UNIVERSIDAD DE MURCIA

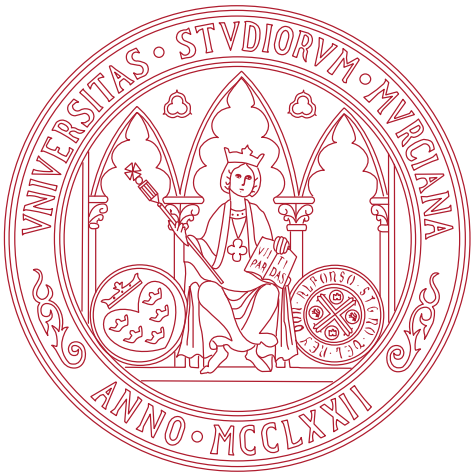
ESCUELA INTERNACIONAL DE DOCTORADO

Cybersecurity on Brain-Computer Interfaces

Ciberseguridad en Interfaces Cerebro-Máquina

D. Sergio López Bernal

2022



UNIVERSITY OF MURCIA

FACULTY OF COMPUTER SCIENCE

Cybersecurity on Brain-Computer Interfaces

Author

Sergio López Bernal

Thesis supervisors

Dr. **Alberto Huertas Celdrán**, *Ph.D.*
Dr. **Gregorio Martínez Pérez**, *Ph.D.*

Murcia, 2022

The following PhD Thesis is a compilation of the next published articles, being the PhD student the main author in all of them:

- Sergio López Bernal, Alberto Huertas Celdrán, Gregorio Martínez Pérez, Michael Taynnan Barros, Sasitharan Balasubramaniam. "**Security in Brain-Computer Interfaces: State-of-the-Art, Opportunities, and Future Challenges.**", *ACM Computing Surveys*, vol. 54, no. 1, pp. 35, 2021.
  DOI: 10.1145/3427376
  JIF 2021: 14.324 (D1)

- Sergio López Bernal, Alberto Huertas Celdrán, Lorenzo Fernández Maimó, Michael Taynnan Barros, Sasitharan Balasubramaniam, Gregorio Martínez Pérez. "**Cyberattacks on Miniature Brain Implants to Disrupt Spontaneous Neural Signaling.**", *IEEE Access*, vol. 8, pp. 152204-152222, 2020.
  DOI: 10.1109/ACCESS.2020.3017394
  JIF 2020: 3.367 (Q2)

- Sergio López Bernal, Alberto Huertas Celdrán, Gregorio Martínez Pérez. "**Neuronal Jamming cyberattack over invasive BCIs affecting the resolution of tasks requiring visual capabilities.**", *Computers & Security*, vol. 112, pp. 102534, 2022.
  DOI: 10.1016/j.cose.2021.102534
  JIF 2021: 5.105 (Q2)

- Sergio López Bernal, Alberto Huertas Celdrán, Gregorio Martínez Pérez. "**Eight Reasons Why Cybersecurity on Novel Generations of Brain-Computer Interfaces Must Be Prioritized.**", *Communications of the ACM*, 2022.
  DOI: 10.1145/3535509
  JIF 2021: 14.065 (D1)

Contents

Acknowledgements

As it could not be otherwise, I want to begin these words by thanking my parents for everything they have done for me. Thank you for the love and the great sacrifice you have always made so that we do not lack anything and allow me to be where I am today. To my father, Pedro, for teaching me since I was a child that effort and dedication are the keys to achieving whatever you set your mind. That any problem, no matter how complicated, always has a solution if we can look at it from the proper perspective.

To my mother, Mariví, for being there every time I have needed it and for the valuable advice you have always given me. If I am writing this acknowledgment today, it is thanks to the perseverance you have always transmitted to me and the many hours you have invested in helping me with my studies since I was a child.

To the rest of my family, for all the signs of affection and support during so many years. In particular, to my brother Eduardo for being there whenever I needed him and for being one of the most important people in my life. I would also like to thank my grandfather Marcial for all the support he has given me since I was a child and for the great advice he always has for any situation. Thank you for teaching me the value of generosity.

To my girlfriend, Bea, for believing in me and supporting me each and every day with your "come on, you can do it all", and your "make the most of it!". Thank you for your patience, for being with me in the worst moments, and for helping me make difficult decisions. Your love makes the days always better. I owe you a lot.

To my classmates at Rinka Koranshin Ryu, my second home, for all the experiences I have had. Thanks to Alex, my sensei, for all the lessons and advice and for transmitting to me that all our experiences are part of the path. Each and every conversation we have had in the last 15 years has made me a better person.

To Rafa, for your countless tips and for going out of your way to make time to talk about any topic. Your way of teaching made me interested in teaching and research from the very first moment. Thank you for recommending me to do my master's degree in France, for all the facilities during that year, and for encouraging me to do my PhD. To each and every one of my CyberDataLab colleagues, from all of you, I learn new things every day. Especially to Leo and Mattia for all the advice and help you have given me since the first day I entered Dibulibu. You have been my big brothers during all these years. To Javi, for all your words of encouragement and for knowing that I can count on you at any time. To Enrique and Mario, for your confidence in the team and in me. You have been my first students, and I hope I have been able to teach you as much as you have taught me. Here's to many more project trips together.

Last but not least, I would like to thank my thesis directors, Alberto and Gregorio, for everything you have done for me. You have supported my work and this line of research from the very beginning. You have been a reference not only for your tireless work and your full dedication to the team but also for your human side. You have taught me that one must fight for one's dreams and do what makes one happy in life. I am very fortunate to have learned so much from you.

To all of you, thank you. This thesis is as much yours as it is mine.

Agradecimientos

Como no podría ser de otra forma, quiero comenzar estas palabras agradeciendo a mis padres todo lo que han hecho por mí. Gracias por el cariño y el gran sacrificio que siempre habéis hecho para que no nos faltase de nada y permitir que hoy esté donde estoy.

A mi padre, Pedro, por enseñarme desde pequeño que el esfuerzo y la dedicación son la clave para conseguir lo que uno se proponga. Que cualquier problema, por complicado que sea, siempre tiene una solución si somos capaces de mirarlo desde la perspectiva correcta.

A mi madre, Mariví, por estar ahí cada vez que lo he necesitado y por los valiosos consejos que siempre me has proporcionado. Si estoy hoy escribiendo este agradecimiento es gracias a la constancia que siempre me has transmitido y a las tantas y tantas horas que has invertido a ayudarme con los estudios desde pequeño.

Al resto de mi familia, por todas las muestras de cariño y apoyo durante tantos años. En especial, a mi hermano Eduardo por estar ahí siempre que lo he necesitado y por ser una de las personas más importantes de mi vida. También agradecer a mi abuelo Marcial todo el apoyo que me ha brindado desde pequeño y los grandes consejos que siempre tiene para cualquier situación. Gracias por enseñarme el valor de la generosidad.

A mi novia, Bea, por creer en mí y apoyarme todos y cada uno de los días con tus "¡venga, que tú puedes con todo!" y tus "¡aprovecha!". Gracias por tu paciencia, por estar conmigo en los peores momentos y por ayudarme a tomar decisiones difíciles. Tu cariño hace que los días sean siempre mejores. Te debo mucho.

A mis compañeros de Rinka Koranshin Ryu, mi segunda casa, por todas las experiencias vividas. Gracias a Alex, mi sensei, por todas las lecciones y consejos, y por transmitirme que todas nuestras experiencias forman parte del camino. Todas y cada una de las conversaciones que hemos tenido en los últimos 15 años me han hecho ser una mejor persona.

A Rafa, por tus incontables consejos y por hacer lo imposible para sacar un rato y charlar de cualquier tema. Tu forma de dar clase hizo que me interesase por la docencia y la investigación desde el primer momento. Gracias por recomendarme hacer el máster en Francia, por todas las facilidades durante ese año, y por animarme a hacer el doctorado.

A todos y cada uno de mis compañeros del CyberDataLab, de todos vosotros aprendo cosas nuevas a diario. En especial, a Leo y Mattia por todos los consejos y ayuda que me habéis brindado desde el primer día que entré a Dibulibu. Habéis sido mis hermanos mayores durante todos estos años. A Javi, por todas tus palabras de ánimo y por saber que puedo contar contigo en cualquier momento. A Enrique y Mario, por vuestra confianza en mí y en el equipo. Habéis sido mis primeros alumnos y espero haber podido enseñaros tanto como vosotros me habéis enseñado a mí. Por muchos más viajes de proyecto juntos.

Finalmente, y no por ello menos importante, quiero agradecer a mis directores de tesis, Alberto y Gregorio, todo lo que habéis hecho por mí. Habéis apostado por mi trabajo y por esta línea de investigación desde el primer momento. Habéis sido referentes no sólo por vuestro trabajo incansable y vuestra dedicación plena al equipo, sino también por vuestra faceta humana. Me habéis enseñado que uno debe luchar por sus sueños y hacer lo que le haga feliz en la vida. Soy muy afortunado por haber podido aprender tanto de vosotros.

A todos vosotros, gracias. Esta tesis es tan vuestra como mía.

Abstract

I Introduction and motivation

Brain-Computer Interfaces (BCIs) are promising systems that enable the interaction between the brain and external devices to acquire neural data or perform neurostimulation actions. Specifically, they aim to measure the status of neurons in terms of their activation (known as an action potential or spike) or to stimulate these neurons to have a particular behavior. Since their creation in the decade of 1970, BCIs have been mainly used in medicine, undergoing a revolution in the 21st century due to new findings in neuroscience. In these scenarios, BCIs are employed for two tasks: medical diagnostics and neurostimulation. Focusing on the first one, BCIs are extremely useful for detecting and evaluating a wide range of neurological conditions, such as epilepsy [1], sleep disorders [2], or anxiety [3]. Additionally, these systems are widely utilized for neuroimaging, where techniques like magnetic resonance allow the visualization of the brain to identify lesions or tumors.

Regarding neurostimulation, BCIs are a promising alternative for specific conditions and diseases when the traditional approach based on the administration of drugs is not effective [4]. Neurostimulation has been proved safe for treating epilepsy, Parkinson's disease, essential tremor, obsessive-compulsive disorder, and dystonia, having clearance from medical organizations such as the FDA in the United States [5, 6]. Furthermore, new conditions and diseases are under research nowadays for their treatment with BCIs, being the case of Alzheimer's disease [7]. Apart from these two main uses, BCIs are successfully utilized to control external devices such as wheelchairs, prosthetic limbs, and exoskeletons, improving the quality of life of rehabilitation patients [8]. Furthermore, BCI technologies can improve cerebral plasticity, memory, reaction ability, or concentration, allowing cognitive improvement in their users.

In the last few years, the expansion and development of these interfaces have reached other sectors outside the medical scenario. There are several reasons for this situation, being the most relevant a reduction of cost and size of technology, an improvement in hardware and software capabilities, better access to technology from end-users, and the application of technology and artificial intelligence to almost any sector. Thanks to these advances, BCIs have gained popularity in entertainment, where users can mentally interact with multimedia systems, such as controlling the volume of a film or changing the TV channel. Moreover, video games are one of the most promising areas for applying BCIs since their combination with virtual reality could control the avatar of the game with the mind, improving the immersive experience [9]. BCIs will also play an essential role in the

metaverse, where users could not only control an avatar but physically feel sensations that occur within the simulation. Apart from recreational uses, BCIs are extremely valuable in marketing research, where these systems help identify the impact that advertisement campaigns have on users from a cognitive and emotional perspective [10]. Moreover, since brain waves are unique for each person, BCIs are also interesting for building robust authentication systems based on thoughts while performing particular tasks, such as visualizing images, imaging limb movements, or mentally recreating a specific song [11].

Based on this technological trend, BCIs are potentially considered Internet of things devices as it is expected that humans will communicate with their minds in the near future. In this direction, futuristic paradigms such as Brain-to-Internet (BtI) and Brain-to-Brain (BtB) communications are expected. The first one involves directly accessing the Internet using a BCI [12], while the latter aims to enable direct communication between brains [13]. An evolution of BtB is brainets, networks of brains that could directly and telepathically communicate information [14]. Although these initiatives are prospecting and particularly ambitious, these directions are being extensively explored in the literature with promising results, indicating that they could be a reality in the following decades. Based on that, numerous companies are also focusing on advancing neurotechnology from both acquisition and neurostimulation perspectives, being an economic sector full of opportunities.

Apart from the separation of BCIs in data acquisition and neurostimulation dimensions, they can also be classified according to their invasiveness. Thus, non-invasive technologies for neural data acquisition are the most common for both medical diagnostics and non-medical scenarios. In this category, electroencephalography (EEG) is the most used due to its simplicity based on electrodes placed on the scalp, portability, reduced cost, and high temporal resolution, although it presents a limited spatial resolution [15]. Magnetic resonance is also included in this category, widely used in hospital diagnostics due to its good spatial resolution, but it presents limited temporal resolution. Additionally, certain medical scenarios require the study of specific neuronal populations with both high temporal and spatial resolutions, typically using invasive techniques such as electrocorticography (ECoG). However, invasive technologies introduce a risk of tissue damage and infection that need to be carefully considered [1].

Focusing on invasive neurostimulation, Deep Brain Stimulation (DBS) [4, 5] and Responsive Neurostimulation (RNS) [16] are the most popular due to their efficacy and safety, having both clearance from the FDA. The former is used to treat neurological conditions such as Parkinson's disease or essential tremor, while the latter is focused on epilepsy. Despite the advantages and benefits these technologies provide, they have considerable limitations. Specifically, they stimulate quite broad areas of the brain, being unable to target individual neurons or even small neuronal populations. Additionally, they are used for particular treatments, being difficult to extend them to other uses based on their inner mechanisms and functioning.

Taking into consideration the limitations of contemporary invasive neurostimulation technologies, new systems have been proposed in the last few years, aiming to target the brain with better temporal and spatial resolution. One of the most relevant initiatives is under development by Neuralink, which aims to provide BCI systems to record and stimulate the brain with a single-neuron resolution using nanoscale electrodes [17]. Additionally, they presented a conceptualization of a wearable system placed on the skull that could be controlled with a smartphone, intending to democratize the access to neurotechnology to the general public. To ease its implantation, the project developed a robot capable of inserting miniaturized electrodes into the brain, minimizing the risk of tissue damage by precisely identifying blood vessels and, thus, determining their best placement. The project

has successfully tested its prototype in pigs and monkeys, highlighting the feasibility of this system.

Besides Neuralink, other systems present interesting features for surpassing the current limitations of invasive neurostimulation technologies. Wireless Optogenetic Nanonetworks (WiOptND) consists of nanodevices implanted in the cerebral cortex (neural dust) that emit light pulses to genetically engineered neurons receptive to these stimuli. This approach permits targeting a tiny population of neurons, allowing their stimulation or external inhibition [18]. Albeit these systems propose interesting functionality for surpassing the limitations of current neurostimulation, they represent concepts and prototypes that need to be evolved in the next years. In this direction, current BCI development tends to create invasive devices with fewer risks for users' heath, the production of wireless devices, improvements of connectivity by linking them to the Internet, a reduction of their size and price, and a better temporal and spatial resolution.

Although these advances envision a future where BCIs would improve human abilities and treat neurological diseases, they also introduce enormous cybersecurity concerns. From the prism of neural data acquisition, several works identified and verified particular cybersecurity issues, as is the case of Martinovic et al. [19]. They documented that attackers presenting malicious visual stimuli to BCI subjects could obtain sensitive information such as bank-related data, living area, emotions, sexual orientation, or religious beliefs. Similarly, Frank et al. [20] presented malicious visual stimuli to subjects, indicating that images not perceptible by the subject (subliminal) could also have a confidentiality impact on BCI users. However, most works in the literature address cybersecurity on BCI from a theoretical perspective, identifying potential risks that cyberattackers could exploit [21, 22, 23, 24]. Nevertheless, it is essential to note that these works are scarce and focus on just particular aspects of the BCI life cycle, with no works performing a comprehensive analysis of cybersecurity aspects of BCIs.

Moving to neurostimulation, literature has focused on cybersecurity applied to implantable medical devices (IMD), identifying risks and possible attacks over neurostimulators [25, 26]. Although these works indicate some potential impacts on the brain, they are quite generic and do not delve into the particularities of the neurological domain. Besides, the development of next-generation neurostimulation systems introduces alarming concerns. First, the generalization of this kind of technology, which would be accessible by the general population, could be an incentive for cyberattackers due to the potential benefit they could obtain in terms of sensitive data. Additionally, the possibility to cause remote harm to BCI users, as is the case of traditional computer systems and networks, could be leveraged by criminals aiming to attack determined public personalities or even the whole country population in terrorist scenarios.

Based on the above, there is an opportunity for works performing a comprehensive analysis of cybersecurity on BCIs, studying each particular BCI technology, the design and implementation of the BCI cycle, and the different application scenarios of these technologies, both existing and prospecting. Furthermore, there are open challenges in neurostimulation systems, where analysis of novel neurostimulation technologies is lacking in the literature. Additionally, there is an opportunity for research addressing possible attacks and impacts of these novel technologies.

Considering these limitations, there is also an opportunity for cyberattacks to affect spontaneous neuronal signaling, which is defined as the neuronal activity that occurs in the brain while no attack is performed. One of these possibilities is targeting specific neurons from individuals using BCIs capable of neural recording and stimulation. Thus, they could stimulate or inhibit neurons of certain cerebral regions to alter spontaneous neural activity,

even executing particular stimulation patterns. This situation is extremely sensitive since attackers targeting a broad coverage of the brain could potentially recreate the effects and behavior of neurodegenerative diseases, causing a fatal impact on users. In addition, the development of these cyberattacks could serve to gain a deeper understanding of the brain and neurodegenerative diseases, contributing to medical research.

Based on the previous considerations, this PhD Thesis first focuses on providing the current state of cybersecurity applied to BCIs. Moreover, this work explores the feasibility of affecting spontaneous neural activity by performing cyberattacks over neurostimulation BCIs, also assessing the impact that they could cause on the brain. In this direction, several research questions arose from the previous challenges, guiding the research process of this PhD Thesis, and are presented as follows:

- RQ1: What is the current status of cybersecurity on BCIs for neural data acquisition and neurostimulation?
- RQ2: What types of cyberattacks and malicious behaviors can affect neural activity and how perform them using BCI systems?
- RQ3: How can neural cyberattacks be tested on a realistic neurological scenario?
- RQ4: What metrics are useful for measuring and comparing the impact caused by neural cyberattacks?

II Objectives

The main goal of this PhD Thesis consists in investigating cybersecurity aspects of BCIs, identifying cyberattacks applicable to different dimensions relevant to BCI, the impact they cause, and possible countermeasures to mitigate them. Additionally, this work aims to study the feasibility of cyberattacks aiming to stimulate or inhibit specific neurons of BCI users in a particular way, analyzing the impact they could cause on spontaneous neural signaling. From this objective, several specific goals are derived as subsequently presented, indicating the research questions related to them:

1. Analyze the current state of the art regarding cybersecurity on BCIs for neural data acquisition and neurostimulation, studying applicable attacks, the impacts that they could cause, and possible countermeasures to reduce or mitigate these impacts (RQ1).
2. Identify vulnerabilities in existing and next-generation neurostimulation technologies that cyberattackers could exploit to cause brain damage to BCI users (RQ1).
3. Propose a taxonomy of neural cyberattacks focused on altering the spontaneous behavior of cerebral activity (RQ2).
4. Implement a set of neural cyberattacks in a biological neural simulator, using a neuronal topology as realistic as possible (RQ3).
5. Define a set of metrics specific to the neuroscience domain based on the analysis of neuronal activity to evaluate the impact caused by neural cyberattacks (RQ4).
6. Analyze the impact that neural cyberattacks could cause on spontaneous neural activity and potentially relate them with the effect of neurodegenerative diseases (RQ4).

III Methodology

This PhD Thesis was conducted following a scientific approach based on the continuous study of the state of the art and the analysis of the results obtained during the different stages of the research. This thesis is defined as a set of four papers published in high-impact journals indexed in the Journal Citation Reports (JCR).

To accomplish its first objective and offer a response to the first research question, this PhD Thesis reviewed the background regarding essential concepts of neuroscience. Additionally, we reviewed relevant aspects of BCIs, their life cycle, their application to different scenarios, and common classifications of these interfaces. After that, we analyzed the state of the art of cybersecurity applied to BCI systems. For that, we first studied the different definitions and versions of the BCI life cycle, both from a neural data acquisition and a neurostimulation perspective, offering an standardized version sufficiently general that could cover any implementation of BCI systems. After that, we identified the applicability of potential cyberattacks over the stages of the BCI cycle and different architectural deployments of BCIs, an analysis of their impact, and a list of possible countermeasures to mitigate these impacts. Finally, trends and future challenges were identified, motivating the development of subsequent publications. All these considerations resulted in the first publication of this PhD Thesis, presented in the first chapter (Survey of Cybersecurity on Brain-Computer Interfaces (Article 1–`ACM_CSUR`)).

After performing the state of the art analysis and identifying the current cybersecurity problems in BCI scenarios, we analyzed potential vulnerabilities in next-generation neurostimulation implants, particularly in Neuralink, neural dust, and wireless optogenetic nanonetworks. Based on this study, we concluded the possibility of performing cyberattacks against these devices to take control over their actions and thus stimulate or inhibit neurons individually (see RQ2). This analysis is aligned with the second objective of the PhD Thesis, as previously presented. Based on these vulnerabilities, the second publication of this PhD Thesis, available in the second chapter of this document, Neuronal Flooding and Neuronal Scanning Cyberattacks (Article 2–`IEEE_Access`), defined the concept of neural cyberattacks as threats able to alter spontaneous neural activity. It also formally presented two neural cyberattacks, Neuronal Flooding (FLO) and Neuronal Scanning (SCA), in charge of performing malicious neurostimulation tasks. These cyberattacks were selected since they represent distinct approaches to affecting neurons by overstimulation, although other approaches are possible, as presented in the last chapter of this thesis. To implement and validate these attacks, we opted for using a neural simulator, Brian2 [27], able to recreate the behavior of neurons as realistically as possible, using the Izhikevich model [28], a neuronal model widely used in neuroscience. This development aligns with the fourth goal of the thesis.

At this point, a limitation in the research line arose. At the moment of elaborating the publication, there was a lack of realistic neuronal topologies modeling the distribution between layers of the cerebral cortex and the connections between neuronal populations. To face this limitation, this thesis had to search for alternatives to model neural activity in a realistic way, as close to the biological scenario as possible. Due to this, we opted for training a Convolutional Neural Network (CNN) [29] in charge of solving the specific problem of a mouse that must find the exit of a particular maze. This decision was justified by existing literature indicating the similarities that CNNs and the visual cortex present in their structure and function. After training, the connections between neurons and their weights were translated into biological terms and introduced into the simulator. Furthermore, the mouse's current position was also used as input to the neuronal model

to simulate what the mouse saw in each moment, differentiating between available cells and walls of the maze. Besides, the model resulting from training the CNN provided the optimal path to exit the maze from any position. Based on that, we only considered the optimal path to reach the exit from the starting cell of the maze to be included in the neuronal simulation, having a simplification of the problem. These considerations intend to answer RQ3.

We tested different numbers of neurons under attack and voltages used to stimulate those neurons for the implementation and subsequent evaluation of these cyberattacks. After implementing both neural cyberattacks in the simulator, we defined three metrics to measure their impact aiming to offer a response to RQ4, also aligning with the fifth objective. First, it is essential to define the concept of a spike, or action potential, as the activation of a neuron and the transmission of the stimuli to subsequent neurons. The first metric, the number of spikes, measured if an attack augmented or reduced the number of action potentials performed by the neurons compared to the spontaneous situation. The second metric, the percentage of shifts, indicated the delay of a spike over time, either forward or backward, compared to the spontaneous case. The dispersion of spikes, measured both in the dimension of time and number of spikes, is the third metric defined and consisted in analyzing the spike patterns to identify changes in their distribution, observing the evolution of the dispersion along the optimal path. Finally, after studying the impact of each attack individually using these metrics, we compared the results between attacks, following the last objective.

Once we verified the effectiveness of FLO and SCA using a neuronal simulator, we defined a new neural cyberattack, Neuronal Jamming (JAM), based on the inhibition of neuronal activity during a temporal windows. Thus, the third publication of this PhD Thesis, presented in the third chapter (Neuronal Jamming Cyberattack (Article 3–`Elsevier_COSE`)), used the same scenario and experimental configuration based on a CNN to implement this cyberattack in Brian2. In contrast to previous work, this publication intended to analyze if there was any relationship between the impact caused by neural cyberattacks on neuronal activity (particularly FLO and JAM) and the impact on the mouse's decision-making ability, assuming that these attacks affect the visual capabilities of the animal. To validate this objective, we first offered a formal description of JAM, followed by an analysis of the impact caused by this cyberattack from a biological perspective using neuronal simulations. After that, we evaluated the CNN model used to build the biological neuronal topology, aiming to determine how JAM could affect the mouse's ability to find the maze exit.

We also evaluated the impact of applying FLO cyberattacks to this scenario. From the biological perspective, the difference with the second chapter of the PhD Thesis is that, in that work, we performed the attack in a particular instant at the beginning of the simulation, and we evaluated its propagation. In contrast, in this third publication, we separately applied an attack in each position of the optimal path, studying the evolution of the impact from both the number of spikes and temporal dispersion metrics. Additionally, we studied the effect of FLO over the artificial network, attending to both the number of steps to reach the exit and the percentage of times the mouse found the exit. For that, we analyzed the impact of the attack when the mouse was placed in each individual position of the optimal path, calculating from that position the performance to exit the maze. As in the case of the biological approach, we obtained the Pearson correlation between variables to understand the relationship between the scenarios. Finally, we compared the results of JAM and FLO, also analyzing the relationship these neural cyberattacks could have on the effects caused by neurodegenerative diseases.

The last work done in the PhD Thesis, presented in the fourth chapter of this document (Taxonomy of Neural Cyberattacks (Article 4–ACM_CACM)) and aligned with the third objective, presented a taxonomy of eight neural cyberattacks comprising stimulation and inhibition of neuronal activity. This work was motivated by a need to propose new neural cyberattacks and offer a categorization of them, according to RQ2. Three of these attacks were already presented in previous publications, being the remaining five novels. For each of these eight cyberattacks, we presented the steps followed by the attack in the proposed implementation to illustrate their functioning better. After that, we individually compared the impact of each neural cyberattack with the spontaneous behavior. Finally, this work contrasted the effect produced between attacks based on the number of spikes metric, studying the damage caused during the first and last five positions of the optimal path. This study aimed to understand the impact induced by these attacks in the short and long term.

In summary, this thesis first reviewed the state of the art of cybersecurity on BCIs, followed by the identification of vulnerabilities in next-generation neurostimulation BCIs. Additionally, this work proposed the definition and implementation of different neural cyberattacks aiming to measure their impact. This methodology allowed for meeting the objectives defined in the thesis, previously presented in Section II.

IV Results

In the first publication of the PhD Thesis, available in (Article 1–ACM_CSUR), we proposed the first standardization of the BCI life cycle, both from neural data acquisition and neurostimulation perspectives, sufficiently general that could cover any implementation of BCI systems. After that, we analyzed potential cybersecurity attacks that could be applied to each stage of the BCI cycle from both approaches, identifying that common cyberattacks applicable to traditional computer systems, such as replay attacks, spoofing attacks, jamming attacks, or malware, could apply to all stages of the BCI cycle. We considered four dimensions to analyze the impacts caused by these cyberattacks: data and service integrity, data confidentiality, data and service availability, and BCI users' safety. Additionally, both countermeasures from the literature and suggested by this work were documented for all attacks to reduce or mitigate the previously presented impacts.

We also analyzed cybersecurity aspects that could affect different architectural deployments of the BCI cycle. For each deployment, we presented a description, a series of examples to better illustrate the concepts, an analysis of cyberattacks that could affect these architectures, and the impact they could cause. In particular, we identified possible cyberattacks impacting the BCI, the device controlling the BCI, or the cloud architecture used to manage users' data. Besides, this work provided a substantial set of potential countermeasures to mitigate the effects of these attacks.

This work was valuable in identifying the trend of current BCI systems, which are moving to BtI and BtB approaches. The goal in these scenarios is to use BCI technologies to interact with other devices, the Internet, and even allow direct communication between brains. However, BCI systems present limitations that will determine their evolution. First, we detected a lack of interoperability between BCI deployments since there is an absence of standards that make it difficult for companies to produce devices compatible with each other. Moreover, their functionalities are difficult to extend as they are manufactured for use in particular application scenarios, complicating the introduction of new cybersecurity capabilities. There is also a lack of data protection mechanisms or regulations in these scenarios, essential for ensuring the correct treatment of health-related sensitive data.

Finally, cybersecurity mechanisms in these systems are missing, requiring an effort to create devices that protect the sensitive data transmitted and the physical integrity of their users. All these previous aspects aim to offer an answer to RQ1.

The second publication (Article 2–IEEE_Access) first identified vulnerabilities in the architecture of prospecting neurostimulation solutions that could allow cyberattackers to control the system and perform malicious actions. For example, an attacker aiming to affect the Neuralink architecture could exploit vulnerabilities in the smartphone connected to the implanted system. Since there are many vulnerabilities and attacks to disrupt these mobile devices, taking control of the smartphone in charge of managing the BCI is feasible. Moreover, the link, an intermediary device between the smartphone and the implanted components, placed under the ear, uses a Bluetooth link that is also susceptible to firmware modification or jamming attacks, among other threats.

Motivated by the previous vulnerabilities, this publication presented two neural cyberattacks: Neuronal Flooding (FLO) and Neuronal Scanning (SCA). Although both cyberattacks stimulate a random set of neurons, FLO aims to stimulate neurons in a determined instant while SCA targets the set of neurons individually and sequentially, avoiding repetitions. Regarding their impact on spontaneous neural behavior (see RQ2), FLO reduced the number of spikes, a difference that increased when the mouse progressed in the maze. Moreover, augmenting the number of neurons under attack generated a more significant decrease in the number of spikes. We also concluded that changing the voltage used to overstimulate the neurons did not significantly impact the metric. Observing the different topology layers, the variation in the mean of spikes was more significant in deeper layers. Attending to the percentage of shifts metric, attacking a higher number of neurons generated a higher percentage of shifts while increasing the voltage had a negligible effect. Finally, regarding the dispersion metric, the temporal dispersion increased compared to spontaneous behavior. Focusing on the dispersion of the number of spikes, the attack generated in the last positions of the optimal path more instants where only one spike occurred, indicating more dispersion as the simulation advanced. These results indicate that FLO can effectively alter spontaneous neural activity, covering the fifth objective of the thesis, as well as offering partial responses for RQ3 and RQ4.

SCA reduced the number of spikes compared to the spontaneous signaling. Moreover, the impact was slightly increased when augmenting the voltage used to attack, but only for low voltages. Thus, and similarly to FLO, the impact of the voltage is negligible. This cyberattack also raised the percentage of spike shifts, degrading the impact when observing deeper layers. Additionally, we observed significant differences in the dispersion metrics compared to the spontaneous behavior. Finally, it is interesting to note that the impact got more aggravating when the mouse progressed in the maze, highlighting the incremental behavior of this cyberattack. Attending the comparison in terms of impact between FLO and SCA, we concluded that the inner mechanisms of each attack generate different behaviors in neuronal activity. FLO is better for altering neural activity in a short period since it affects multiple neurons in a particular instant. In contrast, SCA is more effective in the long term, requiring more time to generate a considerable impact, but after that, the impact is greater than FLO.

In the third publication of the thesis (Article 3–Elsevier_COSE), we presented Neuronal Jamming (JAM) as a neural cyberattack focused on inhibiting the activity of a set of neurons for a determined duration, inspired by neurodegenerative diseases consisting in neuronal malfunction or death, such as Parkinson's and Alzheimer's. This work naturally arose as a continuation of the previous publication with the goal of measuring the impact of inhibition-based cyberattacks, in contrast to previous work focused on stimulation of

neurons. The analysis of JAM from a biological perspective indicated that increasing the number of neurons under attack decreased both the number of spikes and the temporal dispersion. Additionally, we observed an increment in the distribution variability of these metrics when increasing the number of consecutive positions attacked, especially in the number of spikes. From the artificial network, we observed that even attacking a few random nodes dramatically increased the number of steps, not being able to exit the maze in most situations. Comparing the Pearson correlation between biological and artificial metrics, we obtained a low correlation of around 60%. This result was explained by the restrictions on the experimental considerations presented in the article. However, the individual analysis per scenario demonstrated the high impact that JAM presents.

After that, we compared the impact of JAM and FLO cyberattacks. In this context, we first analyzed the individual impact of FLO over both scenarios. In the biological one, the results indicated that performing the attack in later positions had less impact since the neuronal activity remained unaltered most of the time. Additionally, targeting a higher number of neurons generated greater damage. In the artificial network, augmenting the number of nodes under attack increased the impact until a certain position. After that, and since the mouse was closer to the exit cell, the impact decreased as the mouse could find the exit by probability. Comparing both scenarios for FLO, we obtained a correlation of around 80% between the number of steps and the number of spikes and dispersion, concluding a significant relationship between scenarios. Finally, we compared the results of both attacks. As the methodology between attacks differs in this publication, we focused on studying the correlations obtained. Thus, we appreciated a closer relationship between both approaches in FLO but considering the previously stated limitations. This analysis of the impact caused by neural cyberattacks aligns with RQ2.

The last publication of the PhD Thesis (Article 4–ACM_CACM) presented the definition and implementation of a taxonomy of neural cyberattacks, related to RQ4. This work naturally extended the set of neural cyberattacks already presented in the previous two publications of the thesis. Focusing on the novel attacks presented in this work, Neuronal Selective Forwarding (FOR) consists in sequentially inhibiting neurons without repetitions along time, while Neuronal Spoofing (SPO) exactly replicates the activity recorded in a previous temporal window. Neuronal Sybil (SYB) forces a neuron to have the opposite voltage within the natural voltage range of a neuron. In contrast, Neuronal Sinkhole (SIN) consists in stimulating neurons from early cortical layers aiming to affect a particular neuron located in a deeper layer. Finally, Neuronal Nonce (NON) aims to attack a set of neurons in a given instant, deciding randomly for each one to stimulate or inhibit.

This work depicted their behavior, generating an intuition of their dynamics. After that, we empirically measured the impact of the eight cyberattacks on spontaneous activity by attending to the number of spikes metric. Particularly, it studied the impact of the first and last five positions of the optimal path of the maze to highlight which were more harmful in the short and long term. Attending to the short term, NON achieved an approximate 12% reduction, followed by JAM with a 5%. Oppositely, SCA was the most damaging in the long term, offering a reduction of around 9% of spikes, followed by NON with 8%.

V Conclusions and future work

In the last decades, the rapid evolution of BCIs has generated a considerable advance in medicine, allowing better detection of various neurological diseases. They also provide neurostimulation capabilities to treat diseases like Parkinson's when a drug-based treatment results ineffective. This evolution has made them gain popularity in other sectors such

as entertainment or video games. These systems are being investigated as well for their connection to the Internet or even for allowing direct communication between brains. This advance opens a landscape of opportunities for new companies and ideas to dominate a rising sector aiming to reach the general population in the following decades.

Thanks to this variability in application scenarios, there is a wide variety of BCI technologies focusing either on neural data acquisition or neurostimulation, also differentiated based on their cerebral invasiveness. Focusing on invasive neurostimulation BCIs, current techniques with FDA approval for medical purposes are scarce and present limitations, such as having a reduced spatial resolution or being limited to particular diseases and brain regions. Based on that, next-generation BCIs aim to miniaturize electrodes and technology to enable joint neural data recording and stimulation and inhibition of neural activity. Their ultimate goal is to democratize BCI technologies and bring them closer to end-user consumers, separating them from medical scenarios.

However, the previous BCI technologies have not been conceived with the prism of cybersecurity in mind. In particular, these interfaces lack specific standards and regulations, making it difficult to unify the security mechanisms required for their commercial use. There are also no data protection regulations for ensuring the proper use of this sensitive information. Moreover, the trend of these interfaces focused on neurostimulation, in which companies such as Neuralink aim to democratize their access, could have a significant impact on users' safety.

Attending to the previous concerns and limitations, this PhD Thesis has analyzed the state of the art regarding cybersecurity on BCIs, detecting a lack of works addressing this topic. Although some works partially cover certain aspects of cybersecurity in this field, they are scarce and do not offer a comprehensive view of the problem. Based on that, this work first analyzes the attacks, impacts, and countermeasures for both the BCI life cycle and common architectural deployments for these systems. Additionally, this thesis identified trends and challenges that these systems will face in the near future. These findings have offered an answer for RQ1, also allowing to complete the first specific goal of the thesis.

After that, this work proposed the definition of neural cyberattacks as threats that can affect spontaneous neural activity, advancing the literature in terms of cybersecurity on BCIs. They are motivated by vulnerabilities identified in prospecting neurostimulation devices that attackers could exploit to cause harm to BCI users (see the second goal of the thesis). In this direction, this research first presented Neuronal Flooding and Neuronal Scanning as cyberattacks able to maliciously stimulate neurons, analyzing their impact on a neuronal simulation. Since, at that moment, there was a lack of realistic neuronal topologies, this thesis trained a CNN to solve the particular problem of a mouse that has to exit a particular maze, translating the resulting topology to a neuronal simulator. It was motivated by evidence presenting a relationship between some aspects of the functioning and structure of CNNs and the visual cortex. Both cyberattacks were effective in reducing neuronal activity. These results offered an answer to RQ3 and RQ4 and helped advance towards an answer for RQ2 for attacks based on neural stimulation.

With these results in mind, this thesis subsequently presented a third neural cyberattack, Neuronal Jamming, which inhibits the neuronal activity of a set of targeted neurons for a period of time. This work compared its impact with Neuronal Flooding, also considering their relationship with the decision-making ability of the mouse to exit the maze. The results obtained suggested a substantial correlation between the impact of these cyberattacks on neuronal activity and the ability to perform decisions, although further research is required. Based on these results, this work offered new findings for answering RQ2

regarding cyberattacks applying neural inhibition.

Finally, this research presented a taxonomy of eight neural cyberattacks, where five of them were novel. For each one, this thesis provided a definition, a description of their internal functioning, and an analysis of their impact on the short and long term. Based on that, this work indicated which were more suitable to cause an immediate effect and which caused more significant damage in the long term. Thus, these results answered RQ2 since they allowed measuring the impact caused by a broad set of behaviors of neural cyberattacks and helped complete all objectives of the thesis.

In summary, this PhD Thesis has first gathered the existing knowledge in the literature concerning cybersecurity on BCIs. Additionally, this work has substantially advanced the state of the art, proposing novel cyberattacks able to affect spontaneous neural activity, validating their impact in a scenario as realistic as possible to biological neural tissue.

As future work, this thesis first identifies the necessity to comprehensively analyze vulnerabilities existing in both current and prospecting BCI solutions, which will help develop practical cybersecurity solutions for specific products. Additionally, it is necessary to cover the challenges identified in terms of interoperability and extensibility of BCI solutions and fill current opportunities regarding data regulations and security mechanisms.

Moreover, this research detects the need to extend the analysis of neural cyberattacks, studying how other traditional cyberattacks from computer science could be adapted to the neurological scenario. Additionally, this thesis considers it essential to identify aspects of neurodegenerative diseases that could help widen this cybersecurity research area. Besides, this work identifies the necessity to evaluate the impact of neural cyberattacks over more realistic neuronal topologies. Thus, it would first allow measuring the differences between attacking excitatory or inhibitory neuronal populations. Moreover, the increase in the number of neurons and the complexity of the network would provide further conclusions about their effect on natural biological neuronal tissue.

Once a broad understanding of these cyberattacks is obtained, this work highlights an opportunity for detecting and mitigating these cyberattacks. For that, artificial intelligence, such as machine learning and deep learning techniques, could be useful for their implementation in novel generations of BCI devices, helping reduce or even mitigate the harm caused by these threats and even for prospecting ones.

A better intuition of the impact of neural cyberattacks in more realistic conditions could be vital to recreating the behavior and effect of known neurodegenerative diseases. Thus, certain cyberattacks could benefit the effects of particular conditions, establishing a relationship between cyberattacks and diseases. Furthermore, if this milestone is achieved, then research could focus on predicting, based on spontaneous neural activity, the presence of specific neurodegenerative diseases, even in the early stages. These advances could positively benefit medical research and have a massive impact on neurological patients.

Resumen

I Introducción y motivación

Las interfaces cerebro-máquina (BCIs) son sistemas prometedores que permiten la interacción entre el cerebro y dispositivos externos para adquirir datos neurológicos o realizar acciones de neuroestimulación. En concreto, su objetivo es medir el estado de las neuronas en términos de su activación (conocida como potencial de acción o *spike*) o estimular estas neuronas para que tengan un comportamiento determinado. Desde su creación en la década de 1970, las BCIs se han utilizado principalmente en medicina, sufriendo una revolución en el siglo XXI debido a los nuevos descubrimientos en neurociencia. En estos escenarios, las BCIs se emplean para dos tareas: el diagnóstico médico y la neuroestimulación. Centrándonos en la primera, las BCIs son extremadamente útiles para detectar y evaluar una amplia gama de condiciones neurológicas, como la epilepsia [1], los trastornos del sueño [2], o la ansiedad [3]. Además, estos sistemas son ampliamente utilizados para neuroimagen, donde técnicas como la resonancia magnética permiten la visualización del cerebro para identificar lesiones o tumores.

En cuanto a la neuroestimulación, las BCIs son una alternativa prometedora para condiciones y enfermedades específicas cuando el enfoque tradicional basado en la administración de fármacos no es efectivo [4]. La neuroestimulación ha demostrado ser segura para el tratamiento de la epilepsia, la enfermedad de Parkinson, el temblor esencial, el trastorno obsesivo-compulsivo y la distonía, contando con la autorización de organizaciones médicas como la FDA en Estados Unidos [5, 6]. Además, actualmente se están investigando nuevas afecciones y enfermedades para su tratamiento con BCIs, siendo el caso de la enfermedad de Alzheimer [7]. Aparte de estos dos usos principales, las BCIs se utilizan con éxito para controlar dispositivos externos como sillas de ruedas, prótesis y exoesqueletos, mejorando la calidad de vida de los pacientes de rehabilitación [8]. Además, las tecnologías BCI pueden mejorar la plasticidad cerebral, la memoria, la capacidad de reacción o la concentración, permitiendo la mejora cognitiva de sus usuarios.

En los últimos años, la expansión y el desarrollo de estas interfaces han llegado a otros sectores fuera del escenario médico. Son varias las razones que explican esta situación, siendo las más relevantes una reducción del coste y del tamaño de la tecnología, una mejora de las capacidades del hardware y del software, un mejor acceso a la tecnología por parte de los usuarios finales, y la aplicación de la tecnología y la inteligencia artificial a casi cualquier sector. Gracias a estos avances, las BCIs han ganado popularidad en el ámbito del entretenimiento, donde los usuarios pueden interactuar mentalmente con

los sistemas multimedia, como controlar el volumen de una película o cambiar el canal de televisión. Además, los videojuegos son una de las áreas más prometedoras para la aplicación de las BCIs, ya que su combinación con la realidad virtual podría permitir controlar el avatar del juego con la mente, mejorando la experiencia de inmersión [9]. Las BCIs también desempeñarán un papel esencial en el metaverso, donde los usuarios podrían no sólo controlar un avatar sino sentir físicamente las sensaciones que se producen dentro de la simulación. Aparte de los usos recreativos, las BCIs son extremadamente valiosas en la investigación de marketing, donde estos sistemas ayudan a identificar el impacto que las campañas publicitarias tienen en los usuarios desde una perspectiva cognitiva y emocional [10]. Además, dado que las ondas cerebrales son únicas para cada persona, las BCIs también son interesantes para construir sistemas de autenticación robustos basados en los pensamientos mientras se realizan tareas concretas, como visualizar imágenes, imaginar los movimientos de las extremidades o recrear mentalmente una canción específica [11].

Basándose en esta tendencia tecnológica, las BCIs se consideran potencialmente dispositivos del Internet de las cosas, ya que se espera que los humanos se comuniquen con sus mentes en un futuro próximo. En esta dirección, se esperan paradigmas futuristas como las comunicaciones *Brain-to-Internet* (BtI) y *Brain-to-Brain* (BtB). El primero implica acceder directamente a Internet utilizando una BCI [12], mientras que el segundo pretende permitir la comunicación directa entre cerebros [13]. Una evolución de BtB son las *brainets*, redes de cerebros que podrían comunicar información directa y telepáticamente [14]. Aunque estas iniciativas son prospectivas y particularmente ambiciosas, estas direcciones están siendo ampliamente exploradas en la literatura con resultados prometedores, lo que indica que podrían ser una realidad en las próximas décadas. En base a ello, numerosas empresas también se están centrando en el avance de la neurotecnología, tanto desde la perspectiva de la adquisición como de la neuroestimulación, siendo un sector económico lleno de oportunidades.

Aparte de la separación de las BCIs en las dimensiones de adquisición de datos y neuroestimulación, también pueden clasificarse según su carácter invasivo. Así, las tecnologías no invasivas para la adquisición de datos neuronales son las más comunes tanto para los diagnósticos médicos como para los escenarios no médicos. En esta categoría, la electroencefalografía (EEG) es la más utilizada debido a su simplicidad basada en electrodos colocados en el cuero cabelludo, portabilidad, coste reducido y alta resolución temporal, aunque presenta una resolución espacial limitada [15]. También se incluye en esta categoría la resonancia magnética, muy utilizada en el diagnóstico hospitalario por su buena resolución espacial, pero que presenta una resolución temporal limitada. Además, ciertos escenarios médicos requieren el estudio de poblaciones neuronales específicas con resoluciones temporales y espaciales altas, normalmente utilizando técnicas invasivas como la electrocorticografía (ECoG). Sin embargo, las tecnologías invasivas introducen un riesgo de daño de tejidos e infección que debe ser cuidadosamente considerado [1].

Centrándonos en la neuroestimulación invasiva, la estimulación cerebral profunda (DBS) [4, 5] y la neuroestimulación receptiva (RNS) [16] son las más populares debido a su eficacia y seguridad, teniendo ambas la autorización de la FDA. La primera se utiliza para tratar afecciones neurológicas como la enfermedad de Parkinson o el temblor esencial, mientras que la segunda se centra en la epilepsia. A pesar de las ventajas y beneficios que aportan estas tecnologías, tienen considerables limitaciones. En concreto, estimulan áreas bastante amplias del cerebro, siendo incapaces de centrarse en neuronas individuales o incluso en pequeñas poblaciones neuronales. Además, se utilizan para tratamientos particulares, siendo difícil extenderlos a otros usos en función de sus mecanismos internos y su funcionamiento.

Teniendo en cuenta las limitaciones de las tecnologías de neuroestimulación invasivas

contemporáneas, en los últimos años se han propuesto nuevos sistemas que pretenden cubrir el cerebro con una mejor resolución temporal y espacial. Una de las iniciativas más relevantes es la que está desarrollando Neuralink, cuyo objetivo es proporcionar sistemas BCI para obtener actividad neuronal y estimular el cerebro con una resolución de neurona individual utilizando electrodos a nanoescala [17]. Además, presentaron una conceptualización de un sistema *wearable* emplazada en el cráneo que podría ser controlado con un *smartphone*, con la intención de democratizar el acceso a la neurotecnología al público general. Para facilitar su implantación, el proyecto desarrolló un robot capaz de insertar electrodos miniaturizados en el cerebro, minimizando el riesgo de dañar los tejidos al identificar con precisión los vasos sanguíneos y, por tanto, determinar su mejor ubicación. El proyecto ha probado con éxito su prototipo en cerdos y monos, poniendo de manifiesto la viabilidad de este sistema.

Además de Neuralink, otros sistemas presentan características interesantes para superar las limitaciones actuales de las tecnologías de neuroestimulación invasiva. El sistema *Wireless Optogenetic Nanonetworks* (WiOptND) consiste en nanodispositivos implantados en la corteza cerebral (*neural dust*) que emiten pulsos de luz a neuronas genéticamente modificadas y receptivas a estos estímulos. Este enfoque permite dirigirse a una población diminuta de neuronas, permitiendo su estimulación o inhibición externa [18]. Aunque estos sistemas proponen una funcionalidad interesante para superar las limitaciones de la neuroestimulación actual, representan conceptos y prototipos que deben evolucionar en los próximos años. En esta dirección, el desarrollo actual de las BCIs tiende a la creación de dispositivos invasivos con menos riesgos para la salud de los usuarios, a la producción de dispositivos inalámbricos, a la mejora de la conectividad mediante su conexión a Internet, a la reducción de su tamaño y precio, y a una mejor resolución temporal y espacial.

Aunque estos avances vislumbran un futuro en el que las BCIs mejorarían las capacidades humanas y tratarían las enfermedades neurológicas, también introducen enormes problemas de ciberseguridad. Desde el prisma de la adquisición de datos neuronales, varios trabajos identificaron y comprobaron problemas particulares de ciberseguridad, como es el caso de Martinovic et al. [19]. Estos autores documentarion que los atacantes que presentaban estímulos visuales maliciosos a los sujetos de la BCI podían obtener información sensible, como datos relacionados con la banca, la zona en la que viven, emociones, orientación sexual o creencias religiosas. Del mismo modo, Frank et al. [20] presentaron estímulos visuales maliciosos a los sujetos, indicando que las imágenes no perceptibles por el sujeto (subliminales) también podrían tener un impacto de confidencialidad en los usuarios de BCI. Sin embargo, la mayoría de los trabajos en la literatura abordan la ciberseguridad en BCI desde una perspectiva teórica, identificando los riesgos potenciales que los ciberatacantes podrían explotar [21, 22, 23, 24]. Sin embargo, es fundamental señalar que estos trabajos son escasos y se centran sólo en aspectos particulares del ciclo de vida de BCI, no existiendo trabajos que realicen un análisis integral de los aspectos de ciberseguridad de las BCIs.

Pasando a la neuroestimulación, la literatura se ha centrado en la ciberseguridad aplicada a los dispositivos médicos implantables (IMD), identificando riesgos y posibles ataques sobre los neuroestimuladores [25, 26]. Aunque estos trabajos indican algunos impactos potenciales sobre el cerebro, son bastante genéricos y no profundizan en las particularidades del ámbito neurológico. Además, el desarrollo de sistemas de neuroestimulación de nueva generación introduce preocupaciones alarmantes. En primer lugar, la generalización de este tipo de tecnología, que sería accesible para la población en general, podría ser un incentivo para los ciberatacantes debido al potencial beneficio que podrían obtener en términos de datos sensibles. Además, la posibilidad de causar daño a distancia a los usuarios de BCI,

como es el caso de los sistemas y redes informáticas tradicionales, podría ser aprovechada por los delincuentes con el objetivo de atacar a determinadas personalidades públicas o incluso a toda la población de un país en escenarios terroristas.

En base a lo anterior, existe una oportunidad para la realización de trabajos de análisis exhaustivo de la ciberseguridad en las BCIs, estudiando cada tecnología BCI en particular, el diseño e implementación del ciclo de BCI, y los diferentes escenarios de aplicación de estas tecnologías, tanto los existentes como los potencialmente emergentes. Por otra parte, existen retos abiertos en los sistemas de neuroestimulación, donde el análisis de las nuevas tecnologías de neuroestimulación está ausente en la literatura. Además, se presenta una oportunidad para investigar los posibles ataques e impactos de estas nuevas tecnologías.

Teniendo en cuenta estas limitaciones, también existe una oportunidad para realizar ciberataques que afecten a la señalización neuronal espontánea, que se define como la actividad neuronal que se produce naturalmente en el cerebro mientras no se realiza ningún ataque. Una de estas posibilidades es focalizarse en neuronas específicas de los individuos utilizando BCIs capaces de leer y estimular las neuronas. Así, podrían estimular o inhibir neuronas de determinadas regiones cerebrales para alterar la actividad neuronal espontánea, incluso ejecutando patrones de estimulación particulares. Esta situación es extremadamente delicada, ya que atacantes con acceso a amplias zonas del cerebro podrían recrear potencialmente los efectos y el comportamiento de las enfermedades neurodegenerativas, causando un impacto fatal en los usuarios. Además, el desarrollo de estos ciberataques podría servir para conocer mejor el cerebro y las enfermedades neurodegenerativas, contribuyendo a la investigación médica.

Partiendo de las consideraciones anteriores, esta tesis doctoral se centra en primer lugar en proporcionar el estado actual de la ciberseguridad aplicada a las BCIs. Además, este trabajo explora la viabilidad de afectar a la actividad neuronal espontánea mediante la realización de ciberataques sobre BCIs de neuroestimulación, evaluando también el impacto que podrían causar en el cerebro. En esta dirección, varias preguntas de investigación surgieron de los retos anteriores, guiando el proceso de investigación de esta tesis doctoral, tal y como se presentan a continuación:

- RQ1: ¿Cuál es el estado actual de la ciberseguridad en BCIs para adquisición de datos neuronales y neuroestimulación?
- RQ2: ¿Qué tipos de ciberataques y comportamientos maliciosos pueden afectar a la actividad neuronal y cómo aplicarlos usando sistemas BCI?
- RQ3: ¿Cómo se podrían evaluar los ciberataques neuronales en escenarios neurológicos realistas?
- RQ4: ¿Qué métricas son útiles para medir y comparar el impacto causado por ciberataques neuronales?

II Objetivos

El objetivo principal de esta tesis doctoral consiste en investigar los aspectos de ciberseguridad de las BCIs, identificando los ciberataques aplicables a diferentes dimensiones relevantes para las BCIs, el impacto que causan y las posibles contramedidas para mitigarlos. Además, este trabajo pretende estudiar la viabilidad de los ciberataques dirigidos a estimular o inhibir neuronas específicas de los usuarios de BCI de forma particular, analizando el impacto que podrían causar en la señalización neuronal espontánea. De este

objetivo se derivan varias metas específicas que se presentan a continuación, indicando las preguntas de investigación relacionadas con ellas:

1. Analizar el estado del arte actual en materia de ciberseguridad en las BCIs para la adquisición de datos neuronales y la neuroestimulación, estudiando los ataques aplicables, los impactos que podrían causar y las posibles contramedidas para reducir o mitigar estos impactos (RQ1).

2. Identificar las vulnerabilidades en las tecnologías de neuroestimulación existentes y de próxima generación que los ciberatacantes podrían aprovechar para causar daños cerebrales a los usuarios de BCI (RQ1).

3. Proponer una taxonomía de ciberataques neurales centrados en la alteración del comportamiento espontáneo de la actividad cerebral (RQ2).

4. Implementar un conjunto de ciberataques neuronales en un simulador neuronal biológico, utilizando una topología neuronal lo más realista posible (RQ3).

5. Definir un conjunto de métricas específicas del ámbito de la neurociencia basadas en el análisis de la actividad neuronal para evaluar el impacto causado por los ciberataques neuronales (RQ4).

6. Analizar el impacto que los ciberataques neuronales podrían causar en la actividad neuronal espontánea y potencialmente relacionarlos con el efecto de las enfermedades neurodegenerativas (RQ4).

III Metodología

Esta tesis doctoral se ha realizado siguiendo un enfoque científico basado en el estudio continuo del estado del arte y el análisis de los resultados obtenidos durante las diferentes etapas de la investigación. Esta tesis se define como un conjunto de cuatro trabajos publicados en revistas de alto impacto indexadas en el *Journal Citation Reports* (JCR).

Para cumplir su primer objetivo y ofrecer una respuesta a la primera pregunta de investigación, esta tesis doctoral revisó los antecedentes relativos a conceptos esenciales de la neurociencia. Además, revisamos aspectos relevantes de las BCIs, su ciclo de vida, su aplicación a diferentes escenarios y las clasificaciones comunes de estas interfaces. Después, analizamos el estado del arte de la ciberseguridad aplicada a los sistemas BCI. Para ello, primero estudiamos las diferentes definiciones y versiones del ciclo de vida de las BCIs, tanto desde la perspectiva de la adquisición de datos neuronales como de la neuroestimulación, ofreciendo una versión estandarizada lo suficientemente general como para cubrir cualquier implementación de sistemas BCI. A continuación, se identificó la aplicabilidad de los posibles ciberataques a lo largo de las etapas del ciclo de BCI y de los diferentes despliegues arquitectónicos de las BCIs, un análisis de su impacto y una lista de posibles contramedidas para mitigar estos impactos. Por último, se identificaron las tendencias y los retos futuros, lo que motivó el desarrollo de publicaciones posteriores. Todas estas consideraciones dieron lugar a la primera publicación de esta tesis doctoral, presentada en el primer capítulo (Article 1–ACM_CSUR).

Después de realizar el análisis del estado del arte e identificar los problemas actuales de ciberseguridad en los escenarios de BCI, analizamos las posibles vulnerabilidades en los implantes de neuroestimulación de próxima generación, especialmente en Neuralink, *neural dust* y las *wireless optogenetic nanonetworks*. Basándonos en este estudio, concluimos la

posibilidad de realizar ciberataques contra estos dispositivos para tomar el control de sus acciones y así estimular o inhibir neuronas de forma individual (ver RQ2). Este análisis está alineado con el segundo objetivo de la tesis doctoral, presentado anteriormente. En base a estas vulnerabilidades, la segunda publicación de esta tesis doctoral, disponible en el segundo capítulo de este documento, Article 2–IEEE_Access, definió el concepto de ciberataques neuronales como amenazas capaces de alterar la actividad neuronal espontánea. También presentó formalmente dos ciberataques neuronales, Neuronal Flooding (FLO) y Neuronal Scanning (SCA), encargados de realizar tareas de neuroestimulación maliciosa. Estos ciberataques fueron seleccionados ya que representan enfoques distintos para afectar a las neuronas mediante sobreestimulación, aunque son posibles otros enfoques, como se presenta en el último capítulo de esta tesis. Para implementar y validar estos ataques, se optó por utilizar un simulador neuronal, Brian2 [27], capaz de recrear el comportamiento de las neuronas de la forma más realista posible, utilizando el modelo de Izhikevich [28], un modelo neuronal ampliamente utilizado en neurociencia. Este desarrollo se alinea con el cuarto objetivo de la tesis.

En este punto, surgió una limitación en la línea de investigación. En el momento de elaborar la publicación, se carecía de topologías neuronales realistas que modelaran la distribución entre capas de la corteza cerebral y las conexiones entre poblaciones neuronales. Para hacer frente a esta limitación, esta tesis tuvo que buscar alternativas para modelar la actividad neuronal de forma realista, lo más cercana al escenario biológico. Debido a esto, se optó por entrenar una red neural convolucional (CNN) [29] encargada de resolver el problema específico de un ratón que debe encontrar la salida de un determinado laberinto. Esta decisión se justificó por la literatura existente que indica las similitudes que presentan las CNNs y la corteza visual en su estructura y funcionamiento. Tras el entrenamiento, las conexiones entre neuronas y sus pesos se tradujeron en términos biológicos y se introdujeron en el simulador. Además, la posición actual del ratón también se utilizó como entrada al modelo neuronal para simular lo que el ratón veía en cada momento, diferenciando entre las celdas transitables y las paredes del laberinto. Además, el modelo resultante del entrenamiento de la CNN proporcionaba el camino óptimo para salir del laberinto desde cualquier posición. En base a ello, sólo consideramos el camino óptimo para llegar a la salida desde la celda inicial del laberinto para incluirlo en la simulación neuronal, teniendo una simplificación del problema. Estas consideraciones pretenden responder a la RQ3.

Probamos diferentes números de neuronas bajo ataque y voltajes utilizados para estimular esas neuronas para la implementación y posterior evaluación de estos ciberataques. Tras implementar ambos ciberataques neuronales en el simulador, definimos tres métricas para medir su impacto con el fin de ofrecer una respuesta a la RQ4, alineándose también con el quinto objetivo. En primer lugar, es esencial definir el concepto de *spike*, o potencial de acción, como la activación de una neurona y la transmisión del estímulo a las neuronas siguientes. La primera métrica, el número de *spikes*, mide si un ataque aumenta o reduce el número de potenciales de acción realizados por las neuronas en comparación con la situación espontánea. La segunda métrica, el porcentaje de desplazamientos, indicaba el desplazamiento de un *spike* en el tiempo, ya sea hacia delante o hacia atrás, en comparación con el caso espontáneo. La dispersión de los *spikes*, medida tanto en la dimensión del tiempo como del número de *spikes*, es la tercera métrica definida y consistió en analizar los patrones de *spikes* para identificar los cambios en su distribución, observando la evolución de la dispersión a lo largo del camino óptimo. Finalmente, tras estudiar el impacto de cada ataque individualmente utilizando estas métricas, comparamos los resultados entre ataques, siguiendo el último objetivo.

Una vez comprobada la eficacia de FLO y SCA mediante un simulador neuronal, defin-

imos un nuevo ciberataque neuronal, Neuronal Jamming (JAM), basado en la inhibición de la actividad neuronal durante una ventana temporal. Así, la tercera publicación de esta tesis doctoral, presentada en el tercer capítulo (Article 3–`Elsevier_COSE`), utilizó el mismo escenario y configuración experimental basada en una CNN para implementar este ciberataque en Brian2. A diferencia de los trabajos anteriores, en esta publicación se pretendía analizar si existía alguna relación entre el impacto causado por los ciberataques neuronales en la actividad neuronal (particularmente FLO y JAM) y el impacto en la capacidad de decisión del ratón, asumiendo que estos ataques afectan a las capacidades visuales del animal. Para validar este objetivo, primero ofrecimos una descripción formal de JAM, seguida de un análisis del impacto causado por este ciberataque desde una perspectiva biológica utilizando simulaciones neuronales. Después, evaluamos el modelo de la CNN utilizado para construir la topología neuronal biológica, con el objetivo de determinar cómo JAM podría afectar a la capacidad del ratón para encontrar la salida del laberinto.

También evaluamos el impacto de la aplicación de los ciberataques de FLO en este escenario. Desde el punto de vista biológico, la diferencia con el segundo capítulo de la tesis doctoral es que, en ese trabajo, realizamos el ataque en un instante concreto al inicio de la simulación, y evaluamos su propagación. En cambio, en esta tercera publicación, aplicamos por separado un ataque en cada posición del camino óptimo, estudiando la evolución del impacto tanto desde la métrica del número de *spikes* como de la dispersión temporal. Además, estudiamos el efecto de FLO sobre la red artificial, atendiendo tanto al número de pasos para llegar a la salida como al porcentaje de veces que el ratón encontró la salida. Para ello, analizamos el impacto del ataque cuando el ratón se situaba en cada posición individual del camino óptimo, calculando a partir de esa posición el rendimiento para salir del laberinto. Como en el caso del enfoque biológico, obtuvimos la correlación de Pearson entre las variables para entender la relación entre los escenarios. Finalmente, comparamos los resultados de JAM y FLO, analizando también la relación que estos ciberataques neuronales podrían tener sobre los efectos causados por las enfermedades neurodegenerativas.

El último trabajo realizado en la tesis doctoral, que se presenta en el cuarto capítulo de este documento (Article 4–`ACM_CACM`) y que está alineado con el tercer objetivo, presentó una taxonomía de ocho ciberataques neuronales que comprenden la estimulación e inhibición de la actividad neuronal. Este trabajo fue motivado por la necesidad de proponer nuevos ciberataques neuronales y ofrecer una categorización de los mismos, de acuerdo a la RQ2. Tres de estos ataques ya fueron presentados en publicaciones anteriores, siendo los cinco restantes novedosos. Para cada uno de estos ocho ciberataques, presentamos los pasos que sigue el ataque en la implementación propuesta para ilustrar mejor su funcionamiento. Después, comparamos individualmente el impacto de cada ciberataque neural con el comportamiento espontáneo. Por último, este trabajo contrastó el efecto producido entre los ataques en función de la métrica del número de *spikes*, estudiando el daño causado durante las primeras y las últimas cinco posiciones de la trayectoria óptima. Este estudio pretendía comprender el impacto inducido por estos ataques a corto y largo plazo.

En resumen, esta tesis revisó en primer lugar el estado del arte de la ciberseguridad en BCIs, seguido de la identificación de vulnerabilidades en BCIs de neuroestimulación de próxima generación. Además, este trabajo propuso la definición e implementación de diferentes ciberataques neuronales con el objetivo de medir su impacto. Esta metodología permitió cumplir con los objetivos definidos en la tesis, previamente presentados en la sección II.

IV Resultados

En la primera publicación de la tesis doctoral, disponible en (Article 1–ACM_CSUR), propusimos la primera estandarización del ciclo de vida de BCI, tanto desde el punto de vista de la adquisición de datos neuronales como de la neuroestimulación, lo suficientemente general como para poder cubrir cualquier implementación de sistemas BCI. Después, analizamos los posibles ataques de ciberseguridad que podrían aplicarse a cada etapa del ciclo BCI desde ambos enfoques, identificando que los ciberataques comunes aplicables a los sistemas informáticos tradicionales, como los ataques de repetición, los ataques de suplantación, los ataques de interferencia o el malware, podrían aplicarse a todas las etapas del ciclo de la BCI. Consideramos cuatro dimensiones para analizar los impactos causados por estos ciberataques: la integridad de los datos y del servicio, la confidencialidad de los datos, la disponibilidad de los datos y del servicio, y la seguridad física de los usuarios de BCI. Además, para todos los ataques se documentaron tanto las contramedidas procedentes de la literatura como las sugeridas por este trabajo para reducir o mitigar los impactos presentados anteriormente.

También analizamos los aspectos de ciberseguridad que podrían afectar a diferentes despliegues arquitectónicos del ciclo BCI. Para cada despliegue, presentamos una descripción, una serie de ejemplos para ilustrar mejor los conceptos, un análisis de los ciberataques que podrían afectar a estas arquitecturas y el impacto que podrían causar. En particular, identificamos los posibles ciberataques que afectan a la BCI, al dispositivo que controla la BCI o a la arquitectura *cloud* utilizada para gestionar los datos de los usuarios. Además, este trabajo proporcionó un conjunto sustancial de posibles contramedidas para mitigar los efectos de estos ataques.

Este trabajo fue relevante para identificar la tendencia de los sistemas BCI actuales, que están moviéndose a enfoques BtI y BtB. El objetivo en estos escenarios es utilizar las tecnologías BCI para interactuar con otros dispositivos, Internet e, incluso, permitir la comunicación directa entre cerebros. Sin embargo, los sistemas BCI presentan limitaciones que determinarán su evolución. En primer lugar, detectamos una falta de interoperabilidad entre las implantaciones de BCI, ya que existe una ausencia de estándares que dificulta que las empresas produzcan dispositivos compatibles entre sí. Además, sus funcionalidades son difíciles de ampliar, ya que se fabrican para su uso en escenarios de aplicación concretos, lo que complica la introducción de nuevas capacidades de ciberseguridad. También existe una falta de mecanismos o normativas de protección de datos en estos escenarios, esenciales para asegurar el correcto tratamiento de los datos sensibles relacionados con la salud. Por último, faltan mecanismos de ciberseguridad en estos sistemas, lo que exige un esfuerzo para crear dispositivos que protejan los datos sensibles transmitidos y la integridad física de sus usuarios. Todos estos aspectos anteriores pretenden ofrecer una respuesta a la RQ1.

La segunda publicación (Article 2–IEEE_Access) identificó por primera vez vulnerabilidades en la arquitectura de las soluciones de neuroestimulación de nueva generación que podrían permitir a los ciberatacantes controlar el sistema y realizar acciones maliciosas. Por ejemplo, un atacante que pretendiera afectar a la arquitectura de Neuralink podría aprovechar las vulnerabilidades del smartphone conectado al sistema implantado. Dado que existen muchas vulnerabilidades y ataques para perturbar estos dispositivos móviles, tomar el control del smartphone encargado de gestionar la BCI es factible. Además, el *link*, un dispositivo intermediario entre el smartphone y los componentes implantados, colocado bajo la oreja, utiliza un enlace Bluetooth que también es susceptible de modificación de hardware o de sufrir ataques de interferencia, entre otras amenazas.

Motivado por las vulnerabilidades anteriores, esta publicación presentó dos ciberata-

ques neuronales: Neuronal Flooding (FLO) y Neuronal Scanning (SCA). Aunque ambos ciberataques estimulan un conjunto aleatorio de neuronas, FLO pretende estimular las neuronas en un instante determinado mientras que SCA se enfoca en el conjunto de neuronas de forma individual y secuencial, evitando las repeticiones. En cuanto a su impacto en el comportamiento neuronal espontáneo (ver RQ2), FLO redujo el número de *spikes*, una diferencia que aumentó cuando el ratón progresó en el laberinto. Además, el aumento del número de neuronas atacadas generó una disminución más significativa del número de *spikes*. También concluimos que cambiar el voltaje utilizado para sobreestimular las neuronas no tuvo un impacto significativo en la métrica. Observando las diferentes capas de la topología, la variación de la media de *spikes* fue más significativa en las capas más profundas. Atendiendo a la métrica del porcentaje de desplazamientos, atacar un mayor número de neuronas generó un mayor porcentaje de desplazamientos mientras que aumentar el voltaje tuvo un efecto insignificante. Por último, en cuanto a la métrica de la dispersión, la dispersión temporal aumentó en comparación con el comportamiento espontáneo. Centrándonos en la dispersión del número de *spikes*, el ataque generó en las últimas posiciones de la trayectoria óptima más instantes en los que sólo se produjo un *spike*, lo que indica una mayor dispersión a medida que avanzaba la simulación. Estos resultados indican que FLO puede alterar eficazmente la actividad neuronal espontánea, cubriendo el quinto objetivo de la tesis, además de ofrecer respuestas parciales para RQ3 y RQ4.

SCA redujo el número de *spikes* en comparación con la señalización espontánea. Además, el impacto se incrementó ligeramente al aumentar el voltaje utilizado para atacar, pero sólo para voltajes bajos. Así, y de forma similar a FLO, el impacto del voltaje es insignificante. Este ciberataque también aumentó el porcentaje de desplazamientos de los *spikes*, degradando el impacto cuando se observan capas más profundas. Además, identificamos diferencias significativas en las métricas de dispersión en comparación con el comportamiento espontáneo. Por último, es interesante observar que el impacto se agravaba cuando el ratón progresaba en el laberinto, lo que pone de manifiesto el comportamiento incremental de este ciberataque. Atendiendo a la comparación en términos de impacto entre FLO y SCA, concluimos que los mecanismos internos de cada ataque generan comportamientos diferentes en la actividad neuronal. FLO es mejor para alterar la actividad neuronal en un periodo corto ya que afecta a múltiples neuronas en un instante concreto. Por el contrario, SCA es más eficaz a largo plazo, ya que requiere más tiempo para generar un impacto considerable, pero después, el impacto es mayor que FLO.

En la tercera publicación de la tesis (Article 3–`Elsevier_COSE`), presentamos Neuronal Jamming (JAM) como un ciberataque neuronal centrado en la inhibición de la actividad de un conjunto de neuronas durante una duración determinada, inspirado en enfermedades neurodegenerativas consistentes en el mal funcionamiento de las neuronas o su muerte, como el Parkinson y el Alzheimer. Este trabajo surgió naturalmente como continuación de la publicación anterior con el objetivo de medir el impacto de los ciberataques basados en inhibición, en contraste con los trabajos anteriores centrados en la estimulación de las neuronas. El análisis de JAM desde una perspectiva biológica indicó que el aumento del número de neuronas atacadas disminuyó tanto el número de *spikes* como la dispersión temporal. Además, observamos un incremento en la variabilidad de la distribución de estas métricas al aumentar el número de posiciones consecutivas atacadas, especialmente en el número de *spikes*. En la red artificial, detectamos que incluso atacando unos pocos nodos al azar se incrementaba drásticamente el número de pasos, no pudiendo salir del laberinto en la mayoría de las situaciones. Comparando la correlación de Pearson entre las métricas biológicas y artificiales, obtuvimos una baja correlación de alrededor del 60%. Este resultado se explica por las restricciones de las consideraciones experimentales presentadas

en el artículo. Sin embargo, el análisis individual por escenario demostró el alto impacto que presenta JAM.

A continuación, comparamos el impacto de JAM y de FLO. En este contexto, primero analizamos el impacto individual de FLO en ambos escenarios. En el biológico, los resultados indicaron que realizar el ataque en posiciones posteriores tenía un menor impacto ya que la actividad neuronal permanecía inalterada la mayor parte del tiempo. Además, afectar a un mayor número de neuronas generaba un mayor daño. En la red artificial, aumentar el número de nodos atacados incrementaba el impacto hasta una determinada posición. Después de eso, y dado que el ratón estaba más cerca de la celda de salida, el impacto disminuía ya que el ratón podía encontrar la salida por probabilidad. Comparando ambos escenarios para FLO, obtuvimos una correlación de alrededor del 80% entre el número de pasos y el número de *spikes* y la dispersión, concluyendo una relación significativa entre los escenarios. Por último, comparamos los resultados de ambos ataques. Como la metodología entre los ataques difiere en esta publicación, nos centramos en el estudio de las correlaciones obtenidas. Así, apreciamos una relación más estrecha entre ambos enfoques en FLO pero teniendo en cuenta las limitaciones anteriormente expuestas. Este análisis del impacto causado por los ciberataques neurales se alinea con la RQ2.

La última publicación de la tesis doctoral (Article 4–`ACM_CACM`) presentó la definición e implementación de una taxonomía de ciberataques neurales, relacionada con la RQ4. Este trabajo amplió de forma natural el conjunto de ciberataques neuronales ya presentados en las dos publicaciones anteriores de la tesis. Centrándonos en los nuevos ataques presentados en este trabajo, Neuronal Selective Forwarding (FOR) consiste en inhibir secuencialmente neuronas sin repeticiones a lo largo del tiempo, mientras que Neuronal Spoofing (SPO) replica exactamente la actividad registrada en una ventana temporal anterior. Neuronal Sybil (SYB) obliga a una neurona a tener el voltaje opuesto dentro del rango de voltaje natural de una neurona. Por el contrario, Neuronal Sinkhole (SIN) consiste en estimular neuronas de las primeras capas corticales con el objetivo de afectar a una neurona concreta situada en una capa más profunda. Por último, Neuronal Nonce (NON) pretende atacar a un conjunto de neuronas en un instante determinado, decidiendo aleatoriamente por cada una de ellas su estimulación o inhibición.

Este trabajo representó su comportamiento, generando una intuición de su dinámica. Posteriormente, se midió empíricamente el impacto de los ocho ciberataques en la actividad espontánea atendiendo a la métrica del número de *spikes*. En particular, se estudió el impacto de las cinco primeras y últimas posiciones del camino óptimo del laberinto para destacar cuáles eran más dañinas a corto y largo plazo. Atendiendo al corto plazo, NON logró una reducción aproximada del 12%, seguido de JAM con un 5%. Por el contrario, SCA fue el más perjudicial a largo plazo, ofreciendo una reducción de alrededor del 9% de los *spikes*, seguido de NON con un 8%.

V Conclusiones y trabajo futuro

En las últimas décadas, la rápida evolución de las BCIs ha generado un considerable avance en la medicina, permitiendo una mejor detección de diversas enfermedades neurológicas. También proporcionan capacidades de neuroestimulación para tratar enfermedades como el Parkinson cuando un tratamiento basado en fármacos resulta ineficaz. Esta evolución ha hecho que ganen popularidad en otros sectores como el del entretenimiento o los videojuegos. Estos sistemas se están investigando también para su conexión a Internet o incluso para permitir la comunicación directa entre cerebros. Este avance abre un panorama de oportunidades para que nuevas empresas e ideas dominen un sector en alza que aspira a

llegar a la población general en las próximas décadas.

Gracias a esta variabilidad en los escenarios de aplicación, existe una gran variedad de tecnologías BCI centradas en la adquisición de datos neuronales o en la neuroestimulación, diferenciadas también en función de su capacidad de invasividad cerebral. Centrándonos en las BCIs de neuroestimulación invasiva, las técnicas actuales con aprobación de la FDA para fines médicos son escasas y presentan limitaciones, como tener una resolución espacial reducida o estar limitadas a determinadas enfermedades y regiones cerebrales. Partiendo de esta base, las BCIs de nueva generación pretenden miniaturizar los electrodos y la tecnología para permitir el registro conjunto de datos neuronales y la estimulación e inhibición de la actividad neuronal. Su objetivo final es democratizar las tecnologías BCI y acercarlas a los consumidores finales, separándolas de los escenarios médicos.

Sin embargo, las anteriores tecnologías BCI no han sido concebidas bajo el prisma de la ciberseguridad. En concreto, estas interfaces carecen de estándares y reglamentos específicos, lo que dificulta la unificación de los mecanismos de seguridad necesarios para su uso comercial. Tampoco existe una normativa de protección de datos que garantice el buen uso de esta información sensible. Además, la tendencia de estas interfaces centradas en la neuroestimulación, en la que empresas como Neuralink pretenden democratizar su acceso, podría tener un impacto significativo en la seguridad de los usuarios.

Atendiendo a las preocupaciones y limitaciones anteriores, esta tesis doctoral ha analizado el estado del arte en materia de ciberseguridad en las BCIs, detectando una carencia de trabajos que aborden este tema. Aunque algunos trabajos cubren parcialmente ciertos aspectos de la ciberseguridad en este campo, son escasos y no ofrecen una visión integral del problema. En base a ello, este trabajo analiza en primer lugar los ataques, los impactos y las contramedidas tanto para el ciclo de vida de las BCIs como para los despliegues arquitectónicos comunes para estos sistemas. Además, esta tesis ha identificado las tendencias y los retos a los que se enfrentarán estos sistemas en un futuro próximo. Estos hallazgos han ofrecido una respuesta a la RQ1, permitiendo también completar el primer objetivo específico de la tesis.

Posteriormente, este trabajo propuso la definición de ciberataques neuronales como amenazas que pueden afectar a la actividad neuronal espontánea, avanzando en la literatura en términos de ciberseguridad en BCIs. Están motivados por las vulnerabilidades identificadas en dispositivos de neuroestimulación de nueva generación que los atacantes podrían explotar para causar daño a los usuarios de BCI (véase el segundo objetivo de la tesis). En esta dirección, esta investigación presentó primero Neuronal Flooding y Neuronal Scanning como ciberataques capaces de estimular maliciosamente las neuronas, analizando su impacto en una simulación neuronal. Dado que, en ese momento, se carecía de topologías neuronales realistas, esta tesis entrenó una CNN para resolver el problema particular de un ratón que tiene que salir de un laberinto determinado, trasladando la topología resultante a un simulador neuronal. Esta decisión fue motivada por evidencia existente que presenta una relación entre algunos aspectos del funcionamiento y la estructura de las CNNs y la corteza visual. Ambos ciberataques fueron eficaces para reducir la actividad neuronal. Estos resultados ofrecieron una respuesta a la RQ3 y RQ4 y ayudaron a avanzar hacia una respuesta para la RQ2 para los ataques basados en la estimulación neuronal.

Con estos resultados en consideración, esta tesis presentó posteriormente un tercer ciberataque neuronal, Neuronal Jamming, que inhibe la actividad neuronal de un conjunto de neuronas objetivo durante un periodo de tiempo. Este trabajo comparó su impacto con el de Neuronal Flooding, considerando también su relación con la capacidad de decisión del ratón para salir del laberinto. Los resultados obtenidos sugirieron una correlación sustancial entre el impacto de estos ciberataques en la actividad neuronal y la capacidad

de tomar decisiones, aunque se necesita más investigación en esta dirección. A partir de estos resultados, este trabajo ofreció nuevos hallazgos para responder a la RQ2 sobre los ciberataques que aplican inhibición neuronal.

Por último, esta investigación presentó una taxonomía de ocho ciberataques neuronales, de los cuales cinco eran novedosos. Para cada uno de ellos, esta tesis proporcionó una definición, una descripción de su funcionamiento interno y un análisis de su impacto a corto y largo plazo. A partir de ahí, este trabajo indicó cuáles eran más adecuados para causar un efecto inmediato y cuáles causaban un daño más significativo a largo plazo. Así, estos resultados respondieron a la RQ2 ya que permitieron medir el impacto causado por un amplio conjunto de comportamientos de ciberataques neurales y ayudaron a completar todos los objetivos de la tesis.

En resumen, esta tesis doctoral ha recogido en primer lugar el conocimiento existente en la literatura relativa a la ciberseguridad en BCIs. Además, este trabajo ha avanzado sustancialmente el estado del arte, proponiendo nuevos ciberataques capaces de afectar a la actividad neuronal espontánea, validando su impacto en un escenario lo más realista posible al tejido neuronal biológico.

Como trabajo futuro, esta tesis identifica en primer lugar la necesidad de analizar exhaustivamente las vulnerabilidades existentes tanto en las soluciones BCI actuales como en las emergentes, lo que ayudará a desarrollar soluciones prácticas de ciberseguridad para productos específicos. Además, es necesario cubrir los retos identificados en términos de interoperabilidad y extensibilidad de las soluciones BCI y abarcar las oportunidades actuales en cuanto a la regulación de los datos y los mecanismos de seguridad.

Además, esta investigación detecta la necesidad de ampliar el análisis de los ciberataques neuronales, estudiando cómo otros ciberataques tradicionales del ámbito de informática podrían adaptarse al escenario neurológico. Esta tesis también considera fundamental identificar aspectos de las enfermedades neurodegenerativas que puedan ayudar a ampliar esta área de investigación en ciberseguridad. Por otro lado, este trabajo identifica la necesidad de evaluar el impacto de los ciberataques neuronales sobre topologías neuronales más realistas. Así, primero permitiría medir las diferencias entre atacar poblaciones neuronales excitatorias o inhibitorias. Además, el aumento del número de neuronas y de la complejidad de la red permitiría obtener más conclusiones sobre su efecto en el tejido neuronal biológico natural.

Una vez que se obtiene una amplia comprensión de estos ciberataques, este trabajo pone de manifiesto una oportunidad para detectar y mitigar estos ciberataques. Para ello, la inteligencia artificial, como las técnicas de *machine learning* y *deep learning*, podrían ser útiles para su implementación en nuevas generaciones de dispositivos BCI, ayudando a reducir o incluso mitigar el daño causado por estas amenazas e incluso las emergentes.

Una mejor intuición del impacto de los ciberataques neuronales en condiciones más realistas podría ser vital para recrear el comportamiento y efecto de las enfermedades neurodegenerativas conocidas. Así, ciertos ciberataques podrían beneficiar los efectos de condiciones particulares, estableciendo una relación entre ciberataques y enfermedades. Además, si se consigue este hito, la investigación podría centrarse en predecir, basándose en la actividad neuronal espontánea, la presencia de enfermedades neurodegenerativas específicas, incluso en las primeras fases. Estos avances podrían beneficiar positivamente a la investigación médica y tener un impacto masivo en los pacientes neurológicos.

Bibliography

[1] M. A. Lebedev and M. A. L. Nicolelis, "Brain-Machine Interfaces: From Basic Science to Neuroprostheses and Neurorehabilitation," *Physiological Reviews*, vol. 97, no. 2, pp. 767–837, Apr 2017.

[2] W. Zhao, E. J. Van Someren, C. Li, X. Chen, W. Gui, Y. Tian, Y. Liu, and X. Lei, "Eeg spectral analysis in insomnia disorder: A systematic review and meta-analysis," *Sleep Medicine Reviews*, vol. 59, p. 101457, 2021.

[3] G. Giannakakis, D. Grigoriadis, and M. Tsiknakis, "Detection of stress/anxiety state from eeg features during video watching," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 6034–6037.

[4] M. Parastarfeizabadi and A. Z. Kouzani, "Advances in closed-loop deep brain stimulation devices," *Journal of NeuroEngineering and Rehabilitation*, vol. 14, no. 1, p. 79, Aug 2017.

[5] C. J. Hartmann, S. Fliegen, S. J. Groiss, L. Wojtecki, and A. Schnitzler, "An update on best practice of deep brain stimulation in parkinson's disease," *Therapeutic Advances in Neurological Disorders*, vol. 12, p. 1756286419838096, Jan 2019.

[6] C. A. Edwards, A. Kouzani, K. H. Lee, and E. K. Ross, "Neurostimulation Devices for the Treatment of Neurologic Disorders," *Mayo Clinic Proceedings*, vol. 92, no. 9, pp. 1427–1444, 2017.

[7] Y. Luo, Y. Sun, X. Tian, X. Zheng, X. Wang, W. Li, X. Wu, B. Shu, and W. Hou, "Deep brain stimulation for alzheimer's disease: Stimulation parameters and potential mechanisms of action," *Frontiers in Aging Neuroscience*, vol. 13, 2021.

[8] M. A. L. Nicolelis, "Actions from thoughts," *Nature*, vol. 409, no. 6818, pp. 403–407, 2001.

[9] M. Ahn, M. Lee, J. Choi, S. Jun, M. Ahn, M. Lee, J. Choi, and S. C. Jun, "A Review of Brain-Computer Interface Games and an Opinion Survey from Researchers, Developers and Users," *Sensors*, vol. 14, no. 8, pp. 14 601–14 633, Aug 2014.

[10] V. Khurana, M. Gahalawat, P. Kumar, P. P. Roy, D. P. Dogra, E. Scheme, and M. Soleymani, "A survey on neuromarketing using eeg signals," *IEEE Transactions on Cognitive and Developmental Systems*, 2021.

[11] A. Jalaly Bidgoly, H. Jalaly Bidgoly, and Z. Arezoumand, "A survey on methods and challenges in eeg based authentication," *Computers & Security*, vol. 93, p. 101788, 2020.

[12] A. Saboor, F. Gembler, M. Benda, P. Stawicki, A. Rezeika, R. Grichnik, and I. Volosyak, "A Browser-Driven SSVEP-Based BCI Web Speller," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. Miyazaki, Japan: IEEE, Oct 2018. ISBN 978-1-5386-6650-0 pp. 625–630.

[13] M. Pais-Vieira, M. Lebedev, C. Kunicki, J. Wang, and M. A. L. Nicolelis, "A Brain-to-Brain Interface for Real-Time Sharing of Sensorimotor Information," *Scientific Reports*, vol. 3, no. 1, p. 1319, Dec 2013.

[14] M. Pais-Vieira, G. Chiuffa, M. Lebedev, A. Yadav, and M. A. L. Nicolelis, "Building an organic computing device with multiple interconnected brains," *Scientific Reports*, vol. 5, no. 1, p. 11869, Dec 2015.

[15] R. A. Ramadan and A. V. Vasilakos, "Brain computer interface: control signals review," *Neurocomputing*, vol. 223, pp. 26–44, Feb 2017.

[16] B. Jarosiewicz and M. Morrell, "The rns system: brain-responsive neurostimulation for the treatment of epilepsy," *Expert Review of Medical Devices*, vol. 18, no. 2, pp. 129–138, 2021.

[17] E. Musk and Neuralink, "An integrated brain-machine interface platform with thousands of channels," *bioRxiv*, 2019. [Online]. Available: https://www.biorxiv.org/content/early/2019/08/02/703801. DOI: 10.1101/703801

[18] S. A. Wirdatmadja, M. T. Barros, Y. Koucheryavy, J. M. Jornet, and S. Balasubramaniam, "Wireless optogenetic nanonetworks for brain stimulation: Device model and charging protocols," *IEEE Transactions on NanoBioscience*, vol. 16, no. 8, pp. 859–872, 2017.

[19] I. Martinovic, D. Davies, and M. Frank, "On the feasibility of side-channel attacks with brain-computer interfaces," in *Proceedings of the 21st USENIX Security Symposium*. Bellevue, WA: USENIX, 2012. ISBN 978-931971-95-9. ISSN 0733-8716 pp. 143–158.

[20] M. Frank, T. Hwu, S. Jain, R. T. Knight, I. Martinovic, P. Mittal, D. Perito, I. Sluganovic, and D. Song, "Using EEG-Based BCI Devices to Subliminally Probe for Private Information," in *Proceedings of the 2017 on Workshop on Privacy in the Electronic Society - WPES '17*. New York, New York, USA: ACM Press, 2017. ISBN 9781450351751 pp. 133–136.

[21] T. Denning, Y. Matsuoka, and T. Kohno, "Neurosecurity: security and privacy for neural devices," *Neurosurgical Focus*, vol. 27, no. 1, p. E7, 2009.

[22] M. Ienca, "Neuroprivacy, neurosecurity and brain-hacking: Emerging issues in neural engineering," *Bioethica Forum*, vol. 8, no. 2, pp. 51–53, 2015.

[23] M. Ienca and P. Haselager, "Hacking the brain: brain–computer interfacing technology and the ethics of neurosecurity," *Ethics and Information Technology*, vol. 18, no. 2, pp. 117–129, Jun 2016.

[24] Q. Li, D. Ding, and M. Conti, "Brain-Computer Interface applications: Security and privacy challenges," in *2015 IEEE Conference on Communications and Network Security (CNS)*. San Francisco, CA, USA: IEEE, Sep 2015. ISBN 9781467378765 pp. 663–666.

[25] C. Camara, P. Peris-Lopez, and J. E. Tapiador, "Security and privacy issues in implantable medical devices: A comprehensive survey," *Journal of Biomedical Informatics*, vol. 55, pp. 272–289, Jun 2015.

[26] L. Pycroft and T. Z. Aziz, "Security of implantable medical devices with wireless connections: The dangers of cyber-attacks," *Expert Review of Medical Devices*, vol. 15, no. 6, pp. 403–406, Jul 2018.

[27] M. Stimberg, R. Brette, and D. F. Goodman, "Brian 2, an intuitive and efficient neural simulator," *eLife*, vol. 8, p. e47314, Aug. 2019. DOI: 10.7554/eLife.47314

[28] E. M. Izhikevich, "Simple model of spiking neurons," *IEEE Transactions on Neural Networks*, vol. 14, no. 6, pp. 1569–1572, 2003.

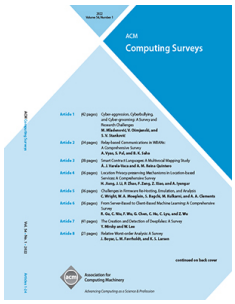
[29] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, aug 2019. ISBN 1492032646

Publications composing

the PhD Thesis

1

Survey of Cybersecurity on Brain-Computer Interfaces



Title:	Security in Brain-Computer Interfaces: State-of-the-Art, Opportunities, and Future Challenges.
Authors:	Sergio López Bernal, Alberto Huertas Celdrán, Gregorio Martínez Pérez, Michael Taynnan Barros, Sasitharan Balasubramaniam.
Journal:	ACM Computing Surveys
JIF:	14.324 D1 (2021)
Publisher:	ACM
Volume:	54
Number:	1
Pages:	35
Year:	2021
Month:	Jan
DOI:	10.1145/3427376
Status:	Published

Abstract

Brain-Computer Interfaces (BCIs) have significantly improved the patients' quality of life by restoring damaged hearing, sight, and movement capabilities. After evolving their application scenarios, the current trend of BCI is to enable new innovative brain-to-brain and brain-to-the-Internet communication paradigms. This technological advancement generates opportunities for attackers, since users' personal information and physical integrity could be under tremendous risk. This work presents the existing versions of the BCI lifecycle and homogenizes them in a new approach that overcomes current limitations. After that, we offer a qualitative characterization of the security attacks affecting each phase of the BCI cycle to analyze their impacts and countermeasures documented in the literature. Finally, we reflect on lessons learned, highlighting research trends and future challenges concerning security on BCIs.

Keywords

Brain-computer interfaces · BCI · Cybersecurity · Privacy · Safety

Security in Brain-Computer Interfaces: State-of-the-Art, Opportunities, and Future Challenges

SERGIO LÓPEZ BERNAL, University of Murcia, Departamento de Ingeniería de la Información y las Comunicaciones
ALBERTO HUERTAS CELDRÁN, Waterford Institute of Technology, Telecommunication Software and Systems Group and Communication Systems Group CSG, Department of Informatics IfI, University of Zurich UZH
GREGORIO MARTÍNEZ PÉREZ, University of Murcia, Departamento de Ingeniería de la Información y las Comunicaciones
MICHAEL TAYNNAN BARROS, University of Essex, School of Computer Science and Electronic Engineering and Tampere University, CBIG/BioMediTech in the Faculty of Medicine and Health Technology
SASITHARAN BALASUBRAMANIAM, Waterford Institute of Technology, Telecommunication Software and Systems Group and RCSI University of Medicine and Health Sciences, FutureNeuro, SFI Research Centre for Chronic and Rare Neurological Diseases

11

Brain-Computer Interfaces (BCIs) have significantly improved the patients' quality of life by restoring damaged hearing, sight, and movement capabilities. After evolving their application scenarios, the current trend of BCI is to enable new innovative brain-to-brain and brain-to-the-Internet communication paradigms. This technological advancement generates opportunities for attackers, since users' personal information and physical integrity could be under tremendous risk. This work presents the existing versions of the BCI life-cycle and homogenizes them in a new approach that overcomes current limitations. After that, we offer a qualitative characterization of the security attacks affecting each phase of the BCI cycle to analyze their impacts and countermeasures documented in the literature. Finally, we reflect on lessons learned, highlighting research trends and future challenges concerning security on BCIs.

This work has been supported by the Irish Research Council under the government of Ireland post-doc fellowship (Grant No. GOIPD/2018/466), by the Science Foundation Ireland (SFI) under Grant No. 16/RC/3948 and co-funded under the European Regional Development Fund and by FutureNeuro industry partners, by the European Union's Horizon 2020 Research and Innovation Programme through the Marie Skłodowska-Curie under Grant Agreement No. 839553, by Armasuisse S+T with project CYD-C-2020003, by the University of Zürich UZH, and by the European Union Horizon 2020 Research and Innovation Program under grant agreement No. 830927, namely the H2020 Concordia Project.
Authors' addresses: S. L. Bernal and G. M. Perez, University of Murcia, Departamento de Ingeniería de la Información y las Comunicaciones, Murcia, Spain; emails: {slopez, gregorio}@um.es; A. H. Celdrán, Waterford Institute of Technology, Telecommunication Software and Systems Group, Waterford, Ireland and Communication Systems Group CSG, Department of Informatics IfI, University of Zurich UZH, CH 8050 Zürich, Switzerland; email: ahuertas@tssg.org; M. T. Barros, University of Essex, School of Computer Science and Electronic Engineering, Essex, UK, Tampere University, CBIG/BioMediTech in the Faculty of Medicine and Health Technology, Tampere, Finland; email: michael.barros@tuni.fi; S. Balasubramaniam, Waterford Institute of Technology, Telecommunication Software and Systems Group, Waterford, Ireland, RCSI University of Medicine and Health Sciences, FutureNeuro, the SFI Research Centre for Chronic and Rare Neurological Diseases, Dublin, Ireland; email: sasib@tssg.org.

0360-0300/2020/12-ART11 $15.00
https://doi.org/10.1145/3427376

ACM Computing Surveys, Vol. 54, No. 1, Article 11. Publication date: December 2020.

CCS Concepts: • **Security and privacy → Domain-specific security and privacy architectures**;

Additional Key Words and Phrases: Brain-computer interfaces, BCI, cybersecurity, privacy, safety

ACM Reference format:
Sergio López Bernal, Alberto Huertas Celdrán, Gregorio Martínez Pérez, Michael Taynnan Barros, and Sasitharan Balasubramaniam. 2020. Security in Brain-Computer Interfaces: State-of-the-Art, Opportunities, and Future Challenges. *ACM Comput. Surv.* 54, 1, Article 11 (December 2020), 35 pages.
https://doi.org/10.1145/3427376

1 INTRODUCTION

Brain-Computer Interfaces (BCI) emerged in the 1970s intending to acquire and process users' brain activity to perform later specific actions over external machines or devices [87]. After several decades of research, this functionality has been extended by enabling not only neural activity recording but also stimulation [167]. Figure 1 describes a simplification of the general components and processes defining a common BCI cycle in charge of recording and stimulating neurons [1, 26, 59], later presented in Section 2. It is important to note that these phases are not standard, so we include the most common ones used in the literature. The clockwise direction, indicated in blue, shows the process of acquiring neural data, and the counterclockwise represents the stimulation one, which is highlighted in red. Regarding the neural data acquisition, neurons interact with each other, producing neural activity, either based on previously agreed actions, such as controlling a joystick, or generated spontaneously (phase 1 of Figure 1). This activity is acquired by the BCI and transformed into digital data (phase 2). After that, data is analyzed by the BCI data processing system to infer the action intended by the user (phase 3). Finally, applications execute the intended action, enabling the control of external devices. These applications can present optional feedback to the users, which allows the generation of new neural activity. However, the counterclockwise direction of Figure 1 starts in phase 4, where applications define the intended stimulation actions to perform. Phase 3 processes this action to determine a firing pattern containing all the essential parameters required by the BCI to stimulate the brain. Finally, the firing pattern is sent to the BCI, which is in charge of stimulating specific neurons belonging to one or more brain regions and is dependent on the technology used. In a nutshell, a BCI can be a unidirectional or bidirectional communication system between the brain and external computational devices. Unidirectional communications are when they either acquire data or stimulate neurons, while bidirectional communications are when they perform both tasks [139].

From the security perspective, BCIs are in an early and immature stage. The literature has not considered security a critical aspect of BCIs until recent years, where terms such as neurosecurity, neuroprivacy, neuroconfidentiality, brain-hacking, or neuroethics have emerged [31, 58, 59]. Existing works of the literature have detected specific security attacks affecting BCI integrity, confidentiality, availability, and safety, but they do not perform a comprehensive analysis and miss relevant concerns [17, 87, 96, 163, 165]. More specifically, the use of neurostimulation BCIs in clinical environments introduces severe vulnerabilities that can have a significant impact on the user's health condition [136]. BCIs already existing on the market would benefit from the implementation of robust security solutions, reducing their impact, particularly in clinical environments. Furthermore, the expansion of BCIs to new markets, e.g., video games or entertainment, generates considerable risks in terms of data confidentiality [87, 96, 163, 165]. In this context, users' personal information, such as thoughts, emotions, sexual orientation, or religious beliefs, are under threats if security measures are not adopted [59, 96, 165]. Besides, contemporary BCI approaches, such as the use of silicon-based interfaces, introduce new security challenges due to the increase in the volume of acquired data and the use of potentially vulnerable technology [121]. The technological revolution

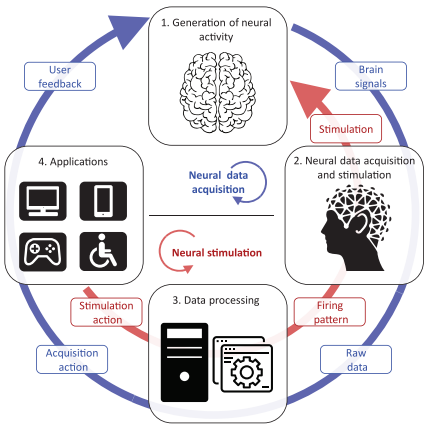


Fig. 1. General functioning of a bidirectional BCI. The clockwise flow indicated with a blue arrow represents the neural data acquisition process, while the counterclockwise flow represented with a red arrow models the brain stimulation.

of recent years, combined with movements such as the Internet of Things (IoT), brings an acceleration in the creation of new devices lacking security standards and solutions based on the concepts of *security-by-design* and *privacy-by-design* [17, 60, 137, 163, 165]. This revolution also brings to reality prospective and disruptive scenarios, where we highlight as examples the direct communications between brains, known as Brain-to-Brain (BtB) or Brainets [67, 126, 127, 184], and brains connected to the Internet (Brain-to-Internet (BtI)), which will require significant efforts from the security prism.

Once summarized the functioning of BCIs and their security status, the scope of this article lies in analyzing the security issues of software components that intervene in the processes, working phases, and communications of BCIs. Besides, this work considers the security concerns of infrastructures, such as computers, smartphones, and cloud platforms, where different BCI architectures are deployed. It is also important to note that, despite this article indicates overall impacts over the brain and the user's physical safety, the main focus of this work is to perform a security analysis from a technological point of view. Aligned with these aspects, and to the best of our knowledge, this article is the first work that exhaustively reviews and analyses the BCI field from the security point of view. Since these aspects have not been studied in depth before and BCI technologies are still immature, this line of work has a particular interest in a medium to long term. However, this area of knowledge is relevant nowadays, since devices already available on the market need to be protected against attacks.

In this context, Section 2 focuses on analyzing the security issues related to the design of the BCI life-cycle. We unify the existing heterogeneous BCI life-cycles in a novel and common approach that integrates recording and stimulation processes. Once proposed the new life-cycle design approach, we review the attacks applicable to each phase of the cycle, the impact generated by the attacks and the countermeasures to mitigate them, both documented in the literature and detected by us. After highlighting the security issues related to the BCI design, Section 3 reviews the inherent cyberattacks, impacts, and countermeasures affecting current BCI deployments scenarios. This section identifies the security issues generated by the devices implementing each life-cycle phase's responsibilities, as well as the communication mechanisms and the application scenarios. The last main contribution of this article is Section 4, where we give our vision regarding the trend

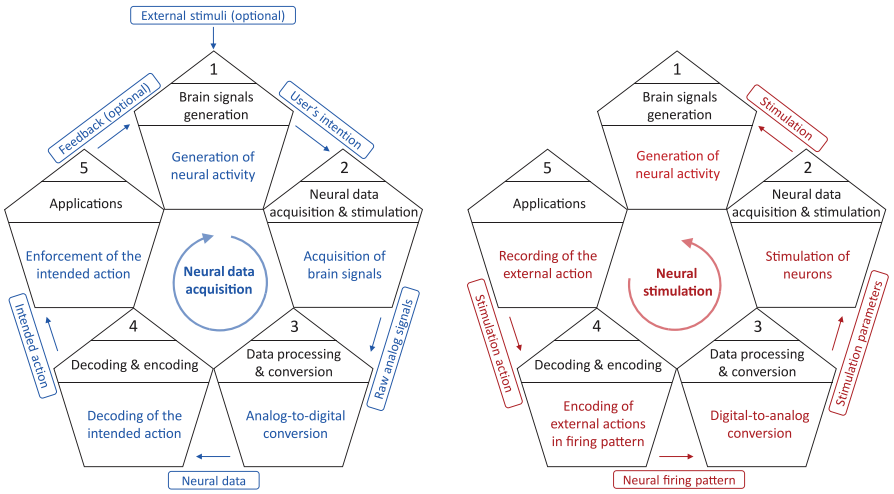


Fig. 2. Bidirectional BCI functioning cycle representing, in black, the common phases for neural data acquisition and brain stimulation. (Left side) Representation, in blue, of the processes performed and the data transferred by each phase of the neural data acquisition process. This cycle can be seen as a closed-loop process, because it starts and ends at the same phase. (Right side) Representation, in red, of the processes and transitions of each phase making up the stimulation process.

of BCI and the security challenges that this evolution will generate in the future. Finally, Section 5 presents some conclusions and future work.

2 CYBERATTACKS AFFECTING THE BCI CYCLE, IMPACTS, AND COUNTERMEASURES

This section reviews the different operational phases of BCIs detected in the literature, known as the BCI cycle, and homogenizes them in a new approach shown in Figure 2. After that, we survey the security attacks affecting each phase of the cycle, their impacts, and the countermeasures documented in the literature. We present as well unexplored opportunities in terms of cyberattacks, and countermeasures affecting each phase.

The literature has proposed different configurations of the BCI cycle. However, the existing versions only consider the signal acquisition process, missing the stimulation of neurons. These solutions present various classifications of the BCI cycle, as some do not consider the generation of brain signals as a phase, or group several phases in only one, without providing information about their roles [26, 59]. Other solutions, as proposed in References [6, 59, 87, 172], are confusing due to they define as new phases, transitions, and data exchanged between different stages. In terms of applications, some authors define a generic stage of applications [1, 26, 87, 148] while others deal with the concept of *commands* sent to external devices [10, 17, 18, 25, 54, 163, 171]. Also, just a few works define the feedback sent by applications to users [10, 17, 18, 25, 59, 87, 163, 171, 172]. To homogenize the BCI cycle and address the previously missing or confusing points, we present a new version of the BCI cycle with five phases (with clearly defined tasks, inputs, and outputs) that consider both acquisition and stimulation capabilities. Figure 2 represents our proposal, where the clockwise direction corresponds to the brain signal acquisition process. The information and tasks concerning this functioning are indicated in blue. In contrast, the stimulation process is indicated

in the counterclockwise direction, starting from phase 5, and, in each phase, the information and tasks are identified in red.

According to the neural acquisition process (clockwise direction in Figure 2), phase 1 focuses on the generation of brain signals. Generated data contain the user's intention to perform particular tasks; for example, controlling an external device. This phase can be influenced by external stimuli, producing modifications in the regular neural activity. In phase 2, the brain waves are captured by electrodes using a wide variety of technologies, such as Electroencephalography (EEG) or Functional Magnetic Resonance Imaging (fMRI). Raw analog signals containing the user's intention are then transmitted to phase 3, where data processing and conversion are required. In particular, this phase performs an analog-to-digital conversion procedure to allow further processing of the data. One of the main goals of this phase is to maximize the Signal-to-Noise Ratio (SNR), which compares the level of the target signal to background noise level to obtain the original signal as accurately as possible. Phase 4 processes the digital neural data to decode the user's intended action, where relevant features are calculated and selected from the neural data. After that, different models (e.g., classifiers, predictors, regressors) or rule-based systems determine the intended action [25, 148]. The action finally arrives at applications in phase 5, which execute the action. Applications can also send optional feedback to the user to generate brain signals and thus new iterations of the cycle.

Regarding the stimulation process (counterclockwise direction in Figure 2), the loop starts in phase 5, where it is specified the stimulation action in a general way (e.g., stimulate a particular brain region to treat Alzheimer's disease). This intended action is transmitted to phase 4, where this input is processed by different techniques, such as Machine Learning (ML), to generate a firing pattern that contains high-level information about the stimulation devices to be activated, the frequencies used and the temporal planning. Phase 3 intends to transform the firing pattern received, indicated in a general fashion, to specific parameters related to the BCI technology used. For example, the identification of neurons to stimulate or the power and voltage required for the process. Phase 2 transmits these stimulation parameters to the stimulation system, that is in charge of the physical stimulation of the brain. After this process, the brain generates neural activity as a response, which can also be acquired by the BCI to measure the state of the brain after each stimulation process. At this point, an alternation between brain stimulation and signal acquisition is possible, moving from one direction of Figure 2 to the other.

Before reviewing the attacks, impacts and countermeasures of each phase of the BCI cycle, it is essential to accurately define the concept of *security*, which refers to the "protection of information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction to provide integrity, confidentiality and availability" [149]. The concepts of integrity, confidentiality and availability, together with the concept of *safety*, are used in this section as metrics to evaluate the impact of security attacks against BCI systems. The standard definitions of these concepts are the following:

- **Integrity:** "protection against unauthorized modification or destruction of information. A state in which information has remained unaltered from the point it was produced by a source, during transmission, storage, and eventual receipt by the destination" [76].
- **Confidentiality:** "preservation of authorized restrictions on access and disclosure, including means for protecting personal privacy and proprietary information" [149].
- **Availability:** "property that data or information is accessible and usable upon demand by an authorized person" [149].
- **Safety:** "freedom from conditions that can cause death, injury, occupational illness, damage to or loss of equipment or property, or damage to the environment" [143]. This work considers the safety concept from the physiological, psychiatric, and psychological perspectives.

Table 1. Definition of the Attacks Detected for the BCI Cycle

Attack	Description
Adversarial attacks [38, 90]	Presentation of intentionally crafted inputs to an ML system to disrupt its normal functioning and output.
Misleading stimuli attacks [40, 79, 96]	Presentation of malicious sensory or motor stimuli to users aiming to generate a specific neural response.
Buffer Overflow attacks [16, 109, 147]	Access to out-of-bounds memory spaces due to insecure software implementations. They take advantage of operations over memory buffers whose boundaries are not well managed.
Cryptographic attacks [58, 59]	Exploit vulnerabilities in the elements that define a system, such as algorithms, protocols or tools. A variety of techniques focused on evading the security measures of cryptographic systems.
Firmware attacks [13, 173]	Extract or modify the firmware of a device, a critical piece of software that controls its hardware.
Battery drain attacks [24, 135]	Consume the battery of a device, reducing its performance or even making it permanently inaccessible.
Injection attacks [105, 134]	Present an input to an interpreter containing particular elements that can modify how it is parsed, taking advantage of a lack of verification of the input.
Malware attacks [77, 154, 177]	Use of hardware, software or firmware aiming to gain access over computational devices to perform malicious actions intentionally.
Ransomware attacks [2, 37]	Encrypt users' data and demand later an economic ransom to decipher it.
Botnet attacks [4, 92]	Use of botnets, networks of infected devices controlled and coordinated by an attacker, to perform particular attacks directed to specific targets.
Sniffing attacks [5]	Acquisition of private information by listening to a communication channel. When the data is not encrypted, attackers have access to the content of the whole communication.
Man-in-the-middle attacks [163]	Alteration of the communication between two entities, making the extremes believe that they are communicating directly between each other.
Replay attacks [77, 166]	Retransmission of previously acquired data to perform a malicious action, such as the impersonation of one of the legitimate participants of the communication.
Social engineering attacks [47, 49]	Psychological manipulation to gain access over restricted resources. An example is phishing attacks, based on the impersonation of a legitimate entity in digital communication.
Spoofing attacks [159, 166]	Masquerade an entity of the communication, transmitting malicious data. Frequent spoofing attacks in network communications are, among others, IP spoofing and MAC spoofing.

At this point, it is essential to note that in this document, the safety concept refers to the preservation of the physical integrity of BCI users, not focusing on the conservation of objects or the environment. To better understand the attacks and countermeasures later discussed in this section, Table 1 offers a brief description of the attacks affecting BCI, whereas Table 2 describes their countermeasures. For each phase of the BCI cycle, we detail the particularities of these attacks and countermeasures.

Figure 3 indicates the attacks, impacts, and countermeasures described in this section. As can be seen, each attack is represented by a color that associates the impacts it generates and the countermeasures to mitigate it. For each impact included in the figure, it includes a simplified version of the BCI cycle. Those phases of the cycle marked in red indicate impacts detected in the literature for that specific phase, whereas the blue color indicates our contribution. Besides, the attacks, impacts and countermeasures marked with references have been proposed in the literature, while those without references are our contribution. It is important to note that this figure highlights the limitations exposed by the literature, as can be appreciated by the volume of our contributions. To simplify the image, we have synthesized most of the safety impacts into a general entry "Cause physical damage," describing the specific impacts over users' health in detail throughout the section.

ACM Computing Surveys, Vol. 54, No. 1, Article 11. Publication date: December 2020.

Table 2. Definition of the Countermeasures Detected for the BCI Cycle

Countermeasure	Description
Training sessions, demos and serious games [59]	Initiatives to increase the awareness of the users about the risks of technology.
User notifications [24]	Alert the users in case an attack is detected, to take part in the defence (e.g., stop using the device).
Directional antennas [186]	Antennas that radiate or receive the energy mainly in particular directions, aiming to reduce interference.
Analysis of the medium [59]	Sensing of the communication medium to detect abnormal behavior.
Low transmission power [170]	Reduction of transmission power to avoid the interception of the communication by malicious entities.
Frequency and channel hopping [46, 186]	Wireless communication models based on pseudo-random hopping patterns previously known by sender and receiver.
Spread spectrum [166, 170, 186]	Transmission of the information in a broader bandwidth to avoid interference in the wireless medium.
Access control mechanisms [24, 164, 165]	Means of detecting and preventing unauthorized access to particular resources.
Privilege management [110–112]	Assign privileges to different groups of users based on roles.
Whitelists and blacklists [106]	List of entities, such as systems or users, that are allowed or forbidden, respectively, to perform specific actions.
Cryptographic mechanisms [8]	Use of encryption and decryption techniques to protect the privacy of data, since unprotected information can be accessed and modified by attackers.
Differential privacy [60, 90]	Cryptographic mechanism based on the addition of noise to the data aiming to suppress sensitive aspects, accessible when combined with a large amount of a user's data.
Homomorphic encryption [90]	Cryptographic mechanism allowing the computation of mathematical operations over ciphered data, generating an encrypted result.
Functional encryption [164, 165]	Cryptographic mechanism where having a secret key allows to learn a function of encrypted data without revealing the data itself.
Authenticity verification [8]	Ensure that the data we are accessing, or the endpoint we are communicating, is who it claims to be.
Legitimacy verification [8]	Review if a malicious software application has replaced a legitimate one.
Feature limitation [123]	Ensure that any software only implements the specific functionality for which it was intended.
Periodic updates [37]	Correct detected vulnerabilities and include new functionalities to reinforce the existing countermeasures.
Robust programming languages [110]	Choose the most adequate languages taking into consideration their strengths and weaknesses.
Compilation techniques and options [111]	Specific capabilities of compilers to protect out of bounds accesses to the device memory or CPU registers.
Application hardening [50]	Modification of an application to make it more resistant against attacks, such as the obfuscation of the application code.
Segmented application architectures [147]	Isolation of architectures and systems, establishing different containers and security groups to communicate with each other.
Sandboxing [104]	Isolate the execution of different programs, allowing its protection against attacks.
Antivirus [159]	Software focused on the prevention, detection, and elimination of malware attacks. Modern antivirus offer protection for a wide variety of threats.
Malware visualization [41]	Technique focused on the analysis of software binaries in a graphical way to detect anomalous malware patterns.

(Continued)

Table 2. Continued

Countermeasure	Description
Quarantine of devices [4]	Isolation of infected or potentially infected software, to avoid further propagation and infection.
Backup plans [3]	Recurrent copy of data stored in a different location to allow its recovery in case of data loss.
Defense distillation [90]	Creation of a second ML model based on the original, with less sensitivity regarding input perturbations and offering smoother and more general results.
Data sanitisation [66]	Rejection of samples that can produce a negative impact on the model, preprocessing and validating all input containing adversarial information.
Adversarial training [44]	Inclusion of adversarial samples in the training process to allow the recognition of attacks in the future.
Monitoring systems [15]	Capture and analyze the behavior of the entities within a system and their communications.
Anomaly detection [24]	Detection of odd behaviors on systems that can potentially correspond to an attack situation.
Firewall [159]	Cybersecurity system that only allows incoming or outgoing network communications previously authorized.
IDS [159]	Analysis of the network activity to identify potentially damaging communications aiming to disrupt the system.
Communication interruption [73]	Detention of an active communication to mitigate the impact of an attack if there is evidence of its presence.
Input validation [134]	Analysis and preprocessing of inputs presented to a system to suppress potential causes of failure.
Randomization [165]	Change of existing data in a way that does not follow a deterministic pattern and prevents privacy leakage.
BCI Anonymizer [17]	Anonymization of brain signals acquired from the brain to be shared without exposing users sensitive information.

2.1 Phase 1. Brain Signals Generation

2.1.1 Attacks. Considering the neural data acquisition flow, this first phase focuses on the brain processes that generate neural activity, which can be influenced by external stimuli. The literature has detected *misleading stimuli attacks* [40, 79, 96], a mechanism to alter the brain signals generation by presenting intentionally crafted stimuli to BCI users. To understand these attacks, it is important to introduce some concepts. *Event-related Potentials (ERP)* are neurophysiological responses to a cognitive, sensory, or motor stimulus, detected as a pattern of voltage variation [26]. Within the different types of Event-related Potentials (ERPs), Evoked Potentials (EP) focus on sensory stimuli and can be divided into two categories, Visual Evoked Potentials (VEPs) and Auditory Evoked Potentials (AEPs), related, respectively, with visual and auditory external stimuli. Specifically, *P300* is a Visual Evoked Potential (VEP) detected as an amplitude peak in the Electroencephalography (EEG) signal about 300ms after a stimulus, extensively used due to its quick response [158].

On the one hand, Martinovic et al. [96] used the P300 potential to obtain private information from test subjects and demonstrated misleading stimuli attacks. Visual stimuli were presented in the form of images, grouped as follows: four-digit PIN codes, bank ATMs and credit cards, the month of birth, and photos of people. The objective of the experiment was to prove that users generate a higher peak in the P300 potential when faced with a known stimulus and, therefore, be able to extract private information. The authors used the Emotiv EPOC 14-channel headset [36], a commercial BCI EEG device, showing that information leakage, measured in information entropy, was 10%–20% of the overall information, and could be increased to approximately 43%. On the other hand, Frank et al. [40] demonstrated the possibility of performing subliminal *misleading*

stimuli attacks. To perform the experiments, the same ERP concept with P300 potentials was used. In this work, the authors showed information hidden within the visual content projected to 29 subjects, in the form of stimuli with a duration of 13.3 milliseconds, imperceptible to the human eye. The study used EEG devices of the brands NeuroSky [118] and Emotiv [34]. We consider that the previous works are relevant to highlight the importance of security in BCI, and additional experiments with a higher number of users are required.

The literature has documented some well-known methods to present stimuli to users and analyze their neural responses [17, 96, 163]. For example, to study the neural activity generated after a question in a lie detection test [79]. Although these methods do not represent attacks themselves, they are an opportunity to develop new misleading stimuli attacks against BCIs, defined as follows:

- *Oddball Paradigm*: specific target stimuli, hidden between a sequence of common non-target stimuli, would generate peaks in ERP. For example, to differentiate a known face among several unknown ones.
- *Guilty Knowledge Test*: the response generated by familiar stimuli can be differentiated from the generated by unfamiliar elements. This principle has been used for lie detection.
- *Priming*: a stimulus can generate an implicit memory effect that later influences other stimuli.

Despite the comprehensive study in the literature on Auditory Evoked Potentials (AEPs), there are no specific works, to the best of our knowledge, describing attacks over auditory stimuli. However, Fukushima et al. [42] described that inaudible high-frequency sounds could affect brain activity. We detect that this scenario generates new opportunities for attackers, since the generation of inaudible auditory stimuli does not require close interaction with the victim, helping the attacker to remain undetected.

Regarding neural stimulation, this phase represents the result of the stimulation process within the brain. Based on a lack of literature defining taxonomies of attacks over the brain, we identify two main attack categories during neurostimulation. The first category consists of taking control of the stimulation process to cause neural tissue damage. These attacks may reproduce or worsen the secondary effects often present during the treatment of neurological conditions, such as Parkinson's disease, either by over-stimulation actions or by preventing the treatment. The feasibility of these attacks is supported by References [48, 128], who indicated that the adverse effects of neurostimulation are related to the parameters and patterns of the stimulation. Additionally, we identify another modality of attack in this category, based on recreating known neurological conditions if there is an existing neurostimulation device with access to the regions naturally affected by those diseases. As an example, we identify the possibility of recreating neurodegenerative diseases, such as Parkinson's and Alzheimer's diseases, based on a deterioration of cerebral tissue, and epileptic seizures. Although these attacks are nowadays just theoretical [11], the advance of prospecting BCI technologies like Neuralink [116], could result in neurostimulation systems that can cover various parts of the brain, thus introducing these threats.

The second category of attacks focuses on inducing an effect or perception in the user. It is well known that neurostimulation can cause multiple psychiatric and psychological impacts, such as mood variations, depression, anxiety, or suicidal thoughts, as later indicated in Section 2.1.2. An attacker could magnify these effects with malicious stimulation parameters to take advantage of the user. As an example, the attack could aim to reduce the patient's inhibition to ease the extraction of private information. This situation introduces the possibility of *social engineering attacks* to BCI, where the attacker would not require sophisticated social techniques to manipulate its victims psychologically.

Table 3. Summary of the Most Common Side Effects During FDA-approved Neurostimulation

Technology	Condition	Brain region	Neurological side effects	Psychiatric/psychological side effects
DBS	Parkinson's disease	STN	Akinesia, cramping in the face or hand, dysarthria, dysphagia, eyelid apraxia, gait disturbance, hypersalivation, impaired vision, incontinence, learning and memory difficulties, paresthesia, postural instability, speech disturbance, lack of verbal fluency, vegetative symptoms, weakness [23, 30, 33, 48, 157]	Anxiety, apathy, cognitive disturbance, confusion, depression, hallucination, submanic state [23, 33, 48]
		GPI	Similar to STN [48]	Anxiety, depression, suicidal thoughts [33, 48]
		VIM	Dysphagia, fine motor disturbance, speech disturbance [157]	
	Essential tremor	VIM	Dysaesthesia, dysarthria, gait disturbance, paresthesia, speech disturbance [23, 33]	
	Dystonia	GPI	Gait disturbance, paresis, speech disturbance, tetanic muscle contractions, visual deficits [23, 33]	Anxiety, cognitive disturbance, confusion, hallucination [23]
	Obsessive-compulsive disorder	VC/VS, NAc		Depression, operant conditioning, reward processing alteration, suicidal thoughts, suicide [102]
RNS	Epilepsy	Seizure origin	Death, change in seizures, hemorrhage, infection [117]	Anxiety, depression, suicide, suicididal thoughts [117]

2.1.2 Impacts. It is important to note that the *misleading stimuli attacks* detailed for this phase have only been conducted against data confidentiality [40, 79], aiming to extract sensitive data from BCI users. However, we consider that they can also affect BCI integrity, availability, and safety. These stimuli can alter the normal functioning of this phase, generating malicious inputs for the next stages that can derive on disruptions of the service or incorrect actions aiming to cause physical damage to users. Specifically, Landau et al. [79] identified that misleading stimuli attacks performed during a medical diagnose, such as a photosensitive epilepsy test in which different visual stimuli are presented, can derive in a misdiagnosis, affecting the users' safety. We also identify as feasible that malicious stimuli, both perceptible or subliminal, can affect the users' mood.

From the perspective of neurostimulation, the attacks above can affect users' health differently according to their previously existing diseases, impacting their physical and psychological safety. The issues related to different BCI technologies are detailed in Section 2.2, indicating general impacts over the brain in this phase. Table 3 presents the most common side effects during particular neurostimulation therapies. As can be seen, performing an attack during the stimulation process can aggravate or even generate a wide range of negative impacts on BCI patients. Additionally, the authors of References [135, 136] highlighted common issues to neurological diseases, such as tissue damage, rebound effects, and denial of stimulation (also affecting the service availability). Besides, they identified that an alteration of voltage, frequency, pulse width, or electrode contact used to stimulate the brain could modify the volume of cerebral tissue activated, inducing non-desired effects in the surrounding structures depending on the electrode location and stimulation technique. Pycroft et al. [135] also indicated that an attack on neurostimulation could induce a patient's thoughts and behavior. In Reference [95], the authors highlighted that attacks on neurostimulation can prevent patients from speaking or moving, cause brain damage or even threaten their life, while the authors of Reference [79] indicated the user's frustration if the result of the process is not adequate.

Pycroft et al. [136] indicated potential attacks and harms against neurostimulation patients. First, they detected that an overstimulation procedure could cause tissue damage, independently of the type of stimulation and medical condition. For Parkinson's disease, an attacker could apply

a ~10Hz stimulation over the STN region to produce hypokinesia or akinesia. In patients with essential tremor, where the ventral intermediate nucleus (VIM) is stimulated, both an increase of voltage and a decrease of frequency could dangerously derive in exacerbated tremor. Finally, a variation in the stimulation parameters during the treatment of obsessive-compulsive disorder could generate alterations of reward processing or operant conditioning.

Based on the above, safety impacts are the most damaging in this phase, presenting a risk of irreversible physical and psychiatric issues. In addition, taking advantage of the victim's psychological status, it could ease social engineering attacks as well. The attacker could aim to reduce or inhibit the patient's mental defense mechanisms, acquiring sensitive information, thus impacting data confidentiality. However, more worrisome would be to take advantage of the victim's mental status, in which the patient unconsciously accedes to undesired acts, such as gambling money, buying unnecessary products, committing a crime, or participating in non-consensual sexual intercourse.

2.1.3 Countermeasures. Focusing on the countermeasures to mitigate misleading stimuli attacks, multiple works [24, 79, 135, 136] identified general measures to raise the awareness of BCI users, such as spreading the risks of these technologies among clinicians and patients and the education of the users in these technologies. This is especially interesting, since humans usually are the weakest element of a security system. In particular, Ienca et al. [59] indicated that specific training sessions could be beneficial to protect users against potentially unsafe stimuli related to authentication methods and banking-related information. Besides, the inclusion of demos and serious games in commercial BCI devices may educate them on the risks of these technologies. However, these countermeasures can only be applied when the user is aware of the stimuli. Because of that, we consider that *misleading stimuli attacks* can be reduced if BCIs are complemented with external systems that monitor the stimuli presented and give users the possibility to evaluate if the content is appropriate. For example, by analyzing if the multimedia contents showed to users, such as images or videos, have been maliciously modified [15, 175], even if they are subliminal. Additionally, we propose using predictive models based on anomaly detection systems, aiming to detect an attack in its early stage and deploy mechanisms to mitigate them.

2.2 Phase 2. Neural Data Acquisition and Stimulation

2.2.1 Attacks. This second phase focuses on the interaction of BCI devices with the brain to acquire neural data or perform its stimulation. Regarding data acquisition, the authors of References [79, 87] identified the use of a combination of *replay and spoofing attacks* in which previous signals from the BCI user, signals from other users, or synthetic signals can impersonate the legitimate brain waves. We detect the applicability of these attacks to stimulation systems, where an attacker can force specific stimulation behaviors based on previous actions. One possible outcome of this control can be an increase in the voltage delivered to the patient's brain [95]. Besides, the authors of References [59, 79] detected the use of *jamming attacks* against the neural data acquisition process, transmitting electromagnetic noise to the medium. Based on Vadlamani et al. [170], we also identify this problem in neural stimulation, where *jamming attacks* can override the legitimate signals emitted by the BCI electrodes if they are transmitted with enough power.

2.2.2 Impacts. Regarding the impacts produced by the previous attacks, Li et al. [87] identified that *replay and spoofing attacks* affect both data integrity and availability, being able to disrupt the acquisition process. Landau et al. [79] highlighted that these attacks could interfere with clinical diagnosis procedures, replacing the legitimate brain signals by malicious ones, concluding in misdiagnosis, and producing either an absence of treatment or an unnecessary one on healthy patients. We identify that these attacks, applied to the stimulation scenario, can disrupt the

stimulation process or acquire and modify the stimulation pattern used by the BCI to maliciously stimulate the neurons, affecting data integrity, data and service availability, and the patient's safety. Focusing on *jamming attacks*, an attacker can aim to prevent the electrodes from capturing brain signals due to the noise transmitted [59, 79], affecting their availability and safety. We detect that jamming attacks can also affect neurostimulation scenarios, where signals with enough power can override the legitimate ones, affecting the integrity and availability of the data, as well as the patient's safety during stimulation actions.

Apart from the impacts derived from the previous attacks, it is important to note that each specific BCI technology presents specific risks according to their invasiveness and functioning, and thus the impact generated by an attack differs. To analyze this situation, we select some of the most used BCI technologies used to acquire neural data or stimulate the brain. For each one of them, we address specific considerations to evaluate their impact.

Regarding the issues related to acquisition technologies, it is necessary to consider both their temporal and spatial resolutions. We identify that a low temporal resolution in acquisition technologies presents concerns on data and service availability, since the devices transmit a reduced amount of data that can be affected more easily by electromagnetic interference and, especially, *jamming attacks*. Besides, this situation can also be beneficial for *replay and spoofing attacks*, since attackers have more time to prepare and send malicious data. A high spatial resolution can impact on data confidentiality, allowing attackers to have access to more sensitive neural data. It is worthy to note that attacks on technologies such as Functional Magnetic Resonance Imaging (fMRI) or Magnetoencephalography (MEG) can potentially have a higher economic impact due to the high cost of these technologies compared to others like EEG [82, 137]. Nevertheless, EEG is the most studied acquisition technology from the security perspective, due to its wide availability outside clinical environments, highlighting the feasibility of attacks such as *misleading stimuli attacks* or *jamming attacks*.

Although the literature has documented some potential security impacts for acquisition technologies, the impact of neurostimulation technologies on patient's health has been studied in a more detailed way, specifically in the field of Implantable Medical Devices (IMDs). Because of that, we first introduce the most common stimulation technologies nowadays to review their specific impact later, mainly addressing safety issues.

Focusing on the specific impacts of neurostimulation technologies, Deep Brain Stimulation (DBS) is the most studied one due to its invasiveness, where Medtronic is one of the most popular brands commercializing open-loop DBS devices [128]. The side effects of this method have been extensively studied in the literature, where some of them have previously been presented in Table 3 for the treatment of particular conditions. According to Pycroft et al. [136], the use of Deep Brain Stimulation (DBS) with high charge densities can cause tissue damage. Furthermore, an increase or decrease in the stimulation frequency can have a considerable impact on its efficacy, even reversing the stimulation effect. Finally, an alteration of emotion and affect processing can occur during DBS as side-effects, such as pathological crying or inappropriate laughter, having a distressing impact.

Moving to Transcranial Magnetic Stimulation (TMS), Polanía et al. [129] indicated that pulses applied to particular areas could induce suppression of visual perception or speech arrest, which serves as an opportunity for attackers. León et al. [84] highlighted that Transcranial Magnetic Stimulation (TMS) could produce side-effects such as headache and neck pain, being epileptic seizures possible but improbable. The side effects of Transcranial Electrical Stimulation (tES) usually are mild, such as skin tingling, itching, and redness [114]. Nevertheless, this technique can have indirect effects on the stimulation of non-neuronal elements, such as peripheral nerves, cranial nerves, or retina. Because of that, the stimulation is limited to maximum tolerable doses [89].

ACM Computing Surveys, Vol. 54, No. 1, Article 11. Publication date: December 2020.

Besides, in patients with depression, Direct Current Stimulation (tDCS) can derive to mania and hypomania cases [99]. It is worthy to note that the side effects described above can naturally arise in controlled environments where clinicians have strict control over the procedure. However, if attackers alter the therapy, they could recreate or amplify malicious conditions, generating a clear impact on patients' health.

The Neuropace RNS is a closed-loop neurostimulation system for treating drug-resistant epilepsy, performing both neural data acquisition and neurostimulation procedures. It presents the advantage of delivering stimulation only when detecting the beginning of seizure activity, reducing secondary effects. Nevertheless, it introduces potential challenges than can be used by an attacker to impact its users' safety [128]. First, we identify that the closed-loop behavior could induce, in both clinicians and patients, a reduction of the perception of risks, assuming that the device is working correctly. Furthermore, since the device presents autonomous capabilities, an attacker could disrupt its behavior, without the knowledge of the user, to generate an impact on data confidentiality, service availability, and safety.

2.2.3 Countermeasures. Regarding the countermeasures to detect and mitigate replay and spoofing attacks, Landau et al. [79] proposed, for data acquisition, the use of anomaly detection mechanisms to detect modified inputs, as well as the accuracy improvement of acquisition devices. Besides, we propose a mechanism able to disable the electrodes not required for the current application usage and avoid potential risks, such as the acquisition of P300 in brain signals. This action could be performed automatically by the BCI system or based on the patient's or clinician's decision. Taking into account neural stimulation, and specifically for IMDs, external devices to authenticate and authorize the stimulation actions can be used [24]. The authors of References [46, 170, 186] documented several detection mechanisms and countermeasures related to the mitigation of jamming attacks. All detection procedures are based on an analysis of the medium to detect abnormal behavior, as identified for neural data acquisition by Ienca et al. [59]. Specifically, Landau et al. [79] proposed using an ensemble of classifiers to detect the addition of noise to the benign input. As proposed countermeasures, Vadlamani et al. [170] identified the use of low transmission power as a possible solution to harden the detection of the legitimate transmission, and the use of directional antennas oriented to the brain to avoid the jamming. The use of frequency hopping [186] and channel hopping [46] after a particular duration of time also aim to reduce the impact of these attacks. We detect that the use of directional antennas is also a possible solution for *replay and spoofing attacks*. Finally, it is worthy to note that the mitigation of the previous impacts focused on user's safety is the consequence of mitigating the attacks spotted against BCI devices.

In the scenario of closed-loop neurostimulation systems, we identify as essential to have information about the behavior of the device, from both acquisition and stimulation procedures. These feedback mechanisms would allow to externally analyze the status of the brain and the stimulation decisions. Another proposal is the use of anomaly detection systems, included in the device, to identify unusual stimulation parameters, or an absence of treatment when a seizure occurs, notifying the user. This second approach could be more energy preserving, and the election of the strategy would depend on the use case.

2.3 Phase 3. Data Processing and Conversion

2.3.1 Attacks. This phase performs the data processing and conversion tasks required to allow neural data and stimulation actions to be ready for subsequent stages. Although the literature has not detected security problems in this phase, according to the aspects indicated by Bonaci et al. in References [17, 18], we identify *malware attacks* as possible against this phase, taking control over the BCI. These attacks are candidates to affect both acquisition and stimulation processes,

impacting the tasks performed in this phase. In particular, we identify that malware can disrupt the analog-to-digital conversion that occurs during neural data acquisition, as well as the translation of firing patterns to particular stimulation devices. We also detect that *jamming attacks* applied to the previous phase for data acquisition can impact this phase, since a distorted input signal with enough noise can be difficult to filter and thus propagate this signal to subsequent phases.

2.3.2 Impacts. In this context, we identify that *malware attacks* have an impact on both neural data acquisition and stimulation, where attackers alter or override the data received from previous phases, generating malicious data sent to subsequent phases. That is, the analog data recorded during neural data acquisition or the firing pattern used in neurostimulation processes. These attacks can gather the sensitive data managed in this phase, both analog and digital, and send it to the attackers, affecting data confidentiality. For example, information about private thoughts or neurological treatments. In terms of data and service availability, both acquisition and stimulation are potentially vulnerable to malware that avoids data transmission to subsequent phases of the cycle. Malware affecting integrity and availability is also a threat against users' physical safety, generating damaging stimulation patterns or dangerous actions sent to applications. Besides, the impacts and countermeasures described in the first phase of the acquisition flow for jamming attacks are also applicable to the current stage.

2.3.3 Countermeasures. Regarding the countermeasures to mitigate attacks affecting data confidentiality, Chizeck et al. [26] defined a U.S. patent application entitled "Brain-Computer Interface Anonymize" that proposes a technology capable of processing neural signals to eliminate all non-essential private information [17, 165]. As a result, sensitive information is never stored in the BCI device or transmitted outside. We identify this method as especially relevant in this phase, as it is the first stage after the BCI's acquisition process. Although the authors do not provide details about techniques or algorithms to understand how raw signals are processed, they indicate that this process can only be performed on hardware or software within the device itself, and not on external networks or computer platforms, as a way to ensure the privacy of the information. Besides, Ienca et al. [60] proposed the use of *differential privacy* to improve the security and transparency of data processing.

The countermeasures to mitigate malware depend on their type and behavior. We consider the use of antivirus software and Intrusion Detection Systems (IDS) as alternatives for the protection of individual devices, based on Reference [79]. Besides, the authors of References [159, 177] considered perimeter security mechanisms, such as *firewalls*, responsible for analyzing all incoming and outgoing communication of the device. We also propose using Machine Learning (ML) anomaly detection systems to identify potential malware threats [24, 141]. Finally, Chakkaravarty et al. [154] reviewed current persistent malware techniques able to bypass common countermeasures and proposed mitigation techniques, such as *sandboxing* [104], *application hardening* [50], and *malware visualization* [41]. It is essential to highlight that the countermeasures applicable for this phase highly depend on the device constraints that implement this phase, which is typically the BCI device (see Section 3).

2.4 Phase 4. Decoding and Encoding

2.4.1 Attacks. Decoding and encoding is the phase focused on identifying the action intended by the users in neural data acquisition or the specification of the neural firing pattern in neurostimulation. *Malware* attacks have been identified in the literature by Bonaci et al. [17, 18] from the signal acquisition perspective. Specifically, they identified that attackers could use *malware* to either override the functioning of this phase or to implement additional malicious algorithms. Besides, we identify that *malware* attacks can also be applied to the stimulation flow, avoiding or

disrupting a firing pattern's generation. Besides, we identify that *adversarial attacks* can also be applied to this phase for both acquisition and stimulation tasks, taking advantage of the classification algorithms used. These attacks affect all types of ML models, and, because of that, they are currently an open challenge [38]. Liu et al. [90] detected the possibility of *poisoning attacks*, where attackers introduce crafted adversarial samples to the data, aiming to change its distribution. *Evasion attacks* aim to create samples that evade detection systems, whereas *impersonate attacks* focus on adversarial samples that derive in incorrect classification of the legitimate ones. Finally, two attack models exist according to the knowledge about the model [44]. In *white-box attacks*, adversaries know the model, while in *black-box attacks*, they only have access to the model through a limited interface.

2.4.2 Impacts. The previously described attacks generate particular impacts on BCI. On the one hand, *malware* has an impact on data integrity and availability, as it can alter or ignore the received data from previous phases, and override the output of the current one. That is, disrupt the intended action sent to BCI applications in the acquisition process, such as preventing the control of a wheelchair or changing its direction, or the firing pattern in neural stimulation, enabling a wide variety of attacks as described in Section 2.1. Besides, *malware* affects the availability of the ML process by the alteration of the trained model or the ML algorithm. From a data confidentiality perspective, *malware* can access the features used in the ML training phase, as well as gather information about the model and the algorithm used. *Malware* also affects users' safety, as the previous integrity and availability impacts derive in malicious actions and firing patterns that affect the integrity of users, such as causing neural damage or inducing particular psychological states. On the other hand, *adversarial attacks* also affect data integrity and availability, as the introduction of malicious samples aiming to disrupt the model can alter or avoid the generation of actions and firing patterns. Shokri et al. [153] demonstrated that ML models are sensitive against *adversarial attacks*, aiming to detect if a sample exists in the model's training dataset. Based on that, an attacker may extract sensitive users' data, such as previous intended actions or used patterns during stimulation actions. Taking into account data confidentiality, Landau et al. [79] detected that a malicious entity taking control of the output of this phase could access the user's intention. Finally, the use of malicious samples, as is the case of *poisoning attacks*, alter the ML system, deriving in safety impacts for both cycle directions.

2.4.3 Countermeasures. To mitigate the attacks on the ML training phase affecting integrity and availability, we have identified several techniques proposed in the literature for generic *adversarial attacks*, that can serve as an opportunity to improve the security of BCI. First, *data sanitization* is useful to reject samples containing adversarial information, thus disrupting the model. Jagielski et al. [66] proposed a similar approach against poisoning attacks applied to regression techniques, where noise and *outliers* are suppressed from the training dataset. Nevertheless, it does not prevent attackers from crafting samples similar to those generated by the legitimate distribution. Countermeasures such as *adversarial training* or *defense distillation* have been presented in this context. However, both have limitations, as they depend on the samples used during the training and can be broken using *black-box attacks* and computationally expensive attacks based on iterative optimization [44, 90]. Goodfellow et al. [44] also proposed *architecture modifications*, based on the improvement of ML models to be more robust, but this derives in models difficult to train that have degradation in the performance when used in non-adversarial situations. Liu et al. [90] documented the integration of techniques to mitigate the attacks, called *ensemble method*. They also indicated two methods that can apply in both training and testing phases: *differential privacy* and *homomorphic encryption* [56, 90, 165]. Finally, it is worthy to note that the countermeasures to mitigate *malware attacks* in the previous phase can apply to the current one.

2.5 Phase 5. Applications

2.5.1 Attacks. From the data acquisition context, applications perform in the physical world the actions intended by users through their neural activity. These actions can range from the interaction with a computer or smartphone, to the control of a robotic limb. From the perspective of neural stimulation, applications are the entry point of the information transmitted to the brain, like sensory stimuli in prosthesis or cognitive enhancement. In this section, we consider attacks on applications, without analyzing their communication with external systems, addressed in Section 3.1.

Considering the issues of this phase, *spoofing attacks* over BCIs have been detected in the literature, where an attacker creates malicious applications identical to the original and make them available in app stores [8]. The authors of References [17, 18, 87] identified *malware attacks* as a threat in BCI. Besides, Pycroft et al. [136] identified that the use of consumer devices, such as smartphones, generates new risks and security problems. Specific considerations about malware are the same as detailed in Sections 2.3 and 2.4. Moreover, we have found several opportunities related to cyberattacks performed against applications. In particular, we detect security misconfiguration issues, Buffer Overflow (BO) attacks, and injection attacks over applications. However, the detailed analysis of these particular attacks is out of the scope of this work, and we only address general aspects related to BCI.

2.5.2 Impacts. Landau et al. [79] identified multiple risks on BCI applications with the independence of any attack. They detected that an attacker could interfere with the user's ability to use the device, impacting its availability. They also detected confidentiality concerns regarding the identification of users by their neural data, illustrating a scenario in which an attacker extracts EEG data from the application and compares it with the EEG database of a hospital, identifying the user and accessing his or her medical records. This identification can derive in a discrimination situation based on the belonging of specific groups, such as religious beliefs. Besides, most BCI development APIs offer full access over the information and do not implement limitations on the stimuli presented to users, generating confidentiality issues [17, 40, 87, 96, 163, 165]. Finally, all the attacks affecting this phase can force applications to send malicious stimuli or actions, causing physical harm [8].

Considering the impact of the previous attacks, applications created by *spoofing attacks* affect both data integrity and confidentiality, as they can present malicious stimuli to obtain sensitive neural information, such as thoughts or beliefs [8]. In neurostimulation scenarios, we identify that these fraudulent applications could entirely modify the firing patterns used to stimulate the patient, generating a high impact over safety. More particularly, these applications could induce psychological states in the victim, making them more willing to gamble, or even generate adverse effects such as anxiety and depression. Based on that, the attacker could take advantage of these mental states, injecting in-app advertisements to earn money from the victim.

Malware attacks impact the integrity of the applications by altering their services and capabilities, such as disabling the encryption of information. Besides, they can compromise applications' confidentiality, gaining access to sensitive information such as medical records and user profiles used during neurostimulation treatments. Concerning the availability of the application, *malware attacks* can derive in denial of service over the application, impacting in processes such as controlling prosthetic limbs or wheelchairs.

We detect that *misconfiguration attacks* present data integrity issues, where attackers take advantage of the system to gain unauthorized access, such as weak access control mechanisms. Data confidentiality issues are also present, for example, on configuration files that have static predefined passwords, allowing attackers to gain access to users' private data. Applications' availability

ACM Computing Surveys, Vol. 54, No. 1, Article 11. Publication date: December 2020.

problems are also possible, as a misconfiguration issue can serve as a first step to disrupt the normal behavior of the BCI application.

Moving to *injection attacks*, they can produce data loss, modification, and corruption, affecting the integrity of applications [105, 134]. In terms of confidentiality, they can produce the disclosure of sensitive information to unauthorized parties [105, 134], such as insurance companies aiming to select the best candidates for their products [8]. Availability can be affected by a denial of access over an authentication system, or producing crash, exit or restart actions on the applications, disrupting vital processes such as clinical neurostimulation [107, 134].

Buffer Overflow (BO) attacks can derive in the execution of unauthorized code or commands, where an attacker can alter the normal functioning of the application or access to sensitive information [110]. Furthermore, they can also aim to bypass protection mechanisms by the execution of code outside the scope of the program's security policy. These actions can affect the data integrity, confidentiality, and availability of the application [111].

2.5.3 Countermeasures. It is necessary to verify the legitimacy of the applications and ensure sufficient control of the app stores to mitigate *spoofing attacks* [8]. In that regard, Landau et al. [79] proposed the use of applications developed by authorized organizations to ensure their trustworthiness. When it comes to *malware attacks*, the same countermeasures proposed for the *Data processing & conversion* phase also apply for applications. That is, the use of antivirus, firewall, Intrusion Detection Systems (IDS), and anomaly detection systems to identify and mitigate the attacks. Furthermore, Takabi et al. [164, 165] proposed the use of access control mechanisms over the information to restrict its access and thus mitigate confidentiality impacts. They also indicated the use of randomization and differential privacy. Besides, they proposed the integration of *homomorphic encryption* to operate with encrypted information combined with *functional encryption* to access only to a subset of the information.

As an opportunity for BCI, we identify some preventive actions against *misconfiguration attacks* defined by the Open Web Application Security Project (OWASP) [123], such as the use of minimal platforms with only necessary features, components, libraries, and software to reduce the probability of misconfiguration issues. Moreover, a periodic review and update of configuration parameters are also beneficial as part of the management process of applications. It is also necessary to create segmented application architectures that offer a division between components and defines different security groups, using Access Control Lists (ACLs).

Concerning *BO*, it is important to use programming languages that protect against these attacks, as well as the use of compilers with detection mechanisms. [147]. Developers must validate all inputs and follow well practice rules when using memory (e.g., verification of the boundaries of buffers). Moreover, sensitive applications must be ran using the lowest privileges possible and even isolated using sandbox techniques [110–112]. To detect *injection attacks*, both static and dynamic analysis of applications' source code have been proposed [134]. For their mitigation, it is necessary to escape all special characters included in the input [107, 134]. Multiple solutions have been proposed, such as the use of whitelists and blacklists [106], the use of safe languages and APIs containing automatic detection mechanisms [105, 134], the use of sandboxing techniques to define strict boundaries between processes [107], the definition of different permissions on the system [106], and error messages with minimal but descriptive details.

3 SECURITY ISSUES AFFECTING THE BCI DEPLOYMENTS

This section reviews the different architectural deployments of the BCI cycle found in the literature. After that, we group them into two main families, characterized by the BCI cycle implementation and its application scenario. In contrast to Section 2, where the security analysis

is independent of the deployment, this section reviews the state of the art of existing attacks affecting the devices implementing each phase of the BCI cycle, as well as their impacts and countermeasures. New opportunities, in terms of attacks and countermeasures, missed by the literature, are also highlighted in this section. Figure 4 represents both architectural deployments defined, Local BCIs, and Global BCIs, indicating the communication between their elements and the phases of the BCI cycle that each element implements according to the type of deployment.

3.1 Local BCI

3.1.1 Architecture Description. Local BCI deployments highlight by managing the neural data acquisition and stimulation processes of single users. This architecture typically deploys the BCI phases between two physical devices, as represented in Figure 4. The first one, identified as *BCI device*, focuses on the neural acquisition and stimulation procedures (phases 1 and 2 of the BCI cycle). In contrast, BCI applications (phase 5) run in a Near Control Device (NCD), a PC or smartphone that controls the BCI device using either a wired or wireless communication link. Phases 3 and 4 of the cycle can be implemented equally in both devices, where manufacturers make the final decision. At this point, it is essential to note that alternative designs can arise due to specific requirements of the deployments, such as the presence of multiple users. Moreover, we consider fully implantable BCIs within this architecture, since they require an external device for its configuration and verification.

3.1.2 Examples of Deployments. This kind of architectural deployment is the most commonly implemented for consumer-grade BCIs, where commercial brands like NeuroSky or Emotiv focus on scenarios such as gaming and entertainment [1, 96, 100]. Neuromedical scenarios also use this approach, where an Near Control Device (NCD) placed in the clinical environment manages the acquisition and stimulation processes. This section specifically addresses the issues detected in physical BCI devices, the inherent problems of the NCD, and those related to the communication between BCI and NCD. At this point, it is important to note that the attacks, impacts, and countermeasures detected for the BCI cycle are also applicable.

3.1.3 Attacks. Focusing on BCI devices, Ballarin et al. [8] identified attacks affecting the device *firmware* throw a configuration link (e.g., USB ports), having an impact on data integrity and confidentiality, also generating disruptions on the system. Pycroft et al. [136] identified the possibility of injecting malicious firmware updates. Moreover, we identify that these attacks can serve as an opportunity to generate safety problems. Ienca et al. [58, 59] documented *cryptographic attacks*, indicating that Cody's Emokit project was able to crack the encryption of data directly from the Emotiv EPOC, a consumer-grade BCI. They detected that these attacks affect data integrity and confidentiality. Marin et al. [95] detected that current Implantable Medical Devices (IMDs) lack robust security mechanisms. Yaqoob et al. [178] identified that neurostimulation devices lack encryption and usually define default passwords, impacting integrity and confidentially, easing unauthorized access to sensitive data. We also identify that they produce service availability and safety issues if they can modify the data.

The authors of References [24, 135] highlighted that attackers could focus on draining the battery of the device and thus affect both service availability and users' physical safety. In neurostimulation systems, losing the battery capacity would result in a loss of treatment, where the disease symptoms would reappear. Due to this, some IMDs include rechargeable batteries, reducing the risks of depleting them, and thus defining more robust solutions. It is also essential to consider that, in non-rechargeable systems, surgery is required to replace the batteries, increasing the risk of both physical and psychological safety issues.

ACM Computing Surveys, Vol. 54, No. 1, Article 11. Publication date: December 2020.

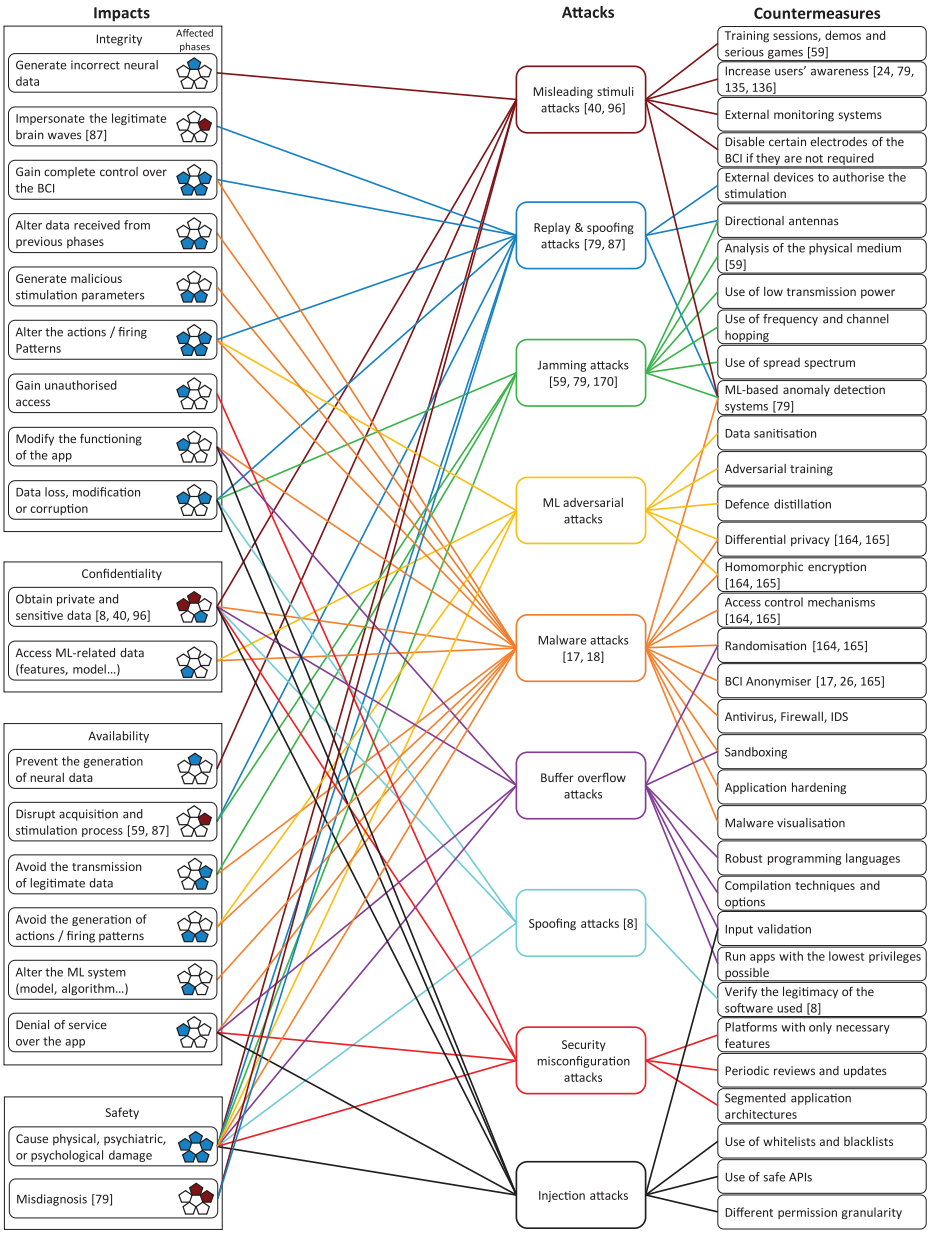


Fig. 3. Relationship between the attacks, impacts, and countermeasures over the BCI cycle. The phases of the cycle colored in red for each impact represent issues documented in the literature, while those marked in blue are our contribution. The attacks, impacts and countermeasures followed by references have been documented in the literature, and those without a cite represent our contribution.

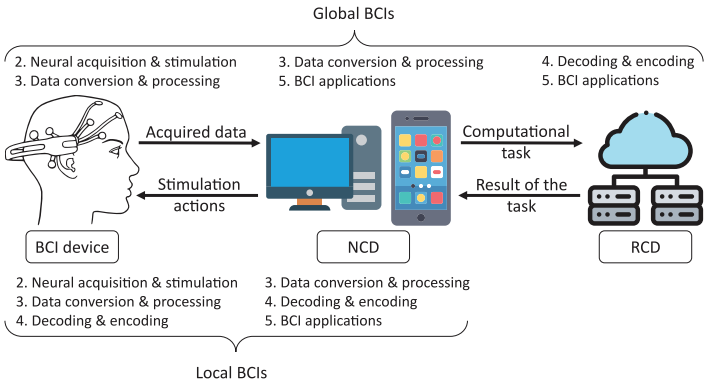


Fig. 4. Representation of Local BCI and Global BCI deployments, indicating the communication between their elements and the stages of the BCI cycle that each element implements according to the architectural deployment.

The authors of References [17, 136] described the possibility of *hijacking attacks*, referred to as *brainjacking*, where the attacker acquires complete access over the device by any means. These attacks generate an impact on all four security impact metrics. Finally, Pycroft et al. [135] identified general confidentiality impacts than can be shared by multiple attacks. They identified that close-loop IMDs use physiological data acquired by the BCI to improve the stimulation procedures or drug delivery. However, this sensitive data can be used by attackers to acquire information about the patient's health condition. Furthermore, an attacker can acquire sensitive information stored in the device, such as stimulation settings, personal data, or battery status, useful to perform new attacks.

Considering NCDs, Ballarin et al. [8] identified *social engineering and phishing attacks* against BCIs, focused on the acquisition of users' authentication credentials, affecting data confidentiality. Although BCI applications do not require a connection to the Internet, the NCD can be connected. Therefore, we detect that these systems can suffer *malware attacks* and, specifically, *ransomware* [2] and those based on *botnets* [74, 77, 159], with an impact on the integrity and availability of data and applications contained in the NCD, as well as users' safety. In particular, *botnets* also generate data confidentiality issues, since attackers have control over the system. Moreover, we detect *sniffing attacks* on NCDs taking advantage of networking configuration and protocols, such as MAC flooding, DHCP attacks, ARP spoofing, or DNS poisoning [5], affecting service and data integrity, confidentiality, and availability.

Focusing on the communication between BCI devices and NCDs, Sundararajan et al. [163] studied the security of the commercial-grade Emotiv Insight, which implemented Bluetooth Low Energy (BLE) in its version 4.0 to communicate with a smartphone that contains the application offered by Emotiv. They successfully performed *man-in-the-middle attacks* over the Bluetooth Low Energy (BLE) link, being able to intercept and modify information, force the BCI to perform unwanted tasks, and conduct *replay attacks* affecting, therefore, integrity, confidentiality, and availability of sensitive data. The literature has documented further integrity and confidentiality impacts, where attackers can intercept and modify sensitive data even using encryption [8, 79, 87, 95, 135, 163, 164]. These attacks are related to the *cryptographic attacks* described above, where weak encryption of the data stored in the device can derive in *man-in-the-middle attacks*. Finally,

it is important to note that the attacks related to user data and credentials have a higher impact if multiple users use the system.

3.1.4 Countermeasures. To some of the previous attacks, different countermeasures have been proposed. Related to *firmware attacks*, Ballarin et al. [8] indicated the encryption of the firmware, as well as an authenticity verification throw hash or signature. Pycroft et al. [136] highlighted periodic firmware updates and the use of authorization mechanisms for these updates. The authors of References [24, 135, 136] identified the use of access control mechanisms placed in external devices with proximity to the patient and anomaly detection systems over the BCI device usage to face potential threats such as *battery drain attacks*. In particular, for these attacks, rechargeable batteries are recommended to avoid a surgical replacement. The authors of Reference [79] proposed, as general countermeasures, the regulation of neurotechnology as a way to standardize its manufacturing processes, as well as a reduction of BCI training process, which tends to frustrate the users, being less willing to cooperate. These measures are complementary with those documented by Reference [135], which considered that BCI devices should keep logs and access events, including mechanisms for reporting bugs.

The use of robust cryptographic mechanisms and the latest protocol versions are determinant to avoid *cryptographic attacks*, *man-in-the-middle attacks*, and *sniffing attacks* [8, 163]. Besides, anonymization of the information transmitted from BCI to NCD is also recommendable against attacks impacting confidentiality, for example, using the BCI Anonymizer [17, 18, 164]. *Social engineering and phishing attacks* focused on credential theft can be reduced by implementing a second authentication factor to access the BCI and proper access control mechanisms [8, 135, 165]. The application of the *malware* countermeasures indicated in Section 2.3 can evade global *malware* threats impacting NCDs, by updating all software to the latest version and implementing periodic backup plans. Moreover, the use of ML techniques, as proposed by Fernández-Maimó et al. [37] for Medical Cyber-Physical Systems (MCPS), can also be used to detect, classify, and mitigate *ransomware attacks*. Concerning *botnets*, a wide variety of detection techniques have been detected by us for the BCI field, like the use of anomaly detection based on ML and signatures, the quarantine of infected devices, and the interruption of particular communication flows [4, 73, 92]. Finally, we consider that the recommendations of the U.S. Food and Drug Administration (FDA) for premarket and postmarket management of security in medical devices apply to BCI [150, 168, 169].

3.2 Global BCI

3.2.1 Architecture Description. Global BCI architectures focus on the management of neural data acquisition and neural stimulation of multiple users through an Internet connection. This architecture considers three devices to deploy the phases making up the BCI cycle, as can be seen in Figure 4. In this family, the BCI device remains focused on data acquisition and stimulation (phase 2), whereas the NCD is in charge of the execution of applications (phase 5), as well as conversion and processing actions (phase 3). Finally, the new element introduced in this architecture is the Remote Control Device (RCD), representing one or more external resources or services accessible via the Internet, such as cloud computing and storage. It typically implements phases 4 and 5 of the BCI cycle, as it has the resources to run more complex applications and information analysis. The main difference between this architecture and the one described for Local BCIs in Section 3.1 is that, in Local BCIs, the NCD does not send user information to external services (e.g., cloud). Finally, this section focuses on the problems associated with the communication between NCD and RCD, and the BCI-related attacks that can apply to RCDs. However, these later attacks are addressed in a general way, as specific cloud computing attacks are outside the scope of this article.

3.2.2 Examples of Deployments. This architectural deployment is the most innovative, as it allows the communication of multiples users with external services and the creation of complex deployments, where the data and information of every user are stored and managed in a shared infrastructure. From a commercial point of view, Emotiv allows users to contrast their data with the data stored by other users, as well as keep users' neural recordings in the cloud to visualize and manipulate them, also offering an API called Emotiv Cortex [35]. Besides, several companies worldwide provide distributed BCI services, as is the case of Lifelines Neuro [88], which offers a continuous EEG acquisition, storage, and visualization in their cloud platform. These scenarios are especially relevant in the context of personalized medicine and early diagnosis.

3.2.3 Attacks and Countermeasures. Considering the attacks on this deployment, the issues documented in Section 3.1 for Local BCIs are also applicable in this architecture. However, Global BCIs present higher risks, since these deployments are an opportunity for remote attacks against interconnected BCI devices, which derives in physical harm for their users. Furthermore, Takabi et al. [165] detected that BCI applications could send raw brain signals to cloud services that execute ML techniques to extract sensitive information and therefore affect confidentiality. We identify that this problem can also be present in Local BCIs if the NCD has an Internet connection. Ballarin et al. [8] identified that *man-in-the-middle attacks* could occur in the communication channel between NCD and RCD, affecting the integrity and confidentiality of the data transmitted as well as the service availability. They also detected that attacks on RCDs could have a higher impact on confidentiality than on Local BCIs, as these platforms store sensitive information from multiple users, that can be stolen or sold to third parties. Ienca et al. [60] detected different issues in Global BCIs in terms of their usage. First, they highlighted that current brands, such as Emotiv [34], indicate in their privacy policy that they can gather personal data, usage information, and interactions with other applications, and that they can infer information from these sources, with potential confidentiality issues. The authors identified as possible the use of big data to extract associations and share the data with third parties. Moreover, they detected that the use of cloud services could derive in a massive database theft with sensitive data, an unclear legal liability in case of breaches.

We identify that this architecture is quite similar to those defined and implemented for Internet of Things (IoT) scenarios, where constrained devices communicate with external services via intermediate systems, especially when multiple devices interact. We detect that most of the security attacks and impacts defined by Stellios et al. [160] are also applicable in this architecture. Moreover, we consider that the issues highlighted by the OWASP in their IoT projects are critical aspects of Global BCIs [125]. This relationship between IoT and external services has been previously studied in cloud computing scenarios [19]. Despite the advantages, attacks on cloud computing can impact integrity, confidentiality, and availability in different cloud architecture levels, such as infrastructure, networking, storage, and software [9, 155]. The evolution of NCDs derives in mobile devices with higher computing capabilities, integrated into mobile cloud computing systems. However, they also have an impact on the security of deployments [113]. We also detect that the improvement of NCDs capabilities can also allow the introduction of fog computing in Global BCIs, where NCDs perform part of the computation, generating new security and trust issues [93, 142, 183]. *Malware attacks* are also present in cloud environments, where ransomware and botnets are common threats [155].

Focusing on general cloud computing countermeasures, Amara et al. [3] identified security threats and attacks, as well as the mitigation techniques against them. The use of honeypots, firewalls, and IDS in cloud scenarios is convenient to reduce the impact of *malware attacks* [142].

ACM Computing Surveys, Vol. 54, No. 1, Article 11. Publication date: December 2020.

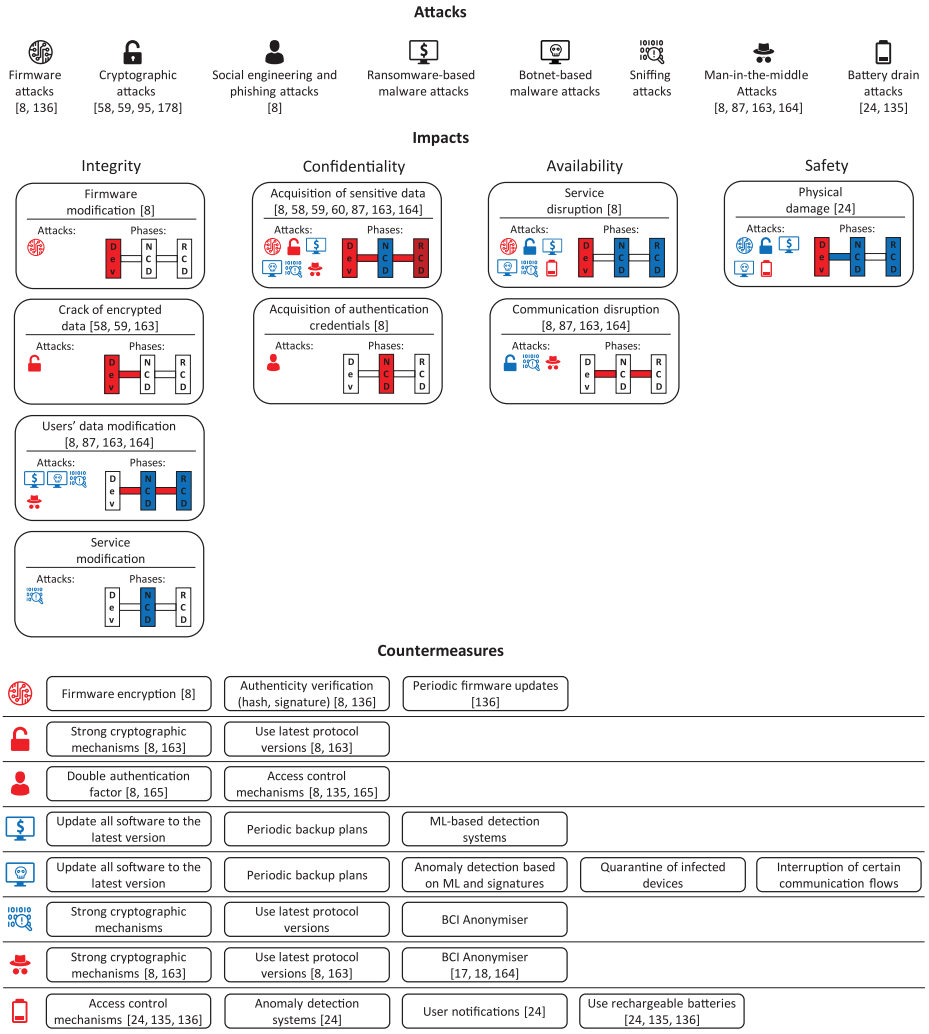


Fig. 5. Attacks, impacts, and countermeasures associated with the BCI architectural deployments. Elements indicated in red represent information detected in the literature, while blue represents our contribution.

Figure 5 summarizes the previous attacks, impacts, and countermeasures. This figure first shows the list of attacks considered in this section, associated with a unique icon, where those attacks with references indicate that they have been detected in the literature, while those without references represent our contribution. After that, we show the impacts that generate the previous attacks, organized by category. For each impact, we indicate the specific attacks that cause the impact, and which elements of the architectural deployments presented in Figure 4 are affected. Moreover, we consider the issues on the communication links between these elements. In particular, the attacks and elements identified in red represent issues detected in the literature, whereas those in blue are

our contributions. Finally, this figure lists countermeasures detected both in the literature and by us, associating each attack with a list of countermeasures. The color and reference criteria used before for the impacts also applies to the countermeasures, where an attack represented with a particular color indicates that all their countermeasures have the same color.

4 BCI TRENDS AND CHALLENGES

One of the first BCI solutions was developed at the end of the 1990s. It supposed a significant advancement in the medical industry, specifically in neurorehabilitation, bringing to the reality the mental control of prosthetic limbs and wheelchairs [119]. During the decade of the 2000s, a new generation of neuroprosthetic devices was developed to restore the mobility of patients severely paralyzed, creating communication links between the brain and a wide variety of actuators, such as robotic exoskeletons [82]. This trend in the field of BCI has resulted in new paradigms and scenarios in the last decade, where acquisition and stimulation procedures are used together to acquire brain activity and deliver feedback to the brain or peripheral nerves, defining the concept of bidirectional, or closed-loop, BCIs. Focusing on these systems, NeuroPace RNS is the only technology clinically approved for closed-loop treatment [33]. DBS is nowadays considered as a unidirectional BCI system, or open-loop, only performing stimulation actions. Nevertheless, current research aims to develop closed-loop DBS systems that are able to automatically identify the best stimulation parameters based on the status of the brain [52]. This evolution is also applicable for neuroprostheses, where the users can mentally control prosthesis while receiving stimulation to recover motor abilities [85].

This evolution allowed the definition of prospect ways of interaction where the BCI acts as an online communication element with other systems and users, based on Global BCI architectures. In particular, we subsequently present several examples of futuristic systems to highlight the importance of security in the progress of BCI technologies. Zhang et al. [182] defined the concept of the Internet of Brain, also known as Brain-to-Internet (BtI), where the BCI uses an NCD to access Internet services, such as search results or social media. Lebedev et al. [82] also described experiments where monkeys controlled remote robotic arms using BCI devices. More recently, Saad et al. [144] identified that 6G technologies could enable the interconnection of BCIs with the Internet. Besides, Martins et al. [97] documented a fusion between neuralnanorobotics and cloud services to acquire knowledge, defining the concept of Human Brain/Cloud Interface (B/CI). Another futuristic approach, Brain-to-Brain (BtB), allows direct communications between two brains, known as BtB [127, 184], where Pais-Vieira et al. [127] documented the real-time exchange of information between the brain of two rats. These systems have also been extended to create networks of interconnected brains, known as Brainet, which can perform collaborative tasks between users and share knowledge, memories, or thoughts through remote brains [67, 126]. Although these systems are in an early research stage, they could be a reality in the next decades, where security aspects will gain enormous importance. To represent this trend, Figure 6 illustrates this evolution of the literature, indicating the years of publication and approaches. Besides, current innovations, such as the use of silicon-based chips, could increase the quantity of information that we can acquire from the brain, and ease the development of electronic devices to improve the resolution of the neural acquisition and sensitivity of the process [121].

The BCI research field has gained relevance in the last few years, where different governments have funded and promoted BCI initiatives. In the United States of America, the DARPA is supporting the BRAIN Initiative (Brain Research through Advancing Innovative Neurotechnologies) [64]. Canada has launched its research line, called the Canadian Brain Research Strategy [63, 162]. On the other side of the Atlantic ocean, the European Union has also supported different projects, such as the Human Brain Project (HBP) [133] or the Brain/Neural Computer Interaction (BNCI)

ACM Computing Surveys, Vol. 54, No. 1, Article 11. Publication date: December 2020.

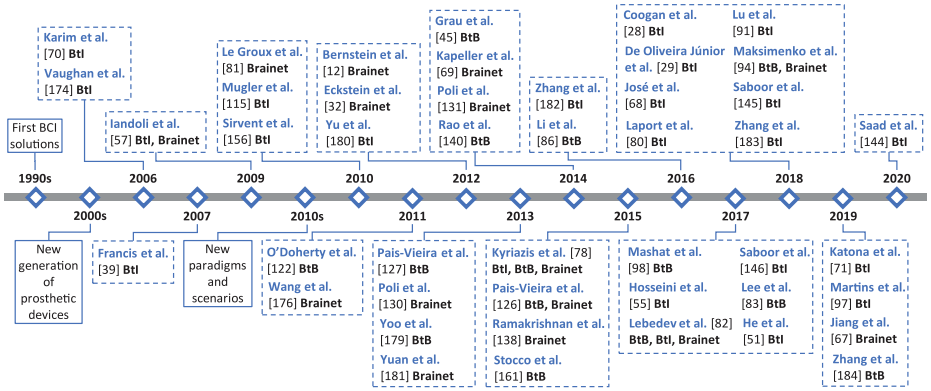


Fig. 6. Timeline of the evolution of BCI research, seen from the perspective of BtI, BtB, and Brainet approaches.

project [21, 22]. Asia has also promoted several initiatives, such as the China Brain Project [132] or the Brain/MINDS project in Japan [20]. All the previous initiatives and projects aim to advance the understanding of the human brain through the use of innovative technologies. As a consequence, emerging technologies offer precise acquisition and stimulation capabilities that enable new BCI application scenarios. The common interest in the study of the human brain and, in particular, on BCI leads to new opportunities for manufacturers, who can increase their competitiveness producing revolutionary BCI services based on growing paradigms such as the IoT, cloud computing, and big data. This development derives in the improvement of the usability, accuracy and safety of the products, together with their expansion to non-medical economic sectors such as entertainment. The result of the above is a trend of BCI toward Global BCI architecture deployments, where multiple BCI devices can communicate between them to perform collaborative tasks, based on the approaches of BtI, BtB, and Brainet. Once summarized the evolution of BCI and its trend, below, we highlight the most relevant current and future challenges concerning security on BCI.

4.1 Interoperability between BCI Deployments

Existing BCI deployments consider isolated devices without standards to provide interoperability in terms of communication and data representation. This is the case of commercial BCI brands and devices, which have been designed to resolve particular problems and are not compatible between them [137]. Moreover, deployments integrating the communication between several BCIs are ad hoc; that is, manufacturers design and implement them, considering only the requirements of a particular scenario. In this context, the current trend of BCIs toward paradigms such as the IoT and cloud computing will require an improvement in interoperability, as it is essential to ensure the future expansion of BCI technologies. Besides, the lack of interoperability limits the definition of global cybersecurity systems and mechanisms that can be applied. In this sense, current BCI solutions are device-oriented and do not offer collaborative mechanisms against cyberattacks. We detect as a future opportunity the use of well-known standardized APIs, communication technologies, and protocols to offer seamless protection on BCI. We also propose the use of ontologies to represent neural information in a formal and standardized fashion. Different companies and products would use a joint representation to ease data interpretation, processing, and sharing. This homogenization would have a positive impact on cybersecurity, enabling the

design and deployment of new protocols and mechanisms for the secure exchange of particular pieces of sensitive data between independent BCI solutions. In particular, the exchange of medical information between different organizations can be accomplished using well-known standards, as is the case of the HL7 standard [53].

4.2 Extensibility of BCI Designs

Extensibility refers to the ability of BCIs to add new functionality and application scenarios dynamically. Nowadays, BCI devices suffer a lack of extensibility, as companies manufacture them to provide particular services on fixed application scenarios. The neural data processing is performed in a fixed way and according to predefined premises. It means that each layer making up BCI architectures performs particular processing tasks, which can not be changed or even modified on demand [163]. Since each application scenario has its requirements and restrictions, the trend toward Global BCI will need new automatic and flexible architectures and processing mechanisms over the acquired neural data. These aspects also affect the security solutions that can be applied, since current constraints of BCI systems prevent the use of reactive and adaptive defensive mechanisms to face the threats described in previous sections. In conjunction with a lack of interoperability, the security responsibilities of each phase of the architecture are predefined and cannot be extended within that element, or delegated to be performed in other systems. As a future line of work, we highlight the design of BCI deployments that allow the implementation of most of the operations performed in software, instead of hardware, allowing developers to change the system's behavior. Another possible solution is a modular design of BCI, including supplementary modules, according to the requirements. However, these modifications introduce new security challenges, since software developments are more prone to errors and attacks, and new modular systems will address specific challenges, such as the verification of their authenticity.

4.3 Data Protection

Current BCI architectures and deployments do not consider the protection of neural data and personal information, as detected in the literature [137, 152, 164]. The evolution of BCIs toward distributed scenarios with heterogeneous and ubiquitous characteristics, such as BtB approaches, will require the storage and management of multiple users' personal and sensitive data. Because of that, future deployments should ensure that this critical information is transmitted and processed securely. Specifically, robust cryptography mechanisms need to be applied over data communication and storage, while techniques such as differential privacy or homomorphic encryption would help to ensure the anonymization of the data. Moreover, users do not have control over their privacy preferences to define who has access to the information and in which particular circumstances. Because of that, there are no specific privacy regulations to ensure that applications and external services can access only to the neural information accepted by users, nor any limitation on manufacturers or third-parties to prevent the processing of sensitive neural data without users authorization. To improve this situation, we propose policy-based solutions that allow users to define their privacy preferences based on their particular context. Besides, we propose the use of user-friendly systems that also help users proposing privacy-preserving recommendations. These initiatives must also align with the data protection law applicable in each country.

4.4 Physical and Architectural BCI Threats

Nowadays, a considerable amount of BCI designs and deployments do not consider cybersecurity issues such as the protection of communications, processing, storage, and applications. Although some solutions include security mechanisms, like Medtronic DBS products, some aspects must be improved. In particular, these devices use proprietary telemetry protocols [101], which recently

has led to vulnerabilities [27]. Nevertheless, companies such as Medtronic or Boston Scientific publish security bulletins when a security vulnerability affecting their devices is detected [103, 151], highlighting the interest that companies have on security. Moreover, the lack of BCI standards and, specifically, cybersecurity standards, prevent the homogenization of the security solutions implemented [17, 137, 163, 165]. The expansion of BCI will require robust dynamic cybersecurity mechanisms to face future challenges. Moreover, the development of more precise BCI devices and the integration of a large number of devices and systems, would result in a massive production of sensitive data. In our opinion, this context could benefit the increase of vulnerable systems and communication links. To address these challenges, manufacturers should evaluate alternatives for the mitigation of cyberattacks from multiple perspectives, aiming to implement seamless cybersecurity solutions. Based on that, we propose using 5G network technologies, since they have been designed to support a significant number of devices, which are necessary for BtB and Brainet scenarios. In particular, we identify that techniques and paradigms associated with 5G, such as Network Function Virtualisation (NFV) and Software-defined Networking (SDN) for the virtualization and dynamic management of network communications, are useful for the development of reactive cybersecurity solutions. Also, technologies such as Blockchain can provide the tracking of the information and ensure that it has not been modified, guaranteeing the integrity of the data. Moreover, we identify the protection of network communications by using protocols such as TLS [62] or IPsec [61] as an opportunity, which offers robust mechanisms against cyberattacks. Moreover, we detect that the application of information risk management standards, such as the ISO 27000 [65], and the NIST Cybersecurity Framework [120] could benefit the creation of homogeneous and robust solutions. Finally, we identify that game theory applied to BCI security strategies can be useful to implement regularly evolving systems. In particular, they can be useful to model how to establish the most appropriate countermeasures against continuously and automatically changing attacks, specifically in distributed scenarios such as BtB [7].

5 CONCLUSION

This article performs a global and comprehensive analysis of the literature of BCIs in terms of security and safety. Mainly, we have evaluated the attacks, impacts and countermeasures that BCI solutions suffer from the software's architectural design and implementation perspectives. Initially, we proposed a unified version of the BCI cycle to include neural data acquisition and stimulation processes. Once having a homogeneous BCI cycle design, we identified security attacks, impacts, and countermeasures affecting each phase of the cycle. It served as a starting point to determine which processes and functioning stages of BCIs are more prone to attacks. The architectural deployments of current BCI solutions have also been analyzed to highlight the security attacks and countermeasures related to each approach to understanding the issues of these technologies in terms of network communications. Finally, we provide our vision regarding BCI trends and depict that the current evolution of BCIs toward interconnected devices is generating tremendous security concerns and challenges, which will increase in the near future.

Among the learned lessons, we highlight the following five: (1) the field of security oriented to BCI technologies is not yet mature, generating opportunities for attackers; (2) even non-sophisticated attacks can have a significant impact on both BCI technologies and users' safety; (3) there is a current opportunity for standardization initiatives to unify BCIs in terms of information security; (4) well-studied fields, such as IMDs and IoT, can define a guide to develop robust security mechanisms for BCIs; (5) users' awareness of BCI security issues is vital.

As future work, we plan to focus our efforts on the design and implementation of solutions able to detect and mitigate attacks affecting the stimulation process in real time. In this context, we are considering using artificial intelligence techniques to detect anomalies in the firing patterns and

neural activity controlled by BCI solutions in charge of stimulating the brain. Besides, we also plan to contribute by improving the interoperability and data protection mechanisms of existing BCI architectures. Finally, another future work is the development of dynamic and proactive systems as an opportunity to mitigate the impacts of the attacks documented in this work.

ACKNOWLEDGMENT

We thank Mattia Zago for his advice during the development of the visual support of the work.

REFERENCES

[1] Minkyu Ahn, Mijin Lee, Jinyoung Choi, Sung Jun, Minkyu Ahn, Mijin Lee, Jinyoung Choi, and Sung Chan Jun. 2014. A review of brain-computer interface games and an opinion survey from researchers, developers and users. *Sensors* 14, 8 (Aug. 2014), 14601–14633.

[2] Bander Ali Saleh Al-rimy, Mohd Aizaini Maarof, and Syed Zainudeen Mohd Shaid. 2018. Ransomware threat success factors, taxonomy, and countermeasures: A survey and research directions. *Comput. Secur.* 74 (May 2018), 144–166.

[3] Naseer Amara, Huang Zhiqui, and Awais Ali. 2017. Cloud computing security threats and attacks with their mitigation techniques. In *Proceedings of the International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC'17)*. IEEE, 244–251.

[4] Pedram Amini, Muhammad Amin Araghizadeh, and Reza Azmi. 2015. A survey on Botnet: Classification, detection and defense. In *Proceedings of the International Electronics Symposium (IES'15)*. IEEE, 233–238.

[5] P. Anu and S. Vimala. 2017. A survey on sniffing attacks on computer networks. In *Proceedings of the International Conference on Intelligent Computing and Control (I2C2'17)*. IEEE, 5.

[6] P. Arico, G. Borghini, G. Di Flumeri, N. Sciaraffa, and F. Babiloni. 2018. Passive BCI beyond the lab: Current trends and future directions. *Physiol. Measure.* 39, 8 (Aug. 2018), 08TR02.

[7] A. Attiah, M. Chatterjee, and C. C. Zou. 2018. A game theoretic approach to model cyber attack and defense strategies. In *Proceedings of the IEEE International Conference on Communications (ICC'18)*. IEEE, 1–7.

[8] Pablo Ballarin Usieto and Javier Minguez. 2018. Avoiding brain hacking—Challenges of cybersecurity and privacy in Brain Computer Interfaces. Retrieved from https://www.bitbrain.com/blog/cybersecurity-brain-computer-interface.

[9] Srijita Basu, Arjun Bardhan, Koyal Gupta, Payel Saha, Mahasweta Pal, Manjima Bose, Kaushik Basu, Saunak Chaudhury, and Pritika Sarkar. 2018. Cloud computing security challenges & solutions-A survey. In *Proceedings of the IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC'18)*. IEEE, 347–356.

[10] Nebia Bentabet and Nasr Eddine Berrached. 2016. Synchronous P300-based BCI to control home appliances. In *Proceedings of the 8th International Conference on Modelling, Identification and Control (ICMIC'16)*. IEEE, 835–838.

[11] S. López Bernal, A. Huertas Celdrán, L. Fernández Maimó, M. T. Barros, S. Balasubramaniam, and G. Martínez Pérez. 2020. Cyberattacks on miniature brain implants to disrupt spontaneous neural signaling. *IEEE Access* 8 (2020), 152204–152222.

[12] Abraham Bernstein, Mark Klein, and Thomas W. Malone. 2012. Programming the global brain. *Commun. ACM* 55, 5 (May 2012), 41.

[13] Meriem Bettayeb, Qassim Nasir, and Manar Abu Talib. 2019. Firmware update attacks and security for IoT devices. In *Proceedings of the ArabWIC 6th Annual International Conference Research Track*. ACM Press, 6.

[14] Marom Bikson, Andre R. Brunoni, Leigh E. Charvet, Vincent P. Clark, Leonardo G. Cohen, Zhi-De Deng, Jacek Dmochowski, Dylan J. Edwards, Flavio Frohlich, Emily S. Kappenman, Kelvin O. Lim, Colleen Loo, Antonio Mantovani, David P. McMullen, Lucas C. Parra, Michele Pearson, Jessica D. Richardson, Judith M. Rumsey, Pejman Sehatpour, David Sommers, Gozde Unal, Eric M. Wassermann, Adam J. Woods, and Sarah H. Lisanby. 2018. Rigor and reproducibility in research with transcranial electrical stimulation: An NIMH-sponsored workshop. *Brain Stimul.* 11, 3 (2018), 465–480.

[15] Gajanan K. Birajdar and Vijay H. Mankar. 2013. Digital image forgery detection using passive techniques: A survey. *Dig. Investigat.* 10, 3 (Oct. 2013), 226–245.

[16] Paul E. Black and Irena Bojanova. 2016. Defeating buffer overflow: A trivial but dangerous bug. *IT Profess.* 18, 6 (Nov 2016), 58–61.

[17] Tamara Bonaci, Ryan Calo, and Howard Jay Chizeck. 2015. App stores for the brain: Privacy and security in brain-computer interfaces. *IEEE Technol. Soc. Mag.* 34, 2 (June 2015), 32–39.

[18] Tamara Bonaci, Jeffrey Herron, Charles Matlack, and Howard Jay Chizeck. 2015. Securing the exocortex: A 21st-century cybernetics challenge. *IEEE Technol. Soc. Mag.* 34, 3 (Sep. 2015), 44–51. arxiv:hep-ph/0011146

[19] Alessio Botta, Walter de Donato, Valerio Persico, and Antonio Pescapé. 2016. Integration of cloud computing and Internet of Things: A survey. *Future Gen. Comput. Syst.* 56 (Mar. 2016), 684–700.

[20] Brain/MINDS project. 2019. Brain/MINDS project. Retrieved from https://brainminds.jp/en/.

[21] Brain/Neural Computer Interaction project. 2015. Brain/Neural Computer Interaction project. Retrieved from http://bnci-horizon-2020.eu/.

[22] Clemens Brunner, Niels Birbaumer, Benjamin Blankertz, Christoph Guger, Andrea Kübler, Donatella Mattia, José del R. Millán, Felip Miralles, Anton Nijholt, Eloy Opisso, Nick Ramsey, Patric Salomon, and Gernot R. Müller-Putz. 2015. BNCI Horizon 2020: Towards a roadmap for the BCI community. *Brain-Comput. Interfaces* 2, 1 (Jan. 2015), 10.

[23] Carsten Buhmann, Torge Huckhagel, Katja Engel, Alessandro Gulberti, Ute Hidding, Monika Poetter-Nerger, Ines Goerendt, Peter Ludewig, Hanna Braass, Chi-un Choe, Kara Krajewski, Christian Oehlwein, Katrin Mittmann, Andreas K. Engel, Christian Gerloff, Manfred Westphal, Johannes A. Köppen, Christian K. E. Moll, and Wolfgang Hamel. 2017. Adverse events in deep brain stimulation: A retrospective long-term analysis of neurological, psychiatric and other occurrences. *PLoS ONE* 12, 7 (July 2017), 1–21.

[24] Carmen Camara, Pedro Peris-Lopez, and Juan E. Tapiador. 2015. Security and privacy issues in implantable medical devices: A comprehensive survey. *J. Biomed. Info.* 55 (June 2015), 272–289.

[25] Debashis Das Chakladar and Sanjay Chakraborty. 2018. Feature extraction and classification in brain-computer interfacing: Future research issues and challenges. In *Natural Computing for Unsupervised Learning*. Springer, Cham, Chapter 5, 101–131.

[26] Howard Jay Chizeck and Tamara Bonaci. 2014. Brain-Computer Interface Anonymizer. U.S. Patent Application. US20140228701A1.

[27] Cybersecurity & Infrastructure Security Agancy (CISA). 2020. ICS Medical Advisory (ICSMA-19-080-01). Retrieved from https://us-cert.cisa.gov/ics/advisories/ICSMA-19-080-01.

[28] Christopher G. Coogan and Bin He. 2018. Brain-computer interface control in a virtual reality environment and applications for the Internet of Things. *IEEE Access* 6 (2018), 10840–10849.

[29] Wilson G. de Oliveira Júnior, Juliana M. de Oliveira, Roberto Munoz, and Victor Hugo C. de Albuquerque. 2020. A proposal for Internet of Smart Home Things based on BCI system to aid patients with amyotrophic lateral sclerosis. *Neural Comput. Appl.* 32, 15 (Aug. 2020), 11007–11017.

[30] Till A. Dembek, Paul Reker, Veerle Visser-Vandewalle, Jochen Wirths, Harald Treuer, Martin Klehr, Jan Roediger, Haidar S. Dafsari, Michael T. Barbe, and Lars Timmermann. 2017. Directional DBS increases side-effect thresholds—A prospective, double-blind trial. *Move. Disord.* 32, 10 (2017), 1380–1388.

[31] Tamara Denning, Yoky Matsuoka, and Tadayoshi Kohno. 2009. Neurosecurity: Security and privacy for neural devices. *Neurosurg. Focus* 27, 1 (2009), E7.

[32] Miguel P. Eckstein, Koel Das, Binh T. Pham, Matthew F. Peterson, Craig K. Abbey, Jocelyn L. Sy, and Barry Giesbrecht. 2012. Neural decoding of collective wisdom with multi-brain computing. *NeuroImage* 59, 1 (Jan. 2012), 94–108.

[33] Christine A. Edwards, Abbas Kouzani, Kendall H. Lee, and Erika K. Ross. 2017. Neurostimulation devices for the treatment of neurologic disorders. *Mayo Clin. Proceed.* 92, 9 (2017), 1427–1444.

[34] Emotiv. 2019. Emotiv. Retrieved from https://www.emotiv.com/.

[35] Emotiv. 2019. Emotiv Cortex API. Retrieved from https://emotiv.github.io/cortex-docs/#introduction.

[36] Emotiv. 2019. EMOTIV EPOC+. Retrieved from https://www.emotiv.com/epoc/.

[37] Lorenzo Fernández Maimó, Alberto Huertas Celdrán, Ángel Perales Gómez, Félix García Clemente, James Weimer, and Insup Lee. 2019. Intelligent and dynamic ransomware spread detection and mitigation in integrated clinical environments. *Sensors* 19, 5 (Mar. 2019), 1114.

[38] Samuel G. Finlayson, John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam, and Isaac S. Kohane. 2019. Adversarial attacks on medical machine learning. *Science* 363, 6433 (Mar. 2019), 1287–1289.

[39] Heylighen Francis. 2007. The global superorganism: An evolutionary-cybernetic model of the emerging network society. *Soc. Evol. Hist.* 6, 1 (2007), 58–119.

[40] Mario Frank, Tiffany Hwu, Sakshi Jain, Robert T. Knight, Ivan Martinovic, Prateek Mittal, Daniele Perito, Ivo Sluganovic, and Dawn Song. 2017. Using EEG-based BCI devices to subliminally probe for private information. In *Proceedings of the on Workshop on Privacy in the Electronic Society (WPES'17)*. ACM Press, New York, New York, 133–136. Retrieved from https://arxiv:1312.6052.

[41] Jianwen Fu, Jingfeng Xue, Yong Wang, Zhenyan Liu, and Chun Shan. 2018. Malware visualization for fine-grained classification. *IEEE Access* 6 (2018), 14510–14523.

[42] Ariko Fukushima, Reiko Yagi, Norie Kawai, Manabu Honda, Emi Nishina, and Tsutomu Oohashi. 2014. Frequencies of inaudible high-frequency sounds differentially affect brain activity: Positive and negative hypersonic effects. *PLoS ONE* 9, 4 (Apr. 2014), e95464.

[43] Joyce Gomes-Osman, Aprinda Indahlastari, Peter J. Fried, Danylo L. F. Cabral, Jordyn Rice, Nicole R. Nissim, Serkan Aksu, Molly E. McLaren, and Adam J. Woods. 2018. Non-invasive brain stimulation: Probing intracortical circuits and improving cognition in the aging brain. *Front. Aging Neurosci.* 10 (2018), 177.

[44] Ian Goodfellow, Patrick McDaniel, and Nicolas Papernot. 2018. Making machine learning robust against adversarial inputs. *Commun. ACM* 61, 7 (July 2018), 56–66.

[45] Carles Grau, Romuald Ginhoux, Alejandro Riera, Thanh Lam Nguyen, Hubert Chauvat, Michel Berg, Julià L. Amengual, Alvaro Pascual-Leone, and Giulio Ruffini. 2014. Conscious brain-to-brain communication in humans using non-invasive technologies. *PLoS ONE* 9, 8 (Aug 2014), e105225.

[46] Kanika Grover, Alvin Lim, and Qing Yang. 2014. Jamming and anti-jamming techniques in wireless networks: A survey. *Int. J. Ad Hoc and Ubiquitous Computing* 17, 4 (2014), 197.

[47] Surbhi Gupta, Abhishek Singhal, and Akanksha Kapoor. 2017. A literature survey on social engineering attacks: Phishing attack. In *Proceedings of the IEEE International Conference on Computing, Communication and Automation (ICCCA'16)*. IEEE, 537–540.

[48] Christian J. Hartmann, Sabine Fliegen, Stefan J. Groiss, Lars Wojtecki, and Alfons Schnitzler. 2019. An update on best practice of deep brain stimulation in Parkinson's disease. *Therap. Adv. Neurol. Disord.* 12 (Jan. 2019), 1756286419838096.

[49] Joseph M. Hatfield. 2018. Social engineering in cybersecurity: The evolution of a concept. *Comput. Secur.* 73 (2018), 102–113.

[50] Vincent Haupert, Dominik Maier, Nicolas Schneider, Julian Kirsch, and Tilo Müller. 2018. Honey, I shrunk your app security: The state of android app hardening. In *Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer International Publishing, Cham, 69–91.

[51] Shenghong He, Tianyou Yu, Zhenghui Gu, and Yuanqing Li. 2017. A hybrid BCI web browser based on EEG and EOG signals. In *Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'17)*. IEEE, 1006–1009.

[52] Franz Hell, Carla Palleis, Jan H. Mehrkens, Thomas Koeglsperger, and Kai Bötzel. 2019. Deep brain stimulation programming 2.0: Future perspectives for target identification and adaptive closed loop stimulation. *Front. Neurol.* 10 (2019), 314.

[53] HL7 International. 2019. Health Level Seven. Retrieved from https://www.hl7.org/.

[54] Keum Shik Hong and Muhammad Jawad Khan. 2017. Hybrid brain-computer interface techniques for improved classification accuracy and increased number of commands: A review. *Front. Neurorobot.* 11 (July 2017), 35.

[55] Mohammad-Parsa Hosseini, Dario Pompili, Kost Elisevich, and Hamid Soltanian-Zadeh. 2017. Optimized deep learning for EEG big data and seizure prediction BCI via Internet of Things. *IEEE Trans. Big Data* 3, 4 (Dec. 2017), 392–404.

[56] Alberto Huertas Celdrán, Ginés Dólera Tormo, Félix Gómez Mármol, Manuel Gil Pérez, and Gregorio Martínez Pérez. 2016. Resolving privacy-preserving relationships over outsourced encrypted data storages. *Int. J. Info. Secur.* 15, 2 (Apr. 2016), 195–209.

[57] Luca Iandoli, Mark Klein, and Giuseppe Zollo. 2009. Enabling on-line deliberation and collective decision-making through large-scale argumentation. *Int. J. Decis. Supp. Syst. Technol.* 1, 1 (Jan. 2009), 69–92.

[58] Marcello Ienca. 2015. Neuroprivacy, neurosecurity and brain-hacking: Emerging issues in neural engineering. *Bioethica Forum* 8, 2 (2015), 51–53.

[59] Marcello Ienca and Pim Haselager. 2016. Hacking the brain: Brain–computer interfacing technology and the ethics of neurosecurity. *Ethics Info. Technol.* 18, 2 (June 2016), 117–129.

[60] Marcello Ienca, Pim Haselager, and Ezekiel J. Emanuel. 2018. Brain leaks and consumer neurotechnology. *Nature Biotechnol.* 36, 9 (2018), 805–810.

[61] IETF. 2011. IP Security (IPsec) and Internet Key Exchange (IKE) Document Roadmap. Retrieved from https://tools.ietf.org/html/rfc6071.

[62] IETF. 2018. The Transport Layer Security (TLS) Protocol Version 1.3. Retrieved from https://tools.ietf.org/html/rfc8446.

[63] Judy Illes, Samuel Weiss, Jaideep Bains, Jennifer A. Chandler, Patricia Conrod, Yves De Koninck, Lesley K. Fellows, Deanna Groetzinger, Eric Racine, Julie M. Robillard, and Marla B. Sokolowski. 2019. A neuroethics backbone for the evolving canadian brain research strategy.*Neuron* 101, 3 (Feb. 2019), 370–374.

[64] The BRAIN Initiative. 2019. The BRAIN Initiative. Retrieved from https://braininitiative.nih.gov/.

[65] ISO. 2018. ISO/IEC 27001 Information security management. Retrieved from https://www.iso.org/isoiec-27001-information-security.html.

[66] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. 2018. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *Proceedings of the IEEE Symposium on Security and Privacy*. IEEE, 19–35.

[67] Linxing Jiang, Andrea Stocco, Darby M. Losey, Justin A. Abernethy, Chantel S. Prat, and Rajesh P. N. Rao. 2019. BrainNet: A multi-person brain-to-brain interface for direct collaboration between brains. *Sci. Rep.* 9, 1 (Dec. 2019), 6115.

[68] Sergio José and Rodríguez Méndez. 2018. Modeling actuations in BCI-O. In *Proceedings of the 8th International Conference on the Internet of Things (IOT'18)*. ACM Press, New York, 6.

[69] Christoph Kapeller, Rupert Ortner, Gunther Krausz, Markus Bruckner, Brendan Z. Allison, Christoph Guger, and Günter Edlinger. 2014. Toward multi-brain communication: Collaborative spelling with a P300 BCI. In *Proceedings of the International Conference on Augmented Cognition*. Springer, Cham, 47–54.

[70] Ahmed A. Karim, Thilo Hinterberger, Jürgen Richter, Jürgen Mellinger, Nicola Neumann, Herta Flor, Andrea Kübler, and Niels Birbaumer. 2006. Neural Internet: Web surfing with brain potentials for the completely paralyzed. *Neurorehab. Neural Repair* 20, 4 (Dec. 2006), 508–515.

[71] Jozsef Katona, Tibor Ujbanyi, Gergely Sziladi, and Attila Kovari. 2019. *Electroencephalogram-based Brain-Computer Interface for Internet of Robotic Things*. Springer International Publishing, Cham, Chapter 12, 253–275.

[72] Elena Khabarova, Natalia Denisova, Aleksandr Dmitriev, Konstantin Slavin, and Leo Verhagen Metman. 2018. Deep brain stimulation of the subthalamic nucleus in patients with parkinson disease with prior pallidotomy or thalamotomy. *Brain Sci.* 8, 4 (Apr. 2018), 66.

[73] G. Kirubavathi and R. Anitha. 2018. Structural analysis and detection of android botnets using machine learning techniques. *Int. J. Info. Secur.* 17, 2 (Apr. 2018), 153–167.

[74] Constantinos Kolias, Georgios Kambourakis, Angelos Stavrou, and Jeffrey Voas. 2017. DDoS in the IoT: Mirai and other botnets. *Computer* 50, 7 (2017), 80–84.

[75] Jan Kubanek. 2018. Neuromodulation with transcranial focused ultrasound.*Neurosurg. Focus* 44, 2 (Feb. 2018), E14.

[76] D. Richard Kuhn, Vincent C. Hu, W. Timothy Polk, and Shu-Jen Chang. 2001. *Introduction to Public Key Technology and the Federal PKI Infrastructure*. Technical Report. National Institute of Standards and Technology, 1–54. Retrieved from https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-32.pdf.

[77] James Kurose and Keith Ross. 2017. *Computer Networking: A Top-Down Approach* (7th ed.). Pearson, London, 852 pages.

[78] Marios Kyriazis. 2015. Systems neuroscience in focus: From the human brain to the global brain? *Front. Syst. Neurosci.* 9 (Feb. 2015), 7.

[79] Ofir Landau, Rami Puzis, and Nir Nissim. 2020. Mind your mind. *Comput. Surveys* 53, 1 (2020), 1–38.

[80] Francisco Laport, Francisco J. Vazquez-Araujo, Paula M. Castro, Adriana Dapena, Francisco Laport, Francisco J. Vazquez-Araujo, Paula M. Castro, and Adriana Dapena. 2018. Brain-computer interfaces for Internet of Things. *Proceedings* 2, 18 (Sep. 2018), 1179.

[81] Sylvain Le Groux, Jonatas Manzolli, Paul F. Verschure, Marti Sanchez, Andre Luvizotto, Anna Mura, Aleksander Valjamae, Christoph Guger, Robert Prueckl, and Ulysses Bernardet. 2010. Disembodied and collaborative musical interaction in the multimodal brain orchestra. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. MIME, 309–314.

[82] Mikhail A. Lebedev and Miguel A. L. Nicolelis. 2017. Brain-machine interfaces: From basic science to neuroprostheses and neurorehabilitation. *Physiol. Rev.* 97, 2 (Apr. 2017), 767–837.

[83] Wonhye Lee, Suji Kim, Byeongnam Kim, Chungki Lee, Yong An Chung, Laehyun Kim, and Seung-Schik Yoo. 2017. Non-invasive transmission of sensorimotor information in humans using an EEG/focused ultrasound brain-to-brain interface. *PLoS ONE* 12, 6 (July 2017), e0178476.

[84] M. León Ruiz, M. L. Rodríguez Sarasa, L. Sanjuán Rodríguez, J. Benito-León, E. García-Albea Ristol, and S. Arce Arce. 2018. Current evidence on transcranial magnetic stimulation and its potential usefulness in post-stroke neurorehabilitation: Opening new doors to the treatment of cerebrovascular disease. *Neurología (English Edition)* 33, 7 (2018), 459–472.

[85] Timothée Levi, Paolo Bonifazi, Paolo Massobrio, and Michela Chiappalone. 2018. Editorial: Closed-loop systems for next-generation neuroprostheses. *Front. Neurosci.* 12 (2018), 26.

[86] Guangye Li and Dingguo Zhang. 2016. Brain-computer interface controlled cyborg: Establishing a functional information transfer pathway from human brain to cockroach brain. *PLoS ONE* 11, 3 (Mar. 2016), e0150667.

[87] Qianqian Li, Ding Ding, and Mauro Conti. 2015. Brain-computer interface applications: Security and privacy challenges. In *Proceedings of the IEEE Conference on Communications and Network Security (CNS'15)*. IEEE, 663–666.

[88] Lifelines Neuro. 2020. Neurodiagnostics Without Boundaries. Retrieved from https://www.lifelinesneuro.com/.

[89] Anli Liu, Mihály Vöröslakos, Greg Kronberg, Simon Henin, Matthew R. Krause, Yu Huang, Alexander Opitz, Ashesh Mehta, Christopher C. Pack, Bart Krekelberg, Antal Berényi, Lucas C. Parra, Lucia Melloni, Orrin Devinsky, and György Buzsáki. 2018. Immediate neurophysiological effects of transcranial electrical stimulation. *Nature Commun.* 9, 1 (Nov. 2018), 5092.

[90] Qiang Liu, Pan Li, Wentao Zhao, Wei Cai, Shui Yu, and Victor C. M. Leung. 2018. A survey on security threats and defensive techniques of machine learning: A data driven view. *IEEE Access* 6 (2018), 12103–12117.

[91] Huimin Lu, Hyoungseop Kim, Yujie Li, and Yin Zhang. 2018. BrainNets: Human emotion recognition using an Internet of Brian Things platform. In *Proceedings of the 14th International Wireless Communications and Mobile Computing Conference (IWCMC'18)*. IEEE, 1313–1316.

[92] Muhammad Mahmoud, Manjinder Nir, and Ashraf Matrawy. 2015. A survey on botnet architectures, detection and defences. *Int. J. Netw. Secur.* 17, 3 (May 2015), 272–289.

[93] Redowan Mahmud, Ramamohanarao Kotagiri, and Rajkumar Buyya. 2018. Fog computing: A taxonomy, survey and future directions. In *Internet of Everything*. Springer, Singapore, 103–130.

[94] Vladimir A. Maksimenko, Alexander E. Hramov, Nikita S. Frolov, Annika Lüttjohann, Vladimir O. Nedaivozov, Vadim V. Grubov, Anastasia E. Runnova, Vladimir V. Makarov, Jürgen Kurths, and Alexander N. Pisarchik. 2018. Increasing human performance by sharing cognitive load using brain-to-brain interface. *Front. Neurosci.* 12 (Dec. 2018), 949.

[95] Eduard Marin, Dave Singelée, Bohan Yang, Vladimir Volski, Guy A. E. Vandenbosch, Bart Nuttin, and Bart Preneel. 2018. Securing wireless neurostimulators. In *Proceedings of the 8th ACM Conference on Data and Application Security and Privacy (CODASPY'18)*. Association for Computing Machinery, New York, NY, 287–298.

[96] Ivan Martinovic, Doug Davies, and Mario Frank. 2012. On the feasibility of side-channel attacks with brain-computer interfaces. In *Proceedings of the 21st USENIX Security Symposium*. USENIX, Bellevue, WA, 143–158.

[97] Nuno R. B. Martins, Amara Angelica, Krishnan Chakravarthy, Yuriy Svidinenko, Frank J. Boehm, Ioan Opris, Mikhail A. Lebedev, Melanie Swan, Steven A. Garan, Jeffrey V. Rosenfeld, Tad Hogg, and Robert A. Freitas. 2019. Human brain/cloud interface. *Front. Neurosci.* 13 (Mar. 2019), 112.

[98] M. Ebrahim M. Mashat, Guangye Li, and Dingguo Zhang. 2017. Human-to-human closed-loop control based on brain-to-brain interface and muscle-to-muscle interface. *Sci. Rep.* 7, 1 (Dec. 2017), 11001.

[99] Hideyuki Matsumoto and Yoshikazu Ugawa. 2017. Adverse events of tDCS and tACS: A review. *Clin. Neurophysiol. Pract.* 2 (2017), 19–25.

[100] M. McMahon and M. Schukat. 2018. A low-cost, open-source, BCI- VR game control development environment prototype for game-based neurorehabilitation. In *Proceedings of the IEEE Games, Entertainment, Media Conference (GEM'18)*. IEEE, 1–9.

[101] Medtronic. 2020. DBS Security Reference Guide. Retrieved from http://manuals.medtronic.com/content/dam/emanuals/neuro/NDHF1550-189563.pdf.

[102] Medtronic. 2020. DBS Theraphy for OCD. Retrieved from https://www.medtronic.com/us-en/patients/treatments-therapies/deep-brain-stimulation-ocd/about/risks-probable-benefits.html.

[103] Medtronic. 2020. Security Bulletins. Retrieved from https://global.medtronic.com/xg-en/product-security/security-bulletins.html.

[104] Najmeh Miramirkhani, Mahathi Priya Appini, Nick Nikiforakis, and Michalis Polychronakis. 2017. Spotless sandboxes: Evading malware analysis systems using wear-and-tear artifacts. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'17)*. IEEE, 1009–1024.

[105] MITRE. 2019. CWE-CWE-74: Improper Neutralization of Special Elements in Output Used by a Downstream Component ("Injection") (3.2). Retrieved from https://cwe.mitre.org/data/definitions/74.html.

[106] MITRE. 2019. CWE-CWE-77: Improper Neutralization of Special Elements used in a Command ("Command Injection") (3.2). Retrieved from https://cwe.mitre.org/data/definitions/77.html.

[107] MITRE. 2019. CWE-CWE-78: Improper Neutralization of Special Elements used in an OS Command ("OS Command Injection") (3.2). Retrieved from https://cwe.mitre.org/data/definitions/78.html.

[108] MITRE. 2019. CWE-CWE-89: Improper Neutralization of Special Elements used in an SQL Command ("SQL Injection") (3.2). Retrieved from https://cwe.mitre.org/data/definitions/89.html.

[109] MITRE. 2019. CWE-119: Improper Restriction of Operations within the Bounds of a Memory Buffer. Retrieved from https://cwe.mitre.org/data/definitions/119.html.

[110] MITRE. 2019. CWE-120: Buffer Copy without Checking Size of Input ("Classic Buffer Overflow") (3.2). Retrieved from https://cwe.mitre.org/data/definitions/120.html.

[111] MITRE. 2019. CWE-121: Stack-based Buffer Overflow (3.2). Retrieved from https://cwe.mitre.org/data/definitions/121.html.

[112] MITRE. 2019. CWE-122: Heap-based Buffer Overflow (3.2). Retrieved from https://cwe.mitre.org/data/definitions/122.html.

[113] Muhammad Baqer Mollah, Md. Abul Kalam Azad, and Athanasios Vasilakos. 2017. Security and privacy challenges in mobile cloud computing: Survey and way ahead. *J. Netw. Comput. Appl.* 84 (Apr. 2017), 38–54.

[114] Ingrid Moreno-Duarte, Nigel Gebodh, Pedro Schestatsky, Berkan Guleyupoglu, Davide Reato, Marom Bikson, and Felipe Fregni. 2014. Chapter 2—Transcranial electrical stimulation: Transcranial Direct Current Stimulation (tDCS), Transcranial Alternating Current Stimulation (tACS), Transcranial Pulsed Current Stimulation (tPCS), and Transcranial Random Noise Stimulation (tRNS). In *The Stimulated Brain*, Roi Cohen Kadosh (Ed.). Academic Press, San Diego, 35–59.

[115] Emily M. Mugler, Carolin A. Ruf, Sebastian Halder, Michael Bensch, and Andrea Kubler. 2010. Design and implementation of a P300-based brain-computer interface for controlling an Internet browser. *IEEE Trans. Neural Syst. Rehab. Eng.* 18, 6 (Dec. 2010), 599–609.

[116] Elon Musk and Neuralink. 2019. An integrated brain-machine interface platform with thousands of channels. *bioRxiv* (2019). Retrieved from arXiv:https://www.biorxiv.org/content/early/2019/08/02/703801.full.pdf.

[117] NeuroPace. 2013. NeuroPace® RNS® System Patient Manual. Retrieved from https://www.accessdata.fda.gov/cdrh_docs/pdf10/p100026c.pdf.

[118] NeuroSky. 2019. NeuroSky. Retrieved from http://neurosky.com/.

[119] Miguel A. L. Nicolelis. 2001. Actions from thoughts. *Nature* 409, 6818 (2001), 403–407.

[120] NIST. 2018. Cybersecurity Framework. Retrieved from https://www.nist.gov/cyberframework.

[121] Abdulmalik Obaid, Mina-Elraheb Hanna, Yu-Wei Wu, Mihaly Kollo, Romeo Racz, Matthew R. Angle, Jan Müller, Nora Brackbill, William Wray, Felix Franke, E. J. Chichilnisky, Andreas Hierlemann, Jun B. Ding, Andreas T. Schaefer, and Nicholas A. Melosh. 2020. Massively parallel microwire arrays integrated with CMOS chips for neural recording. *Sci. Adv.* 6, 12 (2020). Retrieved from arXiv:https://advances.sciencemag.org/content/6/12/eaay2789.full.pdf.

[122] Joseph E. O'Doherty, Mikhail A. Lebedev, Peter J. Ifft, Katie Z. Zhuang, Solaiman Shokur, Hannes Bleuler, and Miguel A. L. Nicolelis. 2011. Active tactile exploration using a brain–machine–brain interface. *Nature* 479, 7372 (Nov. 2011), 228–231.

[123] Open Web Application Security Project. 2017. Top 10-2017 A6-Security Misconfiguration-OWASP. Retrieved from https://www.owasp.org/index.php/Top_10-2017_A6-Security_Misconfiguration.

[124] Open Web Application Security Project. 2017. Top 10-2017 Top 10-OWASP. Retrieved from https://www.owasp.org/index.php/Top_10-2017_Top_10.

[125] Open Web Application Security Project. 2018. OWASP Internet of Things Project. Retrieved from https://www.owasp.org/index.php/OWASP_Internet_of_Things_Project.

[126] Miguel Pais-Vieira, Gabriela Chiuffa, Mikhail Lebedev, Amol Yadav, and Miguel A. L. Nicolelis. 2015. Building an organic computing device with multiple interconnected brains. *Sci. Rep.* 5, 1 (Dec. 2015), 11869.

[127] Miguel Pais-Vieira, Mikhail Lebedev, Carolina Kunicki, Jing Wang, and Miguel A. L. Nicolelis. 2013. A brain-to-brain interface for real-time sharing of sensorimotor information. *Sci. Rep.* 3, 1 (Dec. 2013), 1319.

[128] Mahboubeh Parastarfeizabadi and Abbas Z. Kouzani. 2017. Advances in closed-loop deep brain stimulation devices. *J. NeuroEngineer. Rehab.* 14, 1 (Aug. 2017), 79.

[129] Rafael Polanía, Michael A. Nitsche, and Christian C. Ruff. 2018. Studying and modifying brain function with non-invasive brain stimulation. *Nature Neurosci.* 21, 2 (Feb. 2018), 174–187.

[130] Riccardo Poli, Caterina Cinel, Ana Matran-Fernandez, Francisco Sepulveda, and Adrian Stoica. 2013. Towards cooperative brain-computer interfaces for space navigation. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI'13).* ACM Press, New York, 149.

[131] Riccardo Poli, Davide Valeriani, and Caterina Cinel. 2014. Collaborative brain-computer interface for aiding decision-making. *PLoS ONE* 9, 7 (July 2014), 22.

[132] Mu-ming Poo, Jiu-lin Du, Nancy Y. Ip, Zhi-Qi Xiong, Bo Xu, and Tieniu Tan. 2016. China brain project: Basic neuroscience, brain diseases, and brain-inspired computing. *Neuron* 92, 3 (Nov. 2016), 591–596.

[133] Human Brain Project. 2019. Human Brain Project. Retrieved from https://www.humanbrainproject.eu/en/.

[134] Open Web Application Security Project. 2017. Top 10-2017 A1-Injection-OWASP. Retrieved from https://www.owasp.org/index.php/Top_10-2017_A1-Injection.

[135] Laurie Pycroft and Tipu Z. Aziz. 2018. Security of implantable medical devices with wireless connections: The dangers of cyber-attacks. *Expert Rev. Med. Devices* 15, 6 (July 2018), 403–406.

[136] Laurie Pycroft, Sandra G. Boccard, Sarah L.F. Owen, John F. Stein, James J. Fitzgerald, Alexander L. Green, and Tipu Z. Aziz. 2016. Brainjacking: Implant security issues in invasive neuromodulation. *World Neurosurg.* 92 (Aug. 2016), 454–462.

[137] Rabie A. Ramadan and Athanasios V. Vasilakos. 2017. Brain computer interface: Control signals review. *Neurocomput.* 223 (Feb. 2017), 26–44.

[138] Arjun Ramakrishnan, Peter J. Ifft, Miguel Pais-Vieira, Yoon Woo Byun, Katie Z. Zhuang, Mikhail A. Lebedev, and Miguel A.L. Nicolelis. 2015. Computing arm movements with a monkey Brainet. *Sci. Rep.* 5, 1 (Sep. 2015), 10767.

[139] Rajesh P. N. Rao. 2019. Towards neural co-processors for the brain: Combining decoding and encoding in brain–computer interfaces. *Curr. Opin. Neurobiol.* 55 (Apr. 2019), 142–151.

[140] Rajesh P. N. Rao, Andrea Stocco, Matthew Bryan, Devapratim Sarma, Tiffany M. Youngquist, Joseph Wu, and Chantel S. Prat. 2014. A direct brain-to-brain interface in humans. *PLoS ONE* 9, 11 (Nov. 2014), e111332.

[141] Heena Rathore, Chenglong Fu, Amr Mohamed, Abdulla Al-Ali, Xiaojiang Du, Mohsen Guizani, and Zhengtao Yu. 2020. Multi-layer security scheme for implantable medical devices. *Neural Comput. Appl.* 32, 9 (2020), 4347–4360.

[142] Rodrigo Roman, Javier Lopez, and Masahiro Mambo. 2018. Mobile edge computing, Fog et al.: A survey and analysis of security threats and challenges. *Future Gen. Comput. Syst.* 78 (Jan. 2018), 680–698.

[143] Ron Ross, Victoria Pillitteri, Richard Graubart, Deborah Bodeau, and Rosalie McQuaid. 2019. *Developing Cyber Resilient Systems: A Systems Security Engineering Approach*. Technical Report. National Institute of Standards and Technology. Retrieved from https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-160v2.pdf.

[144] W. Saad, M. Bennis, and M. Chen. 2019. A vision of 6G wireless systems: Applications, trends, technologies, and open research problems. *IEEE Netw.* (2019), 1–9.

[145] Abdul Saboor, Felix Gembler, Mihaly Benda, Piotr Stawicki, Aya Rezeika, Roland Grichnik, and Ivan Volosyak. 2018. A browser-driven SSVEP-based BCI web speller. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC'18)*. IEEE, Miyazaki, Japan, 625–630.

[146] Abdul Saboor, Aya Rezeika, Piotr Stawicki, Felix Gembler, Mihaly Benda, Thomas Grunenberg, and Ivan Volosyak. 2017. SSVEP-based BCI in a smart home scenario. In *Proceedings of the International Work-Conference on Artificial Neural Networks*. Springer, Cham, 474–485.

[147] Takamichi Saito, Ryohei Watanabe, Shuta Kondo, Shota Sugawara, and Masahiro Yokoyama. 2016. A survey of prevention/mitigation against memory corruption attacks. In *Proceedings of the 19th International Conference on Network-based Information Systems (NBiS'16)*. IEEE, 500–505.

[148] Parthana Sarma, Prakash Tripathi, Manash Pratim Sarma, and Kandarpa Kumar Sarma. 2016. Pre-processing and feature extraction techniques for EEG-BCI applications—A review of recent research. *ADBU-J. Eng. Technol.* 5 (2016), 2348–7305.

[149] M. A. Scholl, K. M. Stine, J. Hash, P. Bowen, L. A. Johnson, C. D. Smith, and D. I. Steinberg. 2008. *An Introductory Resource Guide for Implementing the Health Insurance Portability and Accountability Act (HIPAA) Security Rule*. Technical Report. National Institute of Standards and Technology, Gaithersburg, MD. Retrieved from https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-66r1.pdf.

[150] Suzanne B. Schwartz. 2018. Medical device cybersecurity through the FDA lens. In *Proceedings of the 27th USENIX Security Symposium*. USENIX Association, Baltimore, MD.

[151] Boston Scientific. 2020. Product Security Information. Retrieved from https://www.bostonscientific.com/en-US/customer-service/product-security/product-security-information.html.

[152] Diego Sempreboni and Luca Viganò. 2018. Privacy, Security, and Trust in the Internet of Neurons. Retrieved from https://arxiv:cs.CY/1807.06077.

[153] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *Proceedings of the IEEE Symposium on Security and Privacy*. IEEE, 3–18.

[154] S. Sibi Chakkaravarthy, D. Sangeetha, and V. Vaidehi. 2019. A Survey on malware analysis and mitigation techniques. *Comput. Sci. Rev.* 32 (May 2019), 23.

[155] Saurabh Singh, Young-Sik Jeong, and Jong Hyuk Park. 2016. A survey on cloud computing security: Issues, threats, and solutions. *J. Netw. Comput. Appl.* 75 (Nov. 2016), 200–222.

[156] José L. Sirvent, José M. Azorín, Eduardo Iáñez, Andrés Úbeda, and Eduardo Fernández. 2010. P300-based brain-computer interface for Internet browsing. In *Trends in Practical Applications of Agents and Multiagent Systems*. Springer, Berlin, 615–622.

[157] International Neuromodulation Society. 2020. International Neuromodulation Society. Retrieved from https://www.neuromodulation.com/.

[158] Kandhasamy Sowndhararajan, Minju Kim, Ponnuvel Deepa, Se Park, and Songmun Kim. 2018. Application of the P300 event-related potential in the diagnosis of epilepsy disorder: A review. *Scientia Pharmaceutica* 86, 2 (Mar. 2018), 10.

[159] William Stallings. 2017. *Cryptography and Network Security: Principles and Practice* (7th ed.). Pearson, London, 766 pages.

[160] Ioannis Stellios, Panayiotis Kotzanikolaou, Mihalis Psarakis, Cristina Alcaraz, and Javier Lopez. 2018. A survey of IoT-enabled cyberattacks: Assessing attack paths to critical infrastructures and services. *IEEE Commun. Surveys Tutor.* 20, 4 (2018), 3453–3495.

[161] Andrea Stocco, Chantel S. Prat, Darby M. Losey, Jeneva A. Cronin, Joseph Wu, Justin A. Abernethy, and Rajesh P. N. Rao. 2015. Playing 20 questions with the mind: Collaborative problem solving by humans using a brain-to-brain interface. *PLoS ONE* 10, 9 (Sep 2015), e0137303.

[162] Canadian Brain Research Strategy. 2019. Canadian Brain Research Strategy. Retrieved from https://canadianbrain.ca/.

[163] Kaushik Sundararajan. 2017. *Privacy and Security Issues in Brain Computer Interface*. Master's thesis. Auckland University of Technology.

[164] Hassan Takabi. 2016. Firewall for brain: Towards a privacy preserving ecosystem for BCI applications. In *Proceedings of the IEEE Conference on Communications and Network Security (CNS'16)*. IEEE, 370–371.

[165] Hassan Takabi, Anuj Bhalotiya, and Manar Alohaly. 2016. Brain computer interface (BCI) applications: Privacy threats and countermeasures. In *Proceedings of the IEEE 2nd International Conference on Collaboration and Internet Computing*. IEEE, 102–111.

[166] Andrew S. Tanenbaum and David J. Wetherall. 2011. *Computer Networks* (5th ed.). Pearson, London.
[167] William J. Tyler, Joseph L. Sanguinetti, Maria Fini, and Nicholas Hool. 2017. Non-invasive neural stimulation. In *Micro- and Nanotechnology Sensors, Systems, and Applications IX*, Thomas George, Achyut K. Dutta, and M. Saif Islam (Eds.), Vol. 10194. International Society for Optics and Photonics, Anaheim, CA, 280–290.
[168] U.S. Food and Drug Administration. 2016. *Postmarket Management of Cybersecurity in Medical Devices*. Technical Report. U.S. Food and Drug Administration, Rockville, MD.
[169] U.S. Food and Drug Administration. 2018. *Content of Premarket Submissions for Management of Cybersecurity in Medical Devices*. Technical Report. U.S. Food and Drug Administration, Rockville, MD.
[170] Satish Vadlamani, Burak Eksioglu, Hugh Medal, and Apurba Nandi. 2016. Jamming attacks on wireless networks: A taxonomic survey. *Int. J. Prod. Econ.* 172 (Feb. 2016), 76–94.
[171] Swati Vaid, Preeti Singh, and Chamandeep Kaur. 2015. EEG signal analysis for BCI interface: A review. In *Proceedings of the International Conference on Advanced Computing and Communication Technologies (ACCT'15)*. IEEE, 143–147.
[172] Marcel van Gerven, Jason Farquhar, Rebecca Schaefer, Rutger Vlek, Jeroen Geuze, Anton Nijholt, Nick Ramsey, Pim Haselager, Louis Vuurpijl, Stan Gielen, and Peter Desain. 2009. The brain–computer interface cycle. *J. Neural Eng.* 6, 4 (Aug. 2009), 041001.
[173] Sebastian Vasile, David Oswald, and Tom Chothia. 2019. Breaking all the things—A systematic survey of firmware extraction techniques for IoT devices. In *Smart Card Research and Advanced Applications*. Springer, Cham, 171–185.
[174] T. M. Vaughan, D. J. Mcfarland, G. Schalk, W. A. Sarnacki, D. J. Krusienski, E. W. Sellers, and J. R. Wolpaw. 2006. The wadsworth BCI research and development program: At home with BCI. *IEEE Trans. Neural Syst. Rehab. Eng.* 14, 2 (June 2006), 229–233.
[175] Ainuddin Wahid Abdul Wahab, Mustapha Aminu Bagiwa, Mohd Yamani Idna Idris, Suleman Khan, Zaidi Razak, and Muhammad Rezal Kamel Ariffin. 2014. Passive video forgery detection techniques: A survey. In *Proceedings of the 10th International Conference on Information Assurance and Security*. IEEE, 29–34.
[176] Yijun Wang and Tzyy-Ping Jung. 2011. A collaborative brain-computer interface for improving human performance. *PLoS ONE* 6, 5 (May 2011), e20422.
[177] Ping Yan and Zheng Yan. 2018. A survey on dynamic mobile malware detection. *Softw. Qual. J.* 26, 3 (Sep. 2018), 891–919.
[178] T. Yaqoob, H. Abbas, and M. Atiquzzaman. 2019. Security vulnerabilities, attacks, countermeasures, and regulations of networked medical devices—A review. *IEEE Commun. Surveys Tutor.* 21, 4 (2019), 3723–3768.
[179] Seung-Schik Yoo, Hyungmin Kim, Emmanuel Filandrianos, Seyed Javid Taghados, and Shinsuk Park. 2013. Non-invasive brain-to-brain interface (BBI): Establishing functional links between two brains. *PLoS ONE* 8, 4 (Apr. 2013), e60410.
[180] Tianyou Yu, Yuanqing Li, Jinyi Long, and Zhenghui Gu. 2012. Surfing the Internet with a BCI mouse. *J. Neural Eng.* 9, 3 (June 2012), 036012.
[181] Peng Yuan, Yijun Wang, Xiaorong Gao, Tzyy-Ping Jung, and Shangkai Gao. 2013. A collaborative brain-computer interface for accelerating human decision making. In *Proceedings of the International Conference on Universal Access in Human-Computer Interaction*. Springer, Berlin, 672–681.
[182] Lan Zhang, Ker Jiun Wang, Huan Chen, and Zhi Hong Mao. 2016. Internet of brain: Decoding human intention and coupling EEG signals with Internet services. In *Proceedings of the International Conference on Service Science (ICSS'16)*. IEEE, 172–179.
[183] PeiYun Zhang, MengChu Zhou, and Giancarlo Fortino. 2018. Security and trust issues in Fog computing: A survey. *Future Gen. Comput. Syst.* 88 (Nov. 2018), 16–27.
[184] Shaomin Zhang, Sheng Yuan, Lipeng Huang, Xiaoxiang Zheng, Zhaohui Wu, Kedi Xu, and Gang Pan. 2019. Human mind control of rat Cyborg's continuous locomotion with wireless brain-to-brain interface. *Sci. Rep.* 9, 1 (Dec 2019), 1321.
[185] Xiang Zhang, Lina Yao, Shuai Zhang, Salil Kanhere, Michael Sheng, and Yunhao Liu. 2019. Internet of Things meets brain–computer interface: A unified deep learning framework for enabling human-thing cognitive interactivity. *IEEE Internet Things J.* 6, 2 (Apr. 2019), 2084–2092.
[186] Yulong Zou, Jia Zhu, Xianbin Wang, and Lajos Hanzo. 2016. A survey on wireless security: Technical challenges, recent advances, and future trends. *Proc. IEEE* 104, 9 (Sep. 2016), 1727–1765.

Received November 2019; revised August 2020; accepted September 2020

2

Neuronal Flooding and Neuronal Scanning Cyberattacks

Title:	Cyberattacks on Miniature Brain Implants to Disrupt Spontaneous Neural Signaling
Authors:	Sergio López Bernal, Alberto Huertas Celdrán, Lorenzo Fernández Maimó, Michael Taynnan Barros, Sasitharan Balasubramaniam, Gregorio Martínez Pérez
Journal:	IEEE Access
JIF:	3.367 Q2 (2020)
Publisher:	IEEE
Volume:	8
Number:	
Pages:	152204-152222
Year:	2020
Month:	Aug
DOI:	10.1109/ACCESS.2020.3017394
Status:	Published

Abstract

Brain-Computer Interfaces (BCI) arose as systems that merge computing systems with the human brain to facilitate recording, stimulation, and inhibition of neural activity. Over the years, the development of BCI technologies has shifted towards miniaturization of devices that can be seamlessly embedded into the brain and can target single neuron or small population sensing and control. We present a motivating example highlighting vulnerabilities of two promising micron-scale BCI technologies, demonstrating the lack of security and privacy principles in existing solutions. This situation opens the door to a novel family of cyberattacks, called neuronal cyberattacks, affecting neuronal signaling. This article defines the first two neural cyberattacks, Neuronal Flooding (FLO) and Neuronal Scanning (SCA), where each threat can affect the natural activity of neurons. This work implements these attacks in a neuronal simulator to determine their impact over the spontaneous neuronal behavior, defining three metrics: number of spikes, percentage of shifts, and dispersion of spikes. Several experiments demonstrate that both cyberattacks produce a reduction of spikes compared to spontaneous behavior, generating a rise in temporal shifts and a dispersion increase. Mainly, SCA presents a higher impact than FLO in the metrics focused on the number of spikes and dispersion, where FLO is slightly more damaging, considering the percentage of shifts. Nevertheless, the intrinsic behavior of each attack generates a differentiation on how they alter neuronal signaling. FLO is adequate

to generate an immediate impact on the neuronal activity, whereas SCA presents higher effectiveness for damages to the neural signaling in the long-term.

Keywords

Brain-computer interfaces · Security · Artificial neural networks · Biological neural networks

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 17, 2020, accepted August 11, 2020, date of publication August 17, 2020, date of current version August 28, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3017394

Cyberattacks on Miniature Brain Implants to Disrupt Spontaneous Neural Signaling

SERGIO LÓPEZ BERNAL[1], ALBERTO HUERTAS CELDRÁN[2], LORENZO FERNÁNDEZ MAIMÓ[3], MICHAEL TAYNNAN BARROS[4,5], (Member, IEEE), SASITHARAN BALASUBRAMANIAM[2], (Senior Member, IEEE), AND GREGORIO MARTÍNEZ PÉREZ[1], (Member, IEEE)

[1]Departamiento de Ingeniería de la Información y las Comunicaciones, University of Murcia, 30100 Murcia, Spain
[2]Telecommunication Software and Systems Group, Waterford Institute of Technology, X91 K0EK Waterford, Ireland
[3]Departamento de Ingeniería y Tecnología de Computadores, University of Murcia, 30100 Murcia, Spain
[4]School of Computer Science and Electronic Engineering, University of Essex, Essex CO4 3SQ, U.K.
[5]CBIG/BioMediTech, Faculty of Medicine and Health Technology, Tampere University, 33014 Tampere, Finland

Corresponding author: Sergio López Bernal (slopez@um.es)

This work was supported by the Irish Research Council, under the Government of Ireland Postdoctoral Fellowship under Grant GOIPD/2018/466.

ABSTRACT Brain-Computer Interfaces (BCI) arose as systems that merge computing systems with the human brain to facilitate recording, stimulation, and inhibition of neural activity. Over the years, the development of BCI technologies has shifted towards miniaturization of devices that can be seamlessly embedded into the brain and can target single neuron or small population sensing and control. We present a motivating example highlighting vulnerabilities of two promising micron-scale BCI technologies, demonstrating the lack of security and privacy principles in existing solutions. This situation opens the door to a novel family of cyberattacks, called neuronal cyberattacks, affecting neuronal signaling. This article defines the first two neural cyberattacks, Neuronal Flooding (FLO) and Neuronal Scanning (SCA), where each threat can affect the natural activity of neurons. This work implements these attacks in a neuronal simulator to determine their impact over the spontaneous neuronal behavior, defining three metrics: number of spikes, percentage of shifts, and dispersion of spikes. Several experiments demonstrate that both cyberattacks produce a reduction of spikes compared to spontaneous behavior, generating a rise in temporal shifts and a dispersion increase. Mainly, SCA presents a higher impact than FLO in the metrics focused on the number of spikes and dispersion, where FLO is slightly more damaging, considering the percentage of shifts. Nevertheless, the intrinsic behavior of each attack generates a differentiation on how they alter neuronal signaling. FLO is adequate to generate an immediate impact on the neuronal activity, whereas SCA presents higher effectiveness for damages to the neural signaling in the long-term.

INDEX TERMS Brain computer interfaces, security, artificial neural networks, biological neural networks.

I. INTRODUCTION

Brain-computer Interfaces (BCIs) are considered as bidirectional communication systems between the brain and external computational devices. Although BCIs arose as systems focused on controlling external devices such as prosthetic limbs [1], they have gone one step further, enabling *artificial* stimulation and inhibition of neuronal activity [2]. In the last years, neuronal stimulation has already been applied in different scenarios such as the provision of sensory feedback to prosthetic or robotic limbs [3], treatment of neurodegenerative diseases or disorders like Alzheimer's or depression [4], and even futuristic applications such as interconnected networks of brains [5] or brains connected to the Internet [6].

New BCI technologies are emerging, allowing a precise acquisition, stimulation, and inhibition of neuronal signaling. It reduces the brain damage caused by traditional invasive BCI systems and improves the limitations of non-invasive technologies such as attenuation, resolution, and distortion constraints [7], [8]. One of the most recent and promising

The associate editor coordinating the review of this manuscript and approving it for publication was Yassine Maleh.

BCI technique focuses on the use of nanodevices allocated across the brain cortex [9]. Specifically, a relevant task of nanodevices equipped with optogenetic technology is the use of light to stimulate or inhibit engineered neurons according to different firing patterns sent by external transceivers [10]. Promising initiatives such as Neuralink aim to accelerate the development of these technologies [11].

The previous BCI technologies hold the promise of changing our society by improving the cognitive, sensory, and communications skills of their users. However, they also open the door to critical cyberattacks affecting the subjects' safety and data security. In this context, essential vulnerabilities of current non-invasive BCI systems have been documented, exploited, and partially solved in the literature [12]. As an example, the authors of [13], [14] demonstrated the feasibility of presenting malicious visual stimuli to extract subjects' sensitive data like thoughts. Besides, Sundararajan [15] conducted a successful jamming attack over the wireless communication used by the BCI, compromising its availability. However, the irruption of invasive and non-invasive stimulation and inhibition techniques, without security nor privacy capabilities, brings to the reality a novel family of cyberattacks affecting the neuronal activity. We call them *Neural cyberattacks*, and they present a critical number of open challenges like the definition and categorization of the different neural cyberattacks and their neuronal behavior, the impact of each cyberattack to the neuronal behavior, and their consequences in the brain and body.

To improve the previous challenges, the main contributions of this article are the following ones:

- The identification of cybersecurity vulnerabilities on emerging neurostimulation implants.
- To the best of our knowledge, the first description and implementation of neural cyberattacks focused on neuronal stimulation and affecting the activity of neural networks allocated in the human's brain. The proposed cyberattacks, *Neuronal Flooding* and *Neuronal Scanning*, are inspired by the behavior of current well-known cyberattacks in computer networks.
- The definition of three metrics to evaluate the impact of the two neural cyberattacks proposed: number of spikes, percentage of shifts, and dispersion of spikes.
- The implementation of the previous cyberattacks in a neuronal simulator to measure the impact produced by each one of them and the implications that they generate on the neuronal signaling. For that, we model a portion of a mouse's visual cortex based on the implementation of a CNN where the mouse is able to exit a maze.

The paper remainder is organized as follows. Section II gives an overview of the present state-of-the-art of current vulnerabilities, cyberattacks, and countermeasures affecting existing BCIs. After that, Section III illustrates emerging neurostimulation technologies and their cybersecurity concerns. Subsequently, Section IV offers a formal description of the cyberattacks proposed, while Section V describes the implemented use case. Section VI first presents the metrics used to evaluate the impact of these cyberattacks, followed by the analysis of the results and impact that these cyberattacks generate. Finally, Section VII briefly discusses the outcomes and potential future works.

II. RELATED WORK

During the last five years, new concepts such as brain-hacking, or neurocrime have emerged to describe relevant aspects of cybersecurity in BCI [16], [17]. These works highlight that neuronal engineering devices, designed to stimulate targeted regions of the brain, would become a critical cybersecurity problem. In particular, they acknowledge that attackers may maliciously attempt to program the stimulation therapy, affecting the patient's safety. Furthermore, they emphasize that the cyberthreats do not need to be too sophisticated if they only want to cause harm. In this context, as indicated in this article, it is possible to have a high impact on the brain by taking advantage of neurostimulation implants and send malicious electrical signals to the brain. Despite the identification of these risks, there are no studies in the literature defining or implementing neural cyberattacks, where the evaluation of their impact over the brain remains unexplored. However, several vulnerabilities and attacks have been detected in BCI technologies performing neural data acquisition (e.g., EEG), which can serve as a starting point to perform neural cyberattacks. Section III offers additional considerations about vulnerabilities in BCI solutions.

Platforms and frameworks that enable the development of BCI applications also present cybersecurity concerns, as demonstrated in [18], [19]. In this context, the authors of [18] performed an analysis of the privacy concerns of BCI application stores, including Software Development Kits (SDKs), Application Programming Interfaces (APIs), and BCI applications. They discovered that most applications have unrestricted access to subjects' brainwave signals and can easily extract private information about their subjects. Moreover, Cody's Emokit project [17], managed to break the encryption of the Emotiv EPOC device (valid for all models before 2016), having access to all raw data transmitted. The authors of [19] proposed a mechanism to prevent side-channel extraction of subjects' private data, based on the anonymization of neural signals before their storage and transmission.

The majority of the existing BCI systems are oriented to acquire, or record, neural data. Specifically, EEG BCI devices have gained popularity in recent years, due to their low cost and versatility, influencing the number of existing cyberattacks exploiting BCI vulnerabilities. In this context, the authors of [20] studied and analyzed well-known BCI applications and their potential cybersecurity and privacy concerns. Martinovic *et al.* [14] were able to extract users' sensitive information, such as debit cards or PINs, by presenting particular visual stimuli to the users and analyzing their P300 potential response. Another attack, performed by Frank *et al.* [13], focused on presenting subliminal visual stimuli included within a video, aiming to affect the BCI

users' privacy. Finally, in our previous work [21], we studied the feasibility of performing cybersecurity attacks against the stages of the BCI cycle, considering different communication architectures, and highlighting their impact and possible countermeasures.

In conclusion, this section demonstrates that most of the related works are focused on presenting vulnerabilities and cyberattacks affecting the confidentiality, availability, and integrity of private data managed by BCIs. Nevertheless, there is a lack of solutions considering cyberattacks affecting the neuronal activity and, therefore, the subjects' safety. This article proposes two neural cyberattacks affecting the natural behavior of single and population of neurons.

III. CYBERSECURITY VULNERABILITIES OF EMERGING NEUROSTIMULATION IMPLANTS

This section introduces three promising BCI technologies capable of recording and stimulating neuronal activity with single-neuron resolution. For each scenario, we offer a description of its architecture, highlighting the cybersecurity vulnerabilities detected. Although these solutions are in an early stage, and they are still not commercial products, they are contemporary examples of how cybersecurity can affect existing and future implantable BCI solutions, and in particular for solutions that can target small neuron populations. These issues represent the starting point for the cyberattacks illustrated in the next sections of this article. It is important to note that the objective of this section is not to find vulnerabilities in BCI devices or architectures but to justify how the proposed cyberattacks could be performed in realistic BCI systems.

A. NEURALINK

Neuralink aims to record and stimulate the brain using new technologies, materials, and procedures to reduce the impact of implanting electrodes in the brain [11]. The first element of the Neuralink architecture are the *threads*, proposed as an alternative for traditional electrodes due to their biocompatibility, reduced size based on thin threads that are woven into the brain tissue, durability, and the number of electrodes per thread. Groups of threads connect to an *N1 sensor*, a sealed device in charge of receiving the neural recordings from the threads and sending them stimulation impulses. With a simple medical procedure, up to ten N1 implants can be placed in the brain cortex. These devices connect, using tiny wires tunneled under the scalp, to a *coil* implanted under the ear. The coil communicates wirelessly through the skin with a wearable device, or *link*, placed under the ear. The *link* contains a battery that represents the only power source in the architecture, deactivated if the user removes the link. FIGURE 1 represents this architecture.

Although the communication mechanisms between the coil and the link are not provided, the link is managed via Bluetooth from external devices, such as smartphones, using an application. In this sense, Neuralink users can manage and personalize their links, upgrade their firmware, and include

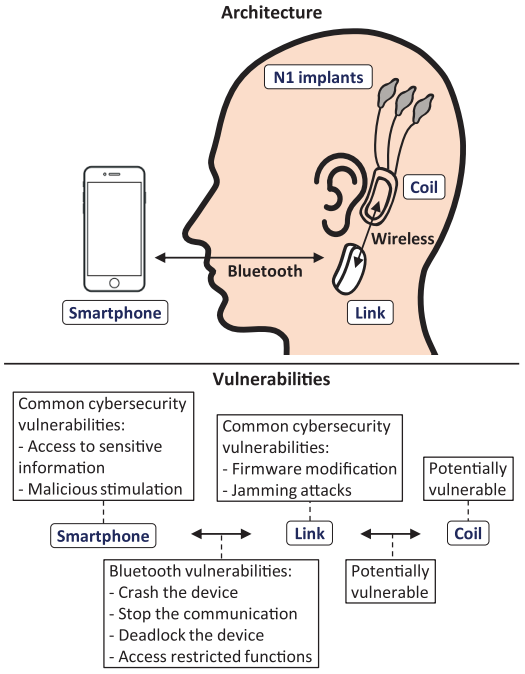


FIGURE 1. Architecture and vulnerabilities of Neuralink.

new security capabilities. We identify that this scenario can be potentially vulnerable as follows. First, the wireless mechanism used in the communication between the coil and the link could be vulnerable, depending on the protocol used [22]. Besides, the Bluetooth communication between the smartphone and the link can also be vulnerable, according to the version used [23], [24]. As an example, we identify SweynTooth, a set of 12 vulnerabilities affecting a large number of devices using Bluetooth Low Energy (BLE) technologies. Based on them, an attacker could crash the device and stop its communications [25], deadlock the device [26], or access functions only available for authorized users [27].

Moreover, the external device manages the logic of both acquisition and stimulation processes, including into these scenarios its inherent risks, and becoming one of the most sensitive elements of the architecture. In particular, Li *et al.* [20] detected that attackers could take total control of a smartphone running a BCI application, getting access to sensitive information, or performing malicious stimulation actions. Furthermore, the *link* is a critical element of the architecture, where attackers can modify the firmware of the device to have a malicious behavior, as identified by [28] for brain implants or to perform jamming attacks to disrupt the communication between devices, described by [29] for wireless networks.

B. NEURAL DUST

This architecture is composed of millions of resource-constrained nanoscale implantable devices, also known as

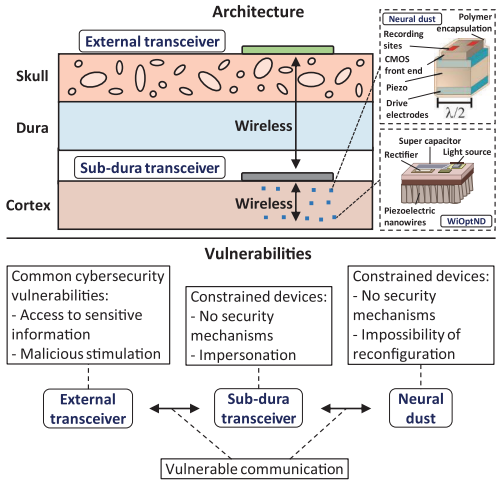


FIGURE 2. Architecture and vulnerabilities of Neural dust.

neural dust, floating in the cortex, able to monitor neural electrophysiological activity precisely [9]. These devices communicate with the sub-dura transceiver, a miniature device (constructed from components that are built from nanomaterials) placed beneath the skull and below the dura mater. This device uses two different transceivers to: (1) power and establish communication links with the neural dust, (2) communicate with external devices. During neural recording, the sub-dura transceiver performs both spatial and frequency discrimination with sufficient bandwidth to power and interrogate each neural dust. The external transceiver is a device without computational and storage restrictions, allocated outside of the patient's head. Wearables, smartphones, or PCs are examples of this device. The main task of the external transceiver is to power and communicate with the sub-dura transceiver and to receive the neuronal behavior from the sensing by the neural dust. FIGURE 2 presents the architecture of this solution, as well as the potential vulnerabilities that it presents.

Nevertheless, this technology has not been conceived following the principle of security and privacy by design. As a consequence, these devices do not implement authentication mechanisms to prevent malicious users from collecting neural sensing data from the neural dust, and they do not protect the transmitted data. In particular, the neural dust are resource-constrained devices without computational and storage capabilities to execute security functionalities like authentication protocols, ciphered communications, or data encryption. In this sense, external attackers could power and communicate to the implants to monitor private neural data. Finally, the sub-dura and external transceivers do not implement authentication protocols nor security mechanisms. An attacker could impersonate the external transceiver to communicate with the sub-dura device, and obtain sensitive neuronal signaling.

C. WIRELESS OPTOGENETIC NANONETWORKS

The Wireless Optogenetic Nanonetworking device (WiOptND) [10] is an extension from the neural dust [9] but with the capability of optogenetically stimulating the neurons. Optogenetic stimulation uses light to stimulate neurons genetically engineered with specific genes that are sensitive to signals at a particular wavelength. This in turn provides targeted stimulation of very small population of neurons that have been engineered, enabling precise targeting of neural circuits within the micro-columns. Similar to the architecture of the neural dust, the WiOptND also receives power that is emitted from the sub-dura, which in turn communicates to the external transceiver. However, since the WiOptND is responsible for stimulating the neurons, the external transceiver will communicate the sequence of firing the neurons to the sub-dura transceiver to synchronize the charging and communication of the WiOptND implants. This is achieved by sending the firing sequence, in the form of a raster plot, to the external transceiver. This opens up new opportunities for attackers to send malicious firing patterns into the external transceiver, which will produce a new sequence of firing patterns for neural stimulation, resulting in detrimental consequences for the brain. Finally, the architecture and vulnerabilities described in FIGURE 2 also apply for WiOptND.

In conclusion, the previous vulnerabilities raise different concerns affecting the integrity, confidentiality and availability of subject's neural data. These vulnerabilities motivate different attack vectors to perform the neural cyberattacks described in subsequent sections.

IV. DEFINITION OF NEURAL CYBERATTACKS

Once demonstrated the feasibility of stimulating individual neurons by attacking different technological solutions, we formally describe two cyberattacks, *Neuronal Scanning* and *Neuronal Flooding*, aiming to maliciously affect the natural activity of neurons during neurostimulation procedures. They are inspired by the behavior and goals of some of the most well-known and dangerous cyberattacks affecting computer networks.

To formalize both cyberattacks, we denote $\mathbb{NE} \subset \mathbb{N}$ as a subset of neurons from the brain, where $n \in \mathbb{NE}$ expresses every single neuron. The voltage of a single neuron in a specific instant of time is denoted as $v_n \in \mathbb{R}$, whereas $vi_n \in \mathbb{R}$ indicates the voltage increase used to overstimulate a neuron n. Moreover, t^{win} represents a temporal window in which the cyberattack is performed, equivalent to the duration of the simulation in Section VI. t^{attk} is the time instant when the cyberattack starts, and Δt the amount of time between evaluations during the process. In the implementation of the cyberattacks, it represents the duration of the steps of the simulation.

1) NEURONAL FLOODING

In the cyberworld, a flooding cyberattack is designed to bring a network or service down by collapsing it with large

amounts of network traffic. Traffic is usually generated by many attackers and forwarded to one or more victims. Extrapolating this network cyberthreat to the brain, a Neuronal Flooding (FLO) cyberattack consists in stimulating multiple neurons in a particular instant of time, changing the normal behavior of the stimulation process and generating an overstimulation impact. The execution of this cyberattack does not require prior knowledge of the status of the affected neurons since the attacker only has to decide what neurons to stimulate and when. This fact makes this cyberattack less complex than other cyberattacks that require prior knowledge of the neuronal behavior.

In particular, FLO performs the overstimulation action at t^{attk}. In that precise moment, a subset of neurons $\mathbb{AN} \subseteq \mathbb{NE}$ is attacked. This cyberattack is formally described in Algorithm 1.

Algorithm 1 FLO Cyberattack Execution

$t = 0$
while $t < t^{win}$ **do**
  if $t == t^{attk}$ **then**
    for all $n \in \mathbb{AN}$ **do**
      $v_n \leftarrow v_n + vi_n$
    end for
  end if
  $t \leftarrow t + \Delta t$
end while

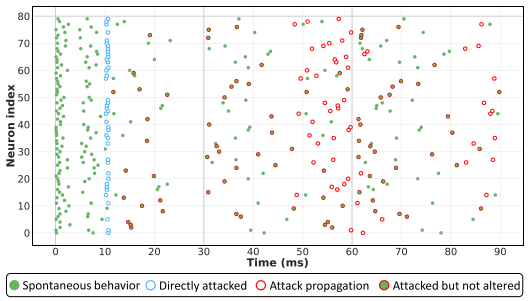


FIGURE 3. Raster plot of a FLO cyberattack when the attack is performed at 10ms.

FIGURE 3 represents an example to appreciate graphically the behavior of a FLO cyberattack, where the details of the neuronal network used in the simulation are not relevant at this point (addressed in Section V). In particular, it represents the comparison of the FLO cyberattack with the spontaneous behavior for a simulation of 80 neurons, a duration of 90ms, and 42 neurons attacked in the instant 10ms. Green dots represent the neuronal spontaneous behavior, blue circles indicate the instant when the neurons are attacked, red circles highlight the propagation of the cyberattack in time, and those dots with a green color and red outline represent spikes common to both spontaneous and under attack situations. In this figure, we can see that all the attacked neurons alter their behavior, having spikes in different moments compared to the spontaneous activity.

2) NEURONAL SCANNING

Port scanning is another well-known cybersecurity technique performed by attackers to discover vulnerabilities in operating systems, programs, and protocols using network communications. In particular, it aims to test every networking port of a machine, checking if it is open and discovering the protocol or service available in that end-point. In the brain context, a Neuronal Scanning (SCA) cyberattack stimulates neurons sequentially, impacting only one neuron per instant of time. Based on that, it is essential to note that attackers do not require prior knowledge of the neuronal state to perform neural scanning cyberattacks. This fact, together with the stimulation of one neuron per instant of time, makes a low attack complexity.

Considering the notation previously defined, Algorithm 2 describes an SCA cyberattack. In particular, it sequentially overstimulates all the neurons included in the set of neurons $\mathbb{NE}$, without repetitions. For each neuron n, its voltage v_n increases by vi_n. It is essential to indicate that the conditional clause limits the instants in which an attack can be performed, where t^{attk} represents the attack over the first neuron of the set, and $t^{attk} + |NE|\Delta t$ the attack over the last neuron.

Algorithm 2 SCA Cyberattack Execution

$t = 0$
while $t < t^{win}$ **do**
  if $t \in [t^{attk}, t^{attk} + |NE|\Delta t]$ **then**
    $n \leftarrow (t - t^{attk})/\Delta t$
    $v_n \leftarrow v_n + vi_n$
  end if
  $t \leftarrow t + \Delta t$
end while

Finally, FIGURE 4 shows, in a visual way, the behavior of an SCA cyberattack. We simulate 80 neurons during 90s, and sequentially attack all neurons, starting in the instant 10ms. The color code followed is the same as in FIGURE 3. As can be seen, the sequential attack of the neurons generates a diagonal line in the spikes. All spikes over the line remain unaltered since those neurons have not yet been affected by the attack. On the contrary, the spikes under the diagonal are affected by the attack.

V. EXPLOITING VULNERABILITIES DUE TO CYBERATTACKS

This section introduces the use case used to implement the cyberattacks defined in Section IV. We present the scenario and the experimental setup implemented to create the neuronal topology required to test the cyberattacks.

A. USE CASE AND EXPERIMENTAL SETUP

The knowledge of precise neocortical synaptic connections in mammalian is nowadays an open challenge [30]. Based on

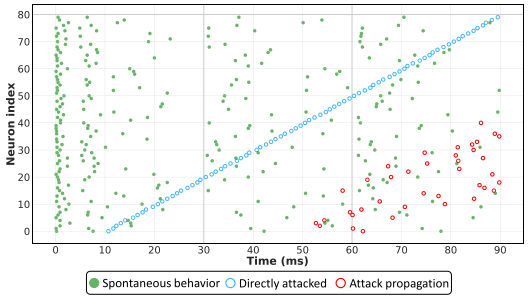


FIGURE 4. Raster plot of an SCA cyberattack, from the instant 10ms to 90ms.

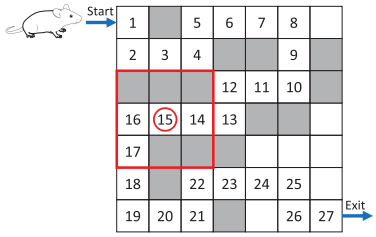


FIGURE 5. Maze used in our use case to model the movement of the mouse, including the optimal path between the starting and final cells. There are nine visible positions from the cell 15, highlighted within a red square.

TABLE 1. Summary of the layers of the CNN.

Layer	Type	Filters	Input size	Output size	Kernel size	Stride	Activation function	Nodes
1	Conv2D	8	7×7×1	5×5×8	3×3	1	ReLU	200
2	Conv2D	8	5×5×8	3×3×8	3×3	1	ReLU	72
3	Dense	-	3×3×8	4	-	-	ReLU	4

this absence of realistic neuronal topologies, we have studied the primary visual cortex of mice and replicated a portion of it, modeled using a Convolutional Neural Network (CNN) [31]. This CNN was trained by means of reinforcement learning [32] to represent a simple system able to make decisions based on a maze and find its exit. As indicated by Kuzovkin *et al.* [33], CNNs, and biological neuronal networks present certain similarities. First, lower layers of a CNN explain gamma-band signals from earlier visual areas, whereas higher layers explain later visual regions. Furthermore, early visual areas are mapped to convolutional layers, where the fully connected layers match the activity of higher visual areas. That is to say, the visual recognition process in both networks is incremental and move from simple to abstract. At this point, it is essential to note that we cannot compare the topology and functionality of a CNN to the complexity of the neuronal connections of a real brain. We only used this technique to provide a simple topology that is then implemented in a neuronal simulator to evaluate how attacks over a simplistic but realistic environment can affect the activity of simulated neurons, as indicated in Section V-C.

In this context, we designed a simple proof of concept based on the idea of a mouse that has to solve the problem of finding the exit of a particular maze, inspired in the code from [34]. The mouse must find the exit with the smallest number of movements and starting from any position. We define a maze of 7 × 7 coordinates, as represented in FIGURE 5. It contains one starting position identified with "1", while the exit is labeled with "27". Moreover, the positions colored in gray represent obstacles, and those in white are accessible positions through which the mouse can move. In this scenario, the mouse can move in all four 2D directions: up, down, left, and right. The numbering from 1 to 27 defines the optimal path determined by the trained CNN to reach the exit position, considering the lowest number of steps. Finally, it is essential to define the concept of *visible position*. From each particular cell of the maze, the mouse can visualize a square of 3 × 3 adjacent positions, including those that represent obstacles. This situation is highlighted in FIGURE 5 with a red square, indicating the visible positions from the cell 15 of the optimal path.

B. CONVOLUTIONAL NEURAL NETWORK

Our objective was to generate a CNN able to exit the maze from any position. We also aimed to define a topology with a reduced number of nodes to be compatible with resource-constrained neuronal simulators since we aim to evaluate this topology in multiple simulators. Nevertheless, for simplicity, this work includes details of the implementation in only one simulator, as described in Section V-C. To solve our maze problem, we implemented a CNN composed of two convolution layers and a dense layer. The ensemble of these three layers defines a complete CNN of 276 neurons, representing a small portion of a mouse primary visual cortex, summarized in Table 1. We implemented this CNN using Keras on top of TensorFlow [35].

FIGURE 6 depicts the architecture of the implemented CNN which is also described in Table 1. In particular, we have included a first 2D convolution layer with a 3 × 3 kernel. This layer takes as input the current status of the maze, focusing each neuron on a square of 9 (3 × 3) adjacent positions. In our experiments we determined that 8 filters of size 3 × 3 in each layer were sufficiently expressive. To represent the maze, each position contains a 1 value if the position is accessible, a 0 value if it is an obstacle, or a 0.5 value in the position of the mouse.

During the training, each filter of the first layer specializes on a particular aspect of the maze. For example, a filter could focus on detecting vertical walls, while another could detect corners. The filters of the second layer can detect more complex scenarios by composing the output of these initial detectors. Since the input is a 7 × 7 maze, and the kernel is 3 × 3, the first convolution process requires 25 neurons (5 × 5 kernel outputs) to cover the new 5 × 5 subset of the maze on the next layer. Since we use 8 different filters, the total number of neurons required to produce the first layer's output of the CNN is 200 (5 × 5 × 8). This is illustrated in FIGURE 6, where each group of neurons has a different color

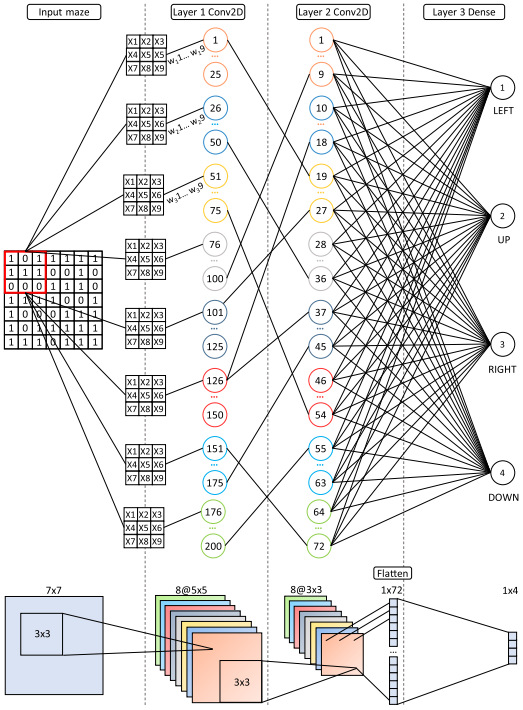


FIGURE 6. Visual representation of the implemented CNN. It introduces a simplifications of the whole topology, indicating how the convolution process is performed and how nodes connect between layers. The color of each node matches the color of its associated filter.

that matches the color of its filter. Therefore, since the first layer generates an output of size $5 \times 5 \times 8$, the application of the 3×3 kernels of the second convolutional layer requires a total of 72 ($3 \times 3 \times 8$) neurons. Finally, this new output is sent through a last dense layer of 4 neurons, one for each possible movement direction on the maze (left, up, right, down). Each output is an estimation of the probability of success with each movement, being selected the direction with the greatest score.

In order to understand Section V-C and Section VI, it is necessary to explain the mapping between the sequential number of each neuron and its position in its associated filter output. FIGURE 6 shows this mapping. Each neuron have associated a 3-dimensional vector, where the third coordinate is its filter and the two first coordinates, the position in that filter output. The order is as follows: the first neuron has the coordinates [0,0,0], corresponding to the first neuron in the first filter output; the eighth neuron corresponds to [0,0,7]; the ninth one is [0,1,0], and so on until the 200th neuron, with coordinates [4,4,7].

C. BIOLOGICAL NEURONAL SIMULATION

After training the CNN, we represented its resulting topology in Brian2, a lightweight neuronal simulator [36]. We selected

TABLE 2. Relationship of parameters between artificial and biological networks.

	CNN	Simulation
Number of neurons	276	
Number of layers	3	
Neuronal topology	200 (Layer 1), 72 (Layer 2), 4 (Layer 3)	
Input data	Maze	
Types of neurons	Artificial	Pyramidal neuron from primary visual cortex
Connection weights	Filter weights	Synaptic weights

Brian2 because it is adequate to run neuronal models in user-grade computers, without the requirement of using multiple machines, or even supercomputers. It also presents a good behavior in the implementation of neuron models with simplified and discontinuous dynamics (such as Leaky Integrate-and-Fire or Izhikevich) [37]. Other alternatives, such as NEURON, present complex solutions to model neurons with fine granularity, offering distributed computation capabilities for high demanding simulations. Nevertheless, this functionality is unnecessary in our particular study.

We maintain in the biological simulation the exact number of layers, the number of neurons per layer, and the topological connections between neurons. However, there is a crucial difference between the implementation of these two approaches. In the CNN, a filter weight represents the importance that a connection between two neurons of different layers have on the topology and, thus, over the solution. In the biological simulation, we transform the CNN weights to synaptic weights, representing the increase of the voltage induced during an action potential. Table 2 summarizes these similarities and differences between both networks.

To represent the behavior of each neuron, we decided to use the Izhikevich neuronal model since it is computationally inexpensive, and it allows us to precisely model different types of neurons within different regions of the brain [38]. This model represents an abstraction of how cortical neurons behave in the brain. In particular, the following set of equations describes the Izhikevich model, whose parameters are indicated in Table 3. This model allows multiple configurations to mimic different regions of the brain. In our scenario, we assigned particular values to the previous parameters to implement a regular spiking signaling from the cerebral cortex, as indicated in [38]. Specifically, we aimed to model pyramidal neurons from the primary visual cortex of a mouse, which correspond to excitatory neurons typically present in the biological visual layers L2/3, L5, and L6 [39]. For simplicity, during the analysis of the results of the simulation, we will refer to these layers in subsequent sections as first layer (L2/3), second layer (L5) and third layer (L6).

$$v' = 0.04v^2 + 5v + 140 + u + I \tag{1}$$

$$u' = a(bv - u) \tag{2}$$

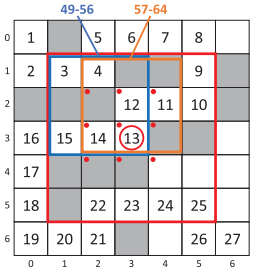
$$if\ v \geqslant 30mV, then \begin{cases} v, & \leftarrow c \\ u, & \leftarrow u + d \end{cases} \tag{3}$$

TABLE 3. Parameters used in the Izhikevich model.

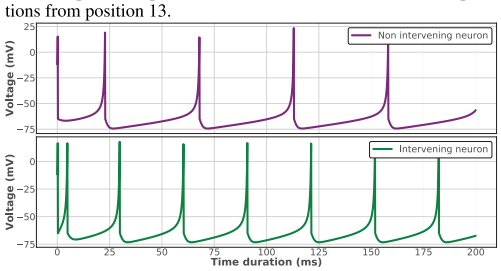
Parameter	Description	Values
v	Membrane potential of a neuron	[-65, 30] mV
u	Membrane recovery variable providing negative feedback to v	(-16, 2) mV/ms
a	Time scale of u	0.02/ms
b	Sensitivity of u to the sub-threshold fluctuations of v	0.2/ms
c	After-spike reset value of v	-65mV
d	After-spike reset value of u	8mV/ms
I	Injected synaptic currents	{10, 15} mV/ms

To create our neuronal topology, we used the weights of the trained CNN as post-synaptic voltage values, normalized within the range between $5mV$ and $10mV$. We selected this range because these values constitute a conservative voltage raise within the range of values of v, indicated in Table 3. At the beginning of the simulation, we assigned the initial voltage of each neuron from a previously generated random list in the range $[-65mV, 0mV)$. This initial value for each neuron is constant between executions to allow their comparison. To define a more realistic use case, we represented in our simulation the movement of the mouse inside the maze (see FIGURE 5), staying one second in each position of the optimal path. To understand this, it is essential to introduce the concept of *intervening neurons*, which defines the set of neurons managing all the visible positions of the mouse when it is placed in a particular position of the maze. FIGURE 7a illustrates the relationship between the position 13 of the optimal path and its intervening neurons, not considering its related visible positions for simplicity. For this position, we define nine 3×3 squares within the surface delimited by the red square, where we represent only the first two squares to improve the legibility of the figure. Focusing on the first square, colored in blue, it comprises eight neurons indexes (49 to 56), obtained from the translation between 3-dimension coordinates previously commented in this section. The second one, highlighted in orange, associates eight different neurons. After applying all nine squares, we obtain the complete list of intervening neurons related to the position 13. This single process is repeated for every visible position from the position 13 (indicated in FIGURE 7a with red dots), obtaining the complete set of intervening neurons. This set of intervening neurons is presented in Table 4, where each visible position from the position 13 is identified by its maze coordinate for simplicity. The last row of the table presents the complete set of intervening neurons for the position 13, obtained as the union of all individual sets of neurons.

The movement of the mouse was implemented by providing external stimuli to the simulation via the I parameter, where a value of $15mV$ was assigned to all intervening neurons from the current location of the mouse. For all non-intervening neurons in a specific instant, we assigned a value of $10mV$. These values align with the range defined in [38]. This information was extracted from the topology of the CNN, which contains the relationship between the neurons of the first layer and the positions of the maze. We took into consideration these aspects in the experimental analysis performed in Section VI. Based on that, we modeled with a higher value of I those intervening neurons, transmitting a more potent visual stimulus to those neurons related to adjacent positions from the current location. Based on Equation 1, an increase in the I parameter will produce a voltage rise in these intervening neurons, generating a raise in the amplitude of the electrical signal. This behavior was modeled taking into consideration the study performed in [40], which indicates that a known visual stimulus generates a voltage amplitude increase. FIGURE 7b graphically compares these differences between values of the I parameter. It highlights that intervening neurons present a higher number of spikes during a particular temporal window, which is interpreted by the brain as the reconnaissance of accessible cells in the maze from the current position.



(a) Relationship between visible positions from the current location of the mouse and their intervening neurons. In this example, the mouse is placed in position 13. Red dots indicate the visible positions from position 13.



(b) Impact of the I parameter on both intervening ($15mV$) and non intervening ($10mV$) neurons.

FIGURE 7. Relationship between positions of the maze and its implication in the modulation of neuronal signaling.

TABLE 4. List of intervening neurons associated to the position 13 of the optimal path of the maze.

Coordinate	List of intervening neurons
(2,2)	[1,24], [41,64], [81,104]
(2,3)	[9,32], [49, 72], [89, 112]
(2,4)	[17,40], [57,80], [97, 120]
(3,2)	[41,64], [81, 104], [121,144]
(3,3)	[49,72], [89,112], [129,152]
(3,4)	[57,80], [97,120], [137,160]
(4,2)	[81,104], [121,144], [161,184]
(4,3)	[89,112], [129,152], [169,192]
(4,4)	[97,120], [137,160], [177,200]
Position 13	[1,200]

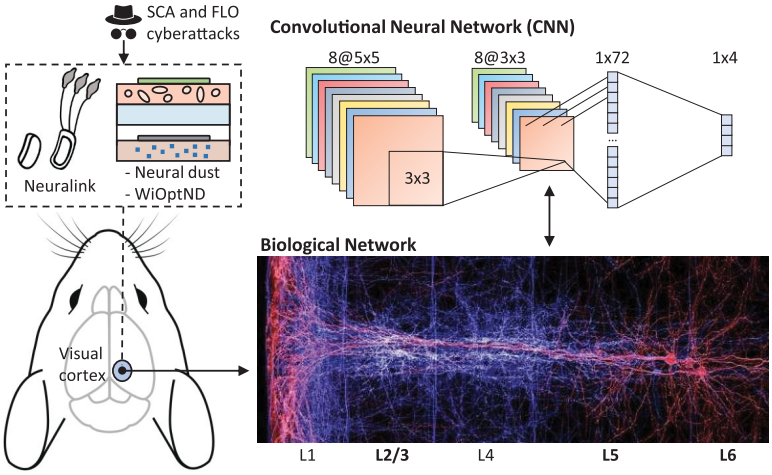


FIGURE 8. Summary of our use case, indicating the translation between the topology of the CNN implemented and the biological network simulated.

Finally, FIGURE 8 introduces a graphical summary of the current use case. It depicts a mouse with a miniature brain implant solution in its primary visual cortex, such as Neuralink or Neural dust. To simulate its biological neuronal network, and based on a lack of realistic cortical topologies, a trained CNN provides the number of nodes and distribution in layers for the biological network. In particular, we modeled pyramidal neurons from visual layers L2/3, L5, and L6, using the Izhikevich model with a regular spiking signaling. Based on this scenario, an external attacker takes advantage of contemporary vulnerabilities in these implantable solutions to alter the behavior of the spontaneous activity of the biological neuronal network.

VI. RESULTS ANALYSIS BASED ON METRICS

In this section, we evaluate the impact that FLO and SCA cyberattacks have on spontaneous neuronal activity of the neuronal topology presented in Section V. To analyze the evolution of the cyberattacks impact while the mouse is moving across the maze, we consider the following three metrics:

- Number of spikes: determine if a cyberattack either increases or reduces the quantity of spikes compared to the spontaneous neuronal signaling.
- Percentage of shifts, being a shift the delay of a spike in time (forward or backward) compared to the spontaneous behavior: study if a cyberattack generates significant delays in the normal activity of the neurons.
- Dispersion of spikes in both dimensions of time and number of spikes: analyze the spiking patterns under attack, aiming to detect if the cyberattack causes a modification on the distribution of the spikes.

For each layer of the topology, and combining all of them, we measured and analyzed the number of spikes and percentage of shifts. Finally, the dispersion of spikes is computed for each position of the optimal path and grouping all layers. Finally, we compared the impact generated by both cyberattacks.

To better understand the impact of FLO and SCA cyberattacks, FIGURE 9 compares the evolution of neuronal spikes for the spontaneous activity, a FLO cyberattack and an SCA cyberattack. We selected three positions of the optimal path to analyze in detail the spiking evolution along with the simulation, presenting only the first 100ms of each position. It is essential to note that this simplification is only for this figure, and all the results subsequently presented consider the complete duration of each position. As can be seen, in the spontaneous signaling, there is a certain natural dispersion caused by the behavior of the neuronal model used, and the movement of the mouse (due to the the modification of the associated I parameter). Specifically, each time the mouse changes from one position to another, the I parameter changes according to the intervening neurons, where a higher value of I is translated to a higher spike rate (see Algorithm 1). Since the mouse periodically changes its position, it modifies the spiking rate of the neurons, generating a natural dispersion in the absence of attacks. Looking at the first position of both spontaneous and FLO, in the instant 50ms, there is a clear difference between them, since we executed the attack in that exact instant. The set of attacked neurons generates spikes before it was intended due to the voltage rise produced by the attack. Consequently, we can see that the dispersion over the following positions (13 and 27) augments, altering the natural pattern of the neurons. Regarding the SCA cyberattack, it also starts in the instant 50ms but, its impact it is not yet present in the first 100ms of the initial position. If we check the subsequent positions, the attack gradually propagates,

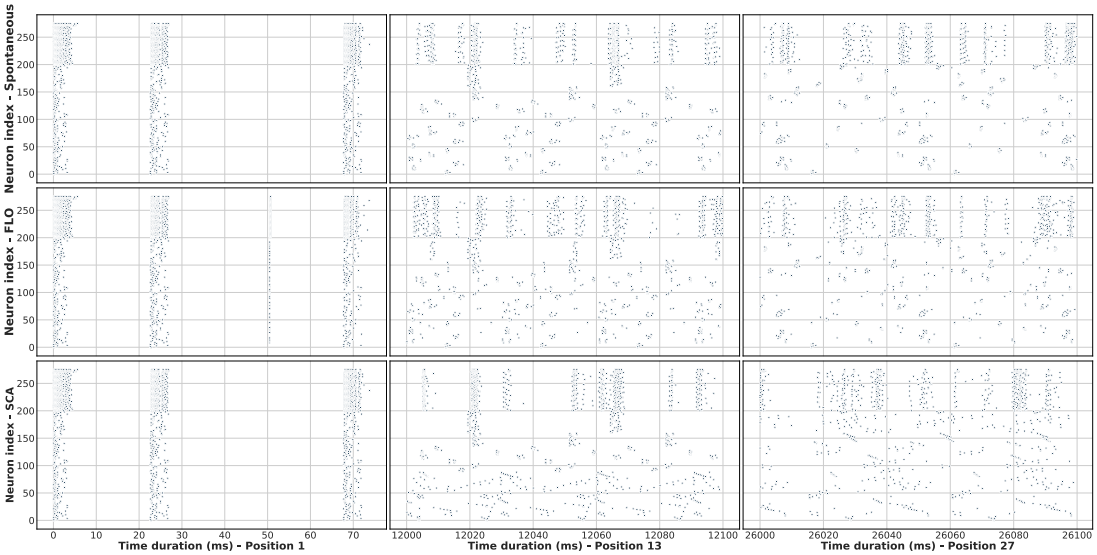


FIGURE 9. Raster plots indicating the evolution of the spontaneous signaling and both FLO and SCA cyberattacks for three positions of the optimal path of the maze.

generating characteristic ascending patterns. Subsequent subsections analyze, in a more detailed way, the information contained in FIGURE 9, extending the analysis to all the positions of the optimal path and using the previous three metrics.

A. NEURONAL FLOODING

In this subsection, we aim to simultaneously attack multiple neurons and analyze its impact using the metrics previously indicated at the beginning of the section. The implementation of this cyberattack is based on the general description indicated in Algorithm 1. We decided to perform only the attacks over the first layer of the topology, from where each target neuron is randomly selected, to evaluate the propagation to deeper layers. Furthermore, we tested a combination of two additional parameters. The first one represents the number of simultaneously attacked neurons, $k \in \{5, 15, \ldots, 95, 105\}$. $\mathbb{AN}$ will contain k neurons randomly selected from $\mathbb{NE}$, the set of neurons in the first layer. It is worthy to note that we reached to attack simultaneously more than half of the neurons of the first layer, which represents a fairly aggressive portion of the neurons. The second parameter of the attack, $\mathbb{VI} = \{20, 40, 60\}$, indicates the different voltage increases in mV used to stimulate the neurons in $\mathbb{AN}$. Its maximum level, $60mV$, approximately represents two-thirds of the voltage range defined by the Izhikevich model. We have executed each combination of parameters 10 times, denoted as $exec = 10$, to ensure that the random selection of neurons performed is representative. The value of t^{sim} is 27s (one second per position of the optimal path), and t^{attk}, is 50ms. Table 5 summarizes the previously indicated parameters.

TABLE 5. Configuration of the implemented FLO cyberattack.

Parameter	Description	Values
t^{sim}	Duration of the simulation	27s
t^{attk}	Instant of the attack	50ms
$\mathbb{NE}$	Set containing all neurons of the first layer	$\{1, 2, ..., 199, 200\}$
$\lvert\mathbb{AN}\rvert$	Number of simultaneously attacked neurons	$\{5, 15, ..., 95, 105\}$
$\mathbb{VI}$	Set containing the voltage used to attack random neurons	$\{20, 40, 60\}mV$
$exec$	Number of executions per combination of parameters	10

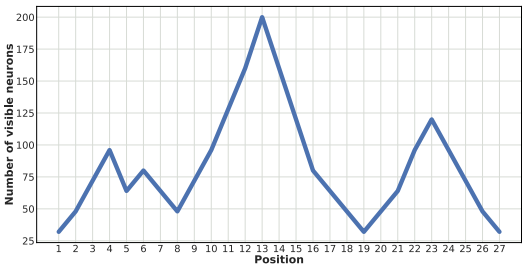


FIGURE 10. Number of intervening neurons related to visible positions from each position of the optimal path.

1) NUMBER OF SPIKES METRIC

To better understand the analysis of this metric, it is necessary to introduce FIGURE 10, which shows, for each position of the optimal path of the maze, the number of intervening neurons involved in the decision-making process of the mouse. Since these intervening neurons are dependent on the number of visible positions from a particular location of the maze, the number of intervening neurons is higher in central cells of the maze compared to those placed near the borders. Moreover, intervening neurons are dependent on the topology

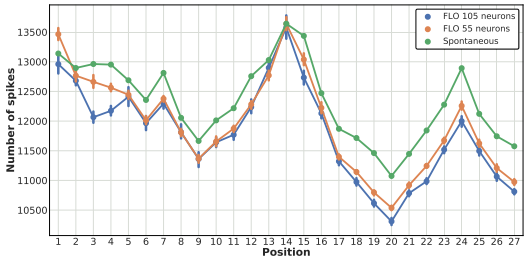


FIGURE 11. Total number of spikes for all neurons of the topology per position of the optimal path, attacking different number of neurons (105 and 55 simultaneous neurons).

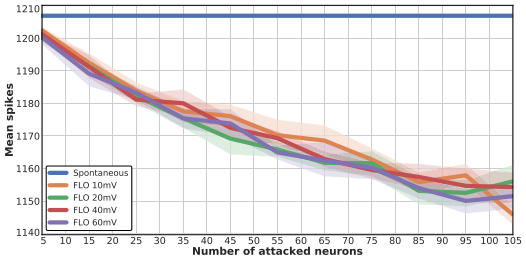


FIGURE 12. Evolution of the mean of spikes with different number of attacked neurons and voltage increases, aggregating all positions of the optimal path.

used and the convolution process of the CNN, as depicted in FIGURE 6.

FIGURE 11 compares, for the spontaneous signaling and two different configurations of FLO, the total number of spikes per position of the optimal path. In particular, the graph plots two different amounts of neurons in $\mathbb{AN}$ (55 and 105 neurons) for all *exec* simulations. In this figure, we fixed v_i to a value of $40mV$ to improve its visualization. As can be seen, both figures share a common tendency, indicating that the higher the number of intervening neurons from a position, the higher the number of spikes. This is a consequence of how the mouse moves across the maze and how neurons and positions are related based on our particular topology. Comparing both figures, FIGURE 10 reaches its highest peaks one position before, since this change of intervening neurons needs to be propagated in time, affecting the number of spikes of its following position.

In FIGURE 11, we can see that, in general, FLO cyberattacks reduce the number of spikes compared to the spontaneous activity, increasing this reduction when the mouse progresses in the maze. Furthermore, increasing the impact of the attack, in terms of the number of attacked neurons, reduces the number of spikes. These aspects are aligned with the results later presented in Section VI-A3, where this reduction is caused by an increase of the dispersion in the attacked neurons. However, it is worth noticing the high number of spikes produced in the first position. The Izhikevich neuronal model for regular spiking generates a quick burst of spikes in a short time, and, after that, it stabilizes its spike rate, explaining this behavior. When we apply a FLO cyberattack, the attacked neurons anticipate their spikes, producing either a raise of spikes if the number of attacked neurons is not so elevated (low dispersion in time), or a reduction of spikes if most of the neurons are attacked (high dispersion). Moreover, the evolution of the simulation after the attack does not tend to come back to the spontaneous signaling, in terms of the number of spikes. In fact, these distances augment over time, reaching a difference of around 700 spikes in position 27, with some variability between both FLO configurations. Based on that, these results indicate that the effect of attacking neurons in a particular instant propagates until the end of the simulation.

After this analysis, we considered relevant to evaluate how the mean of spikes evolved through the three layers of the topology with different configurations of the FLO cyberattack. In particular, we tested different amounts of attacked neurons and voltage increase, with *exec* different executions for each combination of the previous parameters. Using *exec* executions introduces variability in terms of the randomly selected neurons for each execution. We present these results in FIGURE 12, which represents an aggregation of the number of spikes produced during the optimal path of the maze. It indicates that increasing the number of attacked neurons derives in a higher reduction in the number of spikes, while the application of different voltages does not produce a high impact. The dimmed colors surrounding the main lines of the figure indicate the fluctuations between the *exec* simulations. As can be seen, the difference in the mean of spikes compared to the spontaneous signaling grows when the number of attacked neurons raises, having a difference of around 60 spikes for 110 attacked neurons (half of the first layer). These results align with those presented in FIGURE 11 for the positions of the optimal path, where both figures present a clear descending trend when the number of attacked neurons augments. Finally, the use of different increases of voltage during the experiments did not generate a considerable impact on the number of spikes.

To expand the focus on this analysis and to determine whether this descending trend is exclusive to only certain layers, FIGURE 13 analyzes the same parameters but differentiating between the three layers of the topology and focusing only on the last position of the optimal path of the maze. We can see that the variation of the mean of spikes is more significant in deeper layers (2nd and 3rd). This variation is due to the distribution of our topology and the normal behavior of the brain, where initial layers propagate their behavior to subsequent layers, magnifying their activity via synapses. The y-axis range considerably differs between layers, being the difference with the spontaneous signaling of less than one spike in the first layer. The second layer offers a broader range of around 8 spikes in the most damaging situation, whereas the third layer has an approximate separation of between 10 to 25 spikes.

In summary, the previous figures indicate that, under attack, the mean of spikes decreases compared to the

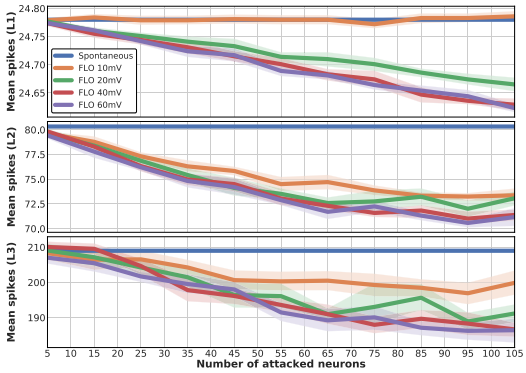


FIGURE 13. Mean of spikes for each layer of the topology, focusing on the last position of the optimal path.

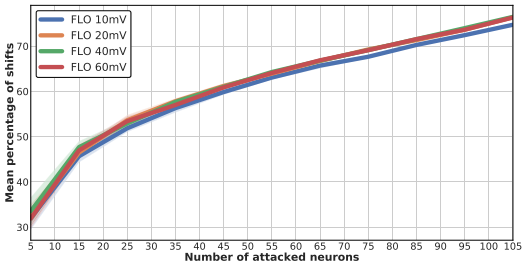


FIGURE 14. Shift percentage mean for an aggregation of all topological layers and positions of the optimal path.

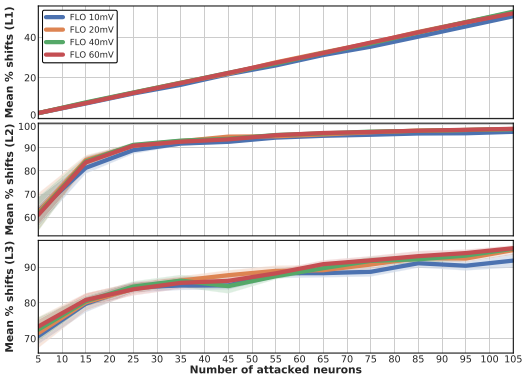


FIGURE 15. Shift percentage mean for each layer of the topology, for the last position of the optimal path.

spontaneous behavior. In particular, we highlight that increasing the number of attacked neurons derives in a higher impact in the mean of spikes. Nevertheless, there are no significant differences in the variation of the voltage used to attack the neurons. Finally, the number of intervening neurons from the visible positions of the optimal path of the maze strongly influences the mean of spikes.

2) PERCENTAGE OF SHIFTS METRIC

For this metric, we first evaluated the percentage of delayed shifts for an aggregation of all three layers. After that, we analyzed the same but combining all the positions of the optimal path of the maze. In this test, we included a different number of attacked neurons and voltage raises. FIGURE 14 describes this situation, where attacking a higher number of neurons produces a higher percentage of shifts. This ascending trend is aligned with the dispersion metric, since an enlargement in the parameters of the attack produces a growth of shifts. As a consequence, it generates a higher dispersion in time and number of spikes.

If we focus on each layer of the topology, FIGURE 15 represents a FLO cyberattack for the last position of the optimal path, where each color line indicates a voltage raise. Focusing on the first layer, we can see a linear growth when we augment the number of attacked neurons since only those neurons shift in the layer. Moving to subsequent layers, we can observe that the growth tendency is more prominent in the second layer. This indicates that, when we advance to the third layer, the effect of the attack gets slightly attenuated.

In conclusion, this metric indicates that attacking more neurons derives in a higher percentage of shifts. Additionally, and similarly to the metric studying the number of spikes, voltage increases have not a high impact on our scenario.

3) DISPERSION METRIC

We first focus on the spike dispersion over time caused by the different number of attacked neurons for each position of the optimal path. This means that, for each position of the maze, we obtain the number of time instants with recorded spikes, independently of the number of spikes. If we take into account that each position of the maze corresponds to one second and that the sampling rate of Brian2, by default, is $0.1ms$, we have a total number of 10 000 instants per position. If a position presents a higher dispersion value than other positions, it indicates that there are more instants with spikes in the former one. We focus on a voltage raise value of $40mV$, since previous analysis indicated that this parameter has a low impact on our scenario.

In FIGURE 16 we can observe that the spontaneous signaling presents some similarities with the trend existing in FIGURE 11 and, specifically, in those positions with the most significant peaks. If a position presents a raise in the number of spikes, the probability of having spikes in FIGURE 16 for a longer period of time also increases. However, the natural dispersion of the simulation attenuates these peaks, where the I parameter changes according to the visible positions of the maze. Considering both FLO configurations, we can appreciate an enlargement in the temporal dispersion compared to the spontaneous behavior. FLO cyberattacks anticipate the spikes of the attacked neurons in a given moment, generating a higher dispersion as the simulation progresses. Specifically, the difference with the spontaneous signaling augments over

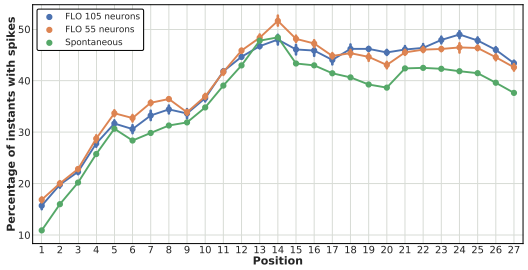


FIGURE 16. Spike dispersion over time for each position of the optimal path.

time, induced by the natural variability of the mouse's movements. Although the attack with 55 neurons presents a higher impact until position 17, from that position until the end, the attack with 105 neurons has a higher impact from this metric. A higher impact over the temporal dispersion when we attack more neurons simultaneously aligns with the results presented in FIGURE 13 for the number of spikes. These results are also related to those presented in Section VI-A2 for the percentage of shifts, where an intensification in these shifts derives in a dispersion growth.

We can also consider this dispersion from the perspective of the number of spikes. For each position of the optimal path, we evaluate the distribution of the number of spikes, setting the voltage increase to a value of $40mV$ and the number of simultaneous attacked neurons to 105. FIGURE 17 illustrates this distribution, where each position contains a violin plot for both the spontaneous and under attack behaviors. It is essential to highlight that this figure represents only one of the *exec* simulations performed for the complete set of experiments to ease the visualization. We can appreciate that the attack in position one reaches a peak of 110 spikes due to the increase of spikes induced by the attack performed at that particular moment. Focusing on the distribution indicated by each violin, the variance progressively reduces when the mouse progresses in the maze, concentrating the distribution of number of spikes around one. That means that in the last positions there are more instants where only one spike occurs, indicating that the attack increases the spike dispersion as the simulation progresses.

This situation aligns with the results presented in FIGURE 11, where a higher number of spikes influence this upper threshold. Nevertheless, it is worth considering the exception in position 13, where this threshold is considerably reduced. To understand this situation, we also have to consider FIGURE 16, which indicates that this position presents the highest percentage of dispersion, with more than 50% of spikes shifted. This position indicates the relationship between these two dispersion approaches, where a high temporal dispersion generates a reduction in the dispersion focused on the number of spikes.

In conclusion, FLO cyberattacks generate a large impact on the spontaneous neuronal activity. In particular, the previous figures highlight how the mouse's natural movement induces particular natural dispersion, both in time and number of spikes. Performing FLO cyberattacks also produces an enlargement in the temporal dispersion, where the neuronal activity is more scattered. This can also be analyzed from the dispersion focused on the number of spikes since this reduction on the aggregation causes the spikes to tend to a low number. It means that there are more instants with a fewer number of spikes compared to the spontaneous behavior.

The previous analysis, based on the number of spikes, percentage of shifts, and dispersion, highlights the impact that FLO cyberattacks can generate over the spontaneous neuronal activity. We subsequently analyze these metrics together since they are strongly dependent between them. In particular, the application of a FLO cyberattack generates a decrease in the number of spikes, where these differences are more prominent in deeper layers of the topology. These results can be explained based on the dispersion induced by the attack, where a growth on the dispersion reduces the probability of multiple action potentials in the first layer. Consequently, the post-synaptic voltage raises arrive at subsequent layers in a more dispersed way, delaying the spikes. The metric focused on the percentage of shifts over the spontaneous signaling is closely related to the dispersion metric. An increase in the percentage of shifts entails a modification in the natural periodicity of the spikes. This change is directly translated to a higher dispersion rate, both in time and number of spikes. Finally, it is essential to note that this behavior and results are dependent on our particular topology. Nevertheless, they can serve as an example of how performing a FLO cyberattack can affect neuronal activity in a particular scenario.

B. NEURONAL SCANNING

This section details the implementation of an SCA cyberattack on our topology, based on the general description of the attack represented by Algorithm 2. For this particular implementation, we have sequentially attacked the 200 neurons that compose the first layer of the topology. We denote as $\mathbb{VI} = \{5, 10, \ldots, 60, 65\}$ the set of voltage raises, in mV, applied separately in each SCA cyberattack. As previously indicated for the FLO cyberattack, the duration of the simulation, t^{sim}, is 27s, staying the mouse one second in each position of the optimal path of the maze. Additionally, the attack initiates in the instant 50ms, represented by t^{attk}. To model the periodicity of attacking the neurons, Δt indicates the temporal separation between two attacks over two consecutive neurons, being 134ms in our particular implementation. Each combination of parameters is executed only once ($exec = 1$) since there is no variability in the selection of neurons, as it is the case of a FLO cyberattack. Finally, Table 6 indicates a summary of the parameters used in the implementation of SCA cyberattacks.

1) NUMBER OF SPIKES METRIC

FIGURE 18 compares the number of spikes per position of the optimal path between the spontaneous neuronal signaling and an SCA cyberattack. In particular, the SCA cyberattack

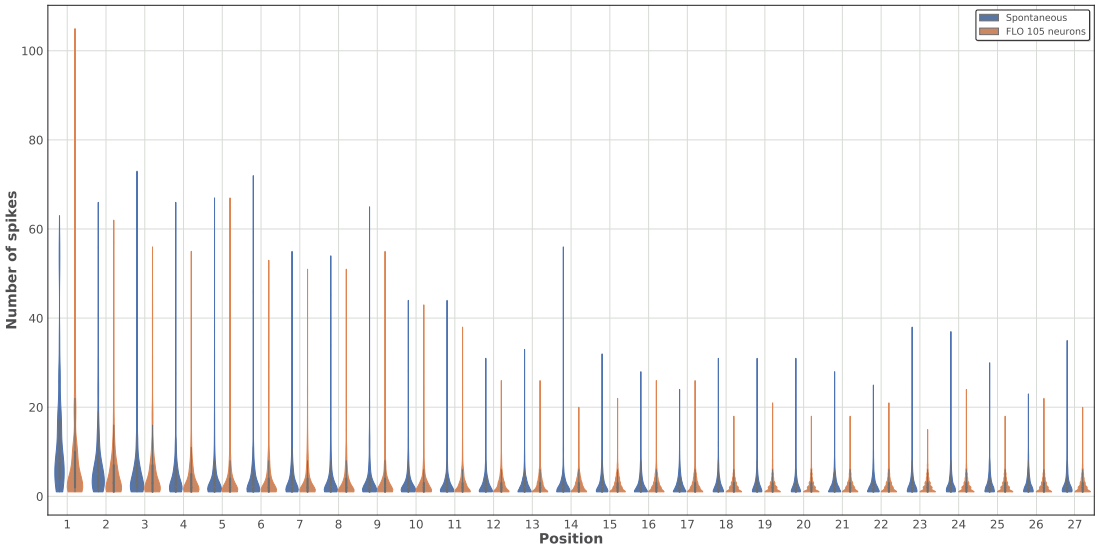


FIGURE 17. Spike dispersion over the number of spikes for each position of the optimal path.

TABLE 6. Configuration of the implemented SCA cyberattack.

Parameter	Description	Values
t^{sim}	Duration of the simulation	27s
t^{attk}	Start of the attack	50ms
t^{step}	Temporal duration between attacking two neurons	134ms
$\mathbb{NE}$	Set containing all neurons of the first layer	$\{1, 2, ..., 199, 200\}$
$\mathbb{VI}$	Set containing the voltage used to attack the neurons	$\{5, 10, ..., 60, 65\}mV$

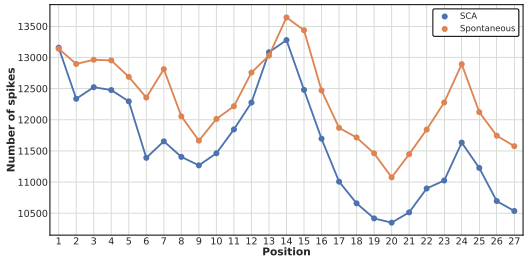


FIGURE 18. Total number of spikes for all neurons of the topology, per position of the optimal path.

establishes a value of $40mV$ from the $\mathbb{VI}$ set and defines an aggregation of all three layers of the neuronal topology. We can appreciate the same trend observed in FIGURE 10 for the intervening neurons from each of the studied positions. The most prominent peaks are, as previously documented for FLO cyberattacks, delayed one position due to the time required to generate an impact over the neurons. These results can be explained based on the sequential behavior of an SCA cyberattack since the number of attacked neurons raises along time. In addition, this progressive reduction in the number of

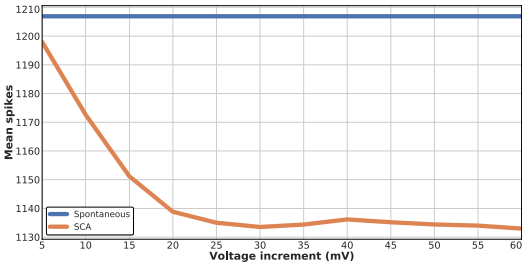


FIGURE 19. Evolution of the spikes mean with different number of attacked neurons and voltage raises, for an aggregation of all positions of the optimal path.

spikes caused by the attack aligns with the results that will be presented in Section VI-B3 for the dispersion metric.

After this analysis, we evaluated in FIGURE 19 the mean of the spikes for the different voltage increases defined in $\mathbb{VI}$, for an aggregation of the three layers of the topology and the positions of the optimal path. We can appreciate that increasing the voltage used to overstimulate the neurons produces a reduction in the number of spikes. It should be noticed that rises higher than $20mV$ do not significantly influence the impact of the attack. Performing an SCA cyberattack with a voltage of $60mV$, the most damaging situation considered, reaches the highest difference in the number of spikes, around 70 spikes compared to the spontaneous behavior.

FIGURE 20 presents a differentiation per layer of the topology for the last position of the optimal path. We can appreciate that, in the first layer, the variation in the number of spikes between different voltage increases is negligible,

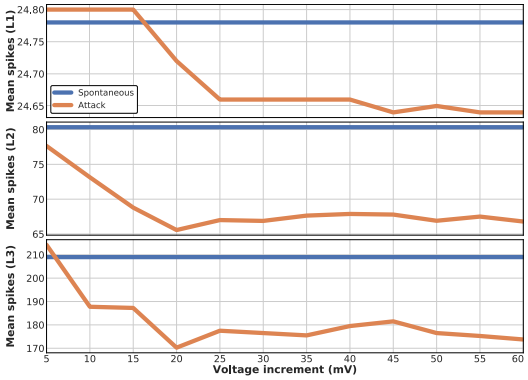


FIGURE 20. Spikes mean for each layer of the topology, focusing on the last position of the optimal path.

being in all cases 24 spikes. Until $15mV$, it presents a small growth of spikes compared to the spontaneous signaling, which benefits of the anticipation of the spikes in time. In more aggressive voltages, the number of spikes gets more reduced than the spontaneous behavior. Moving to the second layer, these differences become more significant, with a number of spikes ranging between 2 and 14 spikes according to the voltage used. This layer presents a general descending trend, reaching the most damaging peak with $20mV$. This trend is common to the third layer, although the range in the number of voltages becomes broader, with a higher difference of 40 spikes compared to the spontaneous signaling. It is interesting to highlight the proliferation of spikes in the third layer when using $5mV$, based on the slight anticipation of spikes in time from the previous layers.

Comparing these results to those presented in FIGURE 19, we can appreciate in the latter specific differences in the evolution of the impact. In this figure, the most damaging voltage is $60mV$, compared to the $20mV$ highlighted for the second and third layers presented in FIGURE 20. This situation is explained by the fact that the analysis focused on differentiating the layers only considers one position and, because of that, some minor differences can arise.

In conclusion, the previous results indicate that performing an SCA cyberattack generates a reduction in the number of spikes, aggravated when the mouse moves across the maze. Increasing the voltage used to overstimulate the neurons does not produce a significant impact with voltages higher than $20mV$. Finally, the number of intervening neurons from each position of the optimal path influences this metric.

2) PERCENTAGE OF SHIFTS METRIC

FIGURE 21 first presents the results concerning the percentage of shifts for different voltage raises. These results represent an aggregation of the three layers and all the positions of the optimal path. In particular, this figure indicates that the percentage of shifts increases when we raise the voltage used to attack the neurons. We can see that an

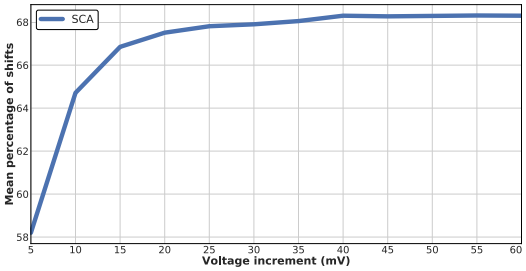


FIGURE 21. Shift percentage mean for an aggregation of all topological layers, aggregating all positions of the optimal path.

overstimulation of $5mV$ generates an approximate 58% of shifts. Slightly increasing this voltage generates considerable impacts, between the range of $5mV$ and $20mV$, reaching a close percentage of 68%. Finally, increasing the stimulation with voltages higher than $20mV$ does not significantly enlarge the percentage of shifts. These thresholds align with those presented in FIGURE 19 for the aggregated number of spikes.

To further explore this metric, we have represented in FIGURE 22, a differentiation of each layer of the topology for just the last position of the optimal path. We can observe that the range of shifts is lower in the first layer compared to deeper layers, based on the influence that the first layer has on the latter due to the transmitted action potentials. Besides, the growth trend existing in the first layer is more prominent, being similar to the one shown in FIGURE 21 for the aggregated analysis of shifts. When we go deeper into the number of layers, we can see that the growth trend is not that aggressive using low voltages, which indicates that the attack progressively loses its effectiveness. It is important to highlight that the ranges shown in FIGURE 21 for the percentage of shifts are much higher than those presented in FIGURE 22. To understand this situation, it is worthy of reflecting on the behavior of SCA cyberattacks. In the first positions of the optimal path, only specific neurons are attacked. When the attack progresses along time, the number of neurons affected by the attack continues increasing. Based on that situation, this last figure focused on the layers presents higher ranges, since they correspond to the last position of the optimal path and, thus, all 200 neurons of the first layer have been affected.

In conclusion, performing an SCA cyberattack generates a raise in the percentage of shifts. This impact becomes more damaging when the mouse moves across the maze since the number of attacked neurons is more abundant along time. Besides, we can observe a degradation of the impact of the attack in deeper layers, where higher voltages are needed to cause a similar impact in terms of shifts.

3) DISPERSION METRIC

Focusing on the temporal dispersion caused by an SCA cyberattack, FIGURE 23 presents its analysis for each position of the optimal path and the aggregation of all the neurons of the

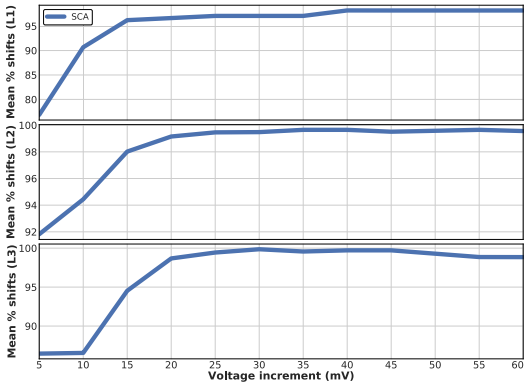


FIGURE 22. Shift percentage mean for each layer of the topology, only focusing on last position of the optimal path.

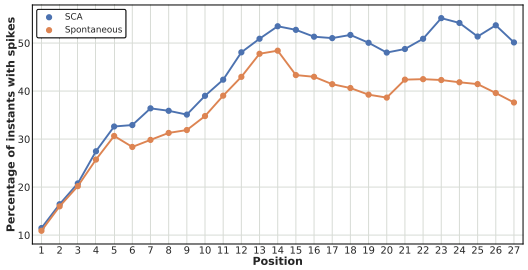


FIGURE 23. Spike dispersion over time for each position of the optimal path.

topology. We can observe that performing an SCA cyberattack progressively augments the temporal dispersion, based on the incremental number of attacked neurons over time. In particular, this dispersion is not significant in the first five positions of the optimal path, due to the number of attacked neurons until that moment and the specific connections of our topology.

After that, we analyze in FIGURE 24 the dispersion from the perspective of the number of spikes. In particular, we represent, for each position of the optimal path, a violin distribution of how the spikes behave. We can observe that, in the first five positions, there are no significant visual differences in the distributions, although the median of the distribution start to slightly decrease. This is justified by the reduced number of neurons affected by the attack until that instant. After that position, the differences with the spontaneous behavior progressively augment, both in the peaks in the number of spikes and the shape of the violins. Focusing on the number of spikes, the maximum number of simultaneous spikes presents a reduction, particularly in the last positions. The shape of the violins progressively changes, due to a reduction in their variance, where the number of spikes concentrates at the value of one only spike. That is to say, the majority of the instants in the last positions had only one spike. These results are aligned to those presented in FIGURE 23 for the analysis of the temporal spike dispersion, since both figures indicate that this dispersion increases when the mouse progresses in the maze.

In summary, this metric indicates that performing an SCA cyberattack disrupts the normal neuronal spiking frequency, inducing dispersion in both temporal and number of spikes dimensions. These differences aggravate when the mouse progresses in the maze, based on the sequential functioning of SCA cyberattacks.

The previous three metrics highlight how SCA cyberattacks can affect the spontaneous neuronal activity on our particular topology. We should consider them as different perspectives to analyze a common issue. As previously indicated, an SCA cyberattack progressively induced a decrease in the number of spikes over time, aggravated in deeper layers of the topology. This decrease is strongly related to both dispersion metrics. The attack generates an alteration in the frequency of spikes in time, producing more instants with spikes in the simulation. Specifically, the previous results indicate that in the last positions of the maze, most of the instant only have one spike, which generates a clear difference with the spontaneous activity. The dispersion metric is strongly related to the percentage of shifts since this dispersion will cause a displacement of the spikes in time. In terms of shifts, the attack gets attenuated in deeper layers.

C. IMPACT COMPARATIVE BETWEEN NEURONAL FLOODING AND SCANNING

This last section compares the results previously discussed for FLO and SCA cyberattacks. Focusing on the total number of spikes (FIGURE 11 and FIGURE 18), we can observe that an SCA cyberattack generates a more impacting reduction in the number of spikes than the most aggressive FLO configuration. The last positions particularly highlight these differences.

When we analyze the number of spikes aggregating both positions and layers (FIGURE 12 and FIGURE 19), we can appreciate one of the main differences between the attacks. In FLO cyberattacks, we can define as parameters of the attack the number of neurons and the voltage used to attack those neurons. In SCA cyberattacks, we can only specify the voltage, since our implementation affects all neurons of the first layer. Based on that, there is not an immediate comparison between these figures in terms of their trend. Nevertheless, we can compare the most aggressive configuration for each attack to determine which produces the highest reduction of spikes. We can see that SCA presents a slightly higher impact than FLO.

Focusing on the distribution of spikes per layer (FIGURE 13 and FIGURE 20), we can observe that there are no significant changes between the attacks. In the second one, SCA presents a slightly lower number of spikes. Finally, the third layer amplifies these differences, where SCA has a more significant reduction of spikes.

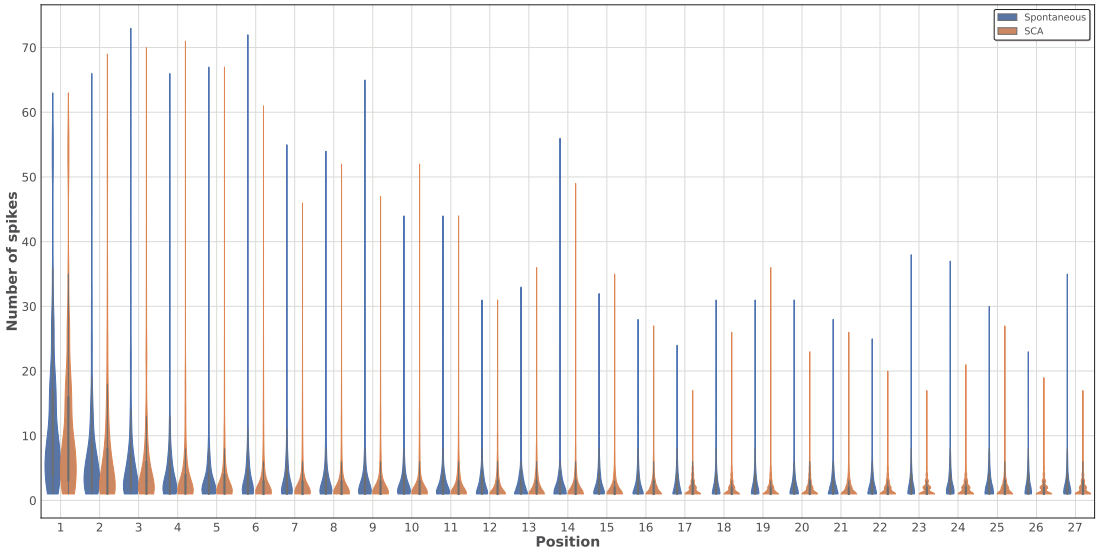


FIGURE 24. Spike dispersion over the number of spikes for each position of the optimal path.

In terms of the percentage of shifts (FIGURE 14 and FIGURE 21), FLO presents a higher impact on this metric. Extending this comparison for each layer of the topology (FIGURE 15 and FIGURE 22), we can see that the main difference lies in the first layer, where SCA duplicates its impact since subsequent layers present similar results. Based on that, we can conclude that FLO presents a higher impact on this metric, although the difference in percentages is slight.

There is a clear difference between both attacks in terms of the temporal dispersion metric (FIGURE 16 and FIGURE 23). FLO has a higher dispersion in the first five positions of the optimal path since the targeted neurons neurons are all attacked in the same instant. After that, SCA evolves in a more damaging way. Focusing on the dispersion based on the number of spikes (FIGURE 17 and FIGURE 24), we can observe that FLO is more effective in the first positions.

This comparative highlights that the inner mechanisms of each attack generates different behaviors in the neuronal activity. FLO is adequate for attacks aiming to disrupt the neuronal activity in a short period of time, affecting multiple neurons in the same instant of time. On the contrary, SCA is a more effective attack for long-term effects, requiring a certain amount of time to reach a significant impact on the neurons. From that threshold, the impact caused on the neurons is more concerning.

VII. CONCLUSION

This work first presents security vulnerabilities of micron-scale BCI to cyberattacks, particularly for implants that can do single-cell or small population sensing and stimulation. Taking these vulnerabilities as a starting point, we describe two novel neural cyberattacks focused on the alteration of neuronal signaling. In particular, we investigated the Neuronal Flooding (FLO) and Neuronal Scanning (SCA), inspired by well-known approaches found in the cybersecurity field. Our investigation is based on a case study of a mouse that learns its navigation within a maze trained by a Convolutional Neural Network (CNN). The CNN was converted into a biological neuronal simulation model representing the workings and functions of real neurons within the brain. The two attacks were applied to the mouse as it migrated through the maze. To evaluate the impact of these attacks on neuronal activity, we proposed three metrics: number of spikes, percentage of shifts, and dispersion of spikes, both over time and number of spikes.

A number of experiments have demonstrated that both attacks can alter the spontaneous neuronal signaling, where the behavior of these attacks generates distinct differences. FLO attacks all targeted neurons in the same instant of time, while SCA presents an incremental behavior, which requires more time to affect the neuronal activity. Focusing on the results, SCA presents a more damaging impact in terms of the number of spikes, which generates a higher reduction than FLO. In terms of shifts, FLO causes more spikes to differ in time than SCA, although these differences are not very significant. Finally, SCA presents a higher impact on the dispersion of the neurons, both in time and number of spikes. These results are highly dependent on the topology used, the neuronal model utilized to represent the neurons, and the types of neurons used (pyramidal from the primary visual cortex). Because of that, this work should be considered as a first step in the study of cyberattacks affecting spontaneous neuronal signaling.

As future work, we plan to define a taxonomy of neuronal cyberattacks affecting not only overstimulation but also neuronal activity inhibition. We aim to explore how neural cyberattacks can affect realistic neuronal tissues and, in particular, various neural circuits within the cortex. Our research lays the groundwork for security countermeasures to also be integrated into BCI systems that utilize miniature implants for small neuronal population stimulation that can have a tremendous effect on the brain.

ACKNOWLEDGMENT

The authors would like to thank Luigi Petrucco, Ethan Tyler, and SciDraw for their publicly-available scientific images [41], [42].

REFERENCES

[1] M. A. Lebedev and M. A. L. Nicolelis, "Brain-machine interfaces: From basic science to neuroprostheses and neurorehabilitation," *Physiol. Rev.*, vol. 97, no. 2, pp. 767–837, Apr. 2017.

[2] L. Yao, X. Sheng, N. Mrachacz-Kersting, X. Zhu, D. Farina, and N. Jiang, "Sensory stimulation training for BCI system based on somatosensory attentional orientation," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 3, pp. 640–646, Mar. 2019.

[3] J. E. O'Doherty, M. A. Lebedev, P. J. Ifft, K. Z. Zhuang, S. Shokur, H. Bleuler, and M. A. L. Nicolelis, "Active tactile exploration enabled by a brain-machine-brain interface," *Nature*, vol. 479, pp. 228–231, 2011.

[4] T. Kaufmann, S. M. Schulz, A. Köblitz, G. Renner, C. Wessig, and A. Käbler, "Face stimuli effectively prevent brain–computer interface inefficiency in patients with neurodegenerative disease," *Clin. Neurophysiol.*, vol. 124, no. 5, pp. 893–900, May 2013.

[5] M. Pais-Vieira, G. Chiuffa, M. Lebedev, A. Yadav, and M. A. L. Nicolelis, "Building an organic computing device with multiple interconnected brains," *Sci. Rep.*, vol. 5, no. 1, Dec. 2015, Art. no. 11869.

[6] D. Sempreboni and L. Viganá, "Privacy, security and trust in the Internet of neurons," 2018, *arXiv:1807.06077*. [Online]. Available: http://arxiv.org/abs/1807.06077

[7] R. A. Ramadan and A. V. Vasilakos, "Brain computer interface: Control signals review," *Neurocomputing*, vol. 223, pp. 26–44, Feb. 2017.

[8] A. Thomson, S. Tielens, T. Schuhmann, T. De Graaf, G. Kenis, B. Rutten, and A. Sack, "The effect of transcranial magnetic stimulation on living human neurons," *Brain Stimulation*, vol. 12, no. 2, p. 524, Mar. 2019.

[9] D. Seo, J. M. Carmena, J. M. Rabaey, E. Alon, and M. M. Maharbiz, "Neural dust: An ultrasonic, low power solution for chronic brain-machine interfaces," 2013, *arXiv:1307.2196*. [Online]. Available: https://arxiv.org/abs/1307.2196

[10] S. A. Wirdatmadja, M. T. Barros, Y. Koucheryavy, J. M. Jornet, and S. Balasubramaniam, "Wireless optogenetic nanonetworks for brain stimulation: Device model and charging protocols," *IEEE Trans. Nanobiosci.*, vol. 16, no. 8, pp. 859–872, Dec. 2017.

[11] E. M. Neuralink. (2019). *An Integrated Brain-Machine Interface Platform With Thousands of Channels*. [Online]. Available: https://www.biorxiv.org/content/early/2019/08/02/703801

[12] P. Ballarin Usieto and J. Minguez. (2018). *Avoiding Brain Hacking—Challenges of Cybersecurity and Privacy in Brain Computer Interfaces*. [Online]. Available: https://www.bitbrain.com/blog/cybersecurity-brain-computer-interface

[13] M. Frank, T. Hwu, S. Jain, R. T. Knight, I. Martinovic, P. Mittal, D. Perito, I. Sluganovic, and D. Song, "Using EEG-based BCI devices to subliminally probe for private information," in *Proc. Workshop Privacy Electron. Soc.*, 2017, pp. 133–136.

[14] I. Martinovic, D. Davies, and M. Frank, "On the feasibility of side-channel attacks with brain-computer interfaces," in *Proc. 21st USENIX Secur. Symp.*, 2012, pp. 143–158.

[15] K. Sundararajan, "Privacy and security issues in brain computer interface," M.S thesis, Auckland Univ. Technol., Auckland, New Zealand, 2017. [Online]. Available: http://orapp.aut.ac.nz/bitstream/handle/10292/11449/SundararajanK.pdf

[16] M. Ienca, "Neuroprivacy, neurosecurity and brain-hacking: Emerging issues in neural engineering," *Bioethica Forum*, vol. 8, no. 2, pp. 51–53, 2015.

[17] M. Ienca and P. Haselager, "Hacking the brain: Brain–computer interfacing technology and the ethics of neurosecurity," *Ethics Inf. Technol.*, vol. 18, no. 2, pp. 117–129, Jun. 2016.

[18] H. Takabi, A. Bhalotiya, and M. Alohaly, "Brain computer interface (BCI) applications: Privacy threats and countermeasures," in *Proc. IEEE 2nd Int. Conf. Collaboration Internet Comput. (CIC)*, Nov. 2016, pp. 102–111.

[19] T. Bonaci, R. Calo, and H. J. Chizeck, "App stores for the brain: Privacy and security in brain-computer interfaces," *IEEE Technol. Soc. Mag.*, vol. 34, no. 2, pp. 32–39, Jun. 2015.

[20] Q. Li, D. Ding, and M. Conti, "Brain-computer interface applications: Security and privacy challenges," in *Proc. IEEE Conf. Commun. Netw. Secur. (CNS)*, Sep. 2015, pp. 663–666.

[21] S. L. Bernal, A. H. Celdrán, G. M. Pérez, M. T. Barros, and S. Balasubramaniam, "Cybersecurity in brain-computer interfaces: State-of-the-art, opportunities, and future challenges," 2019, *arXiv:1908.03536*. [Online]. Available: http://arxiv.org/abs/1908.03536

[22] S. F. Lorenzo, J. A. Benito, P. G. Cardarelli, J. A. Garaia, and S. A. Juaristi, "A comprehensive review of RFID and Bluetooth security: Practical analysis," *Technologies*, vol. 7, no. 1, p. 15, Jan. 2019, doi: 10.3390/technologies7010015.

[23] S. Sevier and A. Tekeoglu, "Analyzing the security of Bluetooth low energy," in *Proc. Int. Conf. Electron., Inf., Commun. (ICEIC)*, Jan. 2019, pp. 1–5.

[24] M. Zubair, D. Unal, A. Al-Ali, and A. Shikfa, "Exploiting Bluetooth vulnerabilities in e-Health IoT devices," in *Proc. 3rd Int. Conf. Future Netw. Distrib. Syst.*, 2019, pp. 1–7, doi: 10.1145/3341325.3342000.

[25] NIST. (2019). *CVE-2019-16336*. [Online]. Available: https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2019-16336

[26] NIST. (2019). *CVE-2019-19192*. [Online]. Available: https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2019-19192

[27] NIST. (2019). *CVE-2019-19194*. [Online]. Available: https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2019-19194

[28] L. Pycroft, S. G. Boccard, S. L. F. Owen, J. F. Stein, J. J. Fitzgerald, A. L. Green, and T. Z. Aziz, "Brainjacking: Implant security issues in invasive neuromodulation," *World Neurosurg.*, vol. 92, pp. 454–462, Aug. 2016.

[29] S. Vadlamani, B. Eksioglu, H. Medal, and A. Nandi, "Jamming attacks on wireless networks: A taxonomic survey," *Int. J. Prod. Econ.*, vol. 172, pp. 76–94, Feb. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S092552731500451X

[30] E. Gal, M. London, A. Globerson, S. Ramaswamy, M. W. Reimann, E. Muller, H. Markram, and I. Segev, "Rich cell-type-specific network topology in neocortical microcircuitry," *Nature Neurosci.*, vol. 20, no. 7, pp. 1004–1013, Jul. 2017, doi: 10.1038/nn.4576.

[31] A. Gáron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Newton, MA, USA: O'Reilly Media, 2019.

[32] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[33] I. Kuzovkin, R. Vicente, M. Petton, J.-P. Lachaux, M. Baciu, P. Kahane, S. Rheims, J. R. Vidal, and J. Aru, "Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex," *Commun. Biol.*, vol. 1, no. 1, p. 107, Aug. 2018, doi: 10.1038/s42003-018-0110-y.

[34] S. Zafrany. (2013). *Deep Reinforcement Learning for Maze Solving*. [Online]. Available: https://www.samyzaf.com/ML/rl/qmaze.html

[35] F. Chollet. (2015). *Keras*. [Online]. Available: https://keras.io

[36] M. Stimberg, R. Brette, and D. F. Goodman, "Brian 2, an intuitive and efficient neural simulator," *eLife*, vol. 8, Aug. 2019, Art. no. e47314.

[37] R. A. Tikidji-Hamburyan, V. Narayana, Z. Bozkus, and T. A. El-Ghazawi, "Software for brain network simulations: A comparative study," *Frontiers Neuroinform.*, vol. 11, p. 46, Jul. 2017. [Online]. Available: https://www.frontiersin.org/article/10.3389/fninf.2017.00046

[38] E. M. Izhikevich, "Simple model of spiking neurons," *IEEE Trans. Neural Netw.*, vol. 14, no. 6, pp. 1569–1572, Nov. 2003.

[39] L. Bachatene, V. Bharmauria, and S. Molotchnikoff, "Adaptation and neuronal network in visual cortex," in *Vision Cortex*, S. Molotchnikoff and J. Rouat, Eds. Rijeka, Croatia: IntechOpen, 2012, ch. 15, doi: 10.5772/46011.

[40] A. M. Norcia, L. G. Appelbaum, J. M. Ales, B. R. Cottereau, and B. Rossion, "The steady-state visual evoked potential in vision research: A review," *J. Vis.*, vol. 15, no. 6, p. 4, May 2015, doi: 10.1167/15.6.4.
[41] L. Petrucco. *Mouse Head Schema*. Accessed: Jul. 17, 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3925903
[42] E. Tyler. *Mouse*. Accessed: Jul. 17, 2020. [Online]. Available: https://zenodo.org/record/3925901

SERGIO LÓPEZ BERNAL received the B.Sc. and M.Sc. degrees in computer engineering from the University of Murcia, and the M.Sc. degree in architecture and engineering for the IoT from IMT Atlantique, France. He is currently pursuing the Ph.D. degree with the University of Murcia. His research interests include ICT security on brain–computer interfaces and network and information security.

ALBERTO HUERTAS CELDRÁN received the M.Sc. and Ph.D. degrees in computer science from the University of Murcia, Spain. He is currently an Irish Research Council Government of Ireland Postdoctoral Research Fellow Associated with the TSSG, Waterford Institute of Technology, Ireland. His scientific interests include cybersecurity, privacy, brain-computer interfaces (BCI), continuous authentication, and computer networks.

LORENZO FERNÁNDEZ MAIMÓ received the M.Sc. and Ph.D. degrees in computer science from the University of Murcia, Spain. He is currently an Associate Professor with the Department of Computer Engineering, University of Murcia. His research interests include machine learning and deep learning applied to cybersecurity and computer vision.

MICHAEL TAYNNAN BARROS (Member, IEEE) received the B.Tech. degree in telematics from the Federal Institute of Education, Science, and Technology of Paraiba, in 2011, the M.Sc. degree in computer science from the Federal University of Campina Grande, in 2012, and the Ph.D. degree in telecommunication software from WIT, in 2016. He is currently a Lecturer with the School of Computer Science and Electronic Engineering, University of Essex, U.K. He was a recipient of the Marie Skłodowska Curie Individual Fellowship (MSCA-IF).

SASITHARAN BALASUBRAMANIAM (Senior Member, IEE) received the B.E. degree in electrical and electronic engineering from the University of Queensland, Brisbane, QLD, Australia, in 1998, the M.E. degree in computer and communication engineering from the Queensland University of Technology, Brisbane, in 1999, and the Ph.D. degree from the University of Queensland, in 2005. He is currently the Director of Research with the TSSG, Waterford Institute of Technology, Ireland. His current research interests include molecular communications, the Internet of (Bio) NanoThings, and terahertz wireless communications.

GREGORIO MARTÍNEZ PÉREZ (Member, IEEE) received the M.Sc. and Ph.D. degrees in computer science from the University of Murcia, Spain. He is currently a Full Professor with the Department of Information and Communications Engineering, University of Murcia. His research interests include cybersecurity, privacy, and networking, working on different national and European IST research projects on these topics.

3

Neuronal Jamming Cyberattack

Title:	Neuronal Jamming cyberattack over invasive BCIs affecting the resolution of tasks requiring visual capabilities
Authors:	Sergio López Bernal, Alberto Huertas Celdrán, Gregorio Martínez Pérez.
Journal:	Computers & Security
JIF:	5.105 Q2 (2021)
Publisher:	Elsevier
Volume:	112
Number:	
Pages:	102534
Year:	2022
Month:	Jan
DOI:	10.1016/j.cose.2021.102534
Status:	Published

Abstract

Invasive Brain-Computer Interfaces (BCIs) are extensively used in medical application scenarios to record, stimulate, or inhibit neural activity with different purposes. An example is the stimulation of some brain areas to reduce the effects generated by Parkinson's disease. Despite the advances in recent years, cybersecurity on BCIs is an open challenge since attackers can exploit the vulnerabilities of invasive BCIs to induce malicious stimulation or treatment disruption, affecting neuronal activity. In this work, we design and implement a novel neuronal cyberattack called Neuronal Jamming (JAM), which prevents neurons from producing spikes. To implement and measure the JAM impact, and due to the lack of realistic neuronal topologies in mammalians, we have defined a use case using a Convolutional Neural Network (CNN) trained to allow a simulated mouse to exit a particular maze. The resulting model has been translated to a biological neural topology, simulating a portion of a mouse's visual cortex. The impact of JAM on both biological and artificial networks is measured, analyzing how the attacks can both disrupt the spontaneous neural signaling and the mouse's capacity to exit the maze. Besides, another contribution of the work focuses on comparing the impacts of both JAM and FLO (an existing neural cyberattack), demonstrating that JAM generates a higher impact in terms of neuronal spike rate. As a final contribution, we discuss whether and how JAM and FLO attacks could induce the

effects of neurodegenerative diseases if the implanted BCI had a comprehensive electrode coverage of the targeted brain regions.

Keywords

Cybersecurity · Safety · Neuronal cyberattacks · Convolutional neural networks · Brain-computer interfaces

COMPUTERS & SECURITY 112 (2022) 102534

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

Neuronal Jamming cyberattack over invasive BCIs affecting the resolution of tasks requiring visual capabilities

Sergio López Bernal[a,*], *Alberto Huertas Celdrán*[b], *Gregorio Martínez Pérez*[a]

[a] *Department of Information and Communications Engineering, University of Murcia, Murcia 30100, Spain*
[b] *Communication Systems Group CSG, Department of Informatics IfI, University of Zurich UZH, CH—8050 Zürich, Switzerland*

ARTICLE INFO

Article history:
Received 19 January 2021
Revised 19 October 2021
Accepted 1 November 2021
Available online 8 November 2021

Keywords:
Cybersecurity
Safety
Neuronal cyberattacks
Convolutional neural networks
Brain-Computer Interfaces

ABSTRACT

Invasive Brain-Computer Interfaces (BCIs) are extensively used in medical application scenarios to record, stimulate, or inhibit neural activity with different purposes. An example is the stimulation of some brain areas to reduce the effects generated by Parkinson's disease. Despite the advances in recent years, cybersecurity on BCIs is an open challenge since attackers can exploit the vulnerabilities of invasive BCIs to induce malicious stimulation or treatment disruption, affecting neuronal activity. In this work, we design and implement a novel neuronal cyberattack called Neuronal Jamming (JAM), which prevents neurons from producing spikes. To implement and measure the JAM impact, and due to the lack of realistic neuronal topologies in mammalians, we have defined a use case using a Convolutional Neural Network (CNN) trained to allow a simulated mouse to exit a particular maze. The resulting model has been translated to a biological neural topology, simulating a portion of a mouse's visual cortex. The impact of JAM on both biological and artificial networks is measured, analyzing how the attacks can both disrupt the spontaneous neural signaling and the mouse's capacity to exit the maze. Besides, another contribution of the work focuses on comparing the impacts of both JAM and FLO (an existing neural cyberattack), demonstrating that JAM generates a higher impact in terms of neuronal spike rate. As a final contribution, we discuss whether and how JAM and FLO attacks could induce the effects of neurodegenerative diseases if the implanted BCI had a comprehensive electrode coverage of the targeted brain regions.

1. Introduction

Brain-Computer Interfaces (BCIs) are devices providing bidirectional communication channels between the brain and external devices. One of the primary uses of BCI technologies is in health scenarios, where clinicians acquire relevant information about the brain for diagnosis purposes (Lebedev and Nicolelis, 2017). Additionally, BCI systems enable artificial stimulation and inhibition of neuronal activity (Yao et al., 2019). In particular, neurostimulation has been used in a wide variety of medical scenarios, ranging from treating neurodegenerative diseases, such as Parkinson's or depression (Hartmann et al., 2019), to provide prosthetic users with feedback (O'Doherty et al., 2011). Within these systems,

* Corresponding author.
E-mail address: slopez@um.es (S.López Bernal).
https://doi.org/10.1016/j.cose.2021.102534

there are two main categories based on their invasiveness. Non-invasive BCIs can externally stimulate the brain without surgery and, although some technologies can target small areas of the brain, non-invasive BCIs cover larger regions of the brain. In contrast, invasive systems can be applied to small areas, even with a single-neuron resolution, but introducing higher physiological risks (Ramadan and Vasilakos, 2017).

Based on the relevance and expansion of BCIs, new technologies and companies have emerged in recent years, focusing on developing new invasive systems to stimulate the brain with neuronal granularity. This is the case of Neuralink (Musk and Neuralink, 2019), a company that has designed disruptive BCI systems to record data at the neuronal level, and it is currently working on covering the stimulation functionality. Besides, Neural Dust (Seo et al., 2013) is an architecture of millions of nanoscale implantable devices located in the cortex that allow neural recording. Evolution of Neural Dust is the Wireless Optogenetic Nanonetworking device (WiOptND) (Wirdatmadja et al., 2017), which uses optogenetics to stimulate the neurons. Although these approaches are promising, the authors of Bernal et al. (2020) have shown that they have vulnerabilities that could allow attackers to control both systems and perform malicious stimulation actions, altering spontaneous neuronal signaling. Depending on the coverage of the attack, in terms of brain regions and number of neurons affected, cyberattackers could inflict permanent brain damage or even cause the death of the patients.

In the same direction, Bernal et al. (2021) identified that the field of cybersecurity in BCI is not mature enough, and non-sophisticated attacks can generate significant damage. In summary, the BCI vulnerabilities could be exploited by attackers to take advantage of these promising neurostimulation technologies. Taking the findings of these works as motivation, this manuscript focuses on the scarce research dealing with cyberattacks aiming to alter neuronal behavior. Additionally, new ways to measure and understand the impact of these attacks are also required. In particular, these issues gain special relevance due to the possibility of attacks being able to worsen or recreate the effects of common neurodegenerative diseases (Bernal et al., 2021).

Intending to improve the previous challenges, the main contribution of this work is the definition and implementation of a novel neuronal cyberattack, *Neuronal Jamming cyberattacks* (JAM), focused on the inhibition of neural activity. The present work aims to explore the impact that inhibitory neuronal cyberattacks can generate on the brain. Nevertheless, there is an absence in the literature of comprehensive neuronal topologies, and therefore, we simulate a portion of the visual cortex of mice, placed in the occipital region of the brain, defining a use case of a mouse trying to exit a given maze. The neuronal topology was built by using a Convolutional Neural Network (CNN) (Géron, 2019) trained to solve this particular use case. The second contribution of this work is the evaluation of the impact caused by JAM cyberattacks over both neuronal and artificial simulation in this specific scenario. To perform the analysis, we have used existing metrics but also defined a subset of new ones, concluding that JAM cyberattacks can alter spontaneous neuronal behavior and force the mouse to perform erratic decisions to escape the maze.

The third main contribution of this work is to compare the impact caused by JAM with an existing cyberattack named Neuronal Flooding (FLO) from the biological and artificial perspectives. We have observed that applying a FLO cyberattack over the last positions of the maze generates a reduction of its effectiveness from both biological and artificial approaches. Additionally, JAM cyberattacks are more damaging when increasing the number of consecutive positions under attack, translated into a reduction in the neural activity and an augmentation in the number of steps to find the exit. The fourth contribution is a comparison between biological and artificial scenarios based on linear correlation analysis between variables. In this sense, FLO presents a high Pearson correlation between experiments, of around 0.8, indicating a strong relationship. On its side, JAM presents worse results, which can be explained due to the particular restrictions during the implementation. Finally, we discuss the relationship that recent neuronal cyberattacks could have with neurodegenerative diseases.

These contributions aim to advance the current state of the art, which is limited to the references presented in this section. Compared to Bernal et al. (2020), which only characterized and measured the impact of two neural cyberattacks (Neural Flooding and Neural Scanning), this work further explores the impact of neural cyberattacks, presenting, for the first time, a comparison between the impact on neuronal and behavioral dimensions.

The remainder of the paper is structured as follows. Section 2 reviews the state of the art in cybersecurity oriented to BCI and neuronal cyberattacks. After that, Section 3 introduces the definition of the Neuronal Jamming cyberattack. Section 4 presents the experimental setup required to implement both JAM and FLO neuronal cyberattacks. Additionally, Section 5 and Section 6 describe, respectively, the results obtained after implementing JAM and FLO cyberattacks over multiple positions of the maze and the impact they cause. These two sections also include a comparison of the relationship between artificial and biological approaches. Subsequently, Section 7 discusses the impact that neuronal cyberattacks can have on neurodegenerative diseases. Finally, Section 8 presents conclusions and future work.

2. Related work

Cybersecurity applied to BCI is relatively recent, emerging in the last five years concepts such as brain-hacking or neurosecurity (Ienca, 2015; Ienca and Haselager, 2016). These publications identify that neurostimulation BCI devices present a high risk in patients' safety since an attacker could disrupt the treatment parameters. Additionally, they highlighted that attacks do not need to be complex to cause brain damage.

During these recent years, the academic literature has widely focused on the study of cybersecurity in health scenarios, aiming to preserve patients' privacy or improving the security of clinical devices (Huertas Celdrán et al., 2017; Huertas Celdrán et al., 2018). However, the literature has focused on particular cybersecurity aspects of BCI, mostly from theoretical and ethical perspectives. Although previous studies have highlighted the applicability of cryptographic and jam-

ming attacks (Ienca and Haselager, 2016), malware strategies (Bonaci et al., 2015), acquisition of sensitive data from neural signals (Quiles Pérez et al., 2021), disruption of neural signals (Martínez Beltrán et al., 2021), or potential attacks over BCI architectures (Ballarin Usieto and Minguez, 2018), these works are scarce and focus on particular privacy and security aspects, not addressing the physical safety dimension. Additionally, the authors of Takabi et al. (2016), Bonaci et al. (2015) identified that the platforms and frameworks used to develop BCI applications could be vulnerable to cyberattacks. Based on that, the authors of Bernal et al. (2021) performed a review of the state of the art in cybersecurity on BCI with a comprehensive analysis of physical safety issues, compiling already documented attacks over the BCI life-cycle, their impacts, and the countermeasures to detect and mitigate them. This work also studied the literature concerning attacks, impacts, and countermeasures from existing and prospecting architectural BCI deployments. Furthermore, they proposed applying well-known attacks, impacts, and countermeasures from the cybersecurity domain to BCI. In a nutshell, they identified an enormous absence of works addressing cybersecurity aspects in BCI technologies.

Regarding cyberattacks altering the behavior of neurons, the authors of Bernal et al. (2020) detected vulnerabilities in emerging neurostimulation technologies. They defined two *neuronal cyberattacks*, Neuronal Flooding (FLO) and Neuronal Scanning (SCA), aiming to disrupt the spontaneous behavior of the targeted zones of the brain. The FLO cyberattack consists in attacking, in a particular instant, a subset of neurons from the brain, while SCA targets one neuron per time instant, imitating the port scanning technique. They also defined several metrics to measure the impact of these attacks compared to spontaneous neuronal activity. In short, they identified that both neuronal cyberattacks induced a considerable alteration in the spontaneous neural signaling.

The neuronal cyberattacks presented in Bernal et al. (2020) demonstrate the feasibility of performing attacks over the brain aiming to disrupt its spontaneous neural activity. However, they do not explore the physiological or psychological consequences that an alteration in neural signaling can generate. In that direction, the authors of Bernal et al. (2021) theoretically proposed recreating the effect of neurodegenerative disorders such as Parkinson's and Alzheimer's diseases. For that, the neurostimulation system would be required to cover the brain regions naturally impacted by these diseases and present vulnerabilities that attackers can exploit. This work highlighted the high impact that recreating neurodegenerative disorders could have on users' physical safety.

To understand how cyberattacks could affect the brain and its relationship with degenerative diseases, it is essential to mention that, from a neurological point of view, most brain disorders are revealed as a dysfunction of communication between neurons or with other organs defining the term of *brain connectivity disorders*. Within this term, we can include neurodegenerative diseases. Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that induces the degradation and death of brain cells. It seems that neurodegenerative diseases spread along structurally connected neural networks, known as *neuronal circuits*, presenting a functional relevance. There is a relationship between AD and changes in neuronal activity in the Default Mode Network circuit (DMN), where parts of the DMN present increased connectivity at the beginning of the disease, indicating compensation for the failure of other regions of the circuit before they degenerate. During the progression of AD, the deactivation of the DMN is gradually more pronounced. Nevertheless, it is not clear if the circuit disruption is a cause or a consequence of the disease (Zott et al., 2018).

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease affecting cortical and spinal neurons, which generates a loss of muscle control and paralysis. ALS is associated with a dysfunction of cortical circuits based on hyperexcitability of neuronal activity. Hyperexcitability can be understood as an exaggerated response to a stimulus, or the response to stimuli that generally do not induce a response. In this sense, ALS presents a perturbation in the excitatory/inhibitory balance, leading to pathological changes in cortical excitability (Brunet et al., 2020).

Despite the current knowledge about the behavior of neurodegenerative diseases, such as AD or ALS, there are no proposed cyberattacks in the literature trying to emulate the neuronal behavior of these conditions. Because of that, the current manuscript explores the possibility of inducing excitatory and inhibitory neuronal behavior to lay the foundation for future research aiming to recreate these conditions in the long term.

3. Neuronal Jamming cyberattack

This section presents the formal definition of the Neuronal Jamming cyberattack (JAM), including algorithmic and graphical representations to ease its understanding.

Jamming is a well-known cyberattack aiming to block the legitimate communication between elements of a system using malicious interference, resulting in the generation of a Denial of Service (DoS) over the communication. From a neurological perspective, we conceive a jamming cyberattack as an inhibition of the spontaneous activity of a set of neurons during a particular duration of time, preventing their interaction with other neurons. This attack does not need previous knowledge by the attacker about the status of the targeted neurons, presenting a low complexity compared to those that could require to study their previous and current status to determine the best instant to attack.

To formalize this attack, we denote $\mathbb{NE} \subset \mathbb{N}$ as a subset of neurons from the brain, where $n \in \mathbb{NE}$ expresses every single neuron. $\mathbf{t}^{\text{attk}}$ is the time instant when the cyberattack starts, and $\mathbf{t}^{\text{pulse}}$ is the duration of the attack. During that particular period, a subset of neurons $\mathbb{AN} \subseteq \mathbb{NE}$ is attacked. The voltage of a single neuron in a specific instant of time is denoted as $v_n \in \mathbb{R}$, whereas $v_{min} \in \mathbb{R}$ indicates the minimum value of the voltage that the neuron can have, directly dependent on the neuronal model used in case of simulations. Moreover, $\mathbf{t}^{\text{win}}$ is the temporal window in which the cyberattack is evaluated, which corresponds to the duration of the simulation presented in subsequent sections. Δt is the amount of time between evaluations during the process, representing the duration of steps of the simulation in the implementation of the cyberattacks.

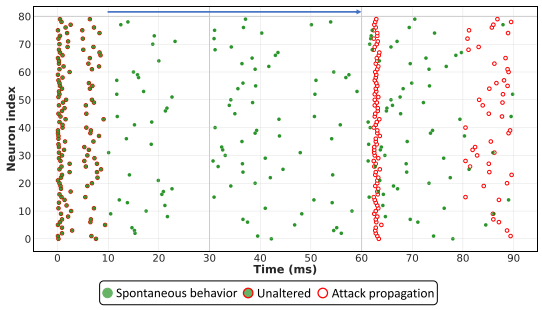


Fig. 1 – Raster plot of a JAM cyberattack when the attack is performed between the instants 10ms and 60ms. This temporal window is represented by a blue arrow for clarity.

Algorithm 1 JAM cyberattack execution.

$t = 0$
while $t < t^{win}$ **do**
  if $t >= t^{attk}$ AND $t < (t^{attk} + t^{pulse})$ **then**
    for all $n \in \mathbb{AN}$ **do**
      $v_n \leftarrow v_{min}$
    end for
  end if
  $t \leftarrow t + \Delta t$
end while

As shown in Algorithm 1, JAM cyberattacks are performed during a continuous duration of time, where the attacked neurons are forced to have their minimum voltage value. In other words, it avoids the targeted neurons to produce spikes, understood as the inhibition of the neurons.

To visually understand the behavior of a JAM cyberattack, Fig. 1 presents the comparison between a JAM cyberattack and the spontaneous neuronal behavior for a simulation of 90ms. Until the instant 10 ms, green dots with a red outline can be appreciated, indicating that the attack has not altered those spikes. This attack, performed between the instants 10 ms and 60 ms, and indicated by a blue arrow, affects all 80 neurons represented in the figure. Because of that, during that temporal window, only green dots are presented, having an absence of neural activity during the application of the attack. After the instant 60ms, white dots with red online appear, indicating the new spikes generated as a consequence of the attack. It is relevant to note that, from that moment until the instant 90ms, the neural signaling generated by the attack is completely different from the spontaneous behavior.

4. Experimental setup

Due to the lack of realistic and precise neuronal topologies in the literature, this section presents the methodology followed to create a neuronal topology used to evaluate the impact of JAM cyberattacks. For simplicity, we have summarized the explanations of this section, where a broader description is available at Bernal et al. (2020).

Nevertheless, it is relevant to indicate that the feasibility of neural cyberattacks was documented in Bernal et al. (2020), where we identified that novel neurostimulation technologies offering recording and neurostimulation capabilities with a single-neuron resolution, such as Neuralink, presented vulnerabilities that cyberattackers could exploit to gain access to the devices and, thus, disrupt the behavior of the brain. This work highlighted the sensitivity of using wireless communications, such as Bluetooth, between the implants and external devices controlling the implant. Thus, attackers could determine the instant (or instants) of attack, the list of targeted neurons, and the voltage used to affect the neurons.

It is essential to highlight that the knowledge of precise neocortical synaptic connections in mammalian is nowadays an open challenge Gal et al. (2017). Although artificial and biological networks cannot be comparable in complexity and functioning, there are works in the literature demonstrating that neurons in the visual cortex present certain similarities with a Convolutional Neural Network (CNN). In this sense, the visual recognition process operates incrementally in both networks, moving from simple to abstract (Kuzovkin et al., 2018). Based of that, we have trained a CNN using Keras on top of TensorFlow (Chollet et al., 2015) to solve a simplistic scenario based on a mouse trying to escape a maze from any position, inspired in the code from Zafrany (0000). The maze has a size of 7x7 positions with fixed obstacles that serve as walls, containing a single starting cell and an exit. Fig. 2 presents the maze, indicating with numbers the optimal path to the exit, which has been determined during the training process of the CNN. It is essential to note that this process does not involve any real mouse since all this testing is based on simulations.

The CNN has been trained employing reinforcement learning (Sutton and Barto, 2018), using a topology consisting in three layers where the first two were convolutional layers, and the third one was dense. After the training process, a topology of interconnected nodes between layers was obtained, where each link had associated a filter weight. These weights represent the relevance that this connection has in the topology to solve the problem. Table 1 summarizes the configuration used to define the CNN, composed of a total number of 276 nodes.

The resulting topology was translated to a biological neuronal network by keeping the exact number of layers and

Table 1 – Summary of the layers of the CNN.

Layer	Type	Filters	Input size	Output size	Kernel size	Stride	Activation function	Nodes
1	Conv2D	8	7×7×1	5×5×8	3×3	1	ReLU	200
2	Conv2D	8	5×5×8	3×3×8	3×3	1	ReLU	72
3	Dense	-	3×3×8	4	-	-	ReLU	4

Table 2 – Parameters used in the Izhikevich model.

Parameter	Description	Values
v	Membrane potential of a neuron	[-65, 30] mV
u	Membrane recovery variable providing negative feedback to v	(-16, 2) mV/ms
a	Time scale of u	0.02/ms
b	Sensitivity of u to the sub-threshold fluctuations of v	0.2/ms
c	After-spike reset value of v	-65mV
d	After-spike reset value of u	8mV/ms
I	Injected synaptic currents	{10, 15} mV/ms

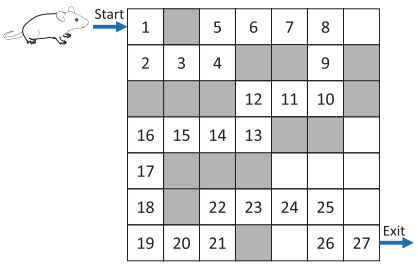


Fig. 2 – Maze used to model the movement of the mouse, including the optimal path between the starting and final cells.

nodes per layer and translating the filter weights to synaptic weights. These synaptic weights represent the influence that the firing of one neuron has on another neuron within a neuronal synapse. Particularly, this topology represents a small section of the visual cortex of a mouse, located in the occipital brain area. Once having the biological topology, we have used the Brian2 neural simulator (Stimberg et al., 2019) to represent the behavior of each individual neuron. In particular, we have implemented the Izhikevich neuronal model (Izhikevich, 2003), whose parameters are presented in Table 2, and Eqs. (1)–(3). It is relevant to highlight the functioning of the I parameter used in the experiments to model the visual stimuli received by the mouse in terms of free cells and walls in the biological simulation. To enclose the problem, we implemented and monitored a neuronal simulation with a total duration of 27 s, where the mouse stayed in one position of the optimal path for one second, and studied its spontaneous behavior and the behavior under attack. When the mouse is in a particular position, the *intervening neurons* associated with each *adjacent position* from the current cell were obtained. The concept of *intervening neurons* can be understood as the set of neurons influenced by the list of adjacent positions from the current cell. For those intervening neurons, the simulation assigns a value of 15mV/ms for the I parameter, keeping a value of 10mV/ms for the rest of the neurons. These particular implementation aspects are presented in-depth in Bernal et al. (2020).

Table 3 – Parameters used in the analysis for JAM cyberattacks.

Parameter	Values
Number of consecutive attacked positions (Bio, CNN)	{1, 2, ..., 27}
Number of neurons/nodes (Bio, CNN)	{5, 35, 55, 75, 105}
Voltage under attack (Bio)	-65 mV
Output importance (CNN)	-1
Number of executions (Bio, CNN)	10

$$v' = 0.04v^2 + 5v + 140 + u + I \tag{1}$$

$$u' = a(bv - u) \tag{2}$$

$$if\ v \geqslant 30mV, \quad then \begin{cases} v \leftarrow c \\ u \leftarrow u + d \end{cases} \tag{3}$$

5. Impact of JAM attacks over biological and artificial neural networks

Once explained the generation of the artificial and biological networks, this section measures and compares the impact generated by Neuronal Jamming cyberattacks (JAM) over biological and artificial networks. In particular, this analysis aims to study if an alteration in neuronal behavior can also impact the mouse's ability to solve the maze based on the evaluation of the CNN model.

Table 3 presents the parameters used to perform the experiments, indicating between parentheses if a parameter is

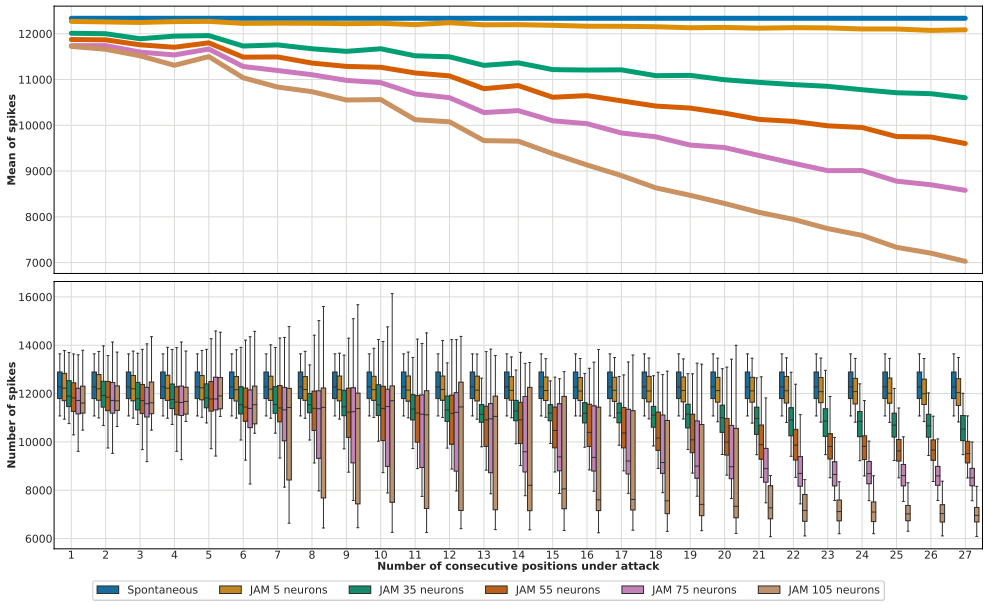


Fig. 3 – Distribution of the number of spikes based on the consecutive number of positions attacked for JAM cyberattacks.

common to both scenarios or specific to one of them. As can be seen, five number of simultaneously attacked neurons (named as nodes in the CNN) have been tested, probing several consecutively attacked positions ranging from one to all the positions of the optimal path of the maze. Additionally, each combination of parameters is executed ten times, where each execution targets a different set of randomly selected neurons. The meaning of these parameters will be presented throughout this section.

5.1. *JAM cyberattacks over the biological network*

Focusing on the biological perspective, attacked neurons are forced to the minimum voltage value of the model, which corresponds to -65 mV, as indicated in Table 3. Fig. 3 presents the experiment consisting in augmenting the number of consecutive positions of the optimal path under attack, always initiating the attack in the first position, and evaluating different numbers of simultaneously attacked neurons. The variability shown corresponds to the ten executions performed per combination of parameters. In particular, this figure highlights how augmenting the number of consecutive positions of the labyrinth under attack impacts in terms of the number of spikes metric. The upper sub-figure depicts that increasing the number of simultaneously attacked neurons considerably reduces the mean of spikes, reaching a difference of 5000 spikes in the most damaging situation compared to spontaneous behavior. The bottom sub-figure shows that the distribution of the number of spikes presents small variability during the first six positions. More consecutive positions under attack generate a progressive reduction in the dispersion, particularly for higher numbers of attacked neurons, indicating that JAM cyberattacks cause an enormous impact on the spike metric. Nevertheless, increasing the number of consecutive positions over more than 20 generates a progressive reduction in the distributions when attacking more than 75 neurons. This situation is explained by many neurons without activity during most of the simulation, decreasing their variability in the number of spikes.

Moving to the temporal dispersion of spikes, Fig. 4 depicts that attacking a higher number of neurons reduces the temporal dispersion. It is relevant to highlight that targeting a reduced number of neurons (up to 35) produces a slightly higher dispersion than the spontaneous behavior, where these peaks can be produced by the slight variations generated by the attack. Nevertheless, increasing the number of selected neurons gets a substantial reduction. In particular, attacking 105 neurons achieves the most damaging configuration, causing a reduction from 36% of instants with spikes to an approximate 28%. It is also important to note that, in the bottom subfigure, the distribution of targeting 105 neurons significantly decreases compared to other numbers of attacked neurons, indicating the importance of this parameter of the attack.

5.2. *JAM cyberattacks over the artificial network*

In the artificial scenario, the attack consists in modifying the targeted nodes of the trained model, affecting their normal functioning. For that, the concept of *output importance* refers to the value used to alter the output of the nodes targeted by the attack, thus affecting their relevance in the network. In JAM, the value used to attack the nodes is -1, which indicates

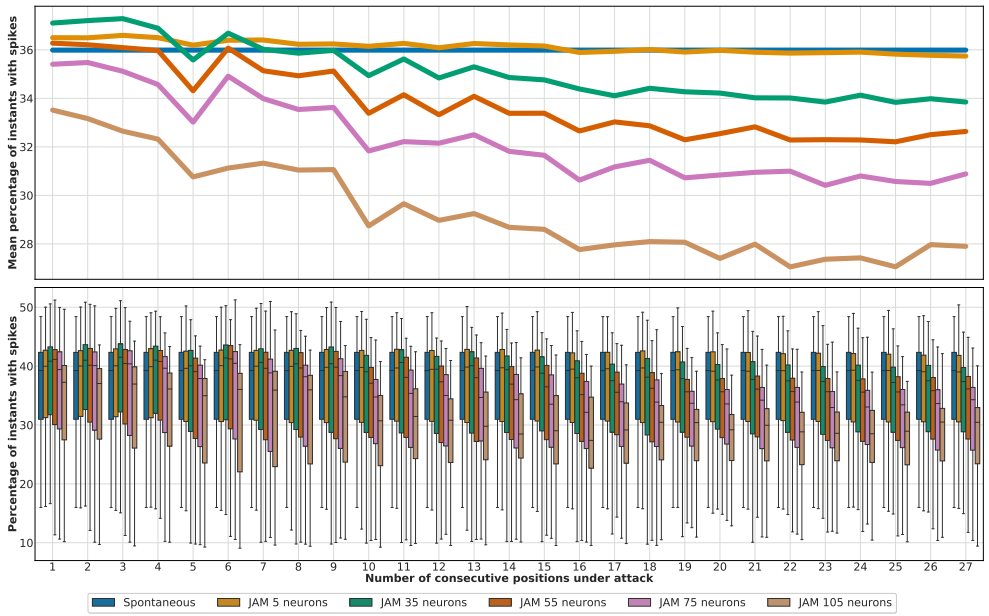


Fig. 4 – Distribution of the temporal dispersion based on the consecutive number of positions attacked for JAM cyberattacks.

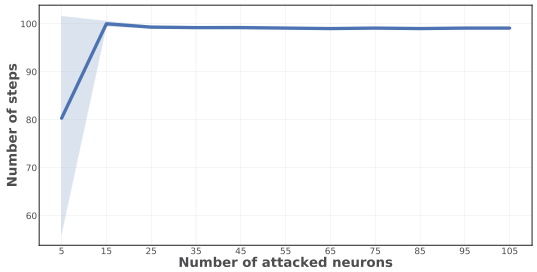


Fig. 5 – Number of steps for different number of neurons between five and 105, with ten executions, for JAM cyberattacks.

that those nodes do not have any relevance in the network, representing their inhibition. This forces the network to find alternative paths to solve the problem, deactivating the paths from the affected nodes to later layers.

The first approach followed was to apply the attacked model for the targeted consecutive positions, restoring it to the non-altered model after the duration of the attack. Although the mouse performed erratic decisions across the maze during the attack, once the model without attacks was restored, the mouse could always find the exit position ultimately. To better measure the impact of this attack in terms of percentage of success and number of steps, we decided to continuously perform the attack for all 27 positions of the maze. These experimentation results are represented in Fig. 5, which indicates that simultaneously attacking more than 15 nodes does not generate any difference since the number of steps gets stabilized in around 100 steps. It is worthy to note that the success percentage is not studied as both variables are highly correlated, with a -0.99 Pearson correlation.

Based on the decision to attack during the whole simulation (27 positions), and compare these results with the biological simulation, we decided to focus the analysis of both scenarios on a number of attacked neurons between one and 20. From the CNN point of view, this decision is motivated by Fig. 6, which indicates that this particular range reflects variations in the number of steps and that further increments in this variable do not offer new variability.

After defining the range, the biological experiments were adapted to be comparable with those from the CNN scenario. For that, a number of attacked neurons between one and 20 were selected, setting the attack to cover all 27 consecutive positions of the optimal path of the maze, starting in the in-

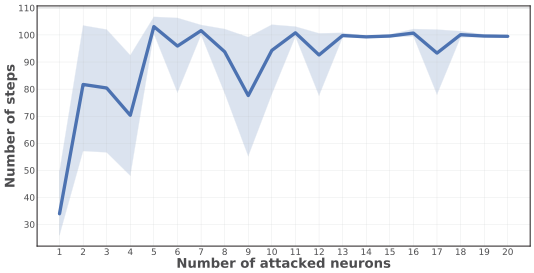


Fig. 6 – Number of steps for a range between one and 20 attacked neurons, with ten executions, for JAM cyberattacks.

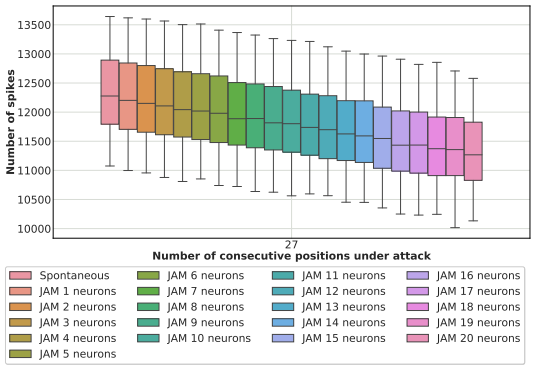


Fig. 7 – Number of spikes for a range of attacked neurons between one and 20 for JAM cyberattacks.

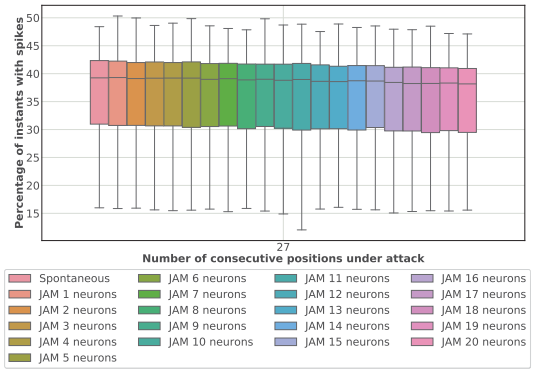


Fig. 8 – Temporal dispersion for a range of attacked neurons between one and 20 for JAM cyberattacks.

stant 50 ms. Figs. 7 and 8 present, respectively, the results for the number of spikes and the temporal dispersion. It is important to highlight that these plots present the same trend as described in Fig. 3 and Fig. 4, respectively, for the analysis between five and 105 attacked neurons.

Finally, Table 4 compares the Pearson correlation between both scenarios, which determines a correlation between the number of steps and the number of spikes of -0.66, which indicates that these variables have a 66% linear correlation in an inversely proportional way. A similar situation happens between the number of steps and the percentage of dispersion, with a -0.59 Pearson correlation. This indicates, in general, a low correlation between scenarios. However, this can be explained due to the reduction in the number of attacked neu-

Table 4 – Correlation of relevant features between CNN and biological experiments for JAM cyber attacks.

	# spikes	% dispersion	# steps	# neurons
# spikes	1.00	0.98	-0.66	-0.99
% dispersion	0.98	1.00	-0.59	-0.98
# steps	-0.66	-0.59	1.00	0.66
# neurons	-0.99	-0.98	0.66	1.00

rons considered. As indicated before, the number of neurons has been limited to a range between one and 20. Although these values offer variability in the CNN, there is not much difference in the distribution between these close sizes in the biological simulation.

Nevertheless, the individual analysis performed in this section for both biological and artificial scenarios presents the high impacts that JAM cyberattacks generate over these scenarios. The Spearman correlation values have also been calculated, studying the non-linearity of the data. Since the values obtained were similar to those presented for the Pearson correlation, we opted to include the latter for concision.

Finally, it is interesting to present the performance of the attacked model in terms of ROC curves. First, it is essential to highlight that the model has four different outputs (up, down, left, right), corresponding to the direction to perform the next step within the maze. Based on that, the ROC curves present the relationship between erroneous and correct predictions when the model is not under attack and when different configurations of the attacks are applied.

Focusing on JAM cyberattacks, and since they affect multiple positions, it is not possible to know the number of steps correctly performed to obtain the True Positive Rate (TPR) and False Positive Rate (FPR). Based on that limitation, we could assume a TPR equal to zero and FPR of 1, according to the configuration of the attack.

6. Comparison of JAM and FLO cyberattacks

This section compares the impact caused by JAM cyberattacks with FLO, a neuronal cyberattack existing in the literature. For that, we first introduce FLO cyberattacks, moving to the analysis of their impacts, and later we compare it with JAM. This section also provides an in-depth study of the results of individually performing FLO cyberattacks in different positions of the optimal path, comparing the results of biological and artificial networks.

6.1. *Definition of Neuronal Flooding cyberattacks*

Neuronal Flooding cyberattacks (FLO) were defined in our previous work (Bernal et al., 2020) as a way to overstimulate targeted neurons. In that work, we just explored the cyberattacks for the first position of the maze, whose behavior is formally represented by Algorithm 2. In particular, it indicates that the attack over the targeted neurons is performed in a particular instant of time $\mathbf{t}^{\text{attk}}$, in contrast to JAM, which is executed within a determined temporal period.

Algorithm 2 FLO cyberattack execution.

```
t = 0
while t < t^win do
  if t == t^attk then
    for all n ∈ 𝔸𝕟 do
      v_n ← v_n + vi_n
    end for
  end if
  t ← t + Δt
end while
```

Table 5 – Parameters used in the analysis for FLO cyberattacks.

Parameter	Values
Positions attacked (Bio, CNN)	{1, 2, ..., 27}
Number of neurons/nodes (Bio, CNN)	{5, 35, 55, 75, 105}
Voltage increment (Bio)	40 mV
Output importance (CNN)	60 %
Number of executions (Bio, CNN)	10

In contrast, the current work performs FLO cyberattacks over each individual position of the optimal maze path, evaluating a different number of simultaneously attacked neurons and multiple increment voltages per position. The parameters used for this experiment are indicated in Table 5, having five different values of simultaneously attacked neurons (or nodes) and a single value of voltage increment. The use of just one voltage value is based on the experiments performed in Bernal et al. (2020), which concluded that, for FLO cyberattacks, the usage of different voltages did not have a substantial impact. Besides, each combination of parameters is executed ten times.

6.2. *FLO cyberattacks over the biological network*

In the biological scenario, we perform a FLO cyberattack individually over each position of the optimal path of the maze, at the instant 50ms after reaching a targeted position, evaluating the impact of the attack during the complete simulation (27 s, until the mouse reaches the exit) based on the number of spikes and temporal dispersion metrics.

Fig. 9 presents the evolution of the number of spikes according to the individual position of the optimal path under attack. As previously indicated, the voltage used to increment the targeted neurons is 40 mV.

It is worthy to note that, for each attacked position, the represented values correspond to the number of spikes over the complete simulation. The upper sub-figure presents the mean of spikes for each position under attack, where each line represents a different number of attacked neurons. The effect of FLO cyberattacks to reduce the temporal dispersion was already documented in Bernal et al. (2020). In Fig. 9, we can observe that performing the attack in later positions of the optimal path generates a lower impact since in the positions before the cyberattack the spikes are not altered and, thus, the spiking behavior is the same as the spontaneous behavior. Particularly, it can be observed that attacking 105 neurons in

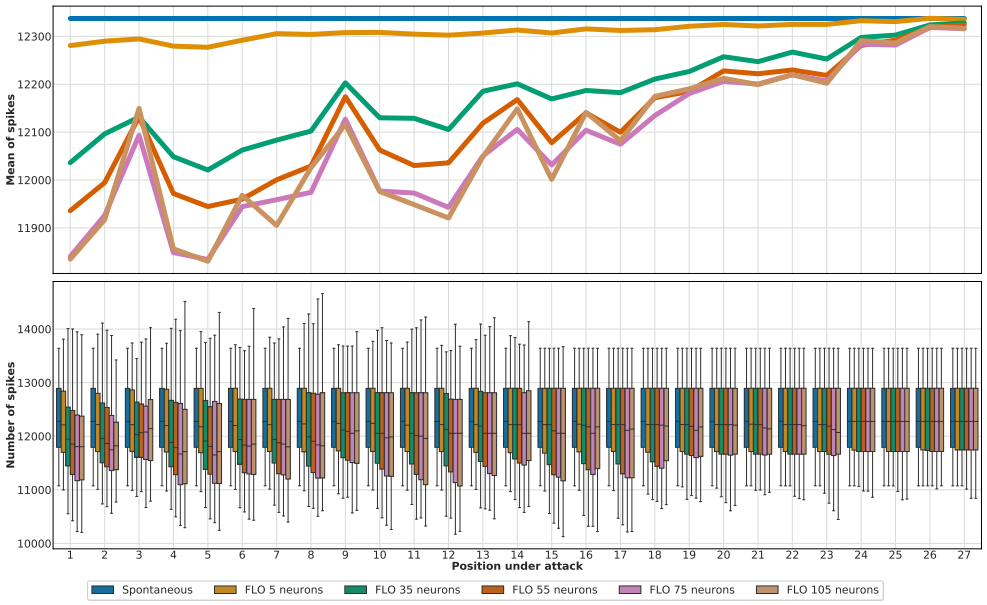


Fig. 9 – Distribution of the number of spikes according to the position under attack for FLO cyberattacks.

the first position generates an approximate reduction of 500 spikes. These results also indicate that this attack causes a desynchronization of neuronal activity over time, presenting a higher variability when the attack is performed in the first positions. This variability is also benefited by the particular model used and the propagation of the spikes.

Additionally, attacking a broader number of neurons produces, in general, a higher reduction in the mean of spikes. Nevertheless, we can observe no significant differences between attacking 75 and 105 simultaneous neurons in terms of the mean of spikes. Regardless of these similarities, there are variations in their maximum and minimum values, indicating variations in their distributions. These data correspond to the mean of the distribution represented in the bottom subfigure, where we can see a higher variability in the number of spikes when the attack is applied in the first positions. This figure also highlights that the maximum and minimum values of the distribution have a significant variability compared to the spontaneous behavior, stabilized when we attack in later positions.

After analyzing the behavior of the FLO cyberattack in terms of the number of spikes, Fig. 10 presents its impact focusing on the temporal dispersion metric. As can be seen, the dispersion is higher when attacking the first positions due to the same reasons addressed for the number of spikes metric. Additionally, attacking a broader number of neurons derives in a higher percentage of instants with spikes. Specifically, simultaneously attacking 75 neurons reaches the highest impact, augmenting the initial 36% of instants with spikes to an approximate 40%. Finally, it is worthy to note that these two metrics are highly related, with a Pearson correlation value of -0.97.

6.3. *FLO cyberattacks over the artificial network*

In terms of attacks over the CNN, it is essential to note that the voltage increment used to attack the biological network has been proportionally adapted to the CNN scenario, corresponding to the output importance indicated in Table 5. Based on that, the value of $40mV$ used in the biological scenario represents a 60% from the voltage range defined by the Izhikevich model used. This 60% is the equivalent value used to increment the importance of the targeted nodes during the attack to the CNN.

Fig. 11 presents the evolution of the mean number of steps among the ten executions per number of consecutively attacked nodes. This figure indicates the impact caused by attacking the mouse when it is placed in each individual position of the optimal path of the maze. When the simulated mouse is placed in a particular position, we obtain the number of steps required to reach the exit from the position attacked. To this resulting number of steps, we add the number of steps correctly performed until the attacked position, which corresponds to correctly performed decisions before the attack. It is essential to note that, once the model is attacked, it is used until the end of that particular execution.

In this figure, each color indicates a different number of simultaneous neurons attacked. It can be appreciated that the number of steps remains constant in the spontaneous behavior of the CNN, requiring 26 steps to find the exit. These 26

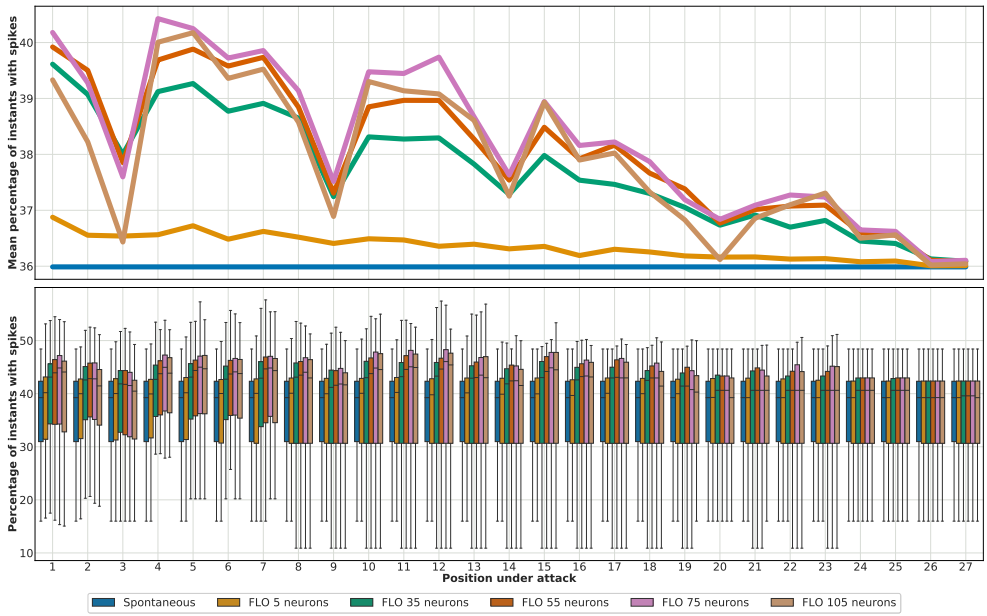


Fig. 10 – Percentage of instants with spikes according to the position under attack for FLO cyberattacks.

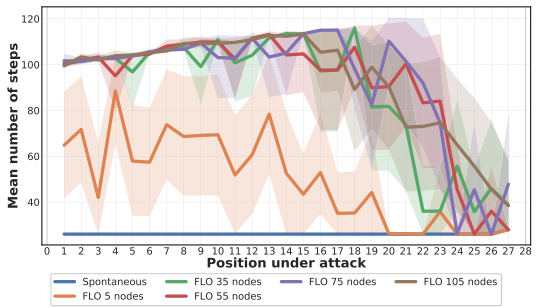


Fig. 11 – Mean of steps when we perform a FLO cyberattack in each position of the optimal path of the maze, considering five different number of simultaneously attacked neurons.

steps are determined by the model resulting from training the CNN, which concluded an optimal path of 27 positions to exit the exit and, thus, 26 steps between them. There is an exception in position 27, where the mouse needs to move to an adjacent cell in the maze to finally reach the exit since the mouse initially started in the exit position. This figure highlights that augmenting the number of attacked neurons increases the number of steps until position 21. From that position, the trend decreases since the closer the mouse is to the exit, the easier it is to solve the maze by probability, even if the mouse suffers an alteration in its decision ability.

Another relevant metric to study this situation is the percentage of times in which the mouse finds the exit. The Pearson correlation has been calculated between the number of steps and the success rate, obtaining a value of -0.99, meaning that they present a trend almost identical in an inversely proportional way. That is to say, we have observed that the number of steps increases when the percentage of success decreases. Based on that, the number of steps will be the sole metric used to evaluate the CNN in this analysis.

It is interesting to consider the relationship between the results obtained from attacking the biological and artificial scenarios to help understand the behavior in the biological network. To perform this comparison, Table 6 presents the Pearson correlation between the relevant features considered in these domains. In particular, we are interested in the relationship between the number of steps and the number of spikes, and between the number of steps and the percentage of dis-

Table 6 – Correlation of relevant features between CNN and biological experiments for FLO cyberattacks.

	position of attack	# spikes	% dispersion	# steps	# neurons
position attack	1.00	0.53	-0.53	-0.42	-0.0
# spikes	0.53	1.00	-0.97	-0.82	-0.66
% dispersion	-0.53	-0.97	1.00	0.81	0.56
# steps	-0.42	-0.82	0.81	1.00	0.65
# neurons	-0.0	-0.66	0.56	0.65	1.00

persion. Based on that, it can be determined that the CNN and biological approaches have a high correlation, with an approximate 80% correlation in both of them.

Based on the above, we can conclude a significant relationship between the results obtained in both experimental dimensions. These results suggest that performing attacks over the brain of the mouse could not only alter its spontaneous neuronal behavior but also affect its decisions to solve the maze, increasing the number of steps to find the exit and decreasing its chances to exit the maze. Nevertheless, these results are limited to our use case, the neuronal topology, and the use of a CNN to model a portion of the mouse's visual cortex.

Once presented the relationship between the biological and artificial scenarios, this section compares the results of both attacks. Since the approaches followed between these attacks are not directly comparable, where FLO focuses on individually attacking different positions and JAM affects multiple consecutive positions, this study focuses on analyzing the correlations obtained for each attack. In FLO, the Pearson correlation obtained was -0.82 for the relationship between the number of steps and number of spikes and 0.81 between steps and temporal dispersion. On the contrary, a value of -0.66 was obtained between the steps and the spikes and -0.59 for the relationship between steps and dispersion for JAM. These values indicate that the relationship between the biological and artificial networks is closer in the FLO situation, despite the analysis for the JAM cyberattack presented some limitations as stated in Section 5.

Finally, and as previously presented for JAM cyberattacks, we offer the performance of the attacked model based on ROC curves. In particular, for FLO cyberattacks, we have obtained two ROC curves. The first curve presents the TPR and FPR for aggregation of positions 24 to 27. We have included this range since in these positions, the mouse is able, on average, to always exit the maze (see Fig. 11). This ROC curve, subsequently presented in Fig. 12, indicates that since the mouse can always find the exit of the maze, the TPR will always be 1. Moreover, the FPR ranges from close to zero (perfect value) when attacking five simultaneous nodes to more than 0.8 when attacking 105. The FPR is determined based on the number of decisions incorrectly taken compared to the decisions performed by the spontaneous behavior.

The second ROC curve obtained for FLO presents an aggregation between positions one to 23 since we can observe in Fig. 11 that performing attacks in those positions is more damaging, and thus, the mouse is not always able, on average, to exit the maze. Because of that, the TPR decreases, where attacking five neurons presents the best TPR. From its part, the FPR is considerably high for a number of simultaneously attacked neurons higher than five, as presented in Fig. 13.

7. Neural cyberattacks and neurodegenerative diseases

This section discusses the results obtained in this work, aiming to understand the impact of these attacks better, their possible consequences in the real world, and defend against them. Additionally, if we could reproduce the effect of neurodegenerative diseases with these attacks, we could generate databases containing multiple attack configurations, study their impact, and propose mechanisms to reduce these impacts.

Previous sections have highlighted the enormous impact that neuronal cyberattacks can cause over spontaneous neural activity, affecting the amount, periodicity, and even the presence of spikes. Additionally, we have observed that these cyberattacks could also alter the simulated mouse's decision ability, forcing it to make mistakes in the resolution of the labyrinth. Furthermore, these cyberattacks possess differences based on their action mechanisms. JAM cyberattacks focus on continuously inhibiting the neuronal activity of the targeted neurons, suppressing this signaling along with the duration of the attack. On the contrary, FLO cyberattacks aim to overstimulate a set of neurons in a particular instant, extending its impact after its application.

Based on these action mechanisms, we identify that the behavior of the previous attacks has similarities with the effects and consequences that certain neurodegenerative diseases generate. As indicated in Section 2, neurodegenerative diseases can be included within the concept of brain connectivity disorders. In particular, for Alzheimer's Disease (AD), the deactivation of the Default Mode Network (DMN) could be reproduced by an attacker able to target individual neurons, reproducing or accelerating the effects of the disease. We identify that JAM, focused on neuronal activity inhibition, could be used for these purposes. On the contrary, Amyotrophic lateral sclerosis (ALS) is based on neuronal activity hyperexcitability, where FLO could be applied to periodically stimulate the targeted neurons and thus produce a perturbation in the excitatory/inhibitory balance of cortical neurons.

Although neuronal cyberattacks are promising mechanisms aiming to extend our knowledge about cybersecurity on BCI, further research is required to study the impact these cyberattacks can cause over neural circuits and cognitive and behavioral functions. The study of neuronal cyberattacks could help identify particular characteristics helping to detect

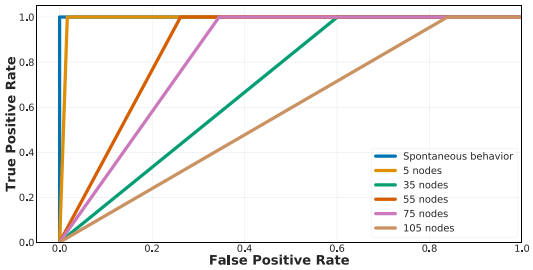


Fig. 12 – ROC curve for an aggregation of positions 24 to 27 of the optimal path of the maze.

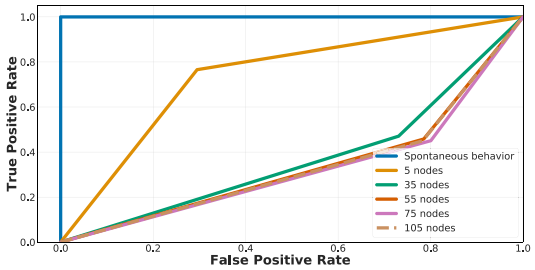


Fig. 13 – ROC curve for an aggregation of positions one to 23 of the optimal path of the maze.

prospect threats on BCI systems. Additionally, the application of neuronal cyberattacks could be beneficial in neurological research, using these cyberattacks to control the spread of the disease in neural models or even in vivo trials.

8. Conclusion

This work introduces the Neuronal Jamming cyberattack (JAM), consisting in the inhibition of neuronal activity. To implement this attack, and due to a lack of realistic neuronal topologies, a Convolutional Neural Network (CNN) has been trained to generate a neuronal topology based on a use case of a mouse trying to exit a maze. Once having both topologies, we analyze the impact that JAM cyberattacks present over biological and artificial scenarios. Additionally, this manuscript offers a comparison between JAM and FLO cyberattacks. For that, we have implemented several configurations of FLO, a cyberattack already existing in the literature aiming to overstimulate neural activity. To measure their impact, we have studied multiple metrics in the biological scenario (number of spikes and temporal dispersion) and in the CNN (number of steps and success rate in solving the problem).

The obtained results highlight that, in JAM cyberattacks, increasing the number of consecutive positions under attack reduces the spikes and temporal dispersion. In the artificial network, attacking up to 20 nodes is enough to prevent the mouse from completing the labyrinth. Moreover, a contribution of this work is the comparison between scenarios based on the study of linear correlation between variables. This analysis indicates that this attack could affect the mouse's ability to escape the maze. We have obtained a Pearson's correlation of 0.6, a low value explained due to the restriction of the number of neurons used to compute the correlations.

Additionally, we have observed for FLO experiments that delaying the instant of attack to later positions reduces the impact from both biological metrics. Moreover, delaying the attack until position 21 generates an increase in the number of steps. From this position, delaying the instant of attack decreases the number of steps since it is more probable to find the exit by probability. Pearson's correlation between variables for this cyberattack was approximately 0.8, highlighting a closer relationship between scenarios. Finally, we have discussed the similarities between neurodegenerative diseases and the neuronal cyberattacks studied.

In future work, we plan to investigate new neuronal cyberattacks with different action mechanisms and impacts. Additionally, since the main limitation of this work is the use of a neuronal topology extracted from a CNN, we aim to explore the possibility of having realistic topologies, which are currently very limited, to simulate existing and prospecting cyberattacks. Finally, as the present work only focuses on the characterization of these cyberattacks, we want to focus our efforts on designing and implementing detection mechanisms to identify the initiation of a neuronal cyberattack and propose mitigation techniques to reduce their impact or even neutralize it.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Sergio López Bernal: Methodology, Writing – original draft, Data curation, Software. **Alberto Huertas Celdrán:** Methodology, Conceptualization, Writing – review & editing. **Gregorio Martínez Pérez:** Supervision, Project administration, Funding acquisition.

Acknowledgment

This work has been partially supported by (a) Bit & Brain Technologies S.L. under the project CyberBrain, associated with the University of Murcia (Spain), by (b) the Swiss Federal Office for Defense Procurement (armasuisse) with the CyberSpec (CYD-C-2020003) project, and by (c) the University of Zürich UZH.

REFERENCES

Ballarin Usieto, P., Minguez, J., 2018. Avoiding brain hacking - challenges of cybersecurity and privacy in Brain Computer Interfaces.

Bernal SLópez, Huertas Celdrán A, Fernández Maimó L, Barros MT, Balasubramaniam S, Martínez Pérez G. Cyberattacks on miniature brain implants to disrupt spontaneous neural signaling. IEEE Access 2020;8:152204–22. doi:10.1109/ACCESS.2020.3017394.

Bernal SLópez, Huertas Celdrán A, Martínez Pérez G, Barros MT, Balasubramaniam S. Security in brain-computer interfaces: state-of-the-art, opportunities, and future challenges. ACM Comput. Surv. 2021;54(1). doi:10.1145/3427376.

Bonaci T, Calo R, Chizeck HJ. App stores for the brain : privacy and security in Brain-Computer Interfaces. IEEE Technol. Soc. Mag. 2015;34(2):32–9. doi:10.1109/ETHICS.2014.6893415.

Brunet A, Stuart-Lopez G, Burg T, Scekic-Zahirovic J, Rouaux C. Cortical circuit dysfunction as a potential driver of amyotrophic lateral sclerosis. Front. Neurosci. 2020;14:363. doi:10.3389/fnins.2020.00363.

Chollet, F., et al., 2015. Keras. https://keras.io.

Gal E, London M, Globerson A, Ramaswamy S, Reimann MW, Muller E, Markram H, Segev I. Rich cell-type-specific network topology in neocortical microcircuitry. Nat. Neurosci. 2017;20(7):1004–13. doi:10.1038/nn.4576.

Géron A. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media; 2019.

Hartmann CJ, Fliegen S, Groiss SJ, Wojtecki L, Schnitzler A. An update on best practice of deep brain stimulation in Parkinson's disease. Ther. Adv. Neurol. Disord. 2019;12 1756286419838096.

Huertas Celdrán A, Gil Pérez M, García Clemente FJ, Martínez Pérez G. Preserving patients' privacy in health scenarios through a multicontext-aware system. Ann. Telecommun. 2017;72(9):577–87. doi:10.1007/s12243-017-0582-7.

Huertas Celdrán A, Gil Pérez M, García Clemente FJ, Martínez Pérez G. Sustainable securing of medical cyber-physical systems for the healthcare of the future. Sustain. Comput. Inform. Syst. 2018;19:138–46. doi:10.1016/j.suscom.2018.02.010.

Ienca M. Neuroprivacy, neurosecurity and brain-hacking: emerging issues in neural engineering. Bioeth. Forum 2015;8(2):51–3.

Ienca M, Haselager P. Hacking the brain: brain-computer interfacing technology and the ethics of neurosecurity. Eth. Inf. Technol. 2016;18(2):117–29. doi:10.1007/s10676-016-9398-9.

Izhikevich EM. Simple model of spiking neurons. IEEE Trans. Neural Netw. 2003;14(6):1569–72. doi:10.1109/TNN.2003.820440.

Kuzovkin I, Vicente R, Petton M, Lachaux J-P, Baciu M, Kahane P, Rheims S, Vidal JR, Aru J. Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. Commun. Biol. 2018;1(1):107. doi:10.1038/s42003-018-0110-y.

Lebedev MA, Nicolelis MAL. Brain-machine interfaces: from basic science to neuroprostheses and neurorehabilitation. Physiol. Rev. 2017;97(2):767–837. doi:10.1152/physrev.00027.2016.

Martínez Beltrán ET, Quiles Pérez M, López Bernal S, Huertas Celdrán A, Martínez Pérez G. Noise-based cyberattacks generating fake p300 waves in brain–computer interfaces. Clust. Comput. 2021. doi:10.1007/s10586-021-03326-z.

Musk E, Neuralink. An integrated brain-machine interface platform with thousands of channels. bioRxiv 2019. doi:10.1101/703801. https://www.biorxiv.org/content/early/2019/08/02/703801.full.pdf

O'Doherty JE, Lebedev MA, Ifft PJ, Zhuang KZ, Shokur S, Bleuler H, Nicolelis MAL. Active tactile exploration enabled by a brain-machine-brain interface. Nature 2011;479:228–31. doi:10.1038/nature10489.

Quiles Pérez M, Martínez Beltrán ET, López Bernal S, Huertas Celdrán A, Martínez Pérez G. Breaching subjects' thoughts privacy: a study with visual stimuli and brain-computer interfaces. J. Healthc. Eng. 2021;2021:5517637. doi:10.1155/2021/5517637.

Ramadan RA, Vasilakos AV. Brain computer interface: control signals review. Neurocomputing 2017;223:26–44. doi:10.1016/J.NEUCOM.2016.10.024.

Seo, D., Carmena, J. M., Rabaey, J. M., Alon, E., Maharbiz, M. M., 2013. Neural dust: An ultrasonic, low power solution for chronic brain-machine interfaces. arXiv:1307.2196.

Stimberg M, Brette R, Goodman DF. Brian 2, an intuitive and efficient neural simulator. eLife 2019;8:e47314. doi:10.7554/eLife.47314.

Sutton RS, Barto AG. Reinforcement Learning: An Introduction. second. The MIT Press; 2018.

Takabi H, Bhalotiya A, Alohaly M. Brain computer interface (BCI) applications: privacy threats and countermeasures. In: Proceedings of the 2016 IEEE 2nd International Conference on Collaboration and Internet Computing, IEEE CIC 2016; 2016. p. 102–11. doi:10.1109/CIC.2016.24.

Wirdatmadja SA, Barros MT, Koucheryavy Y, Jornet JM, Balasubramaniam S. Wireless optogenetic nanonetworks for brain stimulation: device model and charging protocols. IEEE Trans. NanoBiosci. 2017;16(8):859–72. doi:10.1109/TNB.2017.2781150.

Yao L, Sheng X, Mrachacz-Kersting N, Zhu X, Farina D, Jiang N. Sensory stimulation training for BCI system based on somatosensory attentional orientation. IEEE Trans. Biomed. Eng. 2019;66(3):640–6. doi:10.1109/TBME.2018.2852755.

Zafrany, S., Deep reinforcement learning for maze solving.

Zott B, Busche MA, Sperling RA, Konnerth A. What happens with the circuit in Alzheimer's disease in mice and humans? Annu. Rev. Neurosci. 2018;41(1):277–97. doi:10.1146/annurev-neuro-080317-061725. PMID: 29986165

Sergio López Bernal received the B.Sc. and M.Sc. degrees in computer science from the University of Murcia, and the M.Sc. degree in architecture and engineering for the IoT from IMT Atlantique, France. He is currently pursuing the Ph.D. degree with the University of Murcia. His research interests include ICT security on braincomputer interfaces and network and information security.

Alberto Huertas Celdrán received the M.Sc. and Ph.D. degrees in computer science from the University of Murcia, Spain. He is currently a postdoctoral fellow associated with the Communication Systems Group (CSG) at the University of Zurich UZH. His scientific interests include medical cyber-physical systems (MCPS), braincomputer interfaces (BCI), cybersecurity, data privacy, continuous authentication, semantic technology, context-aware systems, and computer networks.

Gregorio Martínez Pérez is Full Professor in the Department of Information and Communications Engineering of the University of Murcia, Spain. His scientific activity is mainly devoted to cybersecurity and networking, also working on the design and autonomic monitoring of real-time and critical applications and systems. He is working on different national (14 in the last decade) and European IST research projects (11 in the last decade) related to these topics, being Principal Investigator in most of them. He has published 160+ papers in national and international conference proceedings, magazines and journals.

4

Taxonomy of Neural Cyberattacks

De: **Communications of the ACM** onbehalfof@manuscriptcentral.com
Asunto: Communications of the ACM - Decision on Manuscript ID CACM-21-06-3996.R1
Fecha: 2 de mayo de 2022, 11:10
Para: slopez@um.es
Cc: slopez@um.es, huertas@ifi.uzh.ch, gregorio@um.es

02-May-2022

Dear Mr. López Bernal:

It is a pleasure to accept your manuscript entitled "Eight Reasons Why Cybersecurity on Novel Generations of Brain-Computer Interfaces Must Be Prioritized" in its current form for publication in the Communications of the ACM. The comments of the reviewer(s) who reviewed your manuscript are included at the foot of this email.

Please do not spend any additional effort on formatting. Your final PDF will serve as the input for the layout of the paper by the CACM editors, and you will receive a copy for approval before publication.

Thank you for this fine article. On behalf of the editors of the Communications of the ACM, we look forward to future contributions from you. You will hear from us when the article is slated for production in a few months.

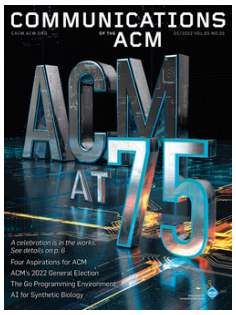
Sincerely,
James R. Larus
Editor-in-Chief, Communications of the ACM (CACM)
Professor, EPFL, Lausanne Switzerland

EIC:
In the interest of moving this along, I am going to accept this article. Please take into account the suggestions from the reviewers and AE when you are revising it for publication.

Co-Chair: Co-Chair, Contributed
Comments to the Author:
(There are no comments.)

Associate Editor: Cleland-Huang, Jane
Comments to the Author:
Thank you for the changes that you have made. Both reviewers are recommending acceptance now. I have marked this as minor revision to give you the opportunity to see and address any of the comments made by reviewer #2 that you are able to address. Addressing these comments is optional, but I wanted to give you the opportunity to do so. Once you resubmit your revision, this will move quickly to 'accept'.

Congratulations!



Title:	Eight Reasons Why Cybersecurity on Novel Generations of Brain-Computer Interfaces Must Be Prioritized
Authors:	Sergio López Bernal, Alberto Huertas Celdrán, Gregorio Martínez Pérez
Journal:	Communications of the ACM
JIF:	14.065 D1 (2021)
Publisher:	ACM
Volume:	
Number:	
Pages:	9
Year:	2022
Month:	May
DOI:	10.1145/3535509
Status:	Accepted

Abstract

Brain-Computer Interfaces (BCIs) enable bidirectional communication between the brain and external devices. These technologies have been mainly used in medical scenarios for diagnosing and treating neurodegenerative diseases. Despite the advances introduced by these systems, they present vulnerabilities that attackers could exploit to cause brain damage. In this context, previous work defined the next three neural cyberattacks altering spontaneous neuronal activity: Neuronal Flooding, Neuronal Scanning, and Neuronal Jamming. In addition, more effort is still needed to detect and characterize new neural cyberattacks with new behaviors. Based on that, this publication presents a taxonomy of eight neural cyberattacks, where the next five are novel: Neuronal Selective Forwarding, Neuronal Spoofing, Neuronal Sybil, Neuronal Sinkhole, and Neuronal Nonce. For each of them, this work offers a formal definition and the conceptualization of their behavior. Finally, it compares them to study their impact on the short and long term. The performed analysis indicated that Neuronal Nonce was the most damaging attack in the short term, with an approximate 12% of neural activity compared to spontaneous neuronal behavior. Finally, Neuronal Scanning was the most effective in the long term, offering a reduction of around 9%.

Keywords

Cybersecurity · Brain-computer interfaces · Neuronal cyberattacks · Taxonomy

Eight Reasons Why Cybersecurity on Novel Generations of Brain-Computer Interfaces Must Be Prioritized

Sergio López Bernal
slopez@um.es
Department of Information and Communications Engineering, University of Murcia
Murcia, Spain

Alberto Huertas Celdrán
huertas@ifi.uzh.ch
Communication Systems Group (CSG), Department of Informatics (IfI), University of Zurich UZH
Zürich, Switzerland

Gregorio Martínez Pérez
gregorio@um.es
Department of Information and Communications Engineering, University of Murcia
Murcia, Spain

CCS CONCEPTS

• **Security and privacy → Domain-specific security and privacy architectures**.

KEYWORDS

Cybersecurity, Brain-Computer Interfaces, Neuronal Cyberattacks, Taxonomy

ACM Reference Format:
Sergio López Bernal, Alberto Huertas Celdrán, and Gregorio Martínez Pérez. 2022. Eight Reasons Why Cybersecurity on Novel Generations of Brain-Computer Interfaces Must Be Prioritized. *Communications of the ACM*, 9 pages. https://doi.org/10.1145/3535509

1 INTRODUCTION

Brain-Computer Interfaces (BCIs) are bidirectional systems that interact with the brain, allowing the acquisition of neural data and neuronal stimulation. BCIs can be classified according to their invasiveness level, being invasive interfaces extensively used in medical therapy. In this sense and as an example, invasive BCIs focused on neural recording have been used to control prosthetic limbs in impaired patients, while BCIs for neuromodulation have been helpful for treating neurodegenerative conditions, such as Parkinson's disease [9]. The second main family of BCIs, in terms of invasiveness, is the non-invasive one. BCIs based on non-invasive principles and, mainly, those focused on neural data acquisition such as electroencephalography (EEG), have gained popularity in recent years, extending their usage from traditional medical scenarios to new domains such as entertainment or video games. However, despite the benefits of non-invasive BCIs, some works in the literature have identified particular cybersecurity issues from a neural data acquisition perspective. In particular, Martinovic et al. [19] demonstrated that an attacker could obtain sensitive personal data from BCI users, taking advantage of their cerebral response (P300 potentials) generated when known visual stimuli are presented to them. Bonaci et al. [1] also described a scenario where attackers could maliciously add or modify software modules defining the BCI to perform dangerous actions over the users. Finally, Takabi et al. [24] highlighted that most APIs used to develop BCI applications offered complete access over the information acquired by the BCI, presenting confidentiality problems.

Cybersecurity of invasive BCIs is also a challenge that has been identified in the literature and whose application is in initial stages [3, 4, 8]. This situation is complicated by the recent introduction of novel BCI designs based on nanotechnology aiming to surpass the limitations of traditional BCIs. One example of these emergent systems is Neuralink [20], which uses nanotechnology to record and stimulate particular brain regions with single-neuron resolution. Despite the advantages of the new generation of invasive BCIs, the literature has already identified that some of these BCIs present vulnerabilities that attackers could exploit to affect neural activity [17]. In particular, the literature has proposed two cyberattacks focused on neural stimulation named Neural Flooding and Neural Scanning [17], as well as a cyberattack focused on neural inhibition [18]. These threats have been defined within the term *neural cyberattacks*, consisting in well-known attacks from computer science, able to disrupt the spontaneous activity of neural networks of the brain, stimulating or inhibiting neurons.

In such a disruptive and novel context, one of the main challenges is formally defining the behavior of different neural cyberattacks affecting the brain. In this direction, studies addressing how neural cyberattacks could recreate the effects induced by certain neurodegenerative diseases are absent in current literature. Furthermore, the analysis of these cyberattacks regarding their impact on spontaneous neural activity is unexplored. Finally, a comparison of the impact caused by distinct neural cyberattacks is required to understand the changes caused over the brain.

With the goal of improving the previous open challenges, this article presents eight neural cyberattacks affecting spontaneous neural activity, inspired by well-known cyberattacks from the computer science domain: Neural Flooding, Neural Jamming, Neural Scanning, Neural Selective Forwarding, Neural Spoofing, Neural Sybil, Neural Sinkhole and Neural Nonce. After presenting their formal definitions, the cyberattacks have been implemented over a simulated biological neural network representing a portion of a mouse's visual cortex, whose topology has been obtained from training a Convolutional Neural Network (CNN). This implementation is based on a lack of realistic neuronal topologies in the literature [7] and existing works indicating the similarities CNNs have with neuronal structures from the visual cortex [11, 13–15]. Finally, a comparison of the impact between each neural cyberattack is presented for the initial and final part of a neural simulation, studying

Accepted in Communications of the ACM, May, 2022

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/3535509

their impact for both the short and long term. In conclusion, Neural Nonce and Neural Jamming are the most suitable cyberattacks for short-term effects, while Neural Scanning and Neural Nonce are the most adequate for long-term effects.

2 THE BRAIN AT RISK DUE TO NOVEL GENERATIONS OF BCI

Although this work focuses on neuronal cyberattacks from a computer science point of view, it is essential to introduce, in a basic and synthesized way, how the brain works to understand their behavior and the current state of neuromodulation technologies able to stimulate and/or inhibit neurons.

The brain is the most complex organ in the human body, managing all major activities of the organism. Its structure is divided into two hemispheres, left and right, controlling the opposite side of the body. Moreover, the cortex of each hemisphere presents four lobes on its surface with differentiated responsibilities. Frontal lobes intervene in reasoning, planning, translating thoughts into words, and defining personality. In contrast, parietal lobes manage sensory perceptions such as taste or touch, additionally to temperature and pain. These lobes also intervene in memory and the understanding of languages. Occipital lobes are in charge of decoding visual information, such as colors or forms, and identify objects, while temporal lobes focus on processing auditory stimuli, also intervening in verbal memory [12].

Within the hemispheres, around 86 billion neurons interact with each other to perform these complex tasks. This interaction is performed by two specific structures of the neuron, the dendrites and the axon. While dendrites receive information from other neurons, axons transmit instructions to neurons. The connection established between these structures is known as a synapse, and it is the base of neuronal communication. In neuronal communication, the dendrites of a given neuron receive stimuli from many neurons (presynaptic neurons) via neurotransmitters, which are molecules that force actions in the receiver neuron (postsynaptic neuron). Presynaptic neurons can be excitatory, producing particular neurotransmitters aiming to initiate an impulse on the postsynaptic neuron or inhibitory, liberating neurotransmitters to prevent its activity. If the sum of these positive and negative impulses exceeds the excitation threshold of the postsynaptic neuron, this neuron will generate a nerve impulse known as action potential (or spike), electrically transmitted along the axon to reach the axon terminals. When the electric stimulus reaches these terminals, they liberate particular neurotransmitters to the synaptic cleft, the space separating the axon from the dendrites of other neurons, aiming to influence their activity in an excitatory or inhibitory way. These electric and chemical processes are repeated neuron after neuron, only if they exceed their excitation threshold.

Neurotechnology plays an essential role in supporting these neuronal communications, used for decades in clinical scenarios to induce or suppress neural activity. There is a wide variety of technologies, both invasive and non-invasive, with different modulation principles such as ultrasounds, electrical currents, magnetic fields, or light pulses (optogenetics) [6]. Despite the differences of these approaches, most of them share common parameters used to adjust the modulation process, such as the amplitude or voltage applied or the duration and periodicity of the pulses. Focusing on invasive BCIs, Deep Brain Stimulation (DBS) represents an excellent example of these technologies used to treat conditions like Parkinson's disease or obsessive-compulsive disorder using neural stimulation [9]. Moreover, most invasive BCIs also offer recording capabilities, enabling the monitoring of the brain to determine the best instant to stimulate or inhibit a particular set of neurons.

In such a scenario, novel solutions such as Neuralink [20] or WiOptND [26] deserve special interest since they have introduced the use of nanotechnology to miniaturize the electrodes implanted in the brain, achieving single-neuron resolution. Particularly, these technologies address neuromodulation from two different perspectives. Neuralink uses electrical currents to stimulate the brain, while WiOptND stimulates or inhibits neuronal activity using optogenetics. Nevertheless, these current initiatives present vulnerabilities in their architectures that attackers could exploit to stimulate or inhibit neurons maliciously [17]. In this direction, Figure 1 introduces the anatomical structure of the head from the scalp to the cerebral cortex, presenting an invasive neuromodulation BCI placed in the cortex that an attacker externally targets. As can be seen, the attacker can execute one of the eight cyberattacks proposed in this work (more details are provided in Section 3). These cyberattacks exploit vulnerabilities existing in current BCIs (see [16]), generating an impact over the BCI, thus stimulating or inhibiting neuronal activity.

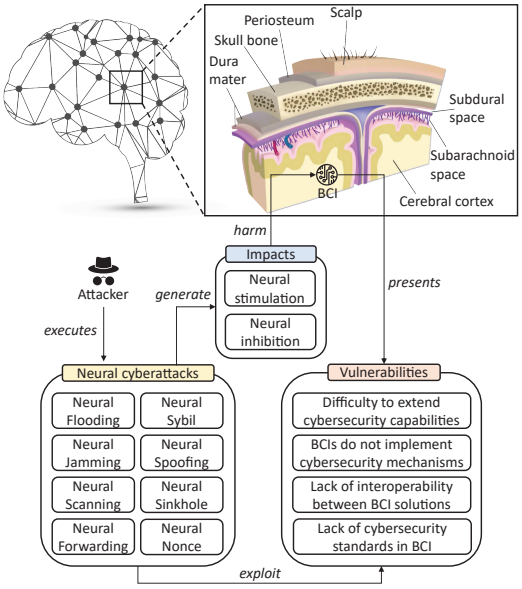


Figure 1: Attacker executing the proposed neuronal cyberattacks that exploit vulnerabilities of invasive neuromodulation BCIs and generate particular impacts on the BCI.

3 EIGHT NEURAL CYBERATTACKS AFFECTING BRAIN BEHAVIOR

Once the vulnerabilities of novel BCIs have been highlighted, it is time to introduce neural cyberattacks exploiting them and affecting neural behavior. In this direction, this work presents eight cyberattacks inspired by well-known threats from digital communications, justified by the potential exploitation of previously highlighted vulnerabilities. Particularly, five of these cyberattacks are new (Neural Selective Forwarding, Neural Spoofing, Neural Sybil, Neural Sinkhole and Neural Nonce, while the remaining three were presented in previous work (Neural Flooding and Neural Scanning in [17], and Neural Jamming in [18]). All these cyberattacks are either based on the stimulation of neurons, their inhibition, or a combination of both. Particularly, and for the sake of simplicity, these cyberattacks assume the usage of technologies able to stimulate or inhibit neuronal behavior.

3.1 Neuronal Flooding

In cybersecurity, flooding cyberattacks focus on collapsing a network by transmitting a high number of data packets, generally directed to particular targets within the network [22]. As a consequence, these endpoints increase their workload, not being able to manage legit communications adequately. Moving to a neurological perspective, Neuronal Flooding (FLO) cyberattacks aim to overstimulate multiple neurons in a particular time instant. This cyberattack does not need previous knowledge about the status of the target neurons, presenting a low complexity compared to other neural cyberattacks.

The general behavior of the FLO cyberattack implemented can be consulted in Figure 2, where green boxes indicate actions performed by the cyberattack, and yellow diamonds are conditional blocks. First, the attacker determines the attacking instant and the list of targeted neurons. During the desired instant, the cyberattack selects each of the neurons and stimulates it. Although the flow chart presented could be interpreted as sequentially affecting these neurons, the attack is performed in a particular instant of time, resulting in attacking the neurons at the same time.

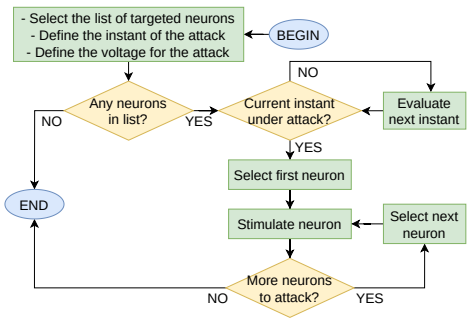


Figure 2: Implemented behavior of Neuronal Flooding.

3.2 Neuronal Jamming

Jamming cyberattacks focus on disrupting legitimate communications by introducing a malicious interference to the medium and preventing the devices from communicating, thus deriving in a denial of service (DoS) [25]. This principle can be translated to the neurological world, where Neuronal Jamming (JAM) consists in the inhibition of the activity of a set of neurons, impeding them from generating or transmitting impulses to adjacent neurons. In contrast to FLO, this cyberattack is performed during a determined temporal window, in which the affected neurons do not generate activity. This cyberattack also presents a low execution complexity, only requiring the selection of the target neurons and the attack duration.

The flow chart depicted in Figure 3 represents a temporal window in which the JAM cyberattack is performed. For each instant between the beginning and the end of the attack, the list of targeted neurons is simultaneously inhibited. This inhibition consists in setting the neurons to their lowest voltage within their natural range of values.

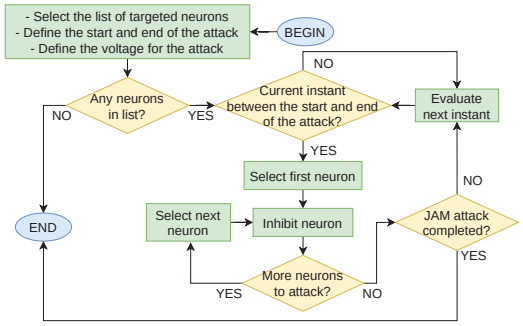


Figure 3: Implemented behavior of Neuronal Jamming.

3.3 Neuronal Scanning

Port scanning is a common cybersecurity technique used to verify if the communication ports of a machine are being used and identify vulnerable services available in those ports [22]. For that, all ports of the machine are sequentially tested. Similarly, Neuronal Scanning (SCA) cyberattacks aim to sequentially stimulate all neurons of a neuronal population, affecting only one neuron per time instant. As in the previous cyberattacks, SCA does not require previous knowledge about the status of the targeted neurons. Nevertheless, it presents a moderate execution complexity since the attacker needs to coordinate the order of the neurons attacked, avoid repetitions between them, and determine the time interval between attacking each neuron.

The SCA cyberattack implemented (see Figure 4) targets one neuron per instant under attack, removing from the list those neurons already attacked to avoid repetitions and ensure a sequential selection. These instants are determined based on the start of the attack and the time that the attacker waits between affecting neurons.

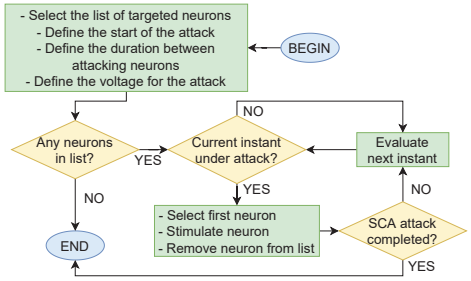


Figure 4: Implemented behavior of Neuronal Scanning.

3.4 Neuronal Selective Forwarding

Selective forwarding is one of the most harmful cyberattacks against communication networks. In this kind of threat, malicious hosts selectively drop some packets instead of forwarding them [2]. The selection of dropping nodes may be random or predefined depending on the attack design. In the brain context, Neuronal Selective Forwarding (FOR) consists in changing the propagation behavior of a set of neurons during a temporal window, inhibiting particular neurons at each instant of the window. This attack is more elaborate than the previous ones because it requires knowledge of the neurons involved in a given neuronal propagation path and their status in each instant. It is achieved by real-time neuronal monitoring or previously knowing the neuronal propagation behavior due to the repetition of actions such as eye blinks or limb movements.

This cyberattack allows a wide variety of different configurations for targeting neurons. It has been followed the same sequential criteria already presented for SCA in this work, inhibiting them instead of performing neural stimulation. Attending to Figure 5, FOR introduces an additional conditional block that verifies if the current voltage of the neuron is suitable for inhibition. Based on the voltage defined for the attack, the implementation verifies if the subtraction between the current voltage and the attacking voltage is lower than the lowest possible value. If so, the attack sets the voltage to the lowest threshold to avoid unrealistic results.

3.5 Neuronal Spoofing

In computer networks, a spoofing cyberattack occurs when a malicious party impersonates a computer or subject to steal sensitive data or launch attacks against other network hosts [22]. In the brain scenario, Neuronal Spoofing (SPO) cyberattacks consist in replicating the behavior of a set of neurons during a given period. After recording the neuronal activity, the attacker uses this pattern to stimulate or inhibit the same or different neurons at a different time. This attack is one of the most sophisticated since it requires recording, stimulation and inhibition capabilities, and deep knowledge of brain functioning. Like most of them, the impact of this cyberattack is high because a malicious attacker could control some vital functions of the subject's body.

The diagram presented in Figure 6 highlights two main processes. First, the attack performs a neuronal recording procedure for the selected neurons during a particular temporal period. For each instant

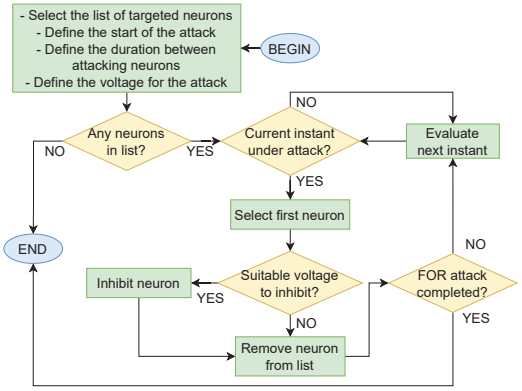


Figure 5: Implemented behavior of Neuronal Selective Forwarding.

within the period, the attacker stores the voltage of each recorded neuron. After that, the second process properly stimulates or inhibits a different neuronal population targeted by the attack, forcing them to have the same behavior that those previously recorded.

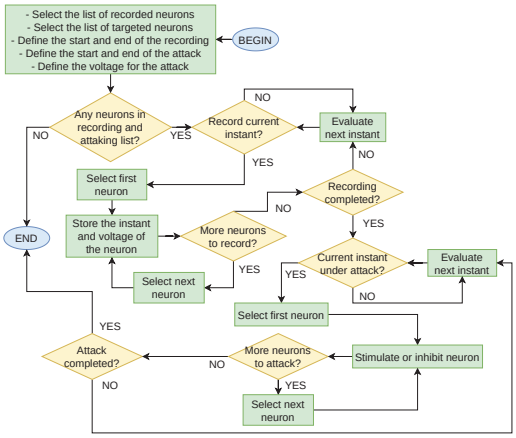


Figure 6: Implemented behavior of Neuronal Spoofing.

3.6 Neuronal Sybil

Sybil cyberattacks happen when a computer is hijacked to claim multiple identities, presenting broad security and safety implications. Having different identities, the behavior of the infected host differs according to the identity acting in each moment [5]. Bringing these cyberattacks to the brain scenario implies that an attacker could alter the operation of one or more neurons doing precisely the opposite as their natural behavior. It means that when a given

neuron is firing, the attacker inhibits the activity, and when it is not firing, the attacker fires it. Neuronal Sybil (SYB) cyberattacks are the most complex of the presented because they require real-time recording (or previous knowledge of the firing pattern) and the functionality of either stimulating or inhibiting a particular neuron in a given instant depending on its natural behavior. The impact of these neural cyberattacks is high, depending on the number of affected neurons.

The implementation of SYB cyberattacks is similar to the one presented for FLO, although the action performed over the neurons is different (see Figure 7). In SYB, the voltage of each targeted neuron is set to the opposite value within its natural range. This is obtained by adding the higher and lower voltage thresholds of the neuron and subtracting the current voltage value.

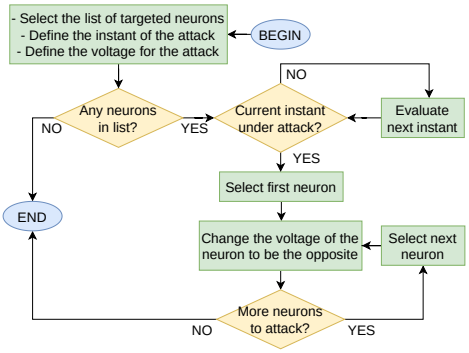


Figure 7: Implemented behavior of Neuronal Sybil.

3.7 Neuronal Sinkhole

Sinkhole cyberattacks are applied to routing protocols, where a node of the network broadcasts that itself is the best path to reach particular destinations. Based on that, the surrounding nodes will transmit their traffic to the malicious node, which could access, modify or discard the received data [21]. From a neurological perspective, Neuronal Sinkhole (SIN) cyberattacks focus on stimulating neurons from superficial layers connected to neurons placed in deeper layers, being the later the main target of the attack. In this regard, SIN cyberattacks present a high complexity since the attacker requires knowledge about the neuronal topology and synapses of a specific area of the brain. Moreover, this cyberattack is performed in a particular instant, stimulating the trigger set of neurons that initiates the attack.

The actions included in the implementation of SIN cyberattacks are the same as the presented for FLO, as shown in Figure 8. The main difference between them lies in the selection of the targeted neurons. SIN cyberattacks directly affect the neurons from early layers connected via synapses with the target neuron located in deep layers. Once identified the neurons to attack, the process of stimulation is the same as FLO.

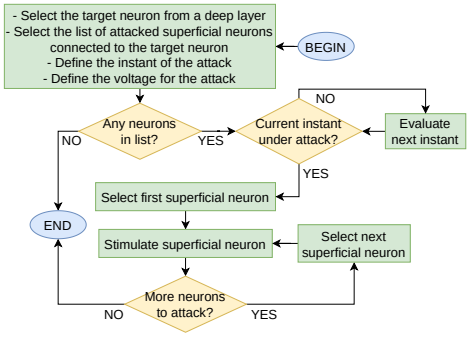


Figure 8: Implemented behavior of Neuronal Sinkhole.

3.8 Neuronal Nonce

Nonce numbers are typically random values utilized in cryptography to secure communications. A nonce is commonly used just once to prevent that old communications are reused and thus perform a replay attack [22]. In the context of neural cyberattacks, Neuronal Nonce (NON) consists in attacking a random set of neurons in a particular instant of time. The action performed could vary based on the interests of the attacker, either producing neural stimulation, neural inhibition, or a combination of both. The next execution of the attack will target a completely different set. Based on this variability, the complexity of the cyberattack is low, just requiring physical access to the target neurons.

This cyberattack has been implemented following the same principles already presented. The main difference (see Figure 9) resides in the selection of the action to apply over each targeted neuron. For each instant under attack and each targeted neuron, the attack randomly determines to stimulate, inhibit or keep its spontaneous behavior. The attacker can also indicate the probability assigned to each action aiming to benefit particular actions.

Once presented the behavior of each neural cyberattack, Table 1 introduces a comparison between them. In particular, the theoretical impact of each attack depends on the aggressiveness of its action mechanism and the knowledge that the attacker has about the target neurons. Nevertheless, these cyberattacks present particular aspects that complicate their comparison, such as their inner behavior, the instants and duration of the cyberattacks, the number of affected neurons, or the voltages used to stimulate those neurons.

4 WHAT IS THE IMPACT OF NEURAL CYBERATTACKS?

To answer this question it is important to mention that biological neural topologies, known as connectomes, are critical to measure the impact caused by cyberattacks. However, there is an absence of realistic neuronal topologies in the literature, being an open challenge of the area [7]. In this context, and to alleviate this limitation, the literature has evidenced that the hierarchy and functioning of neurons in charge of the vision present similarities with the functioning of CNNs [14] [11, 13, 15]. Particularly, the layers in both

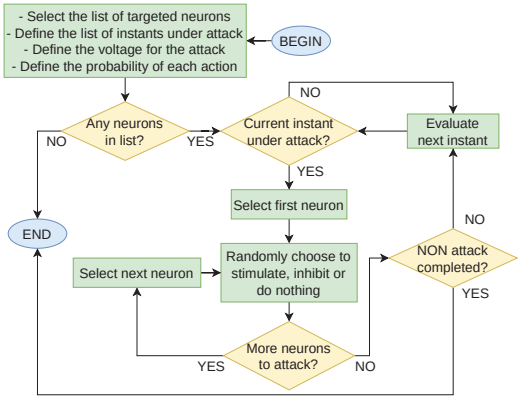


Figure 9: Implemented behavior of Neuronal Nonce.

Table 1: Comparison of proposed neural cyberattacks

Cyberattack	Impact	Neurons involved per instant	Duration	Complexity
Neuronal Flooding	Stimulation	1 - many	One instant	Low
Neuronal Jamming	Inhibition	1 - many	Time window	Low
Neuronal Scanning	Stimulation	1	Time window	Moderate
Neuronal Selective Forwarding	Recording Inhibition	1 - many	Time window	Moderate
Neuronal Spoofing	Recording Stimulation Inhibition	1 - many	Time window	High
Neuronal Sybil	Recording Stimulation Inhibition	1 - many	One instant	High
Neuronal Sinkhole	Stimulation	1 - many	One instant	Low
Neuronal Nonce	Recording Stimulation Inhibition	1 - many	One instant	Low

networks move from simple to abstract, where convolutional layers are related to early visual regions and dense layers present similarities with later visual areas. Furthermore, as stated by [15], CNNs could be good candidates for approximation models of the visual system. Based on that, this work employs a simulated biological network, whose topology is artificially generated from training a CNN, where the resulting CNN weights are transformed to biological synaptic weights, used to represent the voltage increase induced during an action potential. In summary, the CNN is just used to generate a biological topology, while the biological connectome is used to evaluate the impact of neural cyberattacks, representing the effect of attacks over a neurostimulation BCI placed in the brain.

Considering the similarities between CNNs and biological approaches, previous work trained a CNN to solve the problem of a mouse trying to exit a determined maze, modeling a portion of a mouse's visual cortex [17, 18]. This paper also uses this network to generate a simple biological connectome to test the proposed eight attacks. Particularly, this CNN was trained to obtain the optimal path on the maze to find the exit, resulting in 27 positions, whose topology comprises two convolutional layers of 200 and 72 nodes, respectively, and a final dense layer of four nodes. Although this simulated topology is not equivalent to a biological one, it serves to compare the impact that each neural cyberattack has over a common baseline.

Once having the artificial neural topology, it was ported to the Brian2 neuronal simulator [23], modeling the biological behavior of pyramidal neurons from three different layers of the visual cortex of the mouse (L2/3, L5, and L6). For that, Izhikevich's neuronal model [10] was used to represent excitatory neurons with regular spiking dynamics, defining neurons with a voltage range between -65 mV and 30 mV. Finally, a simulation of 27 seconds was defined, simulating a mouse staying one second in each position of the optimal path of the maze previously mentioned. Supplementary information concerning design and implementation aspects can be found in [17].

Table 2 summarizes the parameters used during the experimentation for each neural cyberattack. It is relevant to note that FLO, JAM, SPO, and SYB target random neurons from the first layer, while SCA and FOR sequentially attack all 200 neurons. SIN affects only the neurons related to the target neurons, and NON randomly evaluates the decision over each neuron of the first layer. Finally, NON presents a probability of 20% of stimulating a neuron, a 20% of inhibiting it, and a remaining 60% of keeping its spontaneous behavior until the next instant under attack.

Table 2: Parameters used for each neural cyberattack, where up arrows (↑) indicate a voltage increase, and down arrows (↓) a voltage decrease.

Cyberattack	Attacked neurons	Voltage	Attack start	Attack duration
FLO	100	↑ 40 mV	50 ms	Instantaneous (1 ms)
JAM	100	-65 mV	10 ms	50 ms
SCA	200	↑ 40 mV	10 ms	Whole simulation
FOR	200	↓ 40 mV	10 ms	Whole simulation
SPO	100	Recorded voltages	10 ms	50 ms
SYB	100	Opposite in range	10 ms	Instantaneous (1 ms)
SIN	Up to 200	↑ 40 mV	10 ms	Instantaneous (1 ms)
NON	Up to 200	↑ 40 mV, ↓ 40 mV, or 0 mV	10 ms	Whole simulation

To better understand the behavior of these cyberattacks and the parameters indicated, Figure 10 depicts a raster plot per cyberattack with the evolution of neuronal spikes simulating the biological connectome during a simplified simulation of 215 ms instead of 27 seconds, aiming to improve its visibility. A simulation of 215 ms has been chosen since it is the minimum duration to clearly present SCA and FOR cyberattacks, attacking one neuron per millisecond. Particularly, this figure allows the visual comparison between each cyberattack and the spontaneous behavior. Besides, it is worth noting that this figure does not intend to exhaustively present the impact and evolution of the cyberattacks on neural activity but just illustrate their action mechanisms in a simplified way. Those considerations are later presented in this section.

As can be seen in Figure 10, the first raster plot, representing the spontaneous behavior, presents vertical columns of green dots corresponding to regular spiking from Izhikevich's model. This spontaneous behavior is also included in the plots presenting each cyberattack to compare their behavior easily. Blue dots indicate

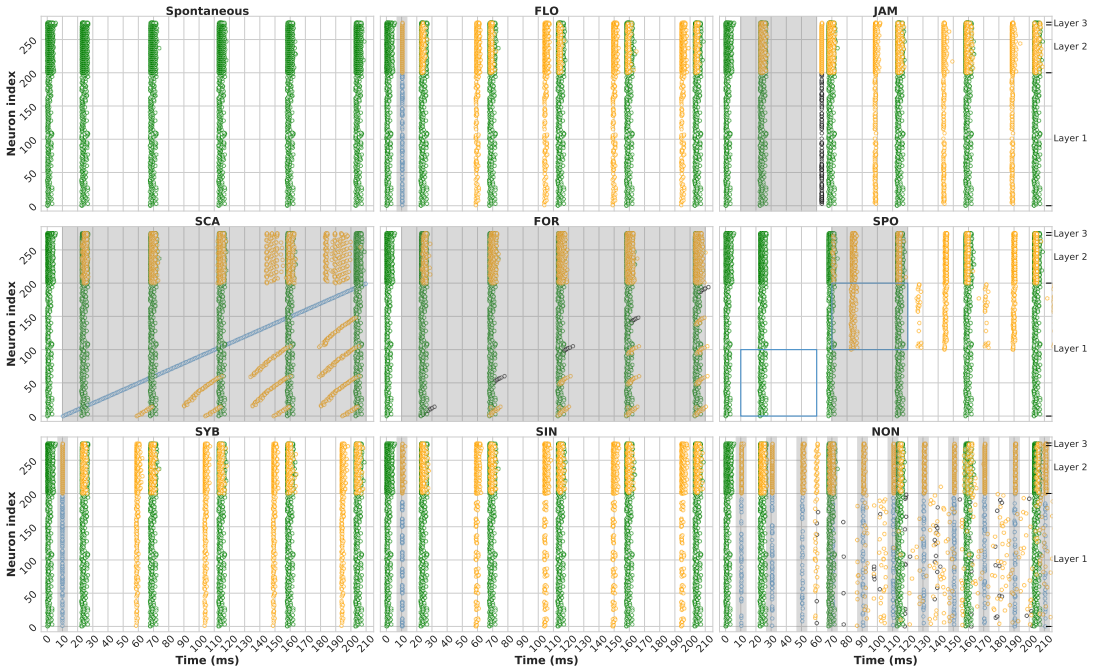


Figure 10: Visual representation of the behavior of each neural cyberattack proposed. Green dots represent neuronal spikes from the spontaneous behavior, blue dots indicate stimulated neurons, black dots inhibited ones, and orange dots highlight the spikes produced as a consequence of the attack. A grey background indicates the duration of the attack.

neurons attacked by neural stimulation, while black dots represent inhibitory actions. Furthermore, orange dots highlight the evolution of each cyberattack. Finally, a grey background indicates the duration of the cyberattack.

Compared to the spontaneous behavior, FLO generates new orange groups of spikes before the spontaneous columns, caused by the stimulation performed at 10 ms. Additionally, orange spikes can be appreciated within the green columns in layers two and three (neurons 200 to 276). These spikes are also a consequence of the attack, applying to subsequent cyberattacks. On the contrary, JAM performs neural inhibition until the instant 60 ms, and it is after that instant when the subset of attacked neurons performs spikes (indicated in black), inducing a delay compared to the spontaneous behavior that is repeated over time as a second column of orange spikes.

Regarding SCA and FOR, both cyberattacks are active during almost all the simulation. However, their impact is quite different. In SCA, a diagonal succession of stimulated neurons can be observed, producing an incremental impact propagated along time. This impact can be appreciated by the apparition of additional diagonal groups of spikes under the diagonal and the anticipation of spikes in the second and third layers. In contrast, FOR only presents small perturbations compared to the spontaneous behavior induced by the implementation considerations already presented in Figure 5. Furthermore, SPO also performs its activity during a temporal window. In this case, there is a clear difference between the behavior of neurons with indexes 100 to 200 compared to the spontaneous behavior caused by the repetition of spikes previously recorded between instants 10 to 60 ms.

Moving to another stimulation cyberattack, SYB presents a similar spikes trend to FLO. This is explained by the voltage range defined by Izhikevich's model, between -65 mV to 30 mV, which introduces a higher probability of stimulating than inhibiting neurons. The instant of attack is also relevant since if a large population of neurons recently performed spikes, the voltage will be low and it will tend to induce stimulation actions. Although the output in terms of spikes is similar, their inner behavior is different.

SIN is another neural cyberattack that also presents similarities with FLO in terms of the visual distribution of spikes. However, it can be seen that there is a particular pattern in the attacked neurons, caused by the real target of the attack: neuron 201, the first neuron of the second layer. In this particular topology, it is determined by the connections between layers of the computational model used. Finally, NON induces a more chaotic behavior when the attack progresses, evaluating the attack condition every 20 ms. As can be

seen, it performs both stimulation and inhibition tasks, randomly selected for each instant under attack and neuron of the first layer.

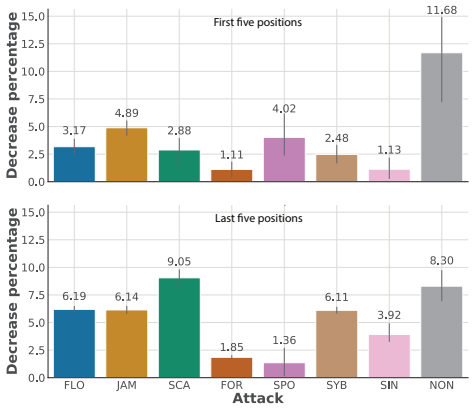


Figure 11: Mean percentage of spikes reduced per neural cyberattack compared with spontaneous behavior, studied over the first and last five positions of the maze in a biological simulation of 27 seconds

.

Once introduced the behavior of each neural cyberattack graphically, Figure 11 depicts the impact caused by each cyberattack compared to spontaneous behavior over a simulation of 27 seconds, indicating the percentage of spikes reduction. This figure shows a differentiation between the first five positions and the last five positions of the optimal path of the maze to find the exit, determining which cyberattacks are more harmful in the short term and which are more suitable for long-term attacks.

The variability presented per cyberattack corresponds to the differences between the five positions considered, either the first positions or the last ones. Moreover, for FLO, JAM, and SYB, which randomly select the target neurons, ten executions are performed to offer variability. Interestingly, the data presented for NON only contains one execution since this attack introduces itself huge randomness and would be difficult to compare.

Regarding these results, NON, due to its random behavior, achieves a reduction of almost 12% over spontaneous activity in the first five positions, being the most damaging cyberattack in the short term, followed by JAM with almost a 5% of reduction. In contrast, SCA is the most impacting attack for the long term, causing a spike reduction of around 9%, followed by NON with a reduction of 8%.

To conclude, it is essential to mention that the metric concerning the number of spikes has been selected for this impact analysis due to its relevance on a wide variety of neurological scenarios. Specifically, the amount of neuronal activity, measured as the number of spikes of a neuronal population, could be helpful to evaluate the impact of certain neurological diseases. As an example, both Amyotrophic Lateral Sclerosis (ALS) and epilepsy naturally generate hyperexcitability of neuronal activity. In this direction, a cyberattack based on neural stimulation, such as FLO, could hypothetically disrupt the natural equilibrium between neuronal excitation and inhibition, recreating or aggravating the disease. On the contrary, neural cyberattacks generating neural inhibition like JAM could recreate conditions such as Alzheimer's disease. Based on that, this publication considers the number of spikes an essential metric to evaluate the damage caused by a cyberattack.

In terms of the generalization of results, these neural cyberattacks have been evaluated over a simplistic and static network with a limited variability compared to the biological visual cortex. In this sense, future work is required to assess their impact on multiple topologies. Moreover, and although the study of the applicability and impact of neural cyberattacks to induce particular neurological conditions is a promising research field, future work is needed to evaluate if our results are consistent with experimentation over realistic biological topologies and even in vivo studies. Additionally, the study of the human-level impact attending to different dimensions, such as psychology or ethics, is out of the scope of this work (for further read, see [4]).

5 CONCLUSION

Novel BCI generations bring countless benefits to society, improving their capabilities to offer better recording and stimulation resolutions. Moreover, the authors envision a future where the reduction in electrode size will derive in a broad coverage of the brain with single-neuron resolution. Although these improvements represent a paradigm change, vulnerabilities in these technologies open the door for cyberattacks to cause physical damage to users.

Based on the previous concerns, this work presents a taxonomy of eight neural cyberattacks aiming to disrupt spontaneous neural activity by maliciously inducing neuronal stimulation or inhibition, exploring the possibility of recreating the effects of particular neurodegenerative conditions. In this sense, two groups of cyberattacks are defined, either based on performing the attack at a particular instant or during a temporal window. These cyberattacks have been evaluated over a neuronal topology modeling a particular region of a mouse's visual cortex. Since there is a lack of realistic neuronal topologies nowadays, and following current literature, a convolutional neural network has been trained to surpass this limitation due to their similarities with biological ones.

The impact of each cyberattack has been measured and compared over a common neural topology, being Neural Nonce and Neural Jamming the most damaging cyberattacks in the short term, causing a spike reduction of around 12% and 5% over spontaneous signaling, respectively. In contrast, Neural Scanning and Neural Nonce are more suitable for long-term damage, causing an approximate spike reduction of 9% and 8%, respectively.

ACKNOWLEDGMENTS

This work has been partially supported by (a) Bit & Brain Technologies S.L. under the project CyberBrain, associated with the University of Murcia (Spain), by (b) the Swiss Federal Office for Defense Procurement (armasuisse) with the CyberSpec (CYD-C-2020003) project, and by (c) the University of Zürich UZH. We thank Blausen Medical[1] and Harryarts[2] from their publicly available images.

[1]https://doi.org/10.15347/wjm/2014.010
[2]https://www.freepik.com/free-vector/lineal-brain-design_841425.htm

REFERENCES

[1] Tamara Bonaci, Ryan Calo, and Howard Jay Chizeck. 2015. App Stores for the Brain : Privacy and Security in Brain-Computer Interfaces. *IEEE Technology and Society Magazine* 34, 2 (Jun 2015), 32–39.

[2] Leela Krishna Bysani and Ashok Kumar Turuk. 2011. A Survey on Selective Forwarding Attack in Wireless Sensor Networks. In *2011 International Conference on Devices and Communications (ICDeCom)*. 1–5.

[3] Carmen Camara, Pedro Peris-Lopez, and Juan E. Tapiador. 2015. Security and privacy issues in implantable medical devices: A comprehensive survey. *Journal of Biomedical Informatics* 55 (2015), 272–289.

[4] Tamara Denning, Yoky Matsuoka, and Tadayoshi Kohno. 2009. Neurosecurity: security and privacy for neural devices. *Neurosurgical Focus FOC* 27, 1 (2009), E7.

[5] John R. Douceur. 2002. The Sybil Attack. In *Peer-to-Peer Systems*, Peter Druschel, Frans Kaashoek, and Antony Rowstron (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 251–260.

[6] Christine A. Edwards, Abbas Kouzani, Kendall H. Lee, and Erika K. Ross. 2017. Neurostimulation Devices for the Treatment of Neurologic Disorders. *Mayo Clinic Proceedings* 92, 9 (Sept. 2017), 1427–1444.

[7] Eyal Gal, Michael London, Amir Globerson, Srikanth Ramaswamy, Michael W. Reimann, Eilif Muller, Henry Markram, and Idan Segev. 2017. Rich cell-type-specific network topology in neocortical microcircuitry. *Nature Neuroscience* 20, 7 (Jul 2017), 1004–1013.

[8] Daniel Halperin, Thomas S. Heydt-Benjamin, Benjamin Ransford, Shane S. Clark, Benessa Defend, Will Morgan, Kevin Fu, Tadayoshi Kohno, and William H. Maisel. 2008. Pacemakers and Implantable Cardiac Defibrillators: Software Radio Attacks and Zero-Power Defenses. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*. 129–142.

[9] Christian J. Hartmann, Sabine Fliegen, Stefan J. Groiss, Lars Wojtecki, and Alfons Schnitzler. 2019. An update on best practice of deep brain stimulation in Parkinson's disease. *Therapeutic Advances in Neurological Disorders* 12 (Jan 2019), 1756286419838096.

[10] E. M. Izhikevich. 2003. Simple model of spiking neurons. *IEEE Transactions on Neural Networks* 14, 6 (2003), 1569–1572.

[11] Ioannis Kalfas, Satwant Kumar, and Rufin Vogels. 2017. Shape Selectivity of Middle Superior Temporal Sulcus Body Patch Neurons. *eNeuro* 4, 3 (2017).

[12] Eric Kandel. 2013. *Principles of neural science*. McGraw-Hill, New York.

[13] Nikolaus Kriegeskorte. 2015. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science* 1, 1 (2015), 417–446.

[14] Ilya Kuzovkin, Raul Vicente, Mathilde Petton, Jean-Philippe Lachaux, Monica Baciu, Philippe Kahane, Sylvain Rheims, Juan R. Vidal, and Jaan Aru. 2018. Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Communications Biology* 1, 1 (Aug 2018), 107.

[15] Grace W. Lindsay. 2021. Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience* 33, 10 (09 2021), 2017–2031.

[16] Sergio López Bernal, Alberto Huertas Celdrán, Gregorio Martínez Pérez, Michael Taynnan Barros, and Sasitharan Balasubramaniam. 2021. Security in Brain-Computer Interfaces: State-of-the-Art, Opportunities, and Future Challenges. *Comput. Surveys* 54, 1, Article 11 (Jan. 2021), 35 pages.

[17] Sergio López Bernal, Alberto Huertas Celdrán, Lorenzo Fernández Maimó, M. T. Barros, S. Balasubramaniam, and G. Martínez Pérez. 2020. Cyberattacks on Miniature Brain Implants to Disrupt Spontaneous Neural Signaling. *IEEE Access* 8 (2020), 152204–152222.

[18] Sergio López Bernal, Alberto Huertas Celdrán, and Gregorio Martínez Pérez. 2022. Neuronal Jamming cyberattack over invasive BCIs affecting the resolution of tasks requiring visual capabilities. *Computers & Security* 112 (2022), 102534.

[19] Ivan Martinovic, Doug Davies, Mario Frank, Daniele Perito, Tomas Ros, and Dawn Song. 2012. On the Feasibility of Side-Channel Attacks with Brain-Computer Interfaces. In *Proceedings of the 21st USENIX Conference on Security Symposium* (Bellevue, WA) *(Security'12)*. USENIX Association, USA, 34.

[20] Elon Musk. 2019. An Integrated Brain-Machine Interface Platform With Thousands of Channels. *Journal of Medical Internet Research* 21, 10 (Oct 2019), e16194.

[21] Aqeel-ur Rehman, Sadiq Ur Rehman, and Haris Raheem. 2019. Sinkhole Attacks in Wireless Sensor Networks: A Survey. *Wireless Personal Communications* 106, 4 (01 Jun 2019), 2291–2313.

[22] William Stallings. 2017. *Cryptography and Network Security: Principles and Practice* (7 ed.). Pearson, London. 766 pages.

[23] Marcel Stimberg, Romain Brette, and Dan FM Goodman. 2019. Brian 2, an intuitive and efficient neural simulator. *eLife* 8 (Aug. 2019), e47314.

[24] Hassan Takabi, Anuj Bhalotiya, and Manar Alohaly. 2016. Brain computer interface (BCI) applications: Privacy threats and countermeasures. In *IEEE 2nd International Conference on Collaboration and Internet Computing*. IEEE, Pittsburgh, PA, USA, 102–111.

[25] Satish Vadlamani, Burak Eksioglu, Hugh Medal, and Apurba Nandi. 2016. Jamming attacks on wireless networks: A taxonomic survey. *International Journal of Production Economics* 172 (2016), 76–94.

[26] Stefanus Arinno Wirdatmadja, Michael Taynnan Barros, Yevgeni Koucheryavy, Josep Miquel Jornet, and Sasitharan Balasubramaniam. 2017. Wireless Optogenetic Nanonetworks for Brain Stimulation: Device Model and Charging Protocols. *IEEE Transactions on NanoBioscience* 16, 8 (2017), 859–872.