On the Feasibility of Side-Channel Attacks with Brain-Computer Interfaces

Ivan Martinovic*, Doug Davies†, Mario Frank†, Daniele Perito†, Tomas Ros‡, Dawn Song†
*University of Oxford** *UC Berkeley†* *University of Geneva‡*

Abstract

Brain computer interfaces (BCI) are becoming increasingly popular in the gaming and entertainment industries. Consumer-grade BCI devices are available for a few hundred dollars and are used in a variety of applications, such as video games, hands-free keyboards, or as an assistant in relaxation training. There are application stores similar to the ones used for smart phones, where application developers have access to an API to collect data from the BCI devices.

The security risks involved in using consumer-grade BCI devices have never been studied and the impact of malicious software with access to the device is unexplored. We take a first step in studying the security implications of such devices and demonstrate that this upcoming technology could be turned against users to reveal their private and secret information. We use inexpensive electroencephalography (EEG) based BCI devices to test the feasibility of simple, yet effective, attacks. The captured EEG signal could reveal the user's private information about, e.g., bank cards, PIN numbers, area of living, the knowledge of the known persons. This is the first attempt to study the security implications of consumer-grade BCI devices. We show that the entropy of the private information is decreased on the average by approximately 15 % - 40 % compared to random guessing attacks.

1 Motivation

Brain-Computer Interfaces (BCIs) enable a non-muscular communication between a user and an external device by measuring the brain's activities. In the last decades, BCIs have been primarily applied in the medical domain with the goal to increase the quality of life of patients with severe neuromuscular disorders. Most BCIs are based on electroencephalography (EEG) as it provides a non-invasive method for recording the electrical fields directly produced by neuronal synaptic activity. The EEG signal is recorded from scalp electrodes by a differential amplifier in order to increase the Signal-to-Noise Ratio of the electrical signal that is attenuated by the skull. This signal is continuously sampled (typically 128 Hz - 512 Hz) to provide a high temporal resolution, making EEG an ideal method for capturing the rapid, millisecond-scale dynamics of brain information processing with a simple setup.

Particular patterns of brain waves have been found to differentiate neurocognitive states and to offer a rich feature space for studying neurological processes of both disabled and healthy users. For example, EEG has not only been used for neurofeedback therapy in attention deficit hyperactivity disorder (ADHD) [20], epilepsy monitoring [6], and sleep disorders [28], but also to study underlying processes of skilled performance in sports and changes in vigilance [14, 31], in estimating alertness and drowsiness in drivers [22] and the mental workload of air-traffic control operators [39].

Besides medical applications, BCI devices are becoming increasingly popular in the entertainment and gaming industries. The ability to capture a user's cognitive activities enables the development of more adaptive games responsive to the user's affective states, such as satisfaction, boredom, frustration, confusion, and helps to improve the gaming experience [26]. A similar trend can be seen in popular gaming consoles such as Microsoft's Xbox 360, Nintendo's Wii, or Sony's Playstation3, which already include different sensors to infer user's behavioral and physiological states through pressure, heartbeat, facial and voice recognition, gaze-tracking, and motion.

In the last couple of years, several EEG-based gaming devices have made their way onto the market and became available to the general public. Companies such as Emotiv Systems [5] and NeuroSky [25] are offering low-cost EEG-based BCI devices (e.g., see Figure 1) and software development kits to support the expansion of

(a) An EPOC device (Emotiv Systems)

(b) A MindSet device (NeuroSky)

Figure 1: Popular consumer-grade BCI devices are available as multi-channel (EPOC) or single-channel (MindSet) wireless headsets using bluetooth transmitters.

tools and games available from their application markets. Currently, there are more than 100 available applications ranging from accessibility tools, such as a mind-controlled keyboard and mouse and hands-free arcade games, to so-called serious games, i.e., games with a purpose other than pure entertainment, such as attention and memory training games. For example, in [2], the authors used the Emotiv BCI device to implement a hands-free brain-to-mobile phone dialing application.

Marketing is another field that has shown increasing interest in commercial applications of BCI devices. In 2008, The Nielsen Company (a leading market research company) acquired NeuroFocus, a company specialized in neuroscience research, and it has recently developed an EEG-based BCI device called Mynd such that *"...market researchers will be able to capture the highest quality data on consumers' deep subconscious responses in real time wirelessly, revolutionizing mobile in-market research and media consumption at home."*[1]

In light of the progress of this technology, we believe that the trend in using EEG-based BCI devices for non-medical applications, in particular gaming, entertainment, and marketing, will continue. Given that this technology provides information on our cognitive processing and allows inferences to be made with regard to our intentions, conscious and unconscious interests, or emotional responses, we are concerned with its security and privacy aspects. More specifically, we are interested in understanding how easily this technology can be turned against its users to reveal their private information, that is, information they would not knowingly or willingly share. In particular, we investigate how third-party EEG applications could infer private information about the users, by manipulating the visual stimuli presented on screen and by analyzing the corresponding responses in the EEG signal.

[1] NeuroFocus Press Release (March 21, 2011): www.neurofocus.com/pdfs/Mynd_NeuroFocus.pdf

Figure 2: Example photo of a videogame controlled with the Emotiv Device.

1.1 Contributions

To justify how crucial the security and privacy concerns of this upcoming technology are, we provide some concrete answers in terms of demonstrating practical attacks using existing low-cost BCI devices. More specifically, the main contributions of this paper are:

- We explore, for the first time, EEG gaming devices as a potential attack vector to infer secret and private information about their users. This attack vector is entirely unexplored and qualitatively different from previously explored side-channels. This calls for research to analyze their potential to leak private information before these devices gain widespread adoption.

- We design and implement BCI experiments that show the possibility of attacks to reveal a user's private and secret information. The experiments are implemented and tested using a Emotiv EPOC BCI device. Since 2009, this consumer-grade device has been available on the market for the entertainment and gaming purposes.

- In a systematic user study, we analyze the feasibility of these attacks and show that they are able to reveal information about the user's *month of birth*, *area of living*, *knowledge of persons known to the user*, *PIN numbers*, *name of the user's bank*, and *the user's preferred bank card*.

2 A Brief Introduction to P300 Event-Related Potentials

In this section, we provide a brief introduction to the specifics of the EEG signal that are required to understand the rationale behind this work.

An important neurophysiological phenomenon used in studies of EEG signals is the Event-Related Potential (ERP). An ERP is detected as a pattern of voltage change after a certain auditory or visual stimulus is presented to a subject. Every ERP is time-locked to the stimulus, i.e., the time frame at which an EEG voltage change is expected to occur is known given the timing of the stimuli.

The most prominent ERP component which is sensitive to complex cognitive processing is the P300, so-called because it can be detected as an amplitude peak in the EEG signal at ≈ 300 ms after the stimulus (see Figure 3). The complexity of the stimulus and individual differences contribute to the variability of the amplitude and latency (e.g., the latency varies between 250 - 500 ms), yet the P300 is considered to be a fundamental physiological component and is reliably measured (for a recent overview of the P300 from a neuroscience perspective, please see, e.g., [27]). While there are two subcomponents of the P300, called P3a and P3b, both are related to complex cognitive processing, such as recognition and classification of external stimuli. In this paper, we take advantage of the subcomponent P3b of the P300, and for the sake of simplicity we will refer to it as the P300, which is also a convention in neuroscience.

The P300 is elicited when subjects discriminate between task-relevant and task-irrelevant stimuli using a so-called "oddball paradigm" (for more information, see, e.g., [16]). During an oddball task the number of task-relevant stimuli (called *target* stimuli) is less frequent than the number of task-irrelevant stimuli (called *non-target*). Probably the most well-known application of the P300 in an oddball task is the P300-Speller. In this application the alphanumeric characters are arranged in a matrix where rows and columns flash on the screen in a rapid succession. The target stimulus is the character that a subject desires to spell and the P300 is evoked each time the target letter is flashed due to a neuronal response triggered by increased attention of recognition. This application has been used to establish a communication channel for patients with locked-in syndrome or with severe neurodegenerative disorders.

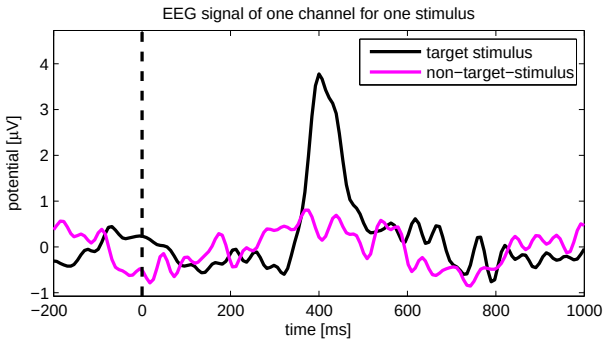


Figure 3: The P300 ERP elicited as a brain response to a target stimuli (in this experiment the non-target stimuli were pictures of unknown faces, while the target stimuli was the picture showing President Obama).

The P300 is seen in response to target stimuli defined by the task, but it has also been observed to be elicited during stimuli that are personally meaningful to participants. For example, if a random sequence of personal names is presented to a subject, the P300 will be the largest during the presentation of the subject's own name [32]. Likewise, it has been shown that the P300 discriminates familiar from unfamiliar faces within randomly presented sequences [24].

3 BCI Attacks: Threat Model and Assumptions

In this section, we explore a number of possible scenarios in which consumer EEG devices could be abused to capture sensitive or private information from users. Currently, both Emotiv and NeuroSky have "App Stores" where the users can download a wide variety of applications. Similarly to application stores for smart phones, the applications are developed by third parties that rely on a common API to access the devices. In the case of the EEG devices, this API provides unrestricted access to the raw EEG signal. Furthermore, such applications have complete control over the stimuli that can be presented to the users.

In this scenario, the attacker is a malicious third-party developer of applications that are using EEG-based BCI devices. Its goal is to learn as much information as possible about the user. Hence, we are neither assuming any malware running on the machine of the victim nor a tampered device, just "brain spyware", i.e., a software intentionally designed to detect private information. Our attacker model cannot access more computer resources than any third party application for the respective BCI device. The attacker can read the EEG signal from the

device and can display text, videos, and images on the screen. Therefore, the attacker can specifically design the videos and images shown to the user to maximize the amount of information leaked while trying to conceal the attacks.

The type of information that could be discovered by such an attack is only bound by the quality of the signal coming from the EEG device and the techniques used to extract the signal. We note that all involved parties (users of BCI devices, their developers, and also attackers) share the same objective: to maximize the signal quality in order to best perform their task. Hence, it is expected that the signal and the measurement processes will improve and, as a result, facilitate the attacks.

In this work we will focus on categorization tasks, in which the mind of the user is probed to detect whether certain stimuli (faces, banks, locations) are familiar to or relevant for the user. However, we note that in the future such attack could be extended to include other sensitive information. For instance, EEG devices have been used, under optimized lab conditions, to study prejudices, sexual orientation, religious beliefs [18], and deviant sexual interests [38, 10].

At the moment, low-cost devices are still very noisy and need a calibration phase to work properly (three minutes in our experiments). However, we note that the attacker could find a natural situation in which to expose the user to target stimuli to extract information and thus gather enough data to succeed in an unnoticed way. Also, such a calibration phase can be concealed in the normal training phase that EEG applications require for proper functioning and that the user is willing to support. Moreover, we expect that BCI devices will become increasingly robust and accurate in the future, resolving many current technical problems.

The experiments presented in this study are meant to show feasibility in favorable conditions. The subjects were partially cooperating in an attack situation and were following our instructions. However, we minimized the interaction between the supervisor and subjects to simulate a realistic environment, where a user is only interacting with his computer (see Appendix A).

4 Experimental Design and Results

The main question, which this paper attempts to answer is: *Can the signal captured by a consumer-grade EEG device be used to extract potentially sensitive information from the users?* In the following, we detail the technical setup, the experimental design, and the analytical methods of our experiments.

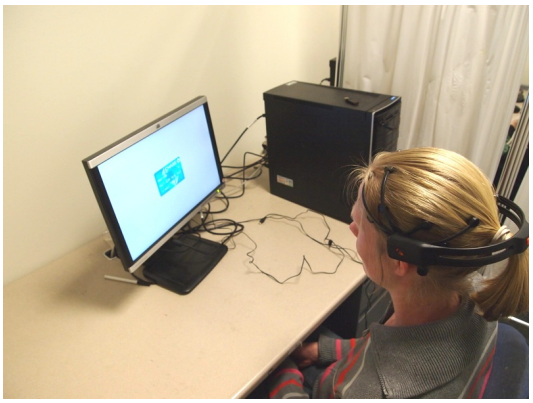


Figure 4: Experimental setup. The instructor sits behind the curtain to minimize interaction during the experiments. In this case, a sequence of credit cards is presented to the user.

4.1 The Setup

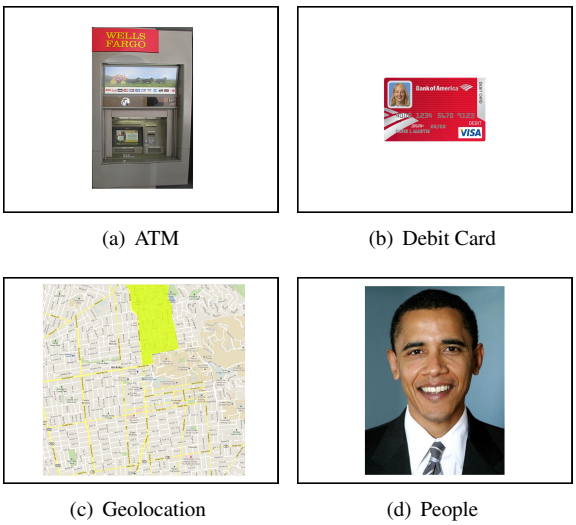
After obtaining the approval of the Institutional Review Board (IRB), we recruited 30 Computer Science students for the experiments. For two participants, the experiments could not be conducted due to faulty equipment (low battery on the EEG device). Of the 28 participants remaining, 18 were male and 10 female. In total, the experiment lasted about 40 minutes. The participants were informed that they were going to participate in an experiment involving the privacy implications of using gaming EEG devices, but we explained neither the details of the experiment nor our objectives. Each participant was seated in front of the computer used for the experiments (see Figure 4). The operator then proceeded to mount the Emotiv EEG device on the participants.

4.2 The Protocol

After the initial setup, the participants were asked to try to remain relaxed for the entire duration of the experiments, as blinking or other face movements cause significant noise. The exact script used during the experiments can be found in Appendix A. The interaction with the participants was kept as short and concise as possible. The order of the experiments was kept fixed in the order found in Appendix A.

Each experiment consisted of three main steps:

1. (Optional) Brief verbal explanation of the task by the operator;
2. (Optional) Message on screen for 2 seconds;
3. Images being flashed in random order for the duration of the experiment.



(a) ATM (b) Debit Card

(c) Geolocation (d) People

Figure 5: Layout of four of the experiments: Bank ATMs, Debit Cards, Geolocation and People. Each frame shows how the stimuli were flashed on the screen.

Each image was shown to the users for a fixed duration of 250ms. On the screen in Figure 4, a photo is being shown to a test participant.

The time of the target and non-target stimuli and the stimulus identifiers were recorded alongside the raw signal coming from the EEG device. After the experiment, we used the classification techniques detailed in Section 4.4 to infer information about the secrets of the participant.

4.3 The Experimental Scenarios

In this section, we describe the calibration of the device and six different experiments. In each experiment, the attacker tries to gain information about a different secret. Each experiment lasted approximately 90 seconds.

4.3.1 Training Phase

This experiment was set up to learn a model to detect the P300 signal from each user. The users were presented with a randomly permuted sequence of numbers from 0 to 9 and were asked by the operator to count the number of occurrences of a target number x. Each number was shown 16 times, with a stimulus duration of 250 ms and a pause between stimuli randomly chosen between 250 ms and 375 ms. At the end of experiment the participants were asked for their count to check for correctness.

We also developed a method to calibrate the classifier without this active training phase. This could be used for a concealed attack in cases where the intended application of the user does not require the detection of P300. We explain this on-the-fly calibration phase in Section 5.

4.3.2 Experiment 1: Pin Code

This experiment has the goal to gather partial information about a user's chosen 4-digit PIN. Given the sensitivity in studying the users' real PINs, we asked the participants to choose and memorize a randomly generated PIN just for the experiment. Furthermore, the participants were asked not to reveal the PIN until after the end of the experiment session. The participants were told that there were no special instructions for the experiment, e.g., no counting numbers. They were just informed that, at the end of the experiment, they would be asked to enter the first digit of their PIN (refer to Appendix A for the exact script). In this way, we bring the information of interest to the attention of the user which makes the subject focus on the desired stimulus without requiring their active support of the classifier. After the instructions were given, the operator started the experiment. There was *no* on-screen message shown at the beginning of the experiment. The experiment images consisted of a sequence of randomly permuted numbers between 0 and 9 that were shown on the screen one by one. Each number was shown 16 times and the experiment lasted approximately 90 seconds.

4.3.3 Experiment 2: Bank Information

The aim of this experiment was to obtain the name of the bank of the participant by reading their response to visual stimuli that involved photos related to banks. The first iteration of this experiment, whose results are not reported, consisted of showing the logo of 10 different banks[2]. The intuition was that the participants would show a higher response when seeing the logo of their bank. However, this attack was unsuccessful. After de-briefing with the early test participants, we realized that they simply recognized the logos of all the banks.

In the second and final iteration of the experiment, we showed two different sets of images: automatic teller machines (ATMs) and credit cards. Rationale for choosing to display ATM or credit card photos, rather than logo images, is that while users might be familiar with all logos, they might be only familiar with the look of their own local bank ATM and debit card. The results are reported in Section 5.

The protocol for this experiment was as follows. Each participant was asked by the operator whether they were a customer of one of the banks in a list. Four participants answered negatively, therefore the experiment was skipped. In case of an affirmative answer, the experiment was started. The screen in front of the participants showed the question "What is the name of your bank?"

[2]List of banks: Bank of America, Chase, Wells Fargo, ING, Barclays, Citi Bank, Postbank, Unicredit, Deutsche Bank

Figure 6: Stimuli for the debit card experiment. Each card was shown separately, full-screen, for the short stimulus duration.

for 2 seconds. Then, for the ATM experiment, images of teller machines were flashed on the screen. For the credit card experiment, images of credit cards were flashed.

4.3.4 Experiment 3: Month of Birth

The operator did not give any specific instructions to the participants and only informed them that the instructions would be provided on the screen. The participants were simply asked in which month they were born by an on-screen message that lasted for 2 seconds, then, a randomly permuted sequence of the names of the months was shown on the screen.

In many access-restricted websites the date of birth or similar information serves as a backup function for resetting a user's password. If an attacker needs this information, the BCI device could provide a potential attack vector.

4.3.5 Experiment 4: Face Recognition

For this experiment, the operator again did not give any specific instruction to the participants and only informed them that the instructions would be provided on the screen. The participants were simply asked "Do you know any of these people?" by an on-screen message that lasted 2 seconds. Then the images of people were randomly flashed for the duration of the experiment.

The goal of this experiment was to understand whether we could infer who the participants knew by reading their EEG response when being showed a sequence of photos of known and unknown people. We used photos of 10 unknown persons and one photo of the current President of the United States of America, Barack Obama. The photo of the president was chosen because, being in a US institution, we were confident that each participant would recognize the President.

One interesting application of such an attack would be scenarios in which the knowledge of particular individual is used as a form of authentication. For example, in recent years, Facebook has started showing photos of friends for the purpose of account verification [3].

4.3.6 Experiment 5: Geographic Location

The purpose of this experiment was to accurately pinpoint the geographic location of the residence of the participants. Each participant was asked if they lived in an area close to campus. Eight participants in total did not live close to campus and did not complete this experiment. In case of an affirmative answer, the participants were shown a sequence of highlighted maps of an area of approximately 4 square kilometers around campus. Each image showed the same area overall, but with a different highlighted zone on the map.

While IP addresses provide a rather accurate way to localize the location of a user, there are cases in which the users actively try to hide their geographic location using proxies. Even though our experiment showed only a predefined map of a rather small geographic area, we envision possible future attacks in which the true geographic location of a user is leaked by showing maps or landmarks with increased accuracy.

While for all the other experiments we did not instruct the user to do particular things except for watching the screen, here we asked the users to count how often their region was highlighted. This experiment was devised to study the influence of active user support, as counting assures a higher attention from the user which is known to improve the detection of P300.

4.4 Analysis Methodology

In this section, we detail how the attacker processes and analyzes the data and provide the specification of the data recorded by the BCI device.

Data characteristics and acquisition The data consists of several parts. The amplitudes of the EEG signal are recorded with 14 different electrodes. Each electrode represents one 'channel' of the signal. According to the standard 10-20 system [19], the 14 channels are called 1: 'AF3', 2: 'F7', 3: 'F3', 4: 'FC5', 5: 'T7', 6: 'P7', 7: 'O1', 8: 'O2', 9: 'P8', 10: 'T8', 11: 'FC6', 12: 'F4', 13: 'F8', and 14: 'AF4'. The location of the channel electrodes can be seen in Figure 7.

Each channel is recorded at a sampling rate of 128Hz. The software for showing stimuli to the user outputs the time stamp for each stimulus and the indicator of the

[3] http://www.facebook.com/help/search/?q=security+verification

stimulus. In this way, the EEG signal can be related to the stimuli.

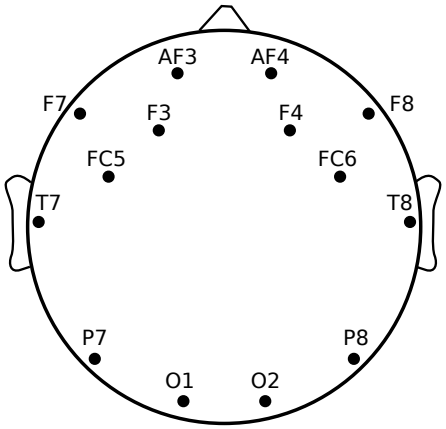


Figure 7: Position of the electrodes of the EPOC device.

As explained in Section 3, our attack vector exploits the occurrence of P300 peaks in the EEG signal triggered by target-stimuli. This requires the attacker to be able to reliably detect these peaks and to discriminate them from all other EEG signals measured on non-target stimuli. This task is very similar to the P300-Speller, where the EEG signal for the intended letter must be discriminated from the signal of unintended letters (as described in Section 2). However, in contrast to the spelling scenario the attacker is dealing with a passive user. This makes an attack much harder than spelling. In our case, the user does not *intend* to provide a discriminative signal for the target stimulus. This means that the user does not support the classifier with increased attention on the target stimulus, as can be achieved, for instance, by counting the number of occurrences of this stimulus. As a consequence, the data available to the attacker is less discriminative between target and non-target stimuli than in the spelling scenario.

An additional challenge for our attack is that the gaming device we are using is not made for detecting P300. For instance, they have more electrodes on the frontal part of the scalp (see Figure 7). This enables them to recognize facial expressions which provide a stronger signal than the EEG signal itself and thus are more robust for controlling games. The P300 is mostly detected at the parietal lobe, optimally with electrodes attached at Pz position, which is a centered on the median line at the top of the head. As we want to investigate the attack in a realistic home-use scenario we did not use other devices optimized for P300 detection and did not adapt the gaming device (for instance by turning it around, which would provide more sampling points in the Pz area).

Classification of target stimuli Detecting P300 in EEG data is a binary classification task. The input is a set of **epochs**. Each epoch is associated with a stimulus. In our setting a stimulus is an image depicted on a computer screen in front of the user. Let n_c be the number of EEG channels and let f be the sampling rate of the device (in our case the signal is sampled with 128 Hz). An epoch consists of n_c time series starting t_p milliseconds prior to the stimulus and ending t_a milliseconds after the stimulus. The number of measurements per time series is then $q = (t_p + t_a)f$. Typically, t_p is a few hundred milliseconds and t_a is between 800 ms and 1500 ms. The signals of all channels are concatenated and each epoch is represented as a real vector $\mathbf{x} \in \mathbb{R}^p$, where $p = qn_c$ is the dimensionality of the vector space.

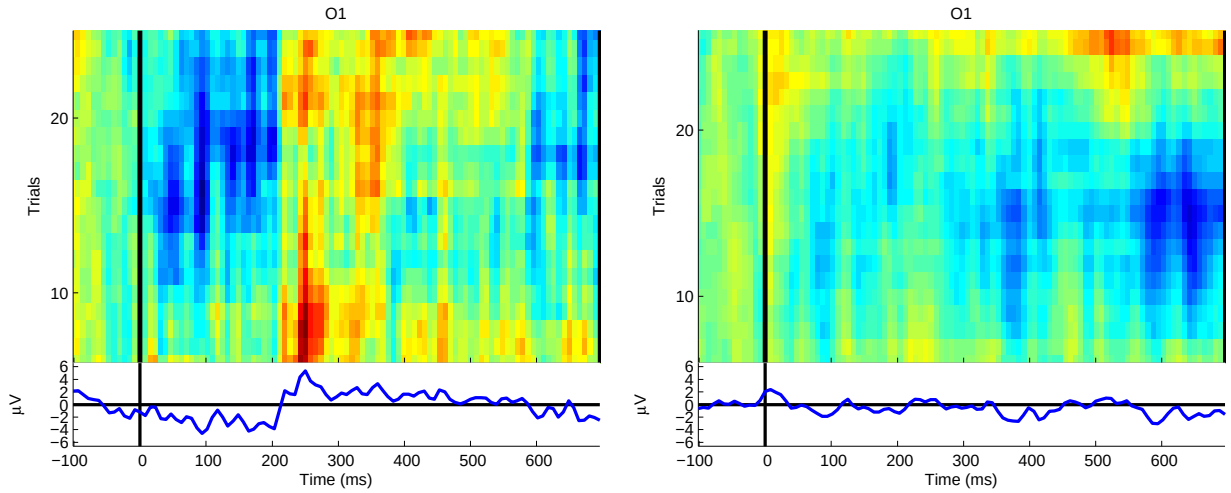
The classification task consists of two phases, the training phase and the classification phase. The input of the training phase is a set of epochs $\mathbf{X}^{\text{tr}} = \{\mathbf{x}_i^{\text{tr}} \in \mathbb{R}^p, i = 1...n_1\}$ and a vector of labels $\mathbf{y} \in \{0,1\}^{n_1}$, where each label y_i indicates whether the epoch $\mathbf{x}_i^{\text{tr}}$ corresponds to a target stimulus ($y_i = 1$) or not ($y_i = 0$). The signal of each epoch has been recorded while the corresponding stimulus was shown to the user on the screen for a short time (we used 500 ms). The stimuli labels $\mathbf{y}$ are known to the classifier as the system knows what it shows to the user. Given this input, the classifier must learn a function g that maps epochs to target stimuli labels:

$$\begin{array}{rcl} g : \mathbb{R}^p & \rightarrow & \{0,1\} \\ \mathbf{x} & \mapsto & y \end{array} \quad (1)$$

In the beginning of Section 5, we explain how to practically carry out the training phase with users that actively support this training phase and with passive users.

In the classification phase the classifier gets a collection of n_2 new epochs $\mathbf{X}^{\text{test}} = \{\mathbf{x}_i^{\text{test}} \in \mathbb{R}^p, i = 1, ..., n_2\}$ as an input and must output an estimate $\hat{\mathbf{y}} = \{\hat{y}_i = g(\mathbf{x}_i^{\text{test}}), i = 1, ..., n_2\}$ of the corresponding labels. This means, for each of the new epochs, the classifier must decide whether the epoch is associated with the target stimulus or not.

The test labels $\hat{\mathbf{y}}$ provide a ranking of the K unique stimuli presented to the user. We sort all stimuli in descending order according to the number of their positive classifications. For stimulus k this number is $N_k^{(+)} = \sum_{i \in E_k} \hat{y}_i$. The set E_k is the set of epoch indices containing all epochs that are associated with stimulus k. In this notation $i \in E_k$ means that we sum over all epochs of stimulus k. For instance, if there are three different stimuli repeatedly shown to the user in random order (three different faces, say), then the classifier would guess that the true face (the one familiar to the user) is the face where the most associated epochs have been classified as



(a) target stimulus

(b) non-target stimulus

Figure 8: Event-related potentials for two different stimuli. Both signals have been recorded on the left back-side of the scalp (Channel 7: 'O1'). The plots have been produced with EEGlab [4]. The scale of the averaged plots (bottom) as well as the colorscale of the heatmap plots (top) are constant over the two stimuli.

target-stimulus. Figure 8 depicts event-related potentials (ERP) for one channel and two different stimuli (target and non-target). In this example one row of one plot represents an epoch and all rows of one plot constitute the set E_k of epochs associated with event k.

The stimulus with the topmost positive classifications is the estimated target-stimulus, the stimulus with the second most positively classified epochs is ranked second, and so on. Most classifiers output a continuous score s_i for each epoch instead of binary labels $\hat{y}_i$. For instance, this could be a probability $s_i = p(y_i = 1)$. In such a case, we sum over all scores of each unique stimulus k to get its vote $N_k^{(+)} = \sum_{i \in E_k} s_i$. In the experiments, we will use this ranking to decide which of the presented stimuli is the target stimulus, that is which of the answers is the true answer for the current user.

In the following we explain two different classifiers that we used in our experiments. The first classifier is a boosting algorithm for logistic regression (bLogReg) and was proposed for P300 spelling in [17]. The second classifier is the publicly available BCI2000 P300 classifier. BCI2000 uses stepwise linear discriminant analysis (SWLDA). In [21] a set of different P300 classifiers, including linear and non-linear support vector machines, was compared and SWLDA performed best.

4.4.1 Boosted logistic regression

This method uses a logistic regression model as the classifier function g. The model is trained on the training data by minimizing the negative Bernoulli log-likelihood of the model in a stepwise fashion as proposed in [11, 12].

As follows, we briefly describe a variant, proposed in [17], where the method has been used to design a P300 speller. The classifier consists of an ensemble of M weak learners. Each weak learner f_m is a regression function minimizing a quadratic cost function:

$$f_m = \operatorname*{argmin}_f \sum_{i=1}^{n_1} \left(\tilde{y}_i - f(\mathbf{x}_i^{\text{tr}}; \mathbf{w})\right)^2 , \qquad (2)$$

where $f(\mathbf{x}_i^{\text{tr}}; \mathbf{w}) = \mathbf{w}^T \mathbf{x}_i^{\text{tr}}$ with coefficients $\mathbf{w} \in \mathbb{R}^p$. The score $\tilde{y}_i$ in Equation (2) is obtained from the first-order condition of maximizing the logarithm of the Bernoulli likelihood

$$L(g_m; \mathbf{X}^{\text{tr}}, \mathbf{y}) = \prod_{i=1}^{n_1} p(y_i = 1 | \mathbf{x}_i^{\text{tr}})^{y_i} (1 - p(y_i = 1 | \mathbf{x}_i^{\text{tr}}))^{1-y_i} \qquad (3)$$

with

$$p(y_i = 1 | \mathbf{x}_i^{\text{tr}}) = \frac{\exp(g_m(\mathbf{x}_i^{\text{tr}}))}{\exp(g_m(\mathbf{x}_i^{\text{tr}})) + \exp(-g_m(\mathbf{x}_i^{\text{tr}}))} \qquad (4)$$

In step m of the algorithm, the current classifier g_{m-1} is updated by adding the new weak classifier f_m: $g_m = g_{m-1} + \gamma_m f_m$. Thereby, the weight γ_m is selected such that the likelihood Eq. (3) is maximized.

The number of weak classifiers M controls the trade-off between overfitting and underfitting. This number is determined by cross-validation on random subsets of the training data $\mathbf{X}^{\text{tr}}$.

Data preprocessing Before training the classifier and prior applying it to each new observation, we process the data in the following way. The input data consists of n_c different time series, whereas n_c is the number of channels. First we epochize the signal with a time frame around the stimuli that starts 200 ms before the respective stimulus and ends 1000 ms after the stimulus. Then, for each epoch, we subtract the mean amplitude of the first 200 ms from the entire epoch as it represents the baseline.

In order to reduce the high-frequency noise, we apply a low-pass FIR filter with a pass band between 0.35 and 0.4 in normalized frequency units. An example of such a preprocessed signal is depicted in Figure 3.

4.4.2 Stepwise Linear Discriminant Analysis

The BCI2000 P300 classifier uses stepwise linear discriminant analysis, an extension of Fisher's linear discriminant analysis. As follows, we briefly explain these two methods.

Fisher's linear discriminant analysis (LDA) LDA was first proposed in [9]. This classifier is a linear hyperplane that separates the observations of the two classes. The hyperplane is parameterized by the coefficient vector $\mathbf{w} \in \mathbb{R}^p$ which is orthogonal to the hyperplane. A new observation $\mathbf{x}_i$ is labeled to belong to either of the two classes by projecting it on the class separation $\mathbf{w}^T\mathbf{x}_i$. LDA assumes observations in both classes to be Gaussian distributed with parameters $(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), j = 1, 2$ and computes the optimally separating coefficients by $\mathbf{w} = (\boldsymbol{\mu}_1^{\text{tr}} - \boldsymbol{\mu}_2^{\text{tr}})(\boldsymbol{\Sigma}_1^{\text{tr}} + \boldsymbol{\Sigma}_2^{\text{tr}})^{-1}$.

Stepwise Linear Discriminant Analysis (SWLDA) SWLDA extends LDA with a feature selection mechanism that sets many of the coefficients in $\mathbf{w}$ to zero. This classifier is supposedly more robust to noise and was first applied to P300 spelling in [7]. The algorithm iteratively adds or removes components of the coefficient vector according to their statistical significance for the label outcome as measured by their p-value. The thresholds $(p_{\text{add}}, p_{\text{rem}})$ for adding or removing features as well as the total number of features must be pre-defined.

In our experiments we used the default configuration of the the BCI2000 P300 classifier with 60 features and $(p_{\text{add}}, p_{\text{rem}}) = (0.1, 0.15)$. The algorithm uses the 800 ms period after the stimulus for classification.

For each stimulus presented, we sum up the scores $\mathbf{w}^T\mathbf{x}_i$ of the corresponding epochs in order to obtain a ranking of the stimuli. Then, the highest ranked stimulus is presumably the target-stimulus.

5 Results

In this section, we evaluate the classification results on each of the experiments described in Section 4.3.

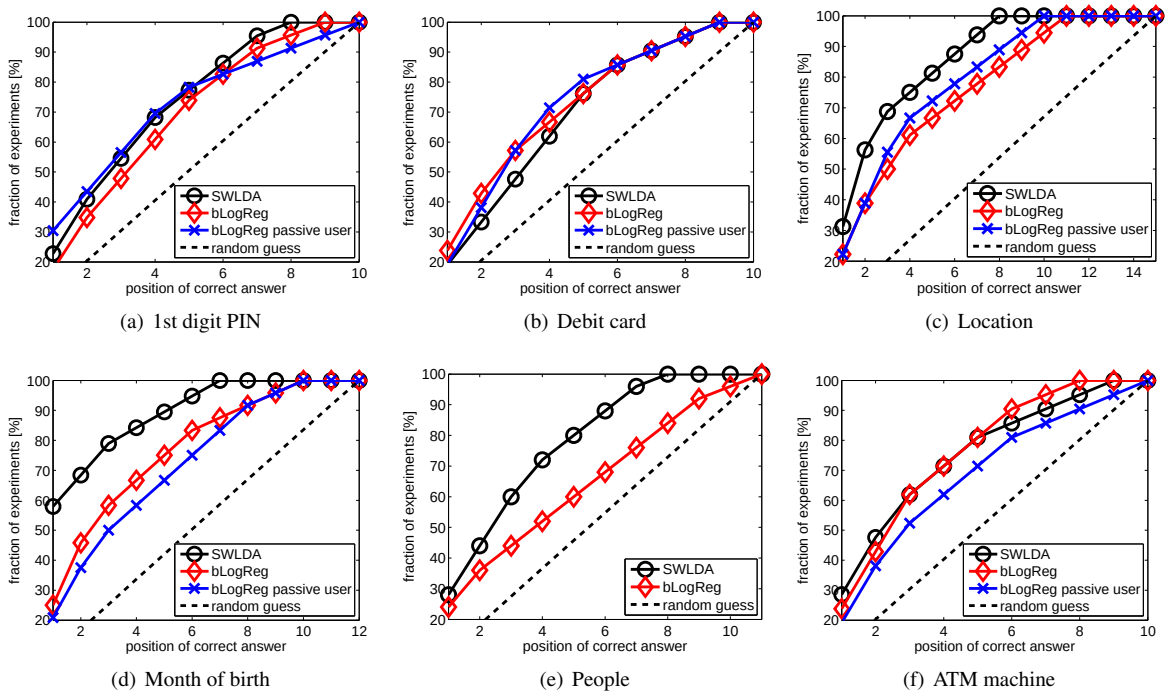
User-supported calibration and on-the fly calibration We calibrate the classifiers on a set of training observations. Thereby, we distinguish two training situations.

In the first situation we have a partially cooperating user, that is, a user who actively supports the training phase of the BCI but then does not actively provide evidence for the target stimulus later. This is a realistic scenario. Each gamer has a strong incentive to support the initial calibration phase of his device, because he will benefit from a high usability and a resulting satisfying gaming experience. The attacker can use the training data to train his own classifier. Despite the user supporting the calibration phase, we do not assume that the user actively supports the detection of target stimuli when the attacker later carries out his attack by suddenly presenting new stimuli on the screen.

In the second training situation, the user is passive. This means that the user does not support the training phase but also does not actively try to disturb it. As a consequence, the attacker must present a set of stimuli where, with high probability, the user is familiar with one of the stimuli and unfamiliar with all other stimuli. In this way the attacker can provide a label vector $\mathbf{y} \in \{0, 1\}^{n_1}$ that can be used for training. We used the people experiment as training data. We showed 10 images of random people to the user as well as one image of President Barack Obama. Assuming that i) every user knows Obama and that ii) it is unlikely that a user knows one of the random face images downloaded from the internet, we can use the Obama image as a target stimulus and the others as non-target stimuli.

Success statistics We report the results of all experiments in Figure 9. Each plot corresponds to one experimental scenario. The black crosses depict the results of the SWLDA classifier used by the BCI2000 P300 speller. The red diamonds are the results of boosted logarithmic regression (bLogReg) trained by the counting experiment, and the blue crosses show the results for bLogReg when trained on the people experiment. The dashed black line depicts the expected result of a random guess.

We depict the results in terms of a cumulative statistic of the rank of the correct answer. This measure provides the accuracy together with a confidence interval at the same time as it includes the probability distribution of the deviation from the optimal rank. The plots read as follows. The x-axis of each plot is the rank of the correct



(a) 1st digit PIN (b) Debit card (c) Location

(d) Month of birth (e) People (f) ATM machine

Figure 9: Cumulative statistics of the ranking of the correct answer according to the classification result. The faster this measure converges towards 100%, the better the classifier. One can directly read the confidence intervals as follows: In more than 20% of the experiments the bLogReg classifier ranked the correct face at the first position. In more than 40% it had the correct face among the first three guesses. Please note that for the passive user, the classifier was trained on the people experiment and the corresponding curve in Fig. 9(e) would depict the training error.

answer as estimated by the respective classifier. For instance, if the correct answer in the month of birth experiment is 'April' and the classifier ranks this month at the third position in the classification output, then x is 3. The y-axis is the fraction (in %) of the users having the correct answer in **at most** ranking position x. In our example with the month of birth, the point $(x; y) = (3; 80\%)$ of the SWLDA classifier means that for 80% of the users the correct bank was among the first three guesses of SWLDA. Please note that we truncated the y-axis at 20% to get a better resolution of the dynamic range.

Overall, one can observe that the attack does not always reveal the correct information on the first guess. However, the classifiers perform significantly better than the random attack. The SWLDA classifier provided the most accurate estimates, except for the experiment on the PIN and the debit card.

The correct answer was found by the first guess in 20% of the cases for the experiment with the PIN, the debit cards, people, and the ATM machine. The location was exactly guessed for 30% of users, month of birth for almost 60% and the bank based on the ATM machines for almost 30%. All classifiers performed consistently good on the location experiment where the users actively concentrated by counting the occurrence of the correct answer. SWLDA performed exceptionally good on the month of birth experiment, even though this experiment was carried out without counting.

Relative reduction of entropy In order to quantify the information leak that the BCI attack provides, we compare the Shannon entropies of guessing the correct answer for the classifiers against the entropy of the random guess attack.

This measure models the guessing attack as a random experiment with the random variable X. Depending of the displayed stimuli, X can take different values. For instance, in the PIN experiment, the set of hypotheses consists of the numbers 0 to 9 and the attack guess would then take one out of these numbers. Now, let's assume we have no other information than the set of hypotheses. Then we would guess each answer with equal probability. This is the random attack. Let the number of possible answers (the cardinality of the set of hypotheses) be K, then the entropy of the random attack is $\log_2(K)$.

More formally, let the ranking of a classifier *clf* be $\mathbf{a}^{(\text{clf})} = \left\{a_1^{(\text{clf})}, ..., a_K^{(\text{clf})}\right\}$, where the first-ranked answer is $a_1^{(\text{clf})}$, the second-ranked answer is $a_2^{(\text{clf})}$, and so on. Let

$p(a_k^{(\text{clf})}) := p(X = a_k^{(\text{clf})}|\mathbf{a}^{(\text{clf})})$ be the probability that the classifier ranks the correct answer at position $k \in K$. Please note that the $p(X = a_k^{(\text{clf})})$ that we will use are empirical relative frequencies obtained from the experiments instead of true probability distributions. Using these probabilities, the empirical Shannon entropy is

$$H(X|\mathbf{a}^{(\text{clf})}) = -\sum_{k=1}^{K} p(a_k^{(\text{clf})}) \log_2 \left(p(a_k^{(\text{clf})})\right) \quad (5)$$

In case of the random attack, the position of the correct answer is uniformly distributed, which results in the said entropy $H(X|\mathbf{a}^{(\text{rand})}) = \log_2(K)$. In case of attacking with a classifier, the attacker would pick a_1, the answer ranked highest, to maximize his success. As our empirical results, depicted in Figure 9, suggest, the rankings are not fully reliable, i.e. the answer ranked highest is not always the correct answer. However, the ranking statistics provide a new non-uniform distribution over the set of possible answers. For instance, we know that for bLogReg the empirical probability that the first-ranked location is the correct one is $p(X = a_1^{(\text{bLogReg})}) = 0.2$, the probability of the second-ranked answer to be correct is also $p(X = a_2^{(\text{bLogReg})}) = 0.2$, and so on.

The redistributed success probabilities reduce the entropy of the guessing experiment. We take the random guess attack as the baseline and compare the entropies of all other attacks against its entropy $H(X|\mathbf{a}^{(\text{rand})})$. We evaluate to what extent a generic classifier *clf* reduces the entropy relative to $H(X|\mathbf{a}^{(\text{rand})})$. The relative reduction of entropy with respect to the random guess attack (in %) is then:

$$\begin{aligned} r(\text{clf}) &:= 100\frac{H(X|\mathbf{a}^{(\text{rand})}) - H(X|\mathbf{a}^{(\text{clf})})}{H(X|\mathbf{a}^{(\text{rand})})} \\ &= 100\left(1 - \frac{H(X|\mathbf{a}^{(\text{clf})})}{\log_2(K)}\right) \end{aligned} \quad (6)$$

A perfect classifier always has the correct answer at the first position, resulting in zero entropy and a relative reduction r of 100%. A poor classifier provides a uniform distribution of the position of the correct rank. As a consequence, its entropy would be maximal and the relative reduction r would be 0%. The entropy difference directly measures the information leaked by an attack. Thereby, comparing the classifier entropies in a relative way enables one to compare results over different experiments with different numbers of possible answers.

We report the relative reduction of entropy for each experimental setting and for each classifier in Figure 10. As one can see, the reduction approximately ranges from 15% to 40% for SWLDA and from 7% to 18% for the two bLogReg variants. Please note that the plot does not report the result of the classifier that has been trained on the people experiment for this very experiment, as this entropy reduction merely refers to the training error of the classifier and provides no information on how well the classifier generalizes to unseen data.

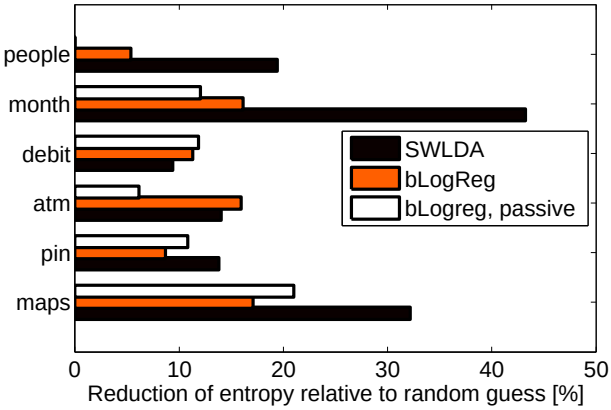


Figure 10: Relative reduction of entropy with respect to the random guess attack. The scale reaches from 0% (no advantage over random guessing) to 100% (correct answer always found). Please note that 'bLogReg, passive' has been trained on the people experiment. We do not report its score on this experiment, as it refers to the training error.

For most scenarios, the information leaked corresponds to approximately 10% to 20% for the best classifier SWLDA with peaks for *maps* (32%) and *month* (43%). The average information leak over all classifiers in the *maps* experiment stands out compared to the other result. The reason for this is that the maps experiment is a counting experiment, in which the users were asked to count the number of occurrences of the target stimulus. This experiment was included to underline the improvement in accuracy with a cooperative user.

Using Prior Knowledge to Improve Accuracy For some secrets there exist global statistics that can improve the success chances of the attack. For instance, often the distribution of customers of different banks in a population is approximately known. Also there might be prior knowledge about the area someone lives in. We did not include such prior knowledge in our experiments. However, such information could improve both the random guess as well as the classifier guesses. Prior probabilities could be included to Bayesian classifiers or could be used for heuristically post-processing classifier output.

For some experiments such as the PINs and the month of birth, the possible answers are approximately uniformly distributed, such that prior knowledge would provide no information. For other experiments prior knowledge might simply be unavailable and thus can not be used for more sophisticated models.

6 Related work

In this section, we overview related papers that use EEG signals in security-relevant applications.

EEG-based identification and authentication EEG signal has successfully been used for user identification (selecting the user identity out of a set of identities) and user authentication (verifying if a claimed user identity is true). In [30], the authors provide an overview of *cognitive biometrics*, an emerging research area that investigates how different biosignals can be used for the purpose of authentication and identification. The authors cover recent papers on biometrics based on EEG, the electrocardiogram (ECG), and the skin conductance, also called electrodermal response (EDR). An **identification** mechanism based on the alpha rhythm has been proposed in [29]. The mechanism uses convex polygon intersections to map new observations to a user identity. The authors report a high true positive rate of 95 % and a true negative rate of 87 % for experiments on 79 users. In method proposed in [23] uses Gaussian mixture models for user **authentication**. The authors test their method with different authentication protocols and report that with increasing temporal distance from the sign-up phase, the accuracy degrades. Using a sign-up phase over several days improves the accuracy. In [36] the authors describe *pass-thoughts*, another **authentication** mechanism that instead of typing a password requires the user to think of a password. The idea is very similar to the conventional P300-Speller scenario we mentioned in Section 2. A matrix containing characters is shown to a user and he focuses on the characters required to spell the password. This way, many shoulder-surfing attacks could be avoided. The main drawback of this authentication method (also mentioned by the authors) is a very low throughput rate of the spelling, which is ≈ 5 characters per minute for the 90% accuracy. Another problem is that the user gets no feedback until the complete passphrase is spelled, and hence the whole process must be repeated if a single character is wrongly classified.

More recently, in [15], the authors introduce a **key-generation technique** resistant against coercion attacks. The idea is to incorporate the user's emotional status through skin conductance measurements into the cryptographic key generation. This way, the generated keys contain a dynamic component that can detect whether a user is forced to grant an access to the system. Skin conductance is used as an indicator of the person's overall arousal state, i.e., the skin conductance of the victim in a stressful scenario significantly changes compared to a situation when the keys were generated.

Another highly related work to ours is described in [37]. The authors exploit an ERP called N400 to detect if a person is actively thinking about a certain stimuli without explicitly looking at it. In contrast to the P300 which is related to attention, the N400 has been associated with semantic processing of words. For example, in an experiment where subjects are shown incongruent sentences like "*I drink coffee with milk and socks*", the amplitude of the N400 would be maximal at the last (incorrect) word. This phenomenon is then used to detect which out of several possible objects the user is actively thinking of. While this paper is not focusing on security issues but rather on **assisting a user in efficient search**, the N400 could serve as another attack vector for similar attacks as those described in this work.

While all listed contributions support our belief that such devices may be used in everyday tasks, they follow an orthogonal approach by considering how to assist users in various tasks like, for instance, authentication. Contrary to that, our objective is to turn the table and to demonstrate that such technology might create significant threats to the security and privacy of the users.

Guilty-Knowledge Test The most closely related work on EEG signals addresses using P300 in lie detection, particularly in the so-called Guilty-Knowledge Test (GKT) [3]. The operating hypothesis of the GKT is that familiar items will evoke different responses when viewed in the context of similar unfamiliar items. It has been shown that the P300 can be used as a discriminative feature in detecting whether or not the relevant information is stored in the subject's memory. For this reason, a GKT based on the P300 has a promising use within interrogation protocols that enable detection of potential criminal details held by the suspect, although some data suggest low detection rates [13]. In contrast, recent GKT experiments based on the P300 have reported detection accuracies as high as 86% [1]. Of course, as with the polygraph-based GKT, the P300-GKT is vulnerable to specific countermeasures, but to a much lesser extent [33, 34].

Such applications in interrogation protocols have quite a number of differences from our work. For instance, we concentrate on consumer-grade devices that have considerably lower signal-to-noise ratios, therefore are more difficult to analyze. The largest difference between our approach and in the GTK is the attacker model. While the GKT-interrogator has full control over the BCI user, in that he can can attach high-precision electrodes in a supportive way and force user to collaborate, our attacker must use the low-cost gaming device selected and attached by the user herself. This makes our attack considerably harder. Moreover, while the GTK victim clearly knows that she is interrogated and can prepare for that, in our case the user does not know that she is attacked. This might increase the validity of revealed information.

7 Discussion and Future Directions

In this section we discuss possible ways to defend against the investigated attacks and describe potential future directions.

Conscious Defenses Users of the BCI devices could actively try to hinder probing by, for instance, concentrating on non-target stimuli. To give a concrete example, users could count the number of occurrences of an unfamiliar face in our people experiment. The effectiveness of such defensive techniques has been tested in the context of guilty knowledge tests, however, there is no definitive conclusion on whether efforts to conceal knowledge are effective [35] or ineffective [8]. It is important to notice that, as we mentioned before, our scenario differs considerably from the GKT scenario. In our case, we assume that the EEG application has control of the user input for extended periods of time and that it conceals the attack in the normal interaction with the application. It would be difficult to imagine a realistic scenario in which a concerned user could try to conceal information from the EEG application for extended periods of normal usage.

An alternative to limiting the scope of the attacks presented in this paper is not to expose the raw data from EEG devices to third-party applications. In this model, the EEG vendor would create a restricted API that could only access certain features of the EEG signal. For example, applications could be restricted to accessing only movement related information (reflected in the spectral power). On the other hand, this poses higher performance demands on the device and limits the potential of developing third-party software.

Another possible way to deal with leaking information through the P300 signal would be adding noise to the EEG raw data before making it available to the applications that must use it. However, it would be difficult to strike a balance between the security of such an approach and the drawbacks in terms of decrease in accuracy of legitimate applications.

Future Directions The overall success of these attacks highly depends on the user's attention to the stimuli. Hence, there are still many open questions concerning the trade-off between obtrusiveness (in order to increase the user's attention during the classification task) and concealment to avoid the discovery of the attacker's true intentions. As part of our future work we intend to explore this trade-off in more detail. Specifically, by asking what is the impact of an uncooperative user who attempts to "lie" during the attack, e.g., similar to guilty-knowledge test settings? How can these attacks be made more stealthy, i.e., to what extent can they be integrated into some benign everyday tasks, games, or videos? How effective is the social engineering approach? For example, by offering fake monetary awards or by simply confusing the user (such as asking him to verify whether his PIN is truly random and telling him to count the number of the PIN occurrences).

8 Conclusion

The broad field of possible applications and the technological progress of EEG-based BCI devices indicate that their pervasiveness in our everyday lives will increase. In this paper, we focus on the possibility of turning this technology against the privacy of its users. We believe that this is an important first step in understanding the security and privacy implications of this technology.

In this paper, we designed and carried out a number of experiments which show the feasibility of using a cheap consumer-level BCI gaming device to partially reveal private and secret information of the users. In these experiments, a user takes part in classification tasks made of different images (i.e., stimuli). By analyzing the captured EEG signal, we were able to detect which of the presented stimuli are related to the user's private or secret information, like information related to credit cards, PIN numbers, the persons known to the user, or the user's area of residence, etc. The experiments demonstrate that the information leakage from the user, measured by the information entropy is 10 %-20 % of the overall information, which can increase up to $\approx 43\,\%$.

The simplicity of our experiments suggests the possibility of more sophisticated attacks. For example, an uninformed user could be easily engaged into "mind-games" that camouflage the interrogation of the user and make them more cooperative. Furthermore, with the ever increasing quality of devices, success rates of attacks will likely improve. Another crucial issue is that current APIs available to third-party developers offer full access to the raw EEG signal. This cannot be easily avoided, since the complex EEG signal processing is outsourced to the application. Consequently, the development of new attacks can be achieved with relative ease and is only limited by the attacker's own creativity.

Acknowledgements

This work was supported in part by the National Science Foundation under grants TRUST CCF-0424422 and grant No. 0842695, by the Intel ISTC for Secure Computing, and by the Carl-Zeiss Foundation.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] ABOOTALEBI, V., MORADI, M. H., AND KHALILZADEH, M. A. A new approach for EEG feature extraction in P300-based lie detection. *Computer Methods and Programs in Biomedicine 94* (April 2009), 48–57.

[2] CAMPBELL, A., CHOUDHURY, T., HU, S., LU, H., MUKERJEE, M. K., RABBI, M., AND RAIZADA, D. Neurophone: brain-mobile phone interface using a wireless EEG headset. In *Proceedings of the Second ACM SIGCOMM Workshop on Networking, Systems, and Applications on Mobile Handhelds* (2010), MobiHeld '10, pp. 3–8.

[3] COMMITTEE TO REVIEW THE SCIENTIFIC EVIDENCE ON THE POLYGRAPH. *The Polygraph and Lie Detection*. Board on Behavioral Cognitive and Sensory Sciences, National Research Council. The National Academies Press, 2003.

[4] DELORME, A., AND MAKEIG, S. EEGLAB: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of Neuroscience Methods 134*, 1 (2004), 9–21.

[5] EMOTIV SYSTEMS. www.emotiv.com. (last accessed: Feb. 12 2012).

[6] ENGEL, J., KUHL, D. E., PHELPS, M. E., AND CRANDALL, P. H. Comparative localization of foci in partial epilepsy by PCT and EEG. *Annals of Neurology 12*, 6 (1982), 529–537.

[7] FARWELL, L., AND DONCHIN, E. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology 70*, 6 (1988), 510 – 523.

[8] FARWELL, L., AND SMITH, S. Using brain mermer testing to detect knowledge despite efforts to conceal. *Journal of Forensic Sciences 46*, 2 (Jan 2001), 135–43.

[9] FISCHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of Human Genetics 7*, 2 (1936), 179–188.

[10] FLOR-HENRY, P., LANG, R., KOLES, Z., AND FRENZEL, R. Quantitative EEG studies of pedophilia. *International Journal of Psychophysiology 10*, 3 (1991), 253 – 258.

[11] FRIEDMAN, J., HASTIE, T., AND TIBSHIRANI, R. Additive logistic regression: a statistical view of boosting. *Annals of Statistics 28* (1998), 2000.

[12] FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics 29* (2000), 1189–1232.

[13] GAMER, M. Does the guilty actions test allow for differentiating guilty participants from informed innocents? a re-examination. *International Journal of Psychophysiology 76* (apr 2010), 19–24.

[14] GRUZELIER, J., EGNER, T., AND VERNON, D. Validating the efficacy of neurofeedback for optimising performance. *Event-Related Dynamics of Brain Oscillations: Progress in Brain Research 159* (2006), 421–431.

[15] GUPTA, P., AND GAO, D. Fighting coercion attacks in key generation using skin conductance. In *Proceedings of the 19th USENIX Conference on Security* (2010), USENIX Security'10, pp. 30–30.

[16] HALGREN, E., MARINKOVIC, K., AND CHAUVEL, P. Generators of the late cognitive potentials in auditory and visual oddball tasks. *Electroencephalography and Clinical Neurophysiology 106*, 2 (1998), 156 – 164.

[17] HOFFMANN, U., GARCIA, G., VESIN, J.-M., DISERENS, K., AND EBRAHIMI, T. A boosting approach to P300 detection with application to brain-computer interfaces. In *2nd International IEEE EMBS Conference on Neural Engineering* (2005), pp. 97 –100.

[18] INZLICHT, M., MCGREGOR, I., HIRSH, J. B., AND NASH, K. Neural markers of religious conviction. *Psychological Science 20*, 3 (2009), 385–392.

[19] J. MALMIVUO AND R. PLONSEY. Bioelectromagnetism: Principles and applications of bioelectric and biomagnetic fields. http://www.bem.fi/book/ (last accessed: Feb. 16 2012).

[20] KROPOTOV, J. D., GRIN-YATSENKO, V. A., PONOMAREV, V. A., CHUTKO, L. S., YAKOVENKO, E. A., AND NIKISHENA, I. S. ERPs correlates of EEG relative beta training in ADHD children. *International Journal of Psychophysiology 55* (2004), 23–34.

[21] KRUSIENSKI, D. J., SELLERS, E. W., CABESTAING, F., BAYOUDH, S., MCFARLAND, D. J., VAUGHAN, T. M., AND WOLPAW, J. R. A comparison of classification techniques for the P300 Speller. *Journal of Neural Engineering 3*, 4 (Dec. 2006), 299–305.

[22] LIN, C.-T., WU, R.-C., LIANG, S.-F., CHAO, W., CHEN, Y.-J., AND JUNG, T.-P. EEG-based drowsiness estimation for safety driving using independent component analysis. *IEEE Transactions On Circuits and Systems. Part I: Regular Papers* (2005), 2726–2738.

[23] MARCEL, S., AND MILLAN, J. Person authentication using brainwaves (EEG) and maximum a posteriori model adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*, 4 (April 2007), 743 –752.

[24] MEIJER, E., SMULDERS, F., AND WOLF, A. The contribution of mere recognition to the P300 effect in a concealed information test. *Applied Psychophysiology and Biofeedback 34* (2009), 221–226.

[25] NEUROSKY INC. www.neurosky.com. (last accessed: Feb. 11 2012).

[26] NIJHOLT, A. BCI for games: A 'state of the art' survey. In *Proceedings of the 7th International Conference on Entertainment Computing* (2009), ICEC '08, pp. 225–228.

[27] POLICH, J. Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology 118*, 10 (2007), 2128 – 2148.

[28] PORTAS, C. M., KRAKOW, K., ALLEN, P., JOSEPHS, O., ARMONY, J. L., AND FRITH, C. D. Auditory processing across the sleep-wake cycle: Simultaneous EEG and fMRI monitoring in humans. *Neuron 28*, 3 (2000), 991 – 999.

[29] POULOS, M., RANGOUSSI, M., CHRISSIKOPOULOS, V., AND EVANGELOU, A. Parametric person identification from the EEG using computational geometry. In *The 6th IEEE International Conference on Electronics, Circuits and Systems* (Sep 1999), vol. 2, pp. 1005 –1008 vol.2.

[30] REVETT, K., AND MAGALHES, S. T. Cognitive biometrics: Challenges for the future. In *Global Security, Safety, and Sustainability*, vol. 92. 2010, pp. 79–86.

[31] ROS, T., MOSELEY, M. J., BLOOM, P. A., BENJAMIN, L., PARKINSON, L. A., AND GRUZELIER, J. H. Optimizing microsurgical skills with EEG neurofeedback. *BMC Neuroscience*, 1 (2009), 10–87.

[32] ROSENFELD, J. P., BIROSCHAK, J. R., AND FUREDY, J. J. P300-based detection of concealed autobiographical versus incidentally acquired information in target and non-target paradigms. *International Journal of Psychophysiology 60*, 3 (2006), 251 – 259.

[33] ROSENFELD, J. P., AND LABKOVSKY, E. New P300-based protocol to detect concealed information: Resistance to mental countermeasures against only half the irrelevant stimuli and a possible ERP indicator of countermeasures. *Psychophysiology 47*, 6 (2010), 1002–1010.

[34] ROSENFELD, J. P., SOSKING, M., BOSH, G., AND RYAN, A. Simple, effective countermeasures to P300-based tests of detection of concealed information. *Psychophysiology 41*, 1 (2004), 205–219.

[35] ROSENFELD, J. P., SOSKINS, M., BOSH, G., AND RYAN, A. Simple, effective countermeasures to P300-based tests of detection of concealed information. *Psychophysiology 41* (mar 2004), 208.

[36] THORPE, J., VAN OORSCHOT, P. C., AND SOMAYAJI, A. Pass-thoughts: authenticating with our minds. In *Proceedings of the 2005 Workshop on New Security Paradigms* (New York, NY, USA, 2005), NSPW '05, ACM, pp. 45–56.

[37] VAN VLIET, M., MHL, C., REUDERINK, B., AND POEL, M. Guessing what's on your mind: Using the n400 in brain computer interfaces. vol. 6334. 2010, pp. 180–191. 10.1007/978-3-642-15314-3_17.

[38] WAISMANN, R., FENWICK, P., WILSON, G., HEWETT, D., AND LUMSDEN, J. EEG responses to visual erotic stimuli in men with normal and paraphilic interests. *Archives of Sexual Behavior 32* (2003), 135–144. 10.1023/A:1022448308791.

[39] WILSON, G. F., AND RUSSELL, C. A. Operator functional state classification using multiple psychophysiological features in an air traffic control task. *The Journal of the Human Factors and Ergonomics 45*, 3 (2003), 381–389.

A Session Script

Preparation. "We will now run a series of experiments. Each one of them takes approximately 1.30 minutes. Please find a comfortable position. Please try to stay still and not move your face." *(Participants are shown EEG feed and show the effects if the participants move their body and face)*

Training. "We will now run through a basic experiment to train our software. The system will display a random sequence of digits zero through nine. Please count the number of times [x] is shown. Please do not count the occurrences of a different number or otherwise attempt to fool the system."

Password. "Please choose and write down a 4 digit PIN and keep it by yourself. Do not show it to me and do not use a PIN code that you normally use."

"There are no special instructions for this experiment. However, at the end of this experiment, you will have to enter the first digit of the PIN you just chose."

Banks ATM. "Are you a customer of any of those ten banks on the list?"
"Are you a customer with just one?"
(If yes to both) "For this experiment, instructions are displayed on-screen"
Message on screen: What is the name of your bank?

Banks Debit Cards. "For this experiment, instructions are displayed on-screen"
Message on screen: What is the name of your bank?

Geographic Location. "Do you live close to campus?"
If yes: "Instructions are displayed on-screen."
Message on screen: Where do you live? Count the number of occurrences.

Month of Birth. "Instructions are displayed on-screen"
Message on screen: When were you born?

People "For this experiment, instructions are displayed on-screen"
Message on screen: Do you know any of these people?