 and Comment 3.1 by Reviewer 3, where similar criticisms are expressed.

We admit that the original manuscript has been very vague on essential elements (hypothesis, research questions and evidences provided), which has inevitably also put into doubt its relevance to and suitability for PLoS Biology.

In the revised manuscript, the hypothesis put forward is clearly and early stated: *“In this study, we investigated the hypothesis that mutual learning is a critical underlying factor for the success of motor imagery (MI) BCI in translational applications. Contrary to a popular trend of focusing almost exclusively on the machine learning aspects of MI training, our hypothesis propounds that a holistic mutual learning training approach grounded symmetrically on all three learning pillars (at the machine, subject and application level) would be the optimal training apparatus for preparing two end-user participants for the Cybathlon BCI race, the first international BCI competition”*

In the absence of information on our competitors strategies our study was originally envisioned as a longitudinal, observational, two-case study. The reviewer seems to agree that such a study has considerable value in itself, despite not providing definite evidences (what a controlled study might do). Indeed, we believe that the within-subject results presented, where we show that mutual learning effects were observed at all levels of the interface (subject, machine, application) and, importantly, with two end-users, represent important findings contributing to the previous literature a quantitative demonstration of subject learning and how it might be influenced by the application. Our emphasis is put on the subject learning effects observed at the neuroimaging level, as we show by means of a thorough literature review that this aspect has been largely neglected in the field of non-invasive MI BCI. Such effects are widely accepted (and reviewed in the revised Discussion) to represent a form of operant conditioning. The topic of “learning to control a BCI” and, in particular, the cortical plasticity that accompanies such learning is of great interest to (Human) Biology and neuroscience audiences and goes beyond engineering. In particular, it is our understanding that the PLoS Biology journal is interested to host this kind of research, as shown by earlier published work in invasive BCI that our manuscript cites: Carmena et al. (2003); Ganguly et al. (2009); Carmena et al. (2013). Our work contributes similar evidence for non-invasive BCI, another research area published in PLoS Biology recently (Chaudhary et al., 2017).

Furthermore, thanks to a complete refocus of our manuscript from the (admittedly, unfortunate and overstated) “demonstration of translational impact of MI BCI” and our victory in Cybathlon, towards the above-stated hypothesis, as well as in the light of new information, namely:

a) a questionnaire answered by most of our competitors (see new supplementary Table S2) allowing us to determine that they have followed training methods strongly oriented towards machine learning, as opposed to our comprehensive mutual learning scheme;

b) the recent publication by the organizers of Cybathlon (Novak et al., 2018), which concludes that the differences in the state-of-the-art machine learning and signal processing methods employed by all teams are too subtle to explain the Cybathlon outcomes;

c) the fact that all, but one, competing pilots suffered from equal or lesser disability than our team’s pilots;

we can assume that the competitive teams provide a crude control group, suggesting that our user training methodology seems to have given a key advantage to our team and justifies the Cybathlon results. Given the clear limitation of such a control, we avoid to present formal statistical testing of the hypothesis (e.g. by grouping the competition’s 15 performances to the equivalent number of the latest races achieved by our pilots, what yields significant differences for both our pilots, Competition: 155.5±18.0 s, P1: 117.6±28.4 s , P2: 117.8±14.4 s, p<0.001 with unpaired Wilcoxon ranksum tests Px vs Competition). Still, we believe this comparison provides evidence towards the confirmation of our hypothesis, on top of the important evidence provided by the within-subject results on mutual learning.

Of note, the same kind of comparison allows us to (reasonably) speculate regarding what are critical features of a mutual learning protocol, where we also compare our framework to similar (but with crucial differences) approaches our laboratory has used in the past. Overall, our revised manuscript makes use of both the competition and previous literature to draw conclusions in favour of confirming our hypothesis with reasonable (though not definite) confidence. Finally, all claims have been toned down and our results are more objectively and critically elaborated in the revised Discussion section.

Chaudhary U, Xia B, Silvoni S, Cohen LG, Birbaumer N (2017). Brain–Computer Interface–Based Communication in the Completely Locked-In State. PLoS Biol 15(1): e1002593.

Comment 4.3: “Several issues may improve the manuscript although i cannot judge what would be an improvement for a more technical journal. Overall the paper is written more like a description of the competition than a scientific paper. The sometimes hyperbolic language distracts from the main message (which in my opinion is the training of motor imagery based BCI in paraplegics)”

We recognize and apologize for the unscientific style, the overstatements and exaggerations in the original manuscript. We believe all these issues have been alleviated in the revised version, where the Introduction and Discussion sections have been re-written from scratch and the text body has been modified accordingly throughout. The focus of the manuscript has been shifted towards what our results really point out: the demonstration that a mutual learning methodology based on all three learning pillars --subject, machine and application-- seems to be the most promising training apparatus for translational non-invasive BCI applications.

Comment 4.4: “what is learned about the brain mechanisms involved or about the effect of interfacing brain and technology on the user?”

The main finding of the paper is the demonstration --with quantitative EEG-based evidence-- of subject learning effects that occur (next to machine learning effects) during longitudinal BCI training with a mutual learning protocol. From the neuroscientific point of view, for both pilots we report emerging SMR patterns directly related to the features exploited for BCI control (Fig. 4). Discriminancy of beta-band SMR activity in medial and bilateral cortical areas positively correlates with the decoding accuracy of the classifier and with the application performances (namely, race completion time).

While neuroplasticity during BCI training has very early been documented for interfaces closely resembling neurofeedback apparatuses and for invasive interfaces, we argue in the sub-section *“Subject learning in MI BCIs”* of the revised Discussion that evidence that such effects may take place in non-invasive MI BCI (and, co-adaptive protocols in particular) were rather scarce and incomplete (not longitudinal, lacking substantiation at the neurophysiological level and gathered on able-bodied populations), notwithstanding very few notable exceptions which also suffer their own limitations. Hence, although this kind of learning has always been hypothesized to happen, this work provides deeper substantiation and translational implications with two end-user participants.

Although we note in the revised manuscript that EEG imaging with 16 channels does not allow for a deeper study of plasticity effects, the reviewer must note that this limitation is essentially a trade-off with the aforementioned translational aspects of our work, as the realization of this study urged for lightweight, portable and minimally obstructive technology. Furthermore, we argue that even in the presence of richer imaging, placing the observed learning effects in the context of instrumental learning inherently confines the main learning evaluation metric to SMR features, (i.e. the same used by the BCI algorithm and fed back to the user through the visual feedback) for which the available imaging is sufficient. Related to this, at the level of mechanisms, in the revised manuscript we offer additional discussion justifying the observed learning effects in the framework of operant conditioning (see also our reply to Comment 4.5).

Lastly, an important further contribution of the revised manuscript is the addition of new results (Fig. 6D) suggesting that, in order to foster such learning effects, an optimization of the control paradigm might be needed.

Comment 4.5: “What is the effect of temporal delay in feedback? this is relevant for assuming that these subjects exhibited operant conditioning and should be discussed”

Thank you for this observation. In the Discussion of the revised manuscript, we present an elaborate justification why the observed learning effects can be justified within an instrumental learning framework, where immediacy and contingency of feedback to the SMR activity is crucial. We argue how these prerequisites are satisfied in our training methodology: *“At the level of mechanisms, our feedback training design has respected the neuropsychological basis of operant conditioning, namely, immediacy and contingency of the visual feedback to the targeted brain rhythms. Indeed, during races, BCI commands always coincide with the presence of SMR, which has to be sufficiently large for the BCI to reach the decision threshold. Thus, although the BCI did not deliver a command to the avatar every time the subject generated a SMR, the opposite holds: whenever the BCI delivered it, the subject was eliciting a SMR. Another clear manifestation of the instrumental nature of subject learning is the fact that, as shown in Table 2, the brain features that responded to training were among those selected for classification and feedback provision.”* Please, note also that given on average short command delivery times (Fig 2), most **reliable** SMR patterns are rewarded during racing, since for the user to reach the decision threshold and elicit a command, the evidence accumulation module requires strong and sustained SMR activity (see *“BCI implementation”* of the Materials and Methods, as well as S1 Movie). Hence, while this system might not reward all SMR patterns, it largely avoids rewarding inadequate or “erroneous” ones.

Of note, also during “online” closed-loop training runs, a continuous visual feedback was provided to the pilots, consisting of a feedback bar moving left/right according to the BCI output (at 16 Hz, see S1 Movie). In this case, all SMR patterns are directly rewarded/punished by the visual feedback, in what is a more classical approach to induce operant conditioning effects.

Comment 4.6: “the residual abilities of the two subjects needs more description. Which movements were still possbible? And will they use their system after thecompetition (‘translation’)?”

We have added the missing information in the Material and Methods (sub-section *“Pilots”*) of the revised manuscript: *“Residual motor abilities included, for both pilots, unaffected bilateral control of shoulder and elbow movements and compromised control of wrist movements, while none of the two maintained control over the fingers.”* We can provide additional information if the reviewer requires so.

Please note that, although the showcased ability of end-users to learn to control a non-invasive MI BCI has clear translational implications, the “translational impact of non-invasive MI BCI” is no more the main claim of our paper. Nevertheless, it is interesting to report that in interviews following the competition, both pilots declared they would be interested in having this technology at home for applications like accessibility and spelling (i.e. where errors are non-critical). While, as of now, we have not made our BCI prototype available to them, this is mostly related to hardware and other technological and practical limitations (e.g. obtrusiveness of current gel-based systems --note though that very reliable dry electrodes have been very recently made available to the market, provision of fully standalone and user-friendly software, etc.) and not to the user training approach, which we believe could be applied off-the-shelf once these practical issues are addressed.

Comment 4.7: “relevant for the generalizability of the findings: could there have been a motivational bias? Meaning that perhaps the participants were highly motivated and therefore trained hard to imrpove. Would less motivated end users also benefit from the reported scheme?”

Motivation has been shown to be a key factor for skill acquisition and it is already well documented that this holds true also for BCI training. Both the original and the revised manuscripts cite the relevant literature experimentally verifying this hypothesis. In the revised manuscript, we have included discussion (sub-section *“Mutual learning: Lessons and recommendations”*) identifying motivation (of training with a game interface, while looking forward to an international competition attracting a lot of media attention) as one of the possible reasons why we have been able to observe strong subject learning effects. However, we do not see why motivation should carry a negative connotation: BCI engineers should simply take these findings into account when designing training protocols, as argued also by Lotte et al. (2013).

However, we believe motivation is not the only reason underlying the learning effects reported here. In the revised Results section, the newly added Fig. 6D illustrates the increase of feature discriminability (the index used to quantify learning effects at the neurophysiological level) occurring with each different control paradigm tested by pilot P1. We found a statistically significant increase of discriminancy only in the case of the last control paradigm. This result supports the idea that learning in BCI is not only related to finding optimal machine learning methods or solely relying on longitudinal operant conditioning to let the subject learn, but also to discovering optimal control paradigm for the BCI-actuated device/application. Furthermore, we believe incremental learning approaches, as recently advocated in the literature (Lotte et al., 2013) and the careful coordination of the machine and subject-learning aspects (e.g. parsimonious, rather than intense and continuous recalibration of the BCI decoder) should also be critical for effective mutual learning BCI protocols.

Comment 4.8: “what does this report mean for MI BCI in paralyzed end users? Is preformance as reported here well enough for people to sometime in the near future buy a system and use it at home? This is suggested by the authors indirectly.”

Please, note that the revised manuscript has been based on a complete shift of focus from the overstated “translational impact of MI BCI” towards the merits of the mutual learning approach we have followed for user training. Please, see also our reply to Comment 1.2 by Reviewer 1.

The performances we report reach almost 100%, but are not really unprecedented for a 2-class MI BCI (i.e. a BCI with few degrees of freedom), even as far as end-users are concerned. Furthemore the problem of robustness is still evident, so that we do not intend to support a claim that non-invasive MI BCI, or our own implementation in particular, are ready to enter the assistive technology market, what will probably require several years of additional, multidisciplinary research.

The actual translational implication of our work is that we show how a suitable training method can allow an end-user with no --or minimal-- ability to spontaneously modulate his/her SMRs to finally acquire (a very high level of) this skill. This is very important because several studies (including our previous work), applying different training protocols, have found that this technology might be completely inaccessible to a large percentage of potential users, while this portion of users has been found to be even greater for paralyzed people, the main group interested in this technology. Our work gives hope that this needs not be the case, pending of course larger studies to verify our findings, examination of the effects on users with different clinical profiles, and further optimizations of the training apparatus along the lines outlined here.

Comment 4.9: “the numbers in figure 2 have no meaningful reference. Is it possible to compute chance levels empirically?”

The “time on pad” (pad crossing time) metric reported in Fig 2, which evaluates simultaneously both command accuracy and command delivery speed, has in fact clear upper and lower bounds imposed by the game interface. Specifically, a perfect input (i.e. one that delivers the correct command as soon as the avatar enters a pad and allows the avatar to immediately cross it at the maximum speed) yields 2 s. A continuously erroneous input (e.g. a BCI that forwards wrong commands immediately and consecutively, forcing the avatar to cross the pad at the lowest speed defined) yields 19 s. Finally, since avatars would proceed at a medium, “baseline” speed in the absence of any input, there exists also a “no delivery” performance that corresponds to 11 s. These limits were in fact already illustrated in Fig 2 of the original manuscript, through dashed lines as coded in the legend.

In the revised manuscript, we have slightly modified the figure to increase the line width and make these limits more evident. Also, we cite Novak et al. (2018) in the revised Introduction, a publication that gives all the necessary details on the Cybathlon BCI race event and the game interface. Finally, please note that both the original and revised manuscripts provide this information in subsection *“Cybathlon BCI race”* of the Materials and Methods. Given the journal’s policy, we tried to keep the Introduction as concise as possible, mentioning there only the basic information needed to understand the results and moving all further details in the Materials and Methods. We are willing of course to move information back to the Introduction upon specific recommendations by the Reviewers.

Finally, note that a “random” performance for this metric, besides being redundant given the definite bounds, cannot be objectively created, as it does not only depend on the distribution of the classifier’s decisions (which could be eventually modeled with a binomial distribution or extracted with random shuffling), but also on various re-configurable hyperparameters of our BCI implementation (rejection threshold, decision threshold, exponential smoothing parameter) that were often manipulated during training.

Comment 4.10: “i am a bit at a loss how mutual learning differs from a straightforward user learning effect. Calibration was optimized only twice over a period of several months (in itslef good news), meaning that the user had a fixed decoder for almost all the time.”

Ou results demonstrate learning/training effects at the subject (discriminancy), machine (BCI accuracy) and application (race completion time) levels. While the merits of machine learning (which we also observe) are regularly highlighted in the literature, those of the subject are rather neglected in non-invasive MI BCI (see subsection *“Subject learning in MI BCIs”* of the revised discussion) and in interfaces that claim to be “co-adaptive” in particular. Our work thus puts the emphasis on showing the existence and extent of subject learning (Fig 4) within our mutual learning scheme, but we do not fail to also demonstrate that machine learning effects also took place (Fig 2 and Fig 3). We believe this is more clear in the revised Discussion.

Furthermore, in the revised version of the manuscript (Discussion section, subsection *“Evidence of mutual learning during training for the Cybathlon”*), we explicitly mention that after each session we analyzed and evaluated possible classifiers based on the new data acquired. However, for both pilots, after the second re-calibration of the classifier the decoding accuracy already converged towards the maximum of 100% (Fig. 3). Thus, no additional “learning” on the machine side was judged necessary. The fact that within a mutual learning framework, the learning pace can be different between the machine and the user is not--in our humble opinion--an argument against the mutual learning concept.

Importantly, as commented in the subsection *“Mutual learning: Lessons and recommendations”*, we hypothesize that parsimonious decoder recalibration might be beneficial to subject learning (while it can also be adequate for the machine learning side, as happened here), because the subject’s learning efforts, assuming those indeed happen in an operant conditioning fashion, could be hindered by frequent re-selection of features and classifier parameters (i.e. constantly changing the “behaviour” that needs to be learned).

Comment 4.11: “the discourse contains many repetitions of aspects of the competition, the fact that the two subjects won the race, and the difficult circumstances under which the subjects had to perform. This could be toned down to improve readability”

In the revised manuscript, all claims have been toned down and the writing style has been changed to reflect scientific standards, avoiding repetitions and flamboyant statements. Furthermore, we have focused on the now clearly stated hypothesis of the study, for the defense of which our victory in the Cybathlon competition is supportive, but is not the main finding.