Reliability of NIRS-Based BCIs: a Placebo-Controlled Replication and Reanalysis of Brainput

Megan Strait
Tufts University
161 Collge Avenue
Medford, MA 02155 USA
megan.strait@tufts.edu

Cody Canning
Tufts University
161 Collge Avenue
Medford, MA 02155 USA
cody.canning@tufts.edu

Matthias Scheutz
Tufts University
161 Collge Avenue
Medford, MA 02155 USA
matthias.scheutz@tufts.edu

CHI 2014, April 26–May 01, 2014, Toronto, Ontario, Canada.

http://dx.doi.org/10.1145/2559206.2578866

Abstract
Previously, we contributed to the development of a brain-computer interface (BCI), Brainput, using functional near infrared spectroscopy (NIRS). Initially Brainput was found to improve performance on a human-robot team task by adapting a robot's autonomy using NIRS-based classifications of the user's multitasking states [15, 16]. However, the failure to find any performance improvements in a follow-up study prompted reinvestigation of the original system via a reanalysis of Brainput's signal processing on a larger NIRS dataset and a placebo-controlled replication using *random* (instead of NIRS-based) state classifications. This reinvestigation revealed confounds in the original study responsible for the initial performance improvements, thus indicating that further work in signal processing is necessary to achieve reliable NIRS-based BCIs.

Author Keywords
Functional near-infrared spectroscopy; brain-computer interfaces; replication

ACM Classification Keywords
H.5.m. [Information Interfaces and Presentation (e.g. HCI)]: Miscellaneous

Introduction

Brain-computer interfaces (BCIs) have been gaining traction in the field of human-computer interaction (HCI) for various application domains (e.g., [6, 19]. Functional near infrared spectroscopy, in particular, has been described as a suitable modality for BCIs (e.g., [3, 9]) given that NIRS is relatively portable and reasonably robust to user-movement [11]. Hence, work in HCI has increasingly focused on the development of NIRS-based BCIs to achieve performance improvements in areas such as human-robot team tasks [15, 16], preference prediction [12], and high-workload situations [1, 7].

Despite the relative portability and robustness to movement, however, the reliability of these NIRS-BCIs remains relatively unexplored (e.g., [2, 3, 17, 18]). Specifically, the NIRS-based passive BCIs that we and others have developed (i.e., [1, 7, 15, 16]) have shown performance improvements on various tasks despite unrestricted participant mobility and task durations. Yet, signal corruption due to artifacts arising from uncontrolled tasks/movement is still considered a major challenge in functional neuroimaging (e.g., [10, 11, 13]). Moreover, of the few investigations that do address system reliability, all indicate low replicability (e.g., [4, 13, 17]).

Thus, in this paper, we present a 2-part reinvestigation of a NIRS-based passive BCI system, Brainput [15], of whose development the authors were part. The Brainput system was designed to react to fluctuations in cognitive workload by adapting a (simulated) robot's level of autonomy [15, 16]. In the initial evaluation of Brainput (reported in [15]), this passive NIRS-based adaptivity was observed to significantly improve performance on a human-robot team task. However, when we attempted a follow-up extension of [15] using real robots, we did not find the improvements that we had observed originally. Hence, we set out to systematically evaluate the reliability of NIRS-based pBCIs through a two-part re-investigation of Brainput via (1) a reanalysis of Brainput's signal processing on a large NIRS dataset and (2) a placebo-controlled replication using *random* (instead of NIRS-based) state classifications. This reinvestigation revealed confounds in the original study responsible for the initial performance improvements and indicated that further signal processing improvements are necessary to achieve robust NIRS-based systems.

Motivation

We previously participated in the development of the passive NIRS-BCI (Brainput), with the aim of classifying a user's multitasking state during a human-robot team task by imaging the anterior prefrontal cortex (PFC) in real time [15]. We hypothesized that adapting the level of a robot's autonomy would lead to better task performance. In the original evaluation ([15]), participants worked with two simulated robots on a *search and report* team task while wearing NIRS sensors which the Brainput system used to classify multitasking states and update a robot's autonomy in real-time. The initial results showed that Brainput system substantially improved task performance (82% of participants successfully completed the task) versus a baseline task performance rate 45%. However, in a follow-up extension to the original evaluation, we did not find *any* performance improvements.

Original Evaluation

A two-channel ISS OxiplexTS was used to image participants' anterior prefrontal cortex. Two rubber-seated sets of sensors were worn on the center of the forehead perpendicular to the nasal bone and above the eyebrows. Sensors were held in place with black cotton headbands. Each probe contained four light sources, each of which

emitted two light wavelengths (690 nm and 830 nm), for a total of sixteen data *channels*. The NIRS data were classified online with a classification model built by the Sequential Minimal Optimization (SMO) algorithm available in the WEKA library [8]. The SMO algorithm was instantiated with a Radial Basis Function kernel with default parameters. All SMO parameters were also set to their default values. For each participant, unique classifiers were trained for each individual NIRS channel on 10 samples of two workload classes (low, high). Each training sample was composed of a feature vector of length (2 probes x 4 sources x 2 signals) x (40s x 6.25 samples/s), for a total of 4000 features (see [15, 16] for more details).

To test the performance improvements due to brain-based adaptivity, we designed a human-robot team task wherein participants simultaneously worked with two robots to locate a goal. Here the participant explored a simulated environment with each robot simultaneously to find a certain level of "field strength" (a particular location in the environment in which information could be transmitted). Participants could interact with each robot with the following commands: "go straight" (robot drives straight), "turn left/right" (robot turns while maintaining forward velocity), "take a reading" (robot stops, assesses field strength for 2s, reports signal strength to participant, continues with previous command), and "go back" (robot drives in reverse). Each robot operated in a unique copy of a simulated 16x16 room with a 2x2 obstacle in the center. The field strength was distributed over each room and values reported by the robots ranged from 1300 to 2500. A strength of at least 2400 was required to transmit the information to home base and the target location occupied 1.25% of the room. Each trial ran until both robots located the transmission location (task success), or until a maximum of 5 minutes had passed (task failure).

During this human-robot team task, Brainput classified participants' hemodynamic activity every 0.16 seconds as either *branching* (high workload) or *non-branching* (low workload) and dynamically adapted the autonomy of one of the robots according to the participant's level of cognitive workload. To test the efficacy of the brain-based adaptivity of Brainput, there were three conditions in which the task was performed: adaptive (autonomy was enabled when high workload was detected), non-adaptive (the robots never acted autonomously), and *mal*-adaptive (autonomy was *disabled* when high workload was detected). When the autonomy mode was enabled, the adaptive robot peformed the following behavioral loop:

1. Stop and take a reading for 2 seconds.
2. Move toward higher field strength for 5 seconds by going forward, turning left, or turning right.

Here we found that rates of successful task completion were substantially moderated by adaptivity. That is, in the adaptive condition, 82% of participants successfully completed the task versus 45% in the non-adaptive (baseline) condition. Moreover, reversing the timing of the autonomy (mal-adaptive autonomy) caused performances rates to significantly worsen (18%), suggesting that the autonomy must be appropriately-timed for it to be effective in human-robot teams.

Follow-up Extension

Inspired by these results, we employed Brainput and the experimental protocol of [15] in a follow-up extension to test whether the performance improvements would extend to real robots (as opposed to simulated robots in simulated environments used originally). This follow-up study was conducted with 24 participants (12 female),

Subject	M	SD	t_{obs}
28	34.0	9.3	-5.44
2	37.5	13.8	-2.86
18	40.5	7.4	-4.91
20	43.1	11.2	-1.94
26	43.7	21.6	-.92
36	45.8	18.8	-.70
19	46.2	13.3	-.90
4	46.2	12.6	-.95
24	47.2	12.2	-.72
29	48.6	14.1	-.31
1	48.9	11.5	-.30
22	50	15.7	-.02
32	51.4	14.0	.31
21	52.1	14.0	.47
38	52.1	14.2	.46
3	52.4	16.5	.46
5	52.8	15.2	.58
35	52.8	14.2	.62
25	53.1	16.5	.59
8	53.8	16.6	.72
6	54.2	16.9	.78
31	54.2	10.8	1.22
40	54.9	13.0	1.1
14	55.6	14.2	1.24
39	56.2	17.8	1.1
34	56.6	13.6	1.53
12	**57.3**	**11.6**	**1.99**
27	57.6	16.5	1.45
37	59.0	17.3	1.65
9	60.4	18.8	1.74
11	**60.8**	**16.3**	**2.09**
17	61.1	23.7	1.48
30	**62.1**	**9.8**	**3.90**
13	**63.9**	**14.9**	**2.95**
23	**65.3**	**17.2**	**2.81**
7	**65.6**	**10.0**	**4.93**
10	**67.0**	**10.8**	**4.97**
33	**71.2**	**18.1**	**3.70**
15	**73.6**	**11.0**	**6.78**
16	**75.7**	**9.0**	**9.03**
	54.5	**14.3**	

Table 1: Average classification accuracies (%). In red: participants with accuracies significantly above chance.

using the same materials and methods as in [15]. Here we replaced one of the two previously-simulated robots with a real robot (Willow Garage PR2) in a real environment, but both were still controlled by the system architecture used in [15]. This *real* robot received commands wirelessly from the participant as it drove around its environment and transmitted a live video feed from a camera attached to its base. The simulated environment was made to be equivalent the real environment, having the same dimensions, obstacles, and lighting.

As in the original evaluation ([15]), each participant did the task three times: once with neither robot autonomous (to measure baseline task performance) and twice with an adaptive robot (once with the simulated and once with the real robot adaptively autonomous; order was counterbalanced). Here we expected to see the same performance improvements as we did in the original evaluation. The results, however, showed *no* performance improvement (for either simulated or real robot) from the baseline performance rate. That is, contrary to the findings of the original evaluation ([15]), the participants equipped with the NIRS-BCI were not aided by the brain-based, dynamically adaptive autonomy of the robots. Given the nearly identical setup (except for the modification to use one real robot), these results suggested some degree of unreliability of the Brainput system. Given that the Brainput framework (for classifying user multitasking states) used high-dimensional feature sets (4000 features), and that SVMs are known to produce poor performance on highly-dimensional data [5], we suspected the Brainput classifier to have limited applicability for a larger population sample. As such, we conducted a re-investigation of the behavior of the Brainput NIRS classifier on a large NIRS dataset as well as on the original data.

Reanalysis of Signal Processing

Based on the results of our follow-up investigation, we had two primary questions: (1) whether the original performance improvements persist over a larger subject sample (i.e., N $>$ 11) and moreover, (2) whether the improvements generalize across variants of the same type of task (i.e., workload-inducing tasks). To evaluate the generalizability of Brainput's classification approach, we first obtained a larger (N=40) NIRS datasest consisting of low and high workload PFC samples induced by a variant of the n-back task [17]. As the n-back task was used in training the Brainput classifier in both [15] and our follow-up study, we expected Brainput to perform similarly as it did in the preliminary analysis of its accuracy ([14]). However, in observing lower-than-expected classification accuracies, we then revisited the dataset from our original investigation ([15]) to identify the factors responsible for the original performance improvements.

Novel Dataset

We obtained the large NIRS dataset from [17], measuring PFC activity associated with workload states. The NIRS equipment and sampling region of interest were the same as that of [15]; however, to test the generalizability of the Brainput approach, the tasks in [17] were arithmetic-based variants of the n-back task (rather than the alphameric tablet task n-back variant used in [15]). This dataset contained fourty participant records, each with eighteen 30-second training samples (nine low and nine high workload).

The new dataset was first preprocessed in MATLAB (MathWorks Inc.) using the same methodology as in [16]. We then trained the Brainput classifier on the nine 30-second samples of each workload class (low, high). The training mirrored the procedure we used originally,

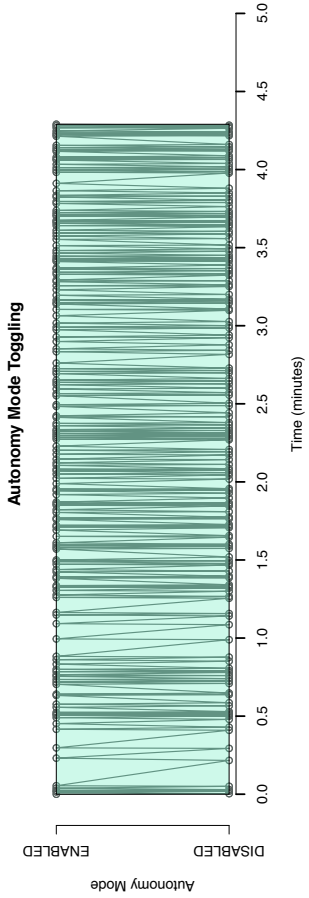

Figure 1: Classification log (subject ID: 109) of *autonomous mode* toggling. Each line indicates a switch in autonomy (e.g. enabled to disabled).

where Brainput was trained on eight to ten 40-second samples each of the low workload and high workload tablet task trials. Following, we ran ten-fold cross validation on the training data for each subject (see [15] for details) in order to predict model performance.

Across all subjects, the average classification accuracy was 54.5% (*SD*=14.3%), with accuracies ranging from 34.0% to 75.7% (see Table 1). Using a right-tailed *t*-test (α <.05), the overall classification accuracy (54.5%) was found to be statistically significant ($t_{crit}(39)$=1.6848; t_{obs}=1.9399) However, of the 40 participants, only 10 participants showed average accuracies that were statistically significantly above chance (right-tailed *t*-test; $t_{crit}(9)$=1.8331, α <.05).

In comparison to the preliminary evaluation of Brainput's classification approach ([14]), Brainput's average rate across participants on this dataset was substantially lower than than what was found originally (54.5% here versus 68.4% in [14]). This discrepancy may have been due to the difference in sample size (N=40 here vs. N=3 of the [14] evaluation) or alternatively, to the difference in task (numeric vs. alphameric), or both.

In either case, the lower performance overall and the low proportion of participants for which Brainput was effective, were particularly worriesome for numerous reasons. First, the realtime task (the human-robot team task) substantially differed from the alphameric training task in the original evaluation [15]. Hence, if the classification schema does not extend well to a variant of the same type of task (numeric vs. alphameric n-back task, as opposed to human-robot interaction task vs. alphameric n-back task), then it suggests that it might not generalize to more realistic applications (such as human-robot interaction tasks). Secondly, with an average classification accuracy only slightly above chance (54.5%), then such large performance improvements in the original evaluation (80% improvement) cannot be explained by the Brainput classifier alone. Specifically, if Brainput is only effective (significantly better than random guessing) for 1/4 of the population, then performance improvements should reflect that (i.e., improvements should be closer to 25%, not 80% as observed).

Original Dataset

Hence, as Brainput's performance on the novel dataset differed substantially (-13.9% in accuracy) from its preliminary offline evaluation [14], we revisited the dataset from the dynamically adaptive task in [15] in order to investigate Brainput's *realtime* behavior and understand how such large performance improvements were originally achieved. Using the logs of the realtime classifications, we constructed plots of the robot's autonomous behavior (autonomy-disable vs. autonomy-enabled; adaptive condition only) over the course of the human-robot team task for each participant (Figures 1, 2, 3, 4, 6 show a subset of 5 out of the 11 participant logs). Enablement of the robot's autonomy corresponded to classification of the NIRS data as *high* workload, whereas disablement indicated the participant was experiencing *low* workload.

For the adaptive autonomy of the robot to be effective, we expected the behavior to show prolonged periods (e.g., 30s in duration) of autonomy enabled/disabled, with a handful of switches between autonomy modes. However, we found instead rapid (sub-2s) classification-switching. Since most of the classification switches took place in under two seconds, this meant that the robot's autonomous behavior rarely proceeded to the second step in its behavioral loop: moving in the direction of greatest field strength. That is, the autonomy amounted primarily

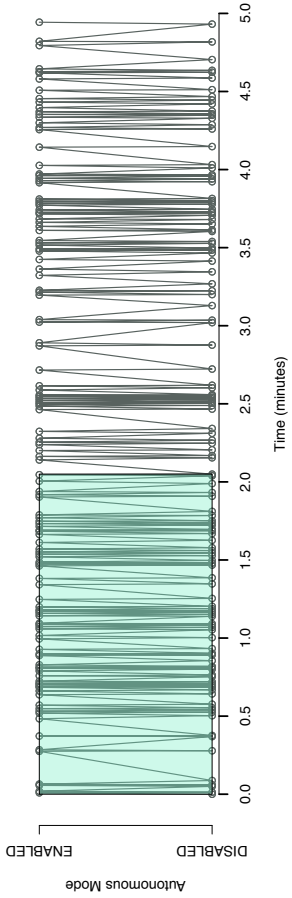

Figure 2: Shaded: time during which subjects interacted with *both* robots. Subject ID: 106.

to the robot only stopping in place, as its autonomy behavioral loop designated first stopping for 2s to take a reading (before moving towards the goal).

However, regardless of the robot's behavioral activity, this rapid oscillation between classifications was even more worriesome than Brainput's performance on the novel dataset. Specifically, the rapid sub-second oscillations were inconsistent with basic hemodynamics – that task-related hemodynamic activity occurs over a period of several seconds (e.g., [3, 11]). These results suggested that, in fact, the Brainput classifier was likely not the primary factor contributing to performance improvements on the human-robot team task in [15]. They suggested instead, perhaps the presence of a placebo effect (e.g., of wearing the NIRS sensors) or a confounding factor (e.g., the mere presence of robot autonomy) in the experimental design. Hence, in order to disentangle the relative effects of the NIRS-based adaptivity versus other experimental factors, we conducted a placebo-controlled replication.

Placebo-Controlled Replication

In order to understand why we did not observe the performance improvements due to adaptive autonomy as expected, we performed a placebo-controlled replication of [15]. The relevant details of [15] are reproduced here or referred to within the Motivation section of this paper. For each of the following subsections more information is available in [15, 16]. For comparison with the original evaluation, we conducted this placebo replication in *precisely the same* fashion as that of the original [15] with only the modification except that *cognitive state classifications from Brainput were replaced by random classifications* in order to test for placebo effects. This design allowed us to explicitly control for effects due to *autonomy* of the robot separate from the effectiveness of the Brainput system in identifying cognitive states.

Participants

Ten participants (4 male) were recruited from the Tufts University area (versus 11 in [15]). The average age was 22.4 (SD = 4.93) with a range of 18 to 35 years old. All participants were right-handed and fluent English speakers with no history of brain trauma. Informed consent was obtained from each participant and the study was approved by the university's institutional review board. Participants were paid for their participation.

Control Modification: Randomly Adaptive Autonomy

In this modified version of the original evaluation ([15]), we did not use Brainput for online classification of multitasking state. Instead we generated *random* classifications (of branching or non-branching) based on the classification distributions from the original study. With this approach, we expected (random) *adaptive* and (random) *maladaptive* conditions to show equal effects on task performance given that the adaptations are drawn randomly from the *same* underlying classification distribution. This allowed us to identify the Brainput-based performance enhancements relative to the performance enhancements from *autonomy*. Greater performance enhancements of Brainput in comparison to random autonomy would indicate that brain-based adaptivity facilitates the task completion better than randomly-initiated autonomy. On the other hand, equal performance enhancements of the two approaches would indicate that the brain-based adaptivity of Brainput *did not siginificantly contribute* to performance improvements.

Procedure

After the participant consented to the experiment, he or she was briefed about the search and report task. The

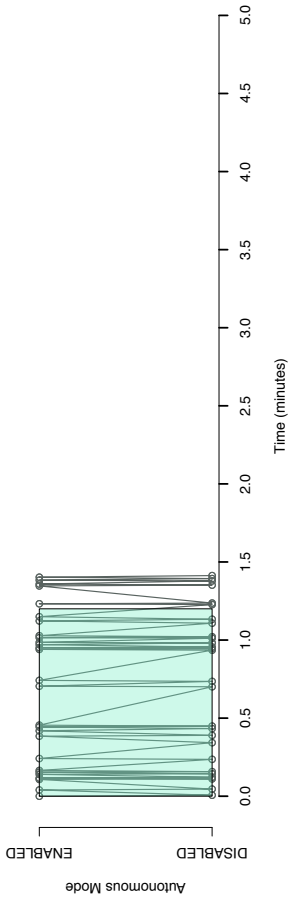

Figure 3: The end of the shaded region marks task success with one of the robots, after which point the subject only interacted with one robot. Subject ID: 111.

participant did a five-minute practice run of the task (in the *non-adaptive* condition). Subsequently the participant was briefed about the tablet task and practiced until he or she scored 80 percent correct or better. If the participant could not reach this performance threshold after a reasonable number of attempts, he or she was compensated and dismissed from the study.

Once practiced in both tasks the participant was fitted with the NIRS probes. The participant was told to move as little as possible and to keep his/her hands in a comfortable, rested position while wearing the probes. The lights were turned off and the participant proceeded with a 1-minute baseline NIRS recording wherein he or she visually fixated on the center of the computer screen. After the baseline the participant did 20 randomly ordered trials of the tablet task while NIRS data were recorded. These trials were separated halfway through by a two-minute break. The experimenter was not present in the room during the task. Data from the tablet task were used as the training set for the classifier (same as [15]). After the tablet task the participant did all three conditions of the search and report task (order was counter-balanced) with short rest periods in between each.

No changes were made to the [15] procedure in this experiment. Specifically, all details of the materials and procedure of this placebo-controlled study were identical to those of the original except where explicitly stated. This includes all scripts, code, programs, and stimuli used in the experiment. Participants still did the alphameric variant of the n-back task (the tablet task) and wore the NIRS probes but the data were not used. Despite the generation of *random* cognitive load classifications, all procedures remained the same: from the participant's perspective, [15] and this experiment were the same.

Experiment	*Adaptive*	*Maladaptive*
Brainput [15]	82%	18%
Brainput [15] (corrected)	75%	0
Random (placebo)	60%	20%

Table 2: Successful task completion with the dynamically autonomous robot, in the *adaptive* and *maladaptive* conditions of the search and report task. In reviewing the original study, it was found that 3 of the 11 participants had different initial conditions and were thus removed (corrected, above).

Results and Discussion

We found similar patterns of task success in comparing the results of this placebo-replication and [15] (see Table 2). Specifically, participants succeeded (in both experiments) more often at locating the goal location with the autonomous robot in the *adaptive* condition than in the *maladaptive* condition. This result was surprising because the random state classifications should have caused the success rates in both the *adaptive* and *maladaptive* conditions to be equal. This disparity in success rate between two conditions that were, in this placebo-controlled experiment, equal in all aspects thus indicated an experimental confound between the execution of the two conditions.

Upon inspection and comparison of the data logs from this replication, we discovered a major confounding details that explained this result. Namely, the transmission location in the environment co-varied with the task condition (in the script that initiated the experiment). In the *maladaptive* condition the robot had significantly further to travel than in the *adaptive* condition; in other words, the task was strictly harder and more time-consuming in the *maladaptive* than the *adaptive* condition. The straight-line distance from the starting location to the transmission location was 9.4 m in the

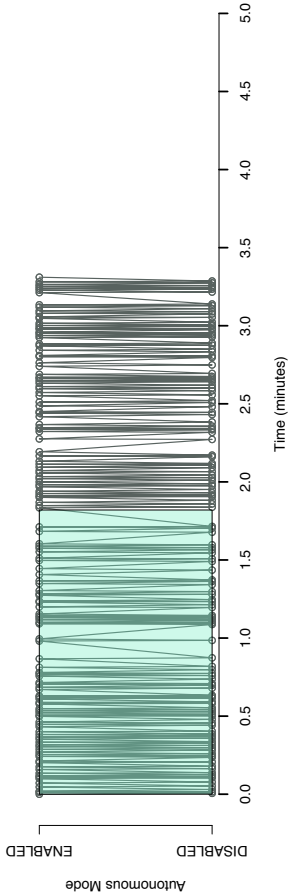

Figure 4: Subject ID: 118.

adaptive compared to 18.4 m in the *maladaptive* condition (see Figure 5). Since task success was determined by the team's ability to locate the goal in five-minutes time, it is clear that the coordinates of the transmission location relative to the robot affected rate of success at the task.

Since no aspects of the underlying system architecture were changed other than the classification approach (from NIRS-based to random) in conducting this placebo-controlled replication of [15], we suspected this affected the original experiment as well. In reviewing the data logs from the original evaluation ([15]), we confirmed that in that experiment as well, the initial starting locations of the robot in the adaptive versus maladaptive condition were those of what we found in this replication. Hence it was unlikely that Brainput alone yielded the performance improvements we observed, as, in the absence of the brain-based adaptivity, we still achieved these improvements. Thus it is more likely that the confounding factor of starting location – which persisted between the original experiment and this replication – was the factor responsible for improved performance.

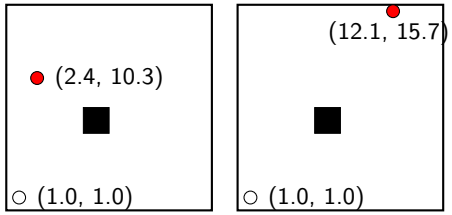

Figure 5: Top-down view of the configurations of the search and report task environment for the *adaptive* condition (left) and the *maladaptive* condition (right). The black square is the 2 x 2 meter obstacle, the unfilled circle is the starting location of the robot, and the red circle is the goal location.

General Discussion

Due to the finding null results in an extension of the original evaluation of Brainput ([15]), we were motivated to conduct a systematic investigation of the reliability of the Brainput framework for NIRS-based BCIs.

We first revisited Brainput's signal processing methods, to test the system's extensibility to a novel dataset and secondly, to look at Brainput's realtime behavior (which was never done previously). In our reanalysis of Brainput's methods, we found the classification approach to perform worse than expected on a novel dataset. This finding is consistent with existing functional neuroimaging literature, which also suggests low reliability of classification accuracy of NIRS data [3, 4, 13, 17]. In addition, Brainput's performance was only significantly better than chance for 10/40 of the subjects, indicating low efficacy for a general population. Moreover, when we looked at Brainput's realtime behavior by plotting the classification logs of [15], we found that Brainput's behavior did not follow basic hemodynamic principles (via observation of sub-1s oscillations). These results thus suggest that further work is necessary to achieve a robust/reliable classification framework of NIRS data, as that of the Brainput system is not effective in the context of larger populations and does not generalize to similar workload-inducing tasks (i.e., numeric n-back variant and the human-robot team task above).

With such unexpected behavior and worse-than-expected classification performance, we then revisited the original evaluation of Brainput's efficacy via a placebo-controlled experiment. There we identified a significant confounding factor responsible for the original performance improvements observed in [15]. Via the placebo-controlled replication of the original study using a distribution of

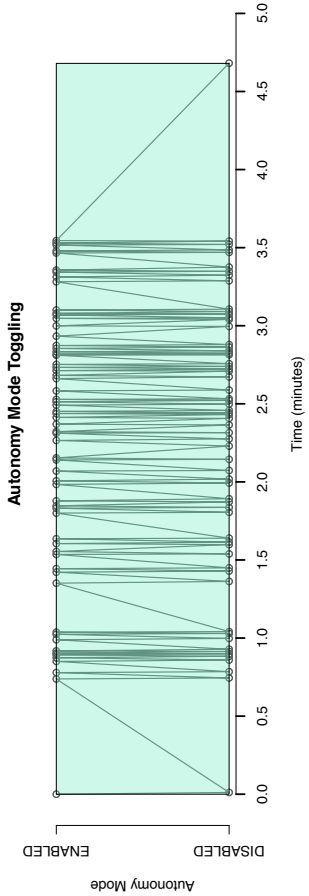

Figure 6: Subject ID: 123.

random classifications to simulate Brainput revealed the same pattern of improved performance at the task. That is, adapting robot behavior to *random* classifications of the human's multitasking state showed the same improvement as adapting it to Brainput's classifications, demonstrating that Brainput is very likely *not* responsible for causing these improvements. Specifically, we successfully replicated the original performance improvements – despite the absence of NIRS-based adaptivity – which suggested a confound within the experimental design.

Hence we investigated further to identify the true cause of these improvements by analyzing logs from the replication and the original study and found a confound in the experimental design. This careful investigation into the design of the study identified a disparity in starting locations between the experiment conditions (i.e., the robot was substantially closer to the target location in the adaptive versus maladaptive condition to begin). That confound resulted in the task being easier in the *adaptive autonomy* condition than in the *maladaptive autonomy* condition. In other words, the key metric of success of the Brainput system was very likely only the result of confounded experimental design.

Although these findings suggest low replicability and extensibility of the Brainput system, the subjective reports in the original experiment indicate the dynamic autonomy improved perceptions of the robot teammates. Specifically, the adaptively autonomous robot was found to be more helpful and cooperative than its non-adaptive counterpart. Thus, there may be still some utility to the NIRS-based adaptivity (despite the confounds), though perhaps not measureable at the scale of task completion rates. However, as Brainput was the earliest demonstration of an effective NIRS-pBCI, its framework has been used in further development of numerous other NIRS-pBCIs (i.e., [1, 7, 12, 16]). Hence, it is increasingly important that we consider the reliability of our systems and the contexts in which they are effective. Moreover, based on the results of this paper, it is important that we revisit our frameworks for NIRS-pBCIs in order to improve general accuracy of the systems.

Conclusions

Functional near-infrared spectroscopy has recently received considerable attention as a tool for real-time adaptive BCIs. However, in a series of reinvestigations of the Brainput NIRS-pBCI, we found significant limitations of its efficacy. First, we found when we increased our sample size from 3 to 40, Brainput's performance was only effective for 1/4 of the population. Moreover, we observed that Brainput's realtime behavior (sub-1s state-switching) is not in accordance with basic hemodynamic principles (slow changes, e.g., 2s+). Further investigation into Brainput's unexpected realtime behavior identified a major confounding factor (different starting locations) in our original evaluation of the system, which was likely responsible for the performance improvements (and not the NIRS-based adaptive autonomy). Hence, it is important that we revisit our NIRS-pBCI frameworks to consider the reliability of our systems. We hope that this systematic reinvestigation will help to identify current obstacles and lead towards more robust realtime adaptive NIRS-BCIs.

References

[1] Afergan, D., Peck, E., Solovey, E., Jenkins, A., Hincks, S., Chang, R., and Jacob, R. Dynamic difficulty using brain metrics of workload. In *CHI* (2014).

[2] Brouwer, A.-M., van Erp, J., Heylen, D., Jensen, O., and Poel, M. Effortless passive BCIs for healthy users. In *HCII* (2013), 615–622.

[3] Canning, C., and Scheutz, M. Function near-infrared spectroscopy in human-robot interaction. *Journal of Human-Robot Interaction 2* (2013), 62–84.

[4] Coffey, E., Brouwer, A.-M., and van Erp, J. Measuring workload using a combination of electroencephalography and near infrared spectroscopy. In *HFES* (2012).

[5] Cortes, C., and Vapnik, V. Support-vector networks. *Machine learning 20*, 3 (1995), 273–297.

[6] Frey, J., Muhl, C., Lotte, F., and Hachet, M. Review of the use of EEG as an evaluation method for human-computer interaction. In *PhyCS* (2014).

[7] Girouard, A., Solovey, E., and Jacob, R. Designing a passive brain computer interface using real time classification of functional near- infrared spectroscopy. *International Journal of Autonomous and Adaptive Communications Systems 6*, 1 (2013).

[8] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter 11*, 1 (2009), 10–18.

[9] Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., and Schultz, T. Mental workload during n-back task - quantified in the prefrontal cortex using fNIRS. *Frontiers in Human Neuroscience 7* (2013).

[10] Hoshi, Y. Functional near-infrared spectroscopy: current status and future prospects. *J. Biomed. Opt. 12*, 6 (2007).

[11] Hoshi, Y. Towards the next generation of near-infrared spectroscopy. *Phil. Trans. of the Royal Society A: Mathematical, Physical and Engineering Sciences 369*, 1955 (2011), 4425–39.

[12] Peck, E., Afergan, D., and Jacob, R. J. K. Brain sensing as input to information filtering systems. In *Augmented Human* (2013).

[13] Plichta, M., Herrmann, M., Baehne, C., Ehlis, A., Richter, M., Pauli, P., and Fallgatter, A. Event-related functional near-infrared spectroscopy (fNIRS) based on craniocerebral correlations: reproducibility of activation? *Human Brain Mapping 28*, 8 (2007), 733–41.

[14] Solovey, E., Lalooses, F., Chauncey, K., Weaver, D., Parasi, M., Scheutz, M., Sassaroli, A., Fantini, S., Girouard, A., and Jacob, R. J. K. Sensing cognitive multitasking for a brain-based adaptive user interface. In *CHI* (2011), 383–392.

[15] Solovey, E., Schermerhorn, P., Scheutz, M., Sassaroli, A., Fantini, S., and Jacob, R. Brainput: enhancing interactive systems with streaming fNIRS brain input. In *CHI* (2012), 2193–2202.

[16] Solovey, E. T. *Real-Time fNIRS Brain Input for Enhancing Interactive Systems*. PhD thesis, Tufts University, 2012.

[17] Strait, M., Canning, C., and Scheutz, M. Limitations of NIRS-based BCI for realistic applications in human-computer interaction. In *BCI Meeting* (2013).

[18] Strait, M., and Scheutz, M. Building a literal bridge between robotics and neuroscience using functional near infrared spectroscopy. In *HRI Workshop on Bridging Robotics and Neuroscience* (2014).

[19] Zander, T. O., and Kothe, C. Towards passive brain-computer interfaces: applying brain-computer interface technology to human-machine systems in general. *J. Neur. Eng. 8*, 2 (2011).