Reviews of 6004 - *"Agency-preserving Action Augmentation: Preemptive Muscle Control using Brain-Computer Interfaces"*

Reviewer 4 (1AC)

Expertise

Expert

Originality (Round 1)

Medium originality

Originality (Round 2)

Medium originality

Significance (Round 1)

Medium significance

Significance (Round 2)

Medium significance

Research Quality (Round 1)

High research quality

Research Quality (Round 2)

Medium research quality

Recommendation (Round 1)

I recommend Revise and Resubmit.

Recommendation (Round 2)

I recommend Reject

1AC: The Meta-Review

This work presents a BCI-based augmentation that is able to stimulate users' muscles at their intent to interact. The system measured readiness potential (RP) from EEG signals to calculate action intention. The authors conducted a user study to measure intentional binding while interacting. The results show a higher level of control working with the system compared to when being passively moved (via EMS).

I find the goal of this work is well-motivated and would fit into CHI. All reviewers appraise the work for its ambitious goal (2AC), has a sensible approach, overall aim interesting (R1), and introduce an original system that can be an important step towards action augmentation using neural input (R2).

However, the paper is well-written and has good-quality research. However, there are some issues that should be addressed in the revision:

** 2AC:
It is unclear if both EMS triggering and EMG measurements were performed in the AUGMENTED condition. If so, how that was done technically should be explained.

The authors did a comparison between three conditions. However, how those conditions are compared to other techniques (as a baseline) in previous studies should be discussed.

Both 2AC and R1 are concerned with a justification of the classification rate (10Hz) and the maximum interaction time reduction that this technique can achieve.

** R1:
I agree with the R1's concern regarding the low number of participants (8 people). As the trials were repeatedly measured and analysed

(independently?), would the means of those trials (i.e., time estimation and sense of control rating) should also be used?

The used timing in the EXTERNAL condition was not particularly extreme and randomly picked from within 5-95% of that user's previous timings. So they weren't faster here, and it's unclear whether they were faster in the augmented condition either.

Justification about the electrode location (Cz): why the main reference throughout is Cz. The paper mentions this electrode earlier together with C3 and C4 as commonly used, but how does Cz stand out from the three?

** R2 has some suggestions about clarifying the definition of action augmentation and providing a brief explanation of how the present study achieves that much earlier in the paper can improve the readability and clarity. A good suggestion also is the inclusion of references to different states (idle and pre-movement state.

I urge the authors to read the reviewers' comments in detail to improve their work in the revision. I believe these can be fixed within the CHI timeline for R&R.

1AC: The Meta-Review (Round 2)

I thank the authors for the revision. The paper was discussed at the PC meeting. All reviewers appraise the effort put into improving the work. However, important issues raised by reviewers have not been addressed or responded to in the revised version. As it stands now, all reviewers agree that the paper is not ready for publication at CHI.

Reviewer 3 (2AC)

Expertise

Knowledgeable

Originality (Round 1)

Medium originality

Originality (Round 2)

Medium originality

Significance (Round 1)

Low significance

Significance (Round 2)

Low significance

Research Quality (Round 1)

Medium research quality

Research Quality (Round 2)

Low research quality

Contribution Compared to Length (Round 1)

The paper length was commensurate with its contribution.

Contribution Compared to Length (Round 2)

The paper length was commensurate with its contribution.

Figure Descriptions

The figure descriptions are adequate and follow the accessibility guidelines.

Recommendation (Round 1)

I can go with either Reject or Revise and Resubmit.

Recommendation (Round 2)

I recommend Reject

Review (Round 1)

I agree with the vision of this research. It deals with a very ambitious goal. The presentation quality of the paper is also not bad. The technical part of classifying EEG signals to trigger EMS was also interesting.

As I read the introduction, I was intrigued, but unfortunately after reading the entire paper, I was disappointed that the findings of this study were not significant enough to be presented at CHI. Below are my comments on the validity of the technique and experimental design.

(1) I wonder if both EMS triggering and EMG measurements were performed in the AUGMENTED condition. If so, I wonder if it is technically possible. Don't the electrical signals of EMS have a significant effect on EMG measurements? The authors stated in the limitations section that situations where the EMS was activated outside of the pre-movement phase were problematic. Couldn't the authors have post-analyzed the EMG signals and separately dealt with only the trials in which the EMS worked properly?

(2) The current baseline (EXTERNAL) condition is too weak. It is a trivial finding that the SoA of the AUGMENTED condition is higher than that of the EXTERNAL condition. Why has no comparison been made with a simpler technique where EMS is triggered by heuristic rules rather than based on EEG signals? Is there a reason the authors did not use as a baseline any technique suggested in previous studies?

So it seems we should consider the INTENTION condition as a more important baseline for this study. And the fact that the AUGMENTED condition shows a lower SoA than the INTENTION condition (Figure 4) is a common problem with EMS-based techniques found in previous studies and is also a key motivation for this study. In other words, it appears that important problems have not been solved even though EEG signals have been utilized.

How does the difference in SoA between the INTENTION condition and the AUGMENTED condition in previous studies compare to the difference in this study?

(3) I agree with the authors' idea that reaction time can be reduced through EMS if we know in advance when the user will move. However, according to the technique proposed by the authors, the maximum amount of reaction time we can reduce is the time interval between detection of the readiness potential (RP) and execution of the movement. Isn't this time interval usually very short (i.e., less than 200 ms)? Furthermore, for more robust classification, wouldn't EEG data sometimes need to be observed for a longer period of time, and as a result, the margin for reducing reaction time could be further reduced? We even have to consider the latency of the EMS itself.

Considering all these factors, I am curious to what maximum reaction time reduction the authors' proposed technique can actually achieve.

The time interval between the detection of RP and the actual movement of the participant's finger is very short (in Figure 5a), so isn't the classification rate of 10 Hz too low?

In addition to the technical aspect, there are several comments regarding the motivation and presentation of the research.

Most of the movements we perform in daily life are continuous rather than discrete, such as tapping. For example, human continuous movement (e.g., pointing) is known to be based on intermittent motor control [1]. In such cases, RP is expected to occur intermittently and repeatedly. Even in such cases, can the technique suggested by the authors be applied?

[1] Do, Seungwon, Minsuk Chang, and Byungjoo Lee. "A simulation model of intermittently controlled point-and-click behaviour." Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021.

There seems to be a trade-off between reducing reaction time (i.e., performance) and increasing SoA. For example, if a quick reaction is important enough to save users' lives (e.g., avoiding people ahead during autonomous driving), what is the point of providing an EMS signal late in consideration of SoA? From a user experience and satisfaction perspective, what are the ultimate benefits of the technique proposed in this study?

The process of processing EEG signals and EMG signals is not familiar to general HCI readers. More evidence should be provided regarding the rigor of the authors' analysis. A diagram summarizing the analysis processes in section 3.4 is also needed.

Please provide more detail as to why one participant was excluded.

Re-review

I appreciate the efforts the authors put into revision. Unfortunately, the authors barely responded to the critical issues I raised. It needs to be made more explicit in the paper that the experimental results have not proven that the authors' technique enables SoA or LoC (Level of Control) similar to the intention condition. Just because it's not statistically significant doesn't mean the difference doesn't exist (for SoA analysis). Aren't even trials that misclassified intent included in the SoA analysis? Even for LoC, the difference between Intention and Augmented conditions was statistically significant. In my opinion, it is a trivial finding that the Augmented condition has a higher SoA than the External condition. This is why a stronger baseline was needed, but the authors did not address the issue critically enough in the revision. How about conducting a more controlled experiment in which a visual stimulus for reaction is explicitly given, as in previous EMS studies? If so, wouldn't it be possible to compare the authors' technique with a strong baseline that is AUGMENT condition but also does not rely on EEG

(i.e., automatically triggering EMS when stimulation is given)?

Additionally, in the Round 1 review, I requested that the authors select only trials that successfully classified movement intent and analyze SoA or LoC limited to those trials, but that request was not addressed in the revision. It is also difficult to agree that the reduction in reaction time cannot be quantified. What if the authors measured the time from immediately after the RP is observed until the visible behavior is produced?

In summary, it seems difficult to change my assessment that, despite revisions, the research covers a narrow topic and did not make meaningful findings enough to be presented at CHI.

Reviewer 1 (reviewer)

Expertise

Knowledgeable

Originality (Round 1)

Medium originality

Originality (Round 2)

Medium originality

Significance (Round 1)

Medium significance

Significance (Round 2)

Medium significance

Research Quality (Round 1)

High research quality

Research Quality (Round 2)

Low research quality

Contribution Compared to Length (Round 1)

The paper length was commensurate with its contribution.

Contribution Compared to Length (Round 2)

The paper length was commensurate with its contribution.

Figure Descriptions

The figure descriptions are adequate and follow the accessibility guidelines.

Recommendation (Round 1)

I recommend Revise and Resubmit.

Recommendation (Round 2)

I recommend Reject

Review (Round 1)

I enjoyed reading this paper and I think the approach is sensible and the overall aim interesting. I'm assuming the authors actually would like to see user augmentation to result in a _decrease_ in reaction times (the introduction mentions an increase), but that goal is relevant and additional work in this direction appreciated. Overall, I think this work is well done, but I was left with some questions, primarily with respect to the analysis, where I think a revision is necessary.

But I'd like to start with something I found a bit odd. The paper positions itself within the human augmentation space, and I think that makes sense. The intro and method sections also discuss this. But did the augmented condition here actually improve reaction times? That does not actually become clear. As stated on p. 6, in the external condition, the used timing was not particularly extreme and randomly picked from within 5-95% of that users previous timings. So they weren't actually faster here, and it's not clear whether they were faster in the augmented condition either. That might well not matter, given that this paper is not about reaction times per se. However, it is a curious aspect of the experimental design, reporting, and framing and thus would be good to discuss a bit more. Given that there probably would be a higher

experienced loss of control with out of the ordinary reaction times, this might well have had an influence on the results, though.

My main worry regarding this paper is the low number of participants and the analysis. There were only 8 participants here, which is rather low given that this is not just exploratory work, but hypotheses are being tested. Based on the model descriptions and Figure 4, I can't help but wonder whether all trials were included separately. I suspect this was the case and even with about a hundred removed, that still means 1698 trials went into this. In my opinion, when a participant repeats a task 75 times, the individual trials are not independent and instead of analyzing those, the means of those trials (i.e., time estimation and sense of control rating) should be used instead. Maybe this is what actually happened anyway and then this is an easy clarification, but if not then the analysis, in my opinion, should be redone.

The analyses also used linear mixed effects models and that itself is a fine approach. But I am wondering whether the data doesn't require some transformation or link function here. Is "estimated time interval" normal distributed? I'm suspecting not and the text later on also notes that not all participants even used continuous values. Similarly, I don't think 7-point Likert ratings are suitable as-is either. Essentially, I'm wondering whether some log or rank transform, or maybe a GLM, should have been used instead. As a side-note: if models are fitted it would be great to just get a table that summarized the model.

A big part of the work here is the classifier that determines whether a participant is idling or just about to move. I'm not an expert here, but I was left with a few questions regarding this classifier. For starters, it's not clear to me why the main reference throughout is Cz. The paper mentions this electrode earlier together with C3 and C4 as commonly used, but how does Cz stand out from the three? All the training is then in relation to the time of movement as detected by the EMG. So not when the brain started to plan the motion, but when the muscle actually is activating. As the whole second up to that moment is training data, it thus contains not just the planning either (but of course _also_ the planning). As Section 3.4.1 notes, 1s signal windows are condensed to slope values, based on the first and last 100ms. Per Section 3.4.3, this window is updated 10 times a second, so essentially shifted by 100ms. However, the predictions are smoothed with two previous ones (btw. I don't get the weighting: 0.3 & 0.5 for the previous predictions, and then presumably only 0.2 as weight for the current one? Why such a low value for the most up-to-date value?). However, doesn't that also mean that prediction of pre-movement is quite laggy? For a slope to be noticeable, the window must likely overlap quite a bit with the increase and then another 200ms of delay come in due to the averaging. With the EMS then active for 500ms, I'm pretty much wondering whether this did preempt the motion and whether the system then remained active longer than needed. As the paper itself states, the performance of the classifier is mixed with several participants reporting issues. So this is probably an aspect to provide a bit more details and explanations for.

Just a few minor comments as well:
- The anonymization is a bit odd at times. We learn participants are paid in Euro, but only get told the setup cost 100 of unknown currency? It's left unclear how the data was processed due to tool anonymization (line 355). And then something happened with the interviews (line 433), but we can't get told what. Maybe translation? I don't think the fact this experiment happened in a certain country endangers anonymity.
- Given that the number of channels used varied by participant, it would be good to provide a bit more detail here. Even if just the range or which channels were prominent.
- That the system "sometimes detected users' intention to interact", does not sound very convincing as does the corresponding part of the limitations. I appreciate the honesty, but it remains a bit unclear to me how well this worked. The F1 score helps, but I'm not sure that's all that intuitive.

Overall, I was left with quite a few questions and I think it would be great to see a revision that tries to make these things clearer.

Re-review

I can keep this short: I think the updated analysis falls short of what was needed. The authors note that "By including a random effect for participants in our LME analyses, we accounted for the dependency of repeated measures within participants. The analysis is correct as is." This is just false. Adding participant ID as an error term, as specified in the paper, indeed fits the model with one intercept per participant. But it has nothing to do with independence and such a formulation still uses _all_ data from the 75 trials per participant in fitting that model. So the independence assumption is still violated. In this case quite severely: instead of 9 people with 3 conditions each, now hundreds of data points are used. Instead of just boldly assuring that the analysis is correct as-is, the authors should have actually redone that analysis. While I was quite positive overall initially, I don't think such a revision is acceptable for publication.

Reviewer 2 (reviewer)

Expertise

Knowledgeable

Originality (Round 1)

Medium originality

Originality (Round 2)

Medium originality

Significance (Round 1)

High significance

Significance (Round 2)

High significance

Research Quality (Round 1)

Medium research quality

Research Quality (Round 2)

Low research quality

Contribution Compared to Length (Round 1)

The paper was too long in addressing its claimed contribution.

Contribution Compared to Length (Round 2)

The paper length was commensurate with its contribution.

Figure Descriptions

The figure descriptions are adequate and follow the accessibility guidelines.

Recommendation (Round 1)

I recommend Revise and Resubmit.

Recommendation (Round 2)

I recommend Reject

Review (Round 1)

*** Originality: what new ideas or approaches are introduced? We want to emphasize that an acceptable paper must make a clear contribution to Human-Computer Interaction;
The researchers introduce an original system that can be an important step towards action augmentation using neural input.

*** Significance: evaluate the paper's contribution to HCI and the benefit that others can gain from the contribution: why do the contribution and benefit matter?
In evaluating the paper’s contribution to HCI and its potential benefits for the research community, it is evident that there is relevance. Improving sense of agency in emerging technologies is a significant challenge, and the researchers approached this research problem by focusing on muscle control.

*** Research quality: how confidently can researchers and practitioners use the results?
Experiments are well done and clearly reported. It is also evident that the researchers put a lot of effort into the research project.
For the pairwise comparisons, the authors did not report whether any correction for multiple comparisons were implemented. It is important to correct the p value for multiple comparisons to deflate the type I error rate.

*** Previous work: is prior work adequately reviewed?
The review of the previous work is well done, however, there is still room for improvements.
On page 3, the first paragraph introduces the terminology regarding different states (idle and pre-movement state) of motor command, and no citations were provided for these distinctions.

*** Presentation clarity: how well is the paper framed and is the argument clear throughout the paper?
Overall, the paper has some organization issues. It has all the necessary information, however they are misplaced in the paper:
For example, clarifying the definition of action augmentation, and providing a brief explanation of how present study achieves that much earlier in the paper can improve the readability and clarity.
Moreover, on page 3, the last paragraph does not refer to the literature in theories of sense of agency, and should not belong to this section.

Re-review

The authors put significant effort in improving the clarity and organization of the paper and it is now clearer to the reader.

On the other hand, the paper still has significant room for improvement in research quality, specifically in their analysis, in response to the reviews raised in the first round:
To do a within-subject analysis, the authors should have a summary statistic (i.e. mean or median) per participant instead of the separate trials, and then fit the data across the conditions. This was a major point that needed to be revised, but the authors did not address it in the second round.

A minor comment for presentation clarity is at line 516: "multiple companion" does not have an established meaning, I assume the authors meant "multiple comparisons"

In summary, the paper is not ready for publication at its current stage.

Return to submission and reviews