Securing the Exocortex: A Twenty-first Century Cybernetics Challenge

Tamara Bonaci[1], Jeffrey Herron[1], Charlie Matlack[1], and Howard Jay Chizeck[1,2]
[1]Department of Electrical Engineering
[2]Department of Bioengineering
University of Washington
{tbonaci, jeffherr, cmatlack, chizeck}@uw.edu

***Abstract*—An exocortex is a wearable (or implanted) computer, used to augment a brain's biological high-level cognitive processes and inform a user's decisions and actions. In this paper, we focus on Brain-Computer Interfaces (BCIs), a special type of exocortex used to interact with the environment via neural signals. BCI use ranges from medical applications and rehabilitation to operation of assistive devices. They can also be used for marketing, gaming, and entertainment, where BCIs are used to provide users with a more personalized experience.**

BCI-enabled technology carries a great potential to improve and enhance the quality of human lives. This technology, however, is not without risk. In this paper, we address a specific class of privacy issues, *brain spyware*, shown to be feasible against currently available non-invasive BCIs. We show this attack can be mapped into a communication-theoretic setting. We then show that the problem of preventing it is similar to the problem of information hiding in communications. We address it in an information-theoretic framework. Finally, influenced by Professor Wiener's computer ethics work, we propose a set of principles regarding appropriate use of exocortex.

***Keywords*:** *cybernetics, exocortex, neural engineering, Brain-Computer Interfaces, privacy, communication system, information theory, game theory*

I. Introduction

The term *exocortex* refers to a wearable computer, used to augment a brain's biological high-level cognitive processes and assist a user's decisions and actions. It stems from computer science and evolutionary psychology, but it has been popularized by science-fiction writers. While the term itself was coined in 2004 by science fiction author Charles Stross [40], the first fictional devices fitting this definition of exocortex were introduced in 1984, by William Gibson [25] and Vernor Vinge [42].

In this paper, we consider Brain-Computer Interfaces (BCIs), a fast-growing class of non-fictional exocortices. Currently, a large variety of BCIs are being proposed and developed for applications ranging from medical to commercial, including advertising, market surveys, focus groups, and gaming. For example, in 2008, the Nielsen Company acquired Neuro-Focus, for the development of neural engineering technologies aimed at better understanding customer needs and preferences [6]. In May 2013, Samsung, in collaboration with the University of Texas, demonstrated how BCIs could be used to control mobile devices [10]. In the same month, the first neurogaming conference attracted more than 50 companies [7].

978-1-4799-4562-7/14/$31.00

Several neural engineering companies (e.g., Emotiv Systems [5] and NeuroSky [8]) currently offer consumer-grade BCIs and software development kits. In addition, the companies have recently introduced the concept of app stores for BCIs, in order to facilitate expansion of BCI applications (as described in [14]). Future BCIs will likely be simpler to use and will require less training time and user effort, while enabling faster and more accurate decoding of users' intentions. It is easy to see that Wiener was correct when he predicted that information technology will give rise to prostheses for able-bodied people which will "give them powers that humans never had before" [15], [45].

To paraphrase Wiener's concerns, this progress poses not only new possibilities for the future, but raises questions about users privacy. In the recent security literature, researchers introduced several BCI-enabled malicious applications, referring to them as *brain spyware* [24], [33]. These applications can be used to extract users' private information, such as credit card PINs, from recorded neural signals.

As BCIs become widespread, more sophisticated spying applications are easy to imagine. Leveraging recent results in neuroscience, it is possible to extract private information about users' memories, prejudices, religious and political beliefs, as well as about their possible neurophysiological disorders [14]. The extracted information could be used for marketing purposes, or to manipulate or coerce users, or otherwise harm them. The impact of BCI malware could be severe in terms of privacy and other important values.

This work is inspired by Wiener's Information Ethics methodology, which can be summarized by the following steps [15]:

1) Identify an ethical problem or positive opportunity regarding the integration of technology into society.
2) If possible, apply existing policies, using precedent and traditional interpretations to resolve the problem or to benefit from the opportunity.
3) If existing policies appear to be ambiguous or vague when applied to the new problem or opportunity, clarify ambiguities and vagueness.
4) If precedent and existing interpretations, including the new clarifications, are insufficient to resolve the problem or to benefit from the opportunity, revise the old policies or create new ones.
5) Apply the new or revised policies to resolve the problem or to benefit from the opportunity.

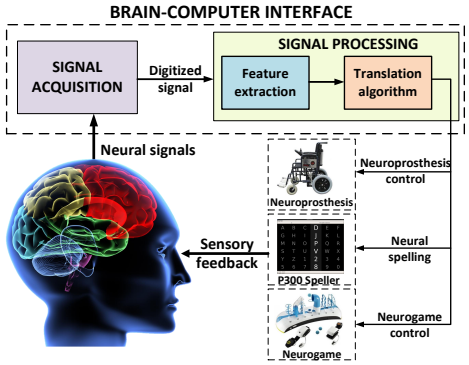

Fig. 1. High-level block diagram of a typical brain-computer interface. A user's neural signals are recorded, processed and decoded in a BCI. Decoded messages are then used to control the environment. In a closed-loop system, a user also receives sensory feedback from the environment, which can be visual, audio, haptic (kinesthetic or tactile), and, in the future, neural.

The time to identify and address privacy and security challenges arising from uses of exocortices is now. In protecting users from *brain spyware*, we encounter challenges similar to those in the information hiding problem [38]. Here, a BCI user is effectively an information hider whose goal is to detect and remove any private information before sending messages to the external environment. The user's opponent is an attacker, seeking to extract private information by distorting the communication channel. Thus, the protection of privacy in BCIs can be thought of as an information hiding game.

II. Preliminaries

The initial development of BCIs was motivated by the needs of individuals having certain physical disabilities, and the recognition of the potential benefits these systems might offer. A variety of BCI platforms exist (and are being developed) for assistance, augmentation, and repair of cognitive and sensorimotor capabilities in conditions such as spinal cord injuries or amyotrophic lateral sclerosis. In addition, BCIs are increasingly being used in non-medical applications, such as gaming, entertainment, and marketing. For example, researchers have proposed a method of predicting success of online games by analyzing a user's electromyographic (EMG) signals (electrical signals produced by a user's skeletal muscles) over the first 45 minutes of the game [17].

A. The Components of a BCI

From an engineering perspective, a BCI is a *communication system* between the brain and the external environment, consisting of inputs (user's neural activity), outputs (external world commands), and components - translating inputs to outputs, referred to as *signal acquisition* and *signal processing*. A high-level block diagram of a typical BCI is depicted in Figure 1.

In this system, communication messages between the brain and the environment are typically encoded as electrophysiological signals [47], [48]. Based on the type of communication signal (and recording location), BCIs can range from completely implanted devices, where electrode arrays are surgically implanted into the brain; to partially invasive, such as electrocorticography (ECoG), where electrodes are typically implanted inside the skull and on top of the dura; to non-invasive devices. Most non-invasive BCIs are based on electroencephalography (EEG), but other biosignals, such as EMG or electrooculograms (EOG) can also be used to control devices. While known to be susceptible to noise and signal distortion, EEG signals are easy to measure. In addition, EEG-based BCIs have relatively low cost and risk, which makes them the most widely used BCI devices [46]. In this paper, we focus on non-invasive BCIs.

A BCI's signal processing component typically consists of a *feature extractor* and *decoding algorithm*. The feature extractor processes recorded signals and extracts signal features, assumed to reflect specific aspects of a user's neural signal. Decoding algorithms take the abstracted features and translate them into application-specific commands. Depending on the application, many different decoding algorithms exist (see e.g., [29]).

B. Emerging Issues Related to the Use of BCIs

With an increase in neural engineering applications, researchers have recognized the need to address emerging ethical and questions arising from the use of neural devices [19], [23], [26]–[28], [41]. Recently Denning et al. [22] summarized the security problem as follows: "the use of standard engineering practices, medical trials, and neuroethical evaluations during the design process can create systems that are safe and that follow ethical guidelines; unfortunately, none of these disciplines currently ensure that neural devices are robust against *adversarial entities* trying to exploit these devices to alter, block, or eavesdrop on neural signals." The authors identified security threats that can be mounted against implanted neural devices, and introduced the term *neurosecurity* as "the protection of the confidentiality, integrity, and availability of neural devices from malicious parties with the goal of preserving the safety of a person's neural mechanisms, neural computation, and free will." [22]

Providing real-world examples of such threats, Martinovic et al. [33] and Frank et al. [24] recently showed that attacks on non-invasive BCIs are indeed feasible. The authors of [33] presented the first malicious software designed to detect a user's private information using a EEG-based BCI, referred to as *brain spyware*. They used a commercially available device to present users with noticeable visual stimuli while recording their EEG signals. The authors analyzed the recorded signals in order to detect a user's: (a) 4-digit PIN, (b) bank information, (c) month of birth, (d) location of residence, and (e) if a user recognized the presented set of faces. In [24], the authors used subliminal visual stimuli (i.e, stimuli too short for conscious perception), to try to determine if a user recognized the presented person.

In both of these papers, the authors focused on the P300 Event Related Potential component [31]. However, it is not hard to imagine *brain spyware* being developed to extract not only private information about users' memories, prejudices and beliefs, but also about their possible neurophysiological disorders. Currently, there does not seem to exist a way to resist these attacks. Moreover, recent results [30] show that attempts at willful deception can themselves be detected from an individual's neural signals. The authors of [30] also show that non-invasive brain stimulators, emitting imperceptible DC electrical stimuli, can be used to make a user's responses

noticeably slower when attempting to lie. This work clearly indicates there is a growing need to address potential privacy and security risks arising from the use of BCIs, in both medical and non-medical applications.

III. From Wiener to Science Fiction to Contemporary Innovations

A. Norbert Wiener's Influence

Norbert Wiener wrote extensively on the applications of cybernetics to the human nervous system (e.g., [44]). Consider his seminal work *Cybernetics: or Control and Communication in the Animal and the Machine* [43]. He identified *cybernetics* as a field encompassing both complex biological and artificial systems. This was because he viewed the human brain as fundamentally an information processing machine. Thus he proposed a framework for the re-examination of the very notion of the brain and conscious mind, which could be observed through the lenses of quantifiable engineering disciplines. Many phenomena could consequently be explained by the same rules that govern the fields of control and communication. One prime example of this was his identification of intention and Parkinsonian tremors as feedback problems [45]. Once identified as such, he and his collaborators implemented example systems in a simple robot analog and were able to replicate the various forms of intentional and postural tremor by simply manipulating the nature of the feedback. Wiener correctly identified that if the neurological processes in our brain follow the same rule set as artificial information processing systems, it would be possible to interface the two synergistically. He had fundamentally given form to the field of brain-computer interfaces long before the desktop computer had come into existence.

B. Influence of Science Fiction

The revolutionary ideas of integrating the biological and artificial through the quantitative engineering disciplines of control and communication theory became a theme of science fiction writers. Science fiction has served as a *speculative sandbox* to test the implications and potentials of new technology, and their implications for society and human relationships. This is especially true for BCI technology, which changes the fundamental capabilities of the human mind. BCI technology has entered science fiction. In particular, the cyberpunk genre extensively examined the possible ramifications of this melding of man, machine and neural engineering. With these concepts, authors created new terminology to describe the technology that would enable this fundamental shift in the human experience. Various names have arisen to describe a new form of the central nervous system that includes extensive cybernetic augmentation, including the term *exocortex*, popularized by author Charles Stross [40].

Some of the most technologically interesting ideas presented in these works include early forms of augmented/virtual realities, brain-to-brain communication through BCI, and how skills and memories can potentially be distributed or downloaded. For example, *Neuromancer* by William Gibson [25] featured characters with extensive cybernetic enhancements and envisioned networked virtual realities accessible through brain-computer interfaces. The Borg from Star Trek [12] are illustrated as a space-faring cyborg society where individuality has been replaced with a hive-mind enabled through widespread brain-to-brain communication. Another example is the animated TV series *Ghost in the Shell*, which features a wide cast of police and combat cyborgs who can utilize an array of downloadable skills as specific as rifle-based mid-range sniping [11]. Much of this body of literature, animation, and film is cautionary, pointing out potential dangers and pitfalls. But some optimistically explore new potential social and societal interactions based on these technologies, through strong characters and plots.

One common thread throughout the genre are the potential problems with the security of these systems, or more interestingly, the possible lack thereof. For example, the storyline of Shirow's *Ghost in the Shell* [39] is driven primarily by a string of cyberbrain hacking where individuals' memories or even identities are altered for malicious purposes. In the recent video game *Deus Ex: Human Revolution* [4], a character involuntarily commits suicide after a hacker half a world away takes control of his body and overrides muscular commands. While the BCI technology presented in these works is incredibly exciting, these attacks are by their nature unsettling.

C. Influence of other Cyber and Cyber-Physical Systems

Applying cybernetics to interface the human mind with machines enables our minds to control our physical environment. At the same time, it opens up the possibility of losing control of our own bodies. This is a terrifying prospect that may be a barrier to end-user adoption of these technologies.

Current medical devices typically rely on the proximity or wired communication between (implanted) devices and practitioners' control interfaces to decrease the potential for malicious attacks. But this is not enough to provide security. Researchers have recently showed that insulin pumps [32], [37] and pacemakers [21] can be remotely compromised.

In the future, it will become increasingly difficult to justify not having devices able to communicate with the outside world. Imagine an implant in our brain with all of the capabilities of a contemporary smart phone (e.g., search, navigation, translation, face recognition, to name a few), and the possibilities that immediately present themselves with such a device. However, the only way for such a device to be safe and secure is to architect it as such from the start. Unfortunately, there is no historical precedent for a secure communication system to spontaneously arise from a technological field. This is something the BCI community needs to be aware of.

Recent technological history teaches us how important it is to include security and privacy considerations into the design process, presenting us with numerous examples of how hard it is to incorporate security methods into a system after it has already been compromised. Even today, the internet, computer, and mobile environments are rife with malware and private information leaks. The foundations of these systems were not designed with security in mind, and as such they have become hostile environments that users have to navigate. If consumer devices with no biological interface cannot be securely developed, how will users have confidence in a system that interfaces with their bodies and brains?

IV. Our Approach

Security and privacy threats arising from the use of exocotices, both future and existing, such as BCI-enabled devices, may not yet pose a critical concern, given their fairly limited deployment outside of research and medical communities. We believe, however, the right time to address these issues is now. Methods to prevent and mitigate these threats should be developed in the early design phase, and embedded throughout the entire life of the technology. We view the development of prevention and mitigation tools as an interdisciplinary effort, but also as a way to *close the loop* around BCI researchers, ethicists and the science fiction community. Influenced by Wiener's work, in particular his account of information ethics, this paper represents an initial step towards facilitating this effort, by analyzing the problem of *brain malware* in EEG-based BCIs. We cast this problem into a well-developed information-theoretic framework. But first, we review how a BCI, operating in a benign environment, can be modeled as communication channel.

A. Modeling BCIs as Communication Channels

The authors of [35] modeled non-invasive EEG-based BCIs as communication channels, based on assumptions that: (1) the main purpose of BCIs is communication with the environment and (2) the environment will wait until a user's intention is clear before taking any action.

1) Intent as a Discrete Random Process: In [35], it was shown that a user's intent can be described as a sequence $W = (W_1, W_2, \dots)$, where each W_i is an element of a finite alphabet, $\mathcal{W} = \{w_1, \dots, w_m\}$. To capture the fact that a user's intent may (and will) differ, depending on a situation, an intent is assumed to be generated by a random process $\mathbb{P}_W(w)$. Thus, for any n, the conditional probability of n-length intent can be computed as [35]:

$$\mathbb{P}_{W^n}(w^n) := \prod_{i=1}^{n} \mathbb{P}_{W^i|W^{i-1}}(w_i|w^{i-1}) \tag{1}$$

2) Coding/Decoding Process as a Discrete Memoryless Channel : In BCIs, classifiers are usually used to determine the intended messages based on recorded neural signals (for example, the intended left- or right-hand motor imagery, or the spelled letter), with some probability of making errors and incorrectly decoding messages. This allows us to think of coding/decoding process as a *discrete memoryless channel*.

For example, in [35] the authors modeled motor-imagery-based BCIs as *binary symmetric channels (BSC)* [20]. The k-th input to the channel is a random variable $X_k \in \{0, 1\}$, representing a user's motor imagery, where $x_k = 0$ corresponds to the left hand movement and $x_k = 1$ to the right. The k-th output of the channel is a random variable $Y_k \in \{0, 1\}$, representing a decoded user's intent. In this channel, $x_k = y_k$ represents a correct decoding (inference), and $x_k \neq y_k$ a case when a decoding error occurs. The probability of an error, $\mathcal{P}_e$, can be computed as [35]:

$$\begin{aligned} \mathcal{P}_e &:= \mathbb{P}_{Y_k|Y^{k-1},X^k}(y_k|y^{k-1}, x^k) = \mathbb{P}_{Y^k|X^k}(y_k|x_k) \\ &= \begin{cases} 1-\epsilon, & \text{if } y_k = x_k, \\ \epsilon, & \text{else} \end{cases} \end{aligned} \tag{2}$$

where ϵ denotes a parameter which can be learned (inferred) from BCI training data.

3) Graphical Display as Noiseless Feedback: A vast majority of non-invasive BCIs operate by providing visual feedback through a graphical display. Such a feedback contains information about the previous channel outputs, $y_1, y_2, \dots, y_k$, which is then used to choose the next channel input, x_{k+1}.

By assuming that a graphical display conveys information about the decoded output without noise and errors, the authors of [35] modeled it as a causal and noiseless feedback channel. Thus, combining ideas described above, an EEG-based BCI can be modeled as a *binary symmetric channel with noiseless feedback* [20].

V. The Problem

In this section, we analyze *brain spyware* attacks. We start by presenting the threat model. We then cast these attacks into the information-theoretic framework, showing how they can be modeled as *multiple access channels with generalized feedback* (MAC-GF). Finally, we analyze possible goals of attackers.

A. Threat Model

We consider an attacker who uses non-invasive BCIs to extract private information about users. BCI manufacturers generally distribute software development kits, technical support, as well as guides and tutorials with their products, all with the goal of promoting faster application development. These *open-development* platforms typically grant application developers full control over the devices' acquisition systems, signal processing components, decoding algorithms, and application rendering and control components. Thus, we assume that an attacker has an access to all of these resources, as well as to the recorded neural signals.

1) Methods of Extracting Private Information: We consider scenarios where an attacker interacts with users by *presenting them with specific stimuli*, and recording their responses to those stimuli. In the current literature, there are several well-established methods to present stimuli to BCI users [14]. We assume an attacker can use any of these presentation methods to facilitate extraction of private information. An attacker can further present malicious stimuli in an *overt (conscious)* fashion, as well as in a *subliminal (unconscious)* way, with subliminal stimulation defined as the process of affecting people by visual or audio stimuli of which they are not aware [13].

2) Types of Attackers: Based on the way attackers analyze recorded neural signals, we distinguish between two types of attackers. The first type extracts users' private information by *hijacking legitimate components of a BCI system*. This attacker exploits feature extraction and decoding algorithms intended for the legitimate BCI applications to mount attacks.

The second type extracts users' private information by *adding or replacing the legitimate BCI components*. This attacker may implement additional feature extraction and decoding algorithms, and either replace or supplement the existing BCI components with the additional malicious code.

B. Brain Spyware Attacks as Communication Channels

Before modeling *brain spyware* attacks as Multiple Access Channel with Generalized Feedback (MAC-GFs), we provide a

TABLE I. A SUMMARY OF NOTATION

Symbol	Definition
$\mathcal{W}_i$	Finite alphabets of source messages W_1, W_2
W_1	Source messsage, modeling a user's intent
W_2	Source messsage, modeling a user's private information
$\mathcal{X}_i$	Finite alphabets of input messages X_1, X_2
X_1	User's intent encoded in neural activty
X_2	User's private information, encoded in neural activty
$\mathcal{Y}$	Finite alphabets of output message Y
Y	User's intent and private information, as contained in a user's recorded neural signals
$\mathcal{Z}_i$	Finite alphabets of feedback messages Z_1, Z_2
Z_1	Graphical display feedback to a user
Z_2	Feedback message about user's private information
$\hat{W}_1$	Decoded user's intent
$\hat{W}_2$	Decoded user's private information
$\mathcal{P}_e$	Error probability
$\mathcal{P}_{success}$	Probability of a successful *brain spyware* attack
$\mathcal{P}_{chance}$	Probability of successfully decoding a user's intent by chance
Δ	Equivovation rate, a measure of an attacker's inability to correctly decode a user's private information

brief overview of these communication channels. More details can be found in seminal papers by Carleial [16], Zhen et al. [50], and Ozarow [36].

1) Multiple Access Channels with Generalized Feedback: A MAC-GF is a communication channel where two or more sources transmit information to a single destination, and each source observes a different feedback message. A block diagram of a two-user MAC-GF is depicted in Figure 2. We can mathematically represent discrete memoryless MAC-GFs as[1]:

$$(\mathcal{X}_1 \times \mathcal{X}_2, \mathbb{P}(y, z_1, z_2|x_1, x_2), \mathcal{Y} \times \mathcal{Z}_1 \times \mathcal{Z}_2) \tag{3}$$

where $\mathcal{X}_1$ and $\mathcal{X}_2$ represent finite input alphabets, $\mathcal{Y}$ a finite output alphabet, and $\mathcal{Z}_1$ and $\mathcal{Z}_2$ finite feedback alphabets. Value $\mathbb{P}(y, z_1, z_2|x_1, x_2)$ represents the conditional probability that output and feedback messages are equal to $(y, z_1, z_2) \in \{\mathcal{Y} \times \mathcal{Z}_1 \times \mathcal{Z}_2\}$ given input messages $(x_1, x_2) \in \{\mathcal{X}_2, \mathcal{X}_2\}$, and it can be computed as [16]:

$$\begin{aligned} & \mathbb{P}(y^n, z_1^n, z_2^n|x_1^n, x_2^n) \\ = & \prod_{i=1}^{n} \mathbb{P}(y_i, z_{1i}, z_{2i}|x_{1i}, x_{2i}, y^{i-1}, z_1^{i-1}, z_2^{i-1}) \\ = & \prod_{i=1}^{n} \mathbb{P}(y_n, z_{1i}, z_{2i}|x_{1i}, x_{2i}) \end{aligned} \tag{4}$$

A typical code for discrete memoryless MAC-GFs consists of:

- Two independent message sources, generating random messages $W_1 \in \{1, 2, \ldots, M_1\}$ and $W_2 \in \{1, 2, \ldots, M_2\}$.
- Two encoding functions:

$$\begin{aligned} x_{1n} &= f_{1n}(W_1, Z_1^{n-1}), \\ x_{2n} &= f_{2n}(W_2, Z_2^{n-1}), n = 1, 2, \ldots, N \end{aligned} \tag{5}$$

where x_{1n}, x_{2n} denote codewords of input messages X_1, X_2; W_1, W_2 source messages; Z_1^{n-1}, Z_2^{n-1} feedback messages, and N the length of input messages X_1, X_2.

[1]Summary of notation used in the paper is provided in Table I.

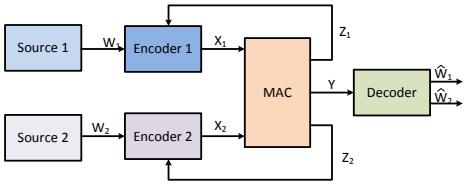

Fig. 2. A block diagram of a two-user Multiple Access Channel with Generalized Feedback (MAC-GF). Signals W_1 and W_2 represent source messages to be transmitted to the receiver. X_1 and X_2 are input messages, processed by encoders, and Y is an output message. Signals Z_1 and Z_2 represent feedback messages to encoders and $\hat{W}_1, \hat{W}_2$ the decoded messages on the receiver side.

- One decoding function:

$$(\hat{W}_1, \hat{W}_2) = g(Y^N) \tag{6}$$

where $\hat{W}_1, \hat{W}_2$ denote decoded source messages and Y^N output message.

For independent and uniformly distributed source messages W_1, W_2, error probability, $\mathcal{P}_e$ is defined as [50]:

$$\mathcal{P}_e := \mathbb{P}[(\hat{W}_1, \hat{W}_2) \neq (W_1, W_2)] \tag{7}$$

The capacity region of a general MAC-GF is not known. However, several coding schemes have been proposed in the literature (e.g., [16], [50]), and achievable regions for those schemes were derived.

2) Modeling Brain Spyware as an MAC-GF: We start by modeling a user's intent as a discrete random sequence W_1 over finite alphabet $\mathcal{W}_1$. Similarly, we model one specific instance of a user's private information (for example, a user's credit card PIN) as a discrete random sequence W_2 over finite alphabet $\mathcal{W}_2$. Building on subsection IV-A, we assume both intent W_1 and instance of a user's private information W_2 are generated by discrete random processes, as described by equation (1).

We next assume intent and private information are both encoded into a user's neural activity, which can be modeled as input messages X_1^N and X_2^N, consisting of N codewords x_{1i} and x_{2i}, $i = 1, \ldots, N$:

$$\begin{aligned} x_{1n} &= f_{1n}(W_1, Z_1^{n-1}) \\ x_{2n} &= f_{2n}(W_2, Z_2^{n-1}), \ n = 1, 2, \ldots, N \end{aligned} \tag{8}$$

Feedback message Z_1 represents feedback provided to a user about his intent. We distinguish three cases:

1) We can have an open-loop BCI, where no feedback about the decoded intent is provided. This case can be modeled as a special case of a MAC-GF, namely the discrete memoryless MAC with one-sided different generalized feedback [50].
2) Next, we can have a closed-loop BCI, where feedback about the decoded intent is assumed to be perfect (causal and noiseless).
3) Finally, we can also have a closed-loop BCI where feedback about user's intent is affected by noise.

In all of the cases, feedback message Z_2 represents an attacker abusing the system, to represent different stimuli to users and infer their private information.

In a BCI, a user's neural activity is recorded using an electrode array. Therefore, in a signal acquisition component of a BCI, both intent and private information are observed as single output signal Y. In a signal processing component of an attacked BCI, signal Y is translated into a decoded intent, $\hat{W}_1$ and a decoded private information, $\hat{W}_2$.

Note 1: In case no attack is mounted against a BCI, the described model easily reduces to a discrete memoryless channel, described in subsection IV-A.

Example 1: Let's consider a user, trying to spell a word `blue`. Let's assume his credit card PIN may be `1234`. In this case, values of random variables W_1 and W_2, representing source messages are:

$$W_1 = \texttt{blue} \text{ and } W_2 = \texttt{1234}$$

These source messages get encoded into inputs X_1 and X_2:

$$X_1 = \{x_{11}, x_{12}, x_{13}, x_{14}\} \text{ and } X_2 = \{x_{21}, x_{22}, x_{23}, x_{24}\}$$

In an EEG-based BCI, these input messages are recorded from a user's skull as an EEG signal $Y = \{y_1, y_2, \ldots, y_8\}$, sampled with some sampling rate f_s and discretized with some quantization resolution, $r_{A/D}$.

Using feature extractor and decoding algorithms, an attacker can decode output signal Y into a user's intent and an instance of private information:

$$\begin{aligned} \hat{W}_1 &= \text{anything} \\ \hat{W}_2 &= \texttt{1234} \text{ or } \hat{W}_2 = \texttt{9876} \end{aligned}$$

C. Modeling Attacker's Goals

In *brain spyware*, an attacker's goal is to correctly infer a user's private information. Thus, an attack is considered *successful* if an attacker can correctly decode a user's private information:

$$\mathbb{P}[\text{Attack successful}] := \mathbb{P}[\hat{W}_2 = W_2] \tag{9}$$

While an attacker, in general, does not care about correctly decoding users' intents, he might be willing to hide presence within a BCI, in order to avoid alarming users, and to keep them engaged with a malicious application for as long as possible. To maintain this *stealthiness*, an attacker will want to correctly decode users' intents, at least as successfully as a benign application would. Keeping this observation in mind, we redefine a *successful attack* as follows.

Definition 1. *A brain spyware is considered successful if an attacker can correctly decode a user's private information, while at the same time hiding his presence within the system. An attacker is considered hidden if he can decode a user's intents equally as good as a benign BCI application would. Thus, success probability of a brain spyware is equal to:*

$$\begin{aligned} \mathcal{P}_{success} &:= \mathbb{P}[\hat{W}_1 = W_1, \hat{W}_2 = W_2] \\ &= \mathbb{P}[(\hat{W}_1, \hat{W}_2) = (W_1, W_2)] \\ &= 1 - \mathcal{P}_e \end{aligned} \tag{10}$$

An attacker's goal can now be formalized as an optimization problem:

$$\begin{aligned} &\text{maximize } \mathcal{P}_{\text{success}} \\ &\text{subject to } \mathbb{P}[\hat{W}_1 \neq W_1] > \mathcal{P}_{\text{chance}} \\ &\qquad \text{Attacker's resources} \end{aligned} \tag{11}$$

where $\mathbb{P}[\hat{W}_1 \neq W_1] > \mathcal{P}_{\text{chance}}$ represents a benign application's constraint on successfully decoding a user's intent, and *attacker's resources* denote all other constraints imposed on the attacker, such as computational power and energy. Value $\mathcal{P}_{\text{chance}}$ denotes the value of successfully decoding the intended message by chance. While this value depends on the source and input alphabets, as well as assumptions made about a communication channel, it represents a lower bound requirement on a decoding algorithm used by a BCI. Using equation (10), an attacker's (11) can be rewritten as:

$$\begin{aligned} &\text{minimize } \mathcal{P}_e \\ &\text{subject to } \mathbb{P}[\hat{W}_1 \neq W_1] > \mathcal{P}_{\text{chance}} \\ &\qquad \text{Attacker's resources} \end{aligned} \tag{12}$$

VI. Our Solution: *BCI Anonymizer*

A block diagram of a BCI with the *BCI Anonymizer* component, under *brain spyware* attacks, is depicted in Figure 3. The basic idea of this component is to pre-process neural signals, before they are stored and transmitted, in order to remove all information except users' specific intents. Unintended information leakage is prevented by *never transmitting* and *never storing* raw neural signals, nor any signal components that are not explicitly needed for the purpose of BCI communication and control.

This approach is similar to those taken in the smart-devices industry, where attackers may attempt to access users' private identifying information (PII), such as users' location or address book entries. In the smartphone industry, such attacks on users' privacy are typically prevented by limiting access to the phone's operating system and users' PII. Neural signals have a similar role as a users' PII data, in that they contain information beyond the intended messages.

The *BCI Anonymizer* can be realized either in hardware or software, as a part of a BCI, but not as part of any external network or computational platform. It thus acts as a secured and trusted software or hardware subsystem that takes the raw neural signal and decomposes it to specific components. Upon request, instead of the complete recorded neural signal, the *BCI Anonymizer* provides an application only with a needed subset of requested signal components [18].

A. Interaction between an Attacker and BCI Anonymizer

The *BCI Anonymizer* is intended to prevent an attacker from gaining access to a user's private data. Thus, *brain spyware* attack on a user of a BCI with the *BCI Anonymizer* can be modeled as an *information hiding game*[2] between two non-cooperative players, the *BCI Anonymizer* and an attacker. The first player tries to maximize a payoff function, while the opponent (an attacker) tries to minimize it. One possible choice for a payoff function is the *equivocation rate* [49], which in this paper we define as follows.

Definition 2. *The equivocation rate, Δ, is a measure of the degree to which an attacker is unable to correctly decode a user's private information. It is equal to the conditional*

[2]The information hiding problem has been extensively studied in recent years. For more details, see for example [34], [38].

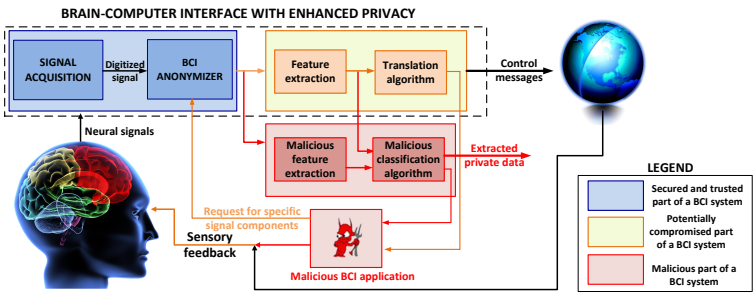

Fig. 3. A simplified diagram of a BCI with the *BCI Anonymizer* subsystem. Legitimate interpretation component (denoted as solid blocks in the diagram) requests data and receives response from the *BCI Anonymizer* (denoted as dashed background blocks in the diagram). Malicious components, added by the adversary (denoted as dotted background blocks in the diagram), may request data, but will not receive response from the *BCI Anonymizer*. In addition, an attacker cannot access states and functionality of the *BCI Anonymizer* components.

entropy of a private information W_2, given output message Y:

$$\begin{aligned}\Delta &:= H(W_2|Y) = \sum_{y\in\mathcal{Y}} \mathbb{P}[y] H(W_2|Y=y) \\ &= -\sum_{y\in\mathcal{Y}} \sum_{w_2\in\mathcal{W}_2} \mathbb{P}[w_2, y] \log\left(\mathbb{P}[w_2|y]\right) \qquad (13)\end{aligned}$$

The information-hiding game can now be formalized as the following optimization problems:

$$\begin{aligned}&\textit{BCI Anonyimizer:} && \text{maximize } \Delta \\ &&& \text{subject to system resources} \qquad (14) \\ &\text{Attacker:} && \text{minimize } \Delta \\ &&& \text{subject to attacker's resources}\end{aligned}$$

VII. Principles of Appropriate Use of Exocortex

We have identified both a positive opportunity and a potential problem in emerging BCI technologies.

In determining principles of appropriate use for BCI-based exocortices, we must begin with an assessment of the current state of policies and practices governing the relationship between information and communication technology and society. This will inform our new principles, because a wide spectrum of mobile information processing technologies exist, of which BCIs can be considered the extreme example of cybernetic integration.

Security features for smartphones provide a convenient example. We mentioned the example of restricting access to PII in the operating system, which is feasible because the OS API is the communication channel for possibly malicious applications. Generally, it is now standard for smartphone operating systems to provide granular user control of access permission to potentially private information. This is done by providing a user option to grant or deny access to specific, atomic resources on the device (e.g. data storage or sensor stream), and relying on the user to make choices informed by her knowledge of the information content of those resources.

This approach to security has failed in several ways. First, the operating system does not exercise sufficient control over all communication channels to enforce security policies. A recent example of this is third-party embedded advertising software broadcasting users' PII over insecure internet connections [1]. Second, some implementations do not give users complete control of resource access at the level of granularity possible. That is, requests for resource access are bundled so that a user may have to choose between compromising a resource exposing private information and not using an application at all. For example, the use of location services on Android phones is offered only on condition of allowing continuous location tracking [9]. Finally, the most fundamental problems may be that users are expected to fully understand what information is encoded in the data resources they grant or deny access to, *and* they cannot filter/hide private information in a resource while simultaneously making it available to an application.

We believe that standard usage of the *BCI Anonymizer* framework proposed here represents a novel policy capable of resolving the insufficiencies of existing security practices for the consumer products in widespread use which are most analogous to the exocortex. We eliminate the first problem above by implementing access restrictions as close to the private information source as possible, when only a single access channel exists. The second problem is a policy problem, not a technical problem, and addressing it in the BCI context is a matter of early and strong policy advocacy to ensure the development of BCIs trusted by consumers. The Electronic Frontier Foundation (EEF) is a well-known nonprofit that works to "...defend free speech online, fight illegal surveillance, advocate for users and innovators, and support freedom-enhancing technologies" [3]. A collaboration between the EEF, IEEE Standards Association, [2] and the BCI developers emerges as a fruitful approach to quickly and efficiently establish the necessary policies. The third problem we explicitly eliminate by removing, as much as possible, the burden on a user to be aware of what private information is available in the communication channel. At the same time, we recognize the remaining limited vulnerability that the *BCI Anonymizer* can only guarantee performance when the private information encoding method is fully known; limited protection may be achievable by changing all properties known not to interfere with the encoding of authorized information.

VIII. Conclusion

It has been five decades since Wiener started the discussion about the nervous system as an information processing unit, and almost three decades since he warned about possible problems when melding human brains with external computational units. A myriad of science fiction writers and enthusiasts followed, depicting a grim picture of exocortices being misused to harm people. And now the technology seems

to be catching up - what only was fiction is rapidly becoming a reality. The first attacks on non-invasive BCIs have already been presented in the security community. Guided by Wiener's principles of justice, we propose that the right moment to act to prevent these issues is now. With exocortices, we have a unique opportunity to include safety, security, privacy, and usability considerations into the system from design and development phases.

ACKNOWLEDGMENT

This work is supported by the NSF Engineering Research Center for Sensorimotor Neural Engineering (Award Number EEC-1028725). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation.

REFERENCES

[1] The Wall Street Journal: Mobile-App Makers Face U.S. Privacy Investigation (last accessed: March 14, 2014).

[2] IEEE Standards Association (last accessed: March 14, 2014).

[3] Electronic Frontier Foundation (last accessed: March 10, 2014).

[4] Deus Ex (last accessed: March 10, 2014).

[5] Emotiv Systems (last accessed: March 6, 2014).

[6] NeuroFocus (last accessed: March 6, 2014).

[7] NeuroGaming 2014 Conference and Expo (last accessed: March 6, 2014).

[8] NeuroSky (last accessed: March 6, 2014).

[9] Your Android Phone is Tracking You (last accessed: March 14, 2014).

[10] Samsung Demos a Tablet Controlled by Your Brain (last accessed: March 6, 2014).

[11] Ghost in the Shell: Stand Alone Complex. Anime TV Series, Episode 14 Season 2.

[12] Star Trek: The Next Generation. American Science Fiction TV Series, Episode 16 Season 2.

[13] R. B. Baldwin. Kinetic Art: On the Use of Subliminal Stimulation of Visual Perception. *Leonardo*, pages 1–5, 1974.

[14] T. Bonaci, R. Calo, and H. J. Chizeck. App Stores for the Brain: Privacy & Security in Brain-Computer Interfaces. In *the Proceedings of the 2014 IEEE International Symposium on Ethics in Engineering, Science, and Technology (to be presented)*, 2014.

[15] T. W. Bynum. Ethical Challenges to Citizens of The Automatic Age: Norbert Wiener on the Information Society. *Journal of Information, Communication and Ethics in Society*, 2(2):65–74, 2004.

[16] A. Carleial. Multiple-Access Channels with Different Generalized Feedback Signals. *IEEE Transactions on Information Theory*, 28(6):841–850, 1982.

[17] Y.-T. Chiu. Mind Reading to Predict the Success of Online Games, February 2013.

[18] H. J. Chizeck and T. Bonaci. Brain-Computer Interfaces Anonymizer. US Patent Application, PCT/US13/67528, February 2014.

[19] J. Contreras-Vidal. Ethical Considerations Behind Brain-Computer Interface Research, December 2012.

[20] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[21] T. Denning, A. Borning, B. Friedman, B. T. Gill, T. Kohno, and W. H. Maisel. Patients, Pacemakers, and Implantable Defibrillators: Human Values and Security for Wireless Implantable Medical Devices. In *the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 917–926. ACM, 2010.

[22] T. Denning, Y. Matsuoka, and T. Kohno. Neurosecurity: Security and Privacy for Neural Devices. *Neurosurgical Focus*, 27(1):1–4, 2009.

[23] N. Farahany. Incriminating Thoughts. *Stanford Law Review*, 64:11–17, 2011.

[24] M. Frank, T. Hwu, S. Jain, R. Knight, I. Martinovic, P. Mittal, D. Perito, and D. Song. SubliminalProbing for Private Information via EEG-Based BCI Devices. *arXiv preprint arXiv:1312.6052*, 2013.

[25] William Gibson. *Neuromancer*. Ace Books, 1984.

[26] J. Illes, M. P. Kirschen, and J. D. E. Gabrieli. From Neuroimaging to Neuroethics. *Nature Neuroscience*, 6(3):205–205, 2003.

[27] J. Illes and E. Racine. Imaging or Imagining? A Neuroethics Challenge Informed by Genetics. *The American Journal of Bioethics*, 5(2):5–18, 2005.

[28] A. R. Jonsen. *The Birth of Bioethics*. Oxford University Press, USA, 2003.

[29] S. Koyama, S. M. Chase, A. S. Whitford, M. Velliste, A. B. Schwartz, and R. E. Kass. Comparison of Brain-Computer Interface Decoding Algorithms in Open-loop and Closed-loop Control. *Journal of Computational Neuroscience*, 29(1-2):73–87, 2010.

[30] B. Luber, C. Fisher, P. S. Appelbaum, M. Ploesser, and S. H. Lisanby. Non-invasive Brain Stimulation in the Detection of Deception: Scientific Challenges and Ethical Consequences. *Behavioral Sciences & the Law*, 27(2):191–208, 2009.

[31] S. J. Luck. *An Introduction to the Event-Related Potential Technique*.

[32] W. H. Maisel and T. Kohno. Improving the Security and Privacy of Implantable Medical Devices. *New England Journal of Medicine*, 362(13), 2010.

[33] 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 *the Proceedings of the 21st USENIX Security Symposium*, pages 143–158, 2012.

[34] P. Moulin and J. A. O'Sullivan. Information-theoretic Analysis of Information Hiding. *IEEE Transactions on Information Theory*, 49(3):563–593, 2003.

[35] C. Omar, A. Akce, M. Johnson, T. Bretl, R. Ma, E. Maclin, M. McCormick, and T. P Coleman. A Feedback Information-theoretic Approach to the Design of Brain-Computer Interfaces. *International Journal of Human-Computer Interaction*, 27(1):5–23, 2010.

[36] L. H. Ozarow. The Capacity of the White Gaussian Multiple Access Channel with Feedback. *IEEE Transactions on Information Theory*, 30(4):623–629, 1984.

[37] N. Paul and T. Kohno. Security Risks, Low-tech user Interfaces, and Implantable Medical Devices: A Case Study with Insulin Pump Infusion Systems. In *the Proceedings of the 3rd USENIX Conference on Health Security and Privacy*. USENIX Association, 2012.

[38] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn. Information Hiding-A Survey. *Proceedings of the IEEE*, 87(7):1062–1078, 1999.

[39] M. Shirow, F. Schoot, and T. Orzechowski. *Ghost in the Shell*. Dark Horse Comics, 2003.

[40] Charles Stross. *Elector*. Asimov's Science Fiction.

[41] The Committee on Science and Law. Are Your Thoughts Your Own?: 'Neuroprivacy' and the Legal Implications of Brain Imaging, 2005.

[42] Vernor Vinge. *The Peace War*. Tor Books, 2003.

[43] N. Wiener. *Cybernetics or Control and Communication in the Animal and the Machine*. John Wiley & Sons, 1948.

[44] N. Wiener. *Cybernetics of the Nervous System*, volume 17. Elsevier, 1965.

[45] N. Wiener. *The Human Use of Human Beings: Cybernetics and Society*. Da Capo Press, 1988.

[46] J. R. Wolpaw, N. Birbaumer, W. J. Heetderks, D. J. McFarland, P. H. Peckham, G. Schalk, E. Donchin, L. A. Quatrano, C. J. Robinson, and T. M. Vaughan. Brain-Computer Interface Technology: A Review of the First International Meeting. *IEEE Transactions on Rehabilitation Engineering*, 8(2):164–173, 2000.

[47] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan. Brain-Computer Interfaces for Communication and Control. *Clinical Neurophysiology*, 113(6):767–791, 2002.

[48] J. R. Wolpaw and E. W. Wolpaw. *Brain-Computer Interfaces: Principles and Practice*. OUP USA, 2012.

[49] A. D. Wyner. The Wire-Tap Channel. *Bell System Technical Journal*, 54(8):1355–1387, 1975.

[50] C.-M. Zeng, F. Kuhlmann, and A. Buzo. Achievability Proof of Some Multiuser Channel Coding Theorems Using Backward Decoding. *IEEE Transactions on Information Theory*, 35(6):1160–1165, 1989.