Adversarial Stimuli: 
Attacking Brain-Computer Interfaces via Perturbed Sensory Events

Bibek Upadhayay¹ and Vahid Behzadan, Ph.D.²

Abstract—Machine learning models are known to be vulnerable to adversarial perturbations in the input domain, causing incorrect predictions. Inspired by this phenomenon, we explore the feasibility of manipulating EEG-based Motor Imagery (MI) Brain Computer Interfaces (BCIs) via perturbations in sensory stimuli. Similar to adversarial examples, these adversarial stimuli aim to exploit the limitations of the integrated brain-sensor-processing components of the BCI system in handling shifts in participants’ response to changes in sensory stimuli. In this paper, we first define adversarial stimuli and enumerate the characteristics of such stimuli. Second, we study the feasibility of adversarial stimuli as an attack vector in a series of human subject experiments, and report the findings on the impact of visual adversarial stimuli in the form of random partial flickers on the MI task of playing a video game. Our findings suggest that adversarial stimuli can significantly deteriorate the performance of MI BCIs across all participants, and also that such attacks are more effective in conditions with induced stress. Furthermore, we provide a preliminary analysis on the underlying dynamics of adversarial stimuli attacks by investigating the variations in Alpha, Beta, and Gamma bands. Lastly, we present a discussion on the impact of our findings, and enumerate directions of further research in this area.

I. INTRODUCTION

In recent years, Brain-Computer Interfaces (BCIs) have enjoyed accelerating rises in accessibility and development. These systems provide a direct medium between the brain and computing devices, thereby facilitating a plethora of applications, ranging from medical monitoring and rehabilitation [1], [2] to enabling seamless modes of communications and control. The foundational mechanisms of such BCI applications are based on exploiting the correlations between measurable electrochemical activities of the brain and the higher-order cognitive functions such as perception and motor intent [3]. While the variety of measurement techniques adopted in BCIs has been increasing, ElectroEncephaloGraphy (EEG) has been dominating the field due to its relative simplicity and accessibility [4]. The increasing availability of commercial EEG devices has facilitated the rise of numerous novel applications, from brain-controlled keyboards [5], neuroprosthetics [2], and controlling wheelchairs [1] to game play [6] and biometric authentication [7].

However, such recent advances in BCI technologies give rise to new concerns about the security of these solutions. In response, the field of Neurosecurity [8] has been established to address the security and privacy issues that may arise from the accelerating adoption of BCI technologies. The growing body of work on this field includes studies on the vulnerabilities of the hardware, software, and machine learning components of BCI systems [9]. Instances of the latter are the studies that propose Man-in-The-Middle (MiTM) attacks, in which the adversary compromises the connection between the EEG headset and the computer, and injects adversarially crafted noise to manipulate the machine learning model processing the correlations between measurements and intent [10], [11], [12], [13], [14], [15]. However, current studies fail to present attacks which exploit vulnerabilities that are specific to BCIs.

In this paper, we aim to address the aforementioned shortcoming by exploring the feasibility of manipulating Motor Imagery (MI) BCIs via perturbing the visual stimuli observed by the BCI user. Inspired by adversarial example attacks against machine learning, we hypothesize that the integration of cognitive, measurement, and machine learning components in EEG-based BCIs may also be vulnerable to minor perturbations that can be induced directly at the sensory level. To examine this hypothesis, we performed a preliminary study on 7 human subjects to measure the impact of visual adversarial stimuli on their performance in MI tasks. The results of these experiments validate the feasibility of adversarial stimuli attacks against EEG-based MI BCIs.

This paper is organized as follows: Section III introduces adversarial stimuli as an attack vector against BCIs. Section IV presents the details of our preliminary experiments on proof-of-concept adversarial stimuli attacks against EEG-based MI BCI tasks. Section V reports our experimental results, and Section VII concludes the paper with a discussion and future research directions.

II. RELATED WORK

BCIs pose various attack surfaces which can be targeted by adversaries (Fig. 1), thus making the safety and security of EEG-based BCIs and their users of paramount concern. According to [16], security refers to the protection of information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction to provide integrity, confidentiality, and availability. EEG-based BCIs are shown to be susceptible to a range of attacks, such as replay, spoofing, and jamming [10], [11], [12], which can affect data integrity and availability during the data acquisition process. Moreover, [10] demonstrate that malware attacks on BCIs can adversely affect the output of BCI-based devices, making it appear as if the user performed

¹Bibek Upadhayay is a graduate student at the University of New Haven, West Haven, CT, USA bupadhayay@newhaven.edu
²Vahid Behzadan, Ph.D. is an assistant professor of computer science at the University of New Haven, West Haven, CT, USA vbehzadan@newhaven.edu
a different task than what they intended. The confidentiality of BCI data is also at risk, as shown by the experimental work of [17], in which private information were extracted from subjects via exposing them to misleading stimuli, thus decreasing the entropy of private information on average by approximately 15 – 40% compared to random guessing attacks. Similarly, [13] demonstrated how an attacker can infer private information even with subliminal brain activity in response to rapid visual stimuli. By reverse-engineering the NeuroSky system framework, Xiao et al. [18] exploited vulnerabilities in the software and wireless components of those systems, which enabled the attacker to steal a user’s brain wave data without even accessing the victim’s device. Another instance of attacks is the work of Meng et al. [19], which demonstrated backdoor attacks by injecting poisoning samples into the training set using a narrow period pulse that does not need to be synchronized with the EEG trails.

In [14], it was shown that EEG-based BCIs using CNN classifiers are vulnerable to adversarial attacks, even with small perturbations. The authors injected perturbations into the wireless link via jamming, which allowed for successfully attacking three different CNN classifiers in white-box, gray-box, and black-box scenarios. However, the authors’ approach has limitations in practical attack models as it requires computing perturbations for each input trial and waiting for the trial to be completed. Liu et al. [15] proposed a universal adversarial perturbation approach that generates perturbations for EEG trials for both targeted and non-targeted attacks in the white-box settings. This approach was tested on three BCI datasets in white-box settings. The authors implemented their attack model against three BCI datasets, namely P300 evoked potentials, Feedback Error Related Negativity (ERN), and Motor imagery (MI). A further work by Jian et al. [20] proposed a black-box adversarial attack model using active learning based on query synthesis. This approach improves query efficiency by searching for informative examples in the input space for substitute model training. These attacks impact the signal generation, acquisition, and processing stages of BCIs, and impact the integrity domain of EEG-based BCIs.

In addition to software and hardware attacks, EEG-based BCIs can be affected by perturbed stimuli. P300 BCIs for assistive vehicles were found to be impacted by auditory stimuli [21]. While one study [22] observed improved user performance with passive distraction compared to active distraction, another study [23] reported a decrease in accuracy from 100% to 87.5% for a P300 speller under active listening conditions. Visual stimuli were found to affect the performance of motor imagery BCIs [24], thus decreasing the mean classification accuracy from 81.5% to 78.3% and increasing physiological and perceived cognitive load [25]. Contrary to the experiments mentioned above, researchers in [26] conducted tests on the performance of SSVEP-based BCIs under different stimuli and observed no significant changes in the BCI performance. Similarly, in [27], a study was conducted on 10 subjects to examine the effect of audio and visual stimuli in SSVEP-based BCIs, and the authors found that the effect of the stimuli on SSVEP-based BCIs was minimal.

III. ADVERSARIAL STIMULI

While an attacker may target and attack each stage of the BCIs as shown in Fig 1, such ad-hoc attacks are not practical. Targeting the software or hardware components may result in changes that render the overall attacks unsuccessful. Additionally, many of these attacks require the attacker to be in close proximity to the victim, making them less likely to succeed in real-life scenarios. In contrast, our threat model allows for attacks without the need to be in close proximity to the user or BCI devices. In our model, the attack surface is the environment, as opposed to the software or hardware.

We consider two scenarios. In the first scenario, the attacker gains access to the screen that the subject employs for controlling/feedback in BCIs. In the second scenario, the attacker gains access to the environment and can introduce stimuli remotely, for instance, by compromising the home automation system. However, in our case, the attacker exploits the environment.

Additionally, previous studies have established that the machine learning models trained on EEG-based MI tasks do not generalize well and are prone to overfitting [28], [29]. Therefore, similar to all machine learning models, such BCI models are also prone to adversarial examples attacks [14], [15], [20]. These findings raise the hypothesis that the EEG-based BCIs may also be vulnerable to perturbations in the sensory domain.

We define Adversarial Stimuli as perturbations introduced by adversaries in sensory events with the intent of manipulating the performance of the BCI. These perturbations are introduced into the environment in the form of stimuli, which can be auditory, visual, or tactile in nature, and are experienced by the subject during observation.

For example, in the case of visual adversarial stimuli, irregular flickering can be added to the scene observed by the user. This flickering effect can be applied to the
entire observable scene, such as flickering in light sources like lamps, or it may be confined to specific regions or observable events, like a segment of the visual interface in the BCI ecosystem. These perturbations can be induced remotely and do not require the attacker to have local access to the environment. We also present the adversarial stimuli mathematically.

We first define the subject behaviour \( \pi \) in the integrated EEG-based BCIs based on which it takes the action.

\[
\pi(o) = a
\]

(1)

Here, \( \pi \) represents the human behaviour or policy, \( a \) represents the action of the integrated BCI systems, and \( o \) represents the observation. In an ideal scenario, the subject’s decision to initiate the mental commands depends on the observation of the environment. Based on the observation of the environment, the subject chooses the action to initiate any one of the mental states, which will be processed by the integrated EEG-based BCIs system and gives the output \( (a) \) accordingly.

Secondly, we add the perturbation \( (\delta) \) in the environment via several means, which changes the observation space in the environment for the subject. We mathematically define the adversarial stimuli in Equation 2.

\[
\pi(o + \delta) \neq \pi(o)
\]

(2)

Additionally, we define several criteria that distinguish the stimuli to be adversarial stimuli.

1) The stimuli that adversaries can use with the intent to significantly manipulate or deteriorate the performance of the BCI.

2) Adversarial attacks that use stimuli to manipulate the BCI’s performance must be executed within the observation environment and not within the hardware or software.

3) The stimuli can be explicit or subliminal.

4) The duration of adversarial stimuli must be shorter than the time required for a single unit of the task, and it should not impede the primary task. For instance, if it takes three steps to reach the target from the starting point, and it takes one second to complete a unit task, the duration of adversarial stimuli should not exceed one second.

IV. EXPERIMENTAL SETUP

To investigate the feasibility of adversarial stimuli attacks, we formulate the following hypotheses: Hypothesis 1 (H1): There are small perturbations in the visual observations that can negatively impact the performance of the EEG-based MI BCIs; and Hypothesis 2 (H2): The effects of such adversarial attacks are more significant in the stressful settings of time-constrained tasks than on the open-time task.

We examined these hypotheses via experiments with seven subjects comprised of 2 males and 5 females aged 23-31 (IRB-2021-066). The subjects were asked to use an MI BCI to play the classic game of Pong displayed on a computer screen. We used the Emotiv Epoc X EEG headset for these experiments with 14 channels (AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4) with a sequential sampling rate of 128 Hz. The EmotivBCI software was used to create individual user profiles with individual training profiles.

In order to familiarize our participants with the BCI settings, the participants were first instructed to train an MI model for the task of moving a cube up or down inside the default environment of EmotivBCI. To facilitate uniformity, the participants were asked to imagine the act of throwing a ball with their right hand to initiate the MI signal for the “up” action. Similarly, in order move the cube in “down” direction, the participants were asked to imagine kicking a ball using their right leg.

After successful training, we evaluated the trained model by asking the participants to follow a random sequence of instructions (i.e., directions) in the cube task. We continued with the experiment only if the participants succeeded in implementing the instructed actions. Otherwise, we restarted the training process.

For the subsequent stages of our experiments, we created two customized versions of Pong, namely: the Warm-up environment and the full game. Each of these environments could be configured to operate in either the normal or adversarial modes. The warm-up environment provides a simplified setting to practice the MI control of the paddle by eliminating the ball movement - that is, the ball remains stationary. The environment provides instructions to participants to move the paddle in up or down directions, and the participants succeed if they manage to move the paddle in instructed direction within 15 seconds. Each participant could score up to 12 points in 3 cycles.

In the full game settings, the participants play the Pong game against an automated player by moving the paddle up or down to hit the ball, and score a point for each time they succeed. The game ends when the participants miss more than 5 balls, or the game duration exceeds 3 minutes. Fig. 2 illustrates this environment.

In the adversarial modes, we simulate a scenario in which an adversary has gained access to the Pong environment and can at any time change the flickering rate of the paddle and the ball. In our experiments, the adversarial flickering rate was set to 20 Hz. We chose this value because it corresponds to beta activity, which is associated with active, task-oriented thinking, busy or anxious mental states, and active concentration. In the normal environment, there are no adversarial stimuli, and the flickering rate remains constant. In contrast, in the adversarial environment, the flickering rate is modified at random intervals to create perturbations. The duration of each flickering stimulus was randomized between 0 and 5 seconds, this satisfied our fourth criteria for adversarial stimuli by not impeding the primary task.

V. RESULTS

For each of the warm-up and full game environments, we measured the total score obtained by participants under both normal and adversarial modes. Furthermore, we also
measured the error rate for each subject, $E_{r} = (S_{n} - S_{a})/S_{n}$, where $S_{n}$ is the score achieved in the normal mode, and $S_{a}$ is the score obtained in the adversarial mode. For experiments in the warm-up mode under normal mode (E1) and adversarial mode (E2), the maximum achievable score was set to 12. In the full game experiments under normal mode (E3) and adversarial mode (E4), no hard limit was set on the maximum score.

In order to test both of our hypotheses, we composed the following null hypotheses: Null hypothesis I states that there is no significant difference in the performance between normal mode and adversarial mode in the warm-up environments; and Null Hypothesis II states that there is no difference between the impact of adversarial stimuli in warm-up environment and full game environment. We performed paired t-tests with significance level $\alpha = 0.05$ for each case, namely E1 vs. E2 and E3 vs. E4. We also performed a paired t-test for the overall normal vs. adversarial environments combining warm-up and full game (i.e., E1, E3 vs. E2, E4). The statistical test results are presented in Table II. We observed that the value of $t \gg t_{c}$, and $p \ll 0.05$ in all of the aforementioned cases, thus rejecting the null hypotheses. The rejection of the first null hypothesis supports the claim that there is significant deterioration of performance between the normal mode and adversarial mode in the Warm-up environment.

VI. DISCUSSION

Impact: In our experiments, we observed that adversarial stimuli can significantly deteriorate the performance of MI tasks in EEG-based BCIs. We also observed that the impact of visual adversarial stimulus is more pronounced in the time-constrained settings, as compared to the warm-up task. This provides evidence that the proposed attack vector can be

TABLE III
ERROR RATE IN Warm-up VS. Full-game ENVIRONMENTS

| Subject | Error Rate (Env 1 to Env 2) | Error Rate (Env 3 to Env 4) |
|---------|----------------------------|----------------------------|
| S1      | 0.16                       | 0.46                       |
| S2      | 0.16                       | 0.62                       |
| S3      | 0.16                       | 0.28                       |
| S4      | 0.25                       | 0.47                       |
| S5      | 0.25                       | 0.47                       |
| S6      | 0.25                       | 0.56                       |
| S7      | 0.25                       | 0.44                       |
used by adversaries to manipulate the performance of EEG-based MI applications. Furthermore, the attack by the adversarial stimuli may affect the neural data generation stage that might affect the later stimuli generation [30], resulting in a prolonged effect on BCIs performance. The practical implications of such attacks are significant - malicious actors can use adversarial stimuli to target the integrity of systems such as MI-based wheelchairs and neuroprosthetics. Similarly, adversarial stimuli can be used for denial of service against EEG-based authentication systems, thereby blocking access to authorized users due to induced authorization failures. We also hypothesize that such attacks are not limited to MI tasks; for instance, changing the flickering frequency of the visual stimuli may affect the performance of SSVEP and VI-based BCIs. We therefore believe that the threat posed by adversarial stimuli warrants further studies on the underlying dynamics and mitigation of such attacks.

**Preliminary Analysis of Causes:** As an initial step towards investigating the dynamics of adversarial stimuli attacks, we further investigated the *warm-up* environment where the state of the paddle and the intent of the subjects were recorded at each timestep. We calculated the beta-to-alpha and beta-high to alpha ratios in the recorded EEG signals, and observed that both ratios increase from normal to adversarial settings. The increase in the beta/alpha ratio suggests high engagement in cognitive power. The alpha synchronization is considered an important component in the selective attention process where it inhibits the unattended positions during visual spatial orienting [31]. However, we observed that the overall alpha and beta power bands decrease from normal to adversarial settings. The flickering stimulus should have elicited its own frequency or frequencies in its multiples (e.g., 20, 40, 60 Hz) however, we did not observe the rise in power band in those frequencies in the adversarial settings. Additionally, we observe that the adversarial stimulus suppresses the amplitude of both the alpha and beta power. The suppression is significantly more in the alpha band than in any other band as depicted in Fig 3. Furthermore, we observed that in the presence of the adversarial stimulus, motor imagery signals based on mu rhythm (8-12 Hz) are also suppressed, which might be a contributing factor for the deterioration of the overall performance. This observation raises a further question on whether or not the MI signals can be fully disentangled from the SSVEP signal in EEG measurements.

Also, the participants in our experiments were focused on the given task and the perturbation was implemented in a smaller region of the screen only. Even though, the participants were asked only to focus on maximizing their score, we can observe that the score decreases significantly in the adversarial environment. One probable explanation for the performance deterioration could be the divided attention. However, our results demonstrate that the visual selective attention based on the beta-to-alpha ratio is higher in the adversarial settings. Hence we did not find sufficient support for this hypothesis. Another probable explanation could be the prediction error. It has been reported that increases in the alpha-to-beta and gamma-band activities are the reflections of predictive coding in visual processing [32], which increases attention to irrelevant cues [33]. However, in our case we only observed the decrease in the power bands in adversarial settings.

**VII. CONCLUSION AND OPEN QUESTIONS**

We introduced adversarial stimuli as an attack vector against MI-based BCIs. To demonstrate the feasibility of these attacks, we performed experiments on human subjects playing a video game via MI through a EEG device, and observed that minor and random flickers in the visual observations of the game results in significant deterioration of their performance in the game. We also presented preliminary analyses of the potential causes of this vulnerability.

We observe many similarities between the adversarial stimuli attacks in our experiments and adversarial example attacks against machine learning models. In deep reinforcement learning, it is shown that malicious actors can manipulate the action policies of agents via perturbing the agent’s observations via adversarial examples [34]. Adversarial stimuli attacks follow a similar process, in which perturbing the perception results in changes in the actions of the integrated BCIs system. This analogy gives rise to further questions: are there optimal or efficient adversarial stimuli, similar to those crafted by adversarial examples generation algorithms [35], that can effectively and persistently induce incorrect actions in BCIs? Moreover, it has been established that training machine learning models on adversarial examples improves their robustness to adversarial examples. We therefore ask whether similar adversarial training procedures in BCI systems and users could enhance robustness against adversarial stimuli? A further line of inquiry arises from the the observed suppression of the mu rhythm under adversarial stimuli: are MI and SSVEP signals separable, or are these sources fundamentally interrelated? And perhaps the most significant question in this direction is about the source of vulnerability to adversarial stimuli: is it the root cause of this vulnerability in the BCI device and software, or does it stem from inherent limitations of the human cognitive system?

While the emphasis of this study is on the security of BCI systems, the authors believe that further investigations of adversarial stimuli and the aforementioned questions may
open a new frontier in understanding the dynamics of decision making and cognition in humans, particularly in areas such as sensorimotor integration, and analysis of robustness in motor functions.

REFERENCES

[1] I. Iturrate, J. Antelis, and J. Minguéz, “Synchronous eeg brain-actuated wheelchair with automated navigation,” in 2009 IEEE International Conference on Robotics and Automation, 2009, pp. 2318–2325.

[2] D. J. McFarland and J. R. Wolpaw, “Brain-computer interface operation of robotic and prosthetic devices,” Computer, vol. 41, no. 10, pp. 52–56, 2008.

[3] N. Kosmyna, J. T. Lindgren, and A. Lécuyer, “Attending to visual stimuli versus performing visual imagery as a control strategy for eeg-based brain-computer interfaces,” Scientific reports, vol. 8, no. 1, pp. 1–14, 2018.

[4] J. I. Ekandem, T. A. Davis, I. Alvarez, M. T. James, and J. E. Gilbert, “Evaluating the ergonomics of bci devices for research and experimentation,” Ergonomics, vol. 55, no. 5, pp. 592–598, 2012.

[5] U. Chaudhary, I. Vlachos, J. B. Zimmermann, A. Espinosa, A. Tonin, A. Jaramillo-Gonzalez, M. Khalili-Ardali, H. Topka, J. Lehnhoff, G. M. Fries et al., “Spelling interface using intracortical signals in a completely locked-in patient enabled via auditory neurofeedback training,” Nature communications, vol. 13, no. 1, pp. 1–9, 2022.

[7] P. Nguyen, D. Tran, X. Huang, and W. Ma, “Motor imagery eeg-based person verification,” in International work-conference on artificial neural networks. Springer, 2013, pp. 430–438.

[10] O. Landau, R. Puzis, and N. Nissim, “Mind your mind: Eeg-based brain-computer interfaces and their security in cyber space,” ACM Computing Surveys (CSUR), vol. 53, no. 1, pp. 1–38, 2020.

[12] X. Zhang and D. Wu, “On the vulnerability of cnn classifiers in eeg-based bcis,” IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 27, no. 5, pp. 814–825, 2019.

[13] Z. Liu, L. Meng, X. Zhang, W. Fang, and D. Wu, “Universal adversarial perturbations for cnn classifiers in eeg-based bcis,” Journal of Neural Engineering, vol. 18, no. 4, p. 046044, 2021.

[14] M. A. Scholl, K. Stine, J. Hash, P. Bowen, L. A. Johnson, C. D. Smith, D. Steinberg et al., “An introductory resource guide for implementing the health insurance portability and accountability act (hipaa) security rule,” 2008.

[15] I. Martinovic, D. Davies, M. Frank, D. Perito, T. Ros, and D. Song, “On the feasibility of side-channel attacks with brain-computer interfaces,” in 21st {USENIX} Security Symposium ({USENIX} Security 12), 2012, pp. 143–158.

[16] Y. Xiao, Y. Jia, X. Cheng, J. Yu, Z. Liang, and Z. Tian, “I can see your brain: Investigating home-use electroencephalography system security,” IEEE Internet of Things Journal, vol. 6, no. 4, pp. 6681–6691, 2019.

[17] L. Meng, J. Huang, Z. Zeng, X. Jiang, S. Yu, T.-P. Jung, C.-T. Lin, R. Chavarriga, and D. Wu, “Eeg-based brain-computer interfaces are vulnerable to backdoor attacks,” arXiv preprint arXiv:2011.00101, 2020.