UNIVERSIDAD DE MURCIA

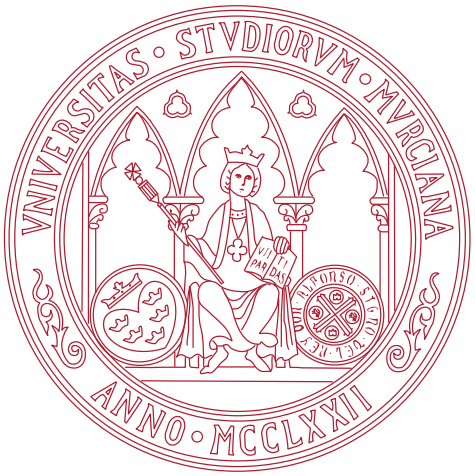
ESCUELA INTERNACIONAL DE DOCTORADO

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.

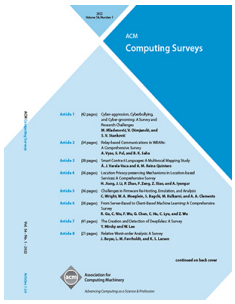
[29] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, aug 2019. ISBN 1492032646

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

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

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

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.

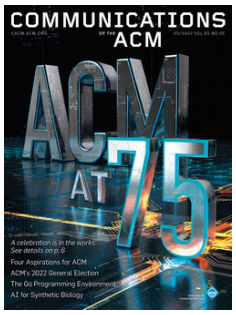
Sincerely,
James R. Larus
Editor-in-Chief, Communications of the ACM (CACM)
Professor, EPFL, Lausanne Switzerland

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