





Análisis del impacto de ciberataques neuronales aplicados a la visión

Victoria Magdalena López Madejska *, Sergio López Bernal *, Gregorio Martínez Pérez*,
Alberto Huertas Celdrán†

*Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, 30100 Murcia, España
{victoriamadgalena.lopezm, slopez, gregorio}@um.es

†Communication Systems Group CSG, Department of Informatics IfI, at the University of Zurich UZH,
CH 8050 Zürich, Switzerland
huertas@ifi.uzh.ch

Resumen—Las interfaces cerebro-máquina (BCIs) son sistemas que interactúan con el cerebro para obtener información cerebral o realizar neuroestimulación, usados en medicina y videojuegos. Pese a sus ventajas, las BCIs invasivas pueden ser vulneradas para realizar ciberataques neuronales, alterando la actividad neuronal. Sin embargo, estos ataques solo se han validado en simulaciones con pocas neuronas, sin considerar su impacto real. Este trabajo analiza el impacto de dos ciberataques existentes en la literatura, Neuronal Flooding y Neuronal Jamming, sobre una topología neuronal compleja de la corteza visual, con unas 230.000 neuronas. Se descubre que ambos causan gran impacto en el número de spikes, siendo Neuronal Jamming más efectivo. Incluso un pequeño número de neuronas atacadas puede provocar un impacto significativo. También se observa que las neuronas son influenciadas por el estímulo visual incluso después de los ciberataques. Además, se verifica si las conclusiones de trabajos menos realistas coinciden con las obtenidas en este estudio.

Index Terms—Interfaces Cerebro-Máquina, Ciberseguridad, Seguridad, Neurociencia, Estímulos visuales, Ciberataques Neuronales

Tipo de contribución: *Investigación en desarrollo*

I. INTRODUCCIÓN

Las interfaces cerebro-máquina, o BCIs, (del inglés, Brain-computer interfaces) son sistemas bidireccionales capaces de interactuar directamente con el cerebro, permitiendo monitorizar las señales cerebrales producidas por el individuo y estimular o inhibir la actividad neuronal. Estas tecnologías se clasifican según si es necesario una cirugía para implementar electrodos directamente en el cerebro (invasivas) o si se utiliza una malla de electrodos sobre el cuero cabelludo (no invasivas). La primera es útil en la adquisición de información cerebral en entornos médicos específicos. Por ejemplo, para el tratamiento de enfermedades neurodegenerativas, como en el caso de la enfermedad de Parkinson. En cambio, la segunda ha ganado popularidad en ámbitos fuera del diagnóstico médico, en concreto en escenarios de entretenimiento y videojuegos. Este último tipo utiliza técnicas como la encefalografía (EEG) para recopilar datos cerebrales. Las BCIs han ganado gran popularidad en los últimos años, siendo un ejemplo a destacar Neuralink [1], una compañía centrada en la creación de interfaces implantables en miniatura capaces de ofrecer una cobertura amplia de neuronas con riesgo menor durante la cirugía. Así, esta empresa investiga interfaces capaces de leer y estimular las neuronas de forma individual, no solo para

tratar a pacientes con desórdenes neurológicos, sino con el objetivo de democratizar el acceso a la neurotecnología y permitir que cualquier persona pueda acceder a estos sistemas. Neuralink ha implantado exitosamente su prototipo en un humano, demostrando la proyección futura de estas tecnologías. Pese a las grandes ventajas que ofrecen las BCIs, no están exentas de problemas de seguridad y privacidad. Por otro lado, los fabricantes están abaratando los costes de producción y disminuyendo su tamaño, dejando al descubierto vulnerabilidades en estos dispositivos. A raíz de ello, la literatura se ha centrado en documentar las vulnerabilidades que afectan a la integridad, la confidencialidad, la disponibilidad y la seguridad física de la persona. Estos problemas pueden perjudicar el correcto funcionamiento de los sistemas, la recopilación de datos, o permiten que el atacante tome el control sobre estos dispositivos. No obstante, estos estudios están centrados en determinados aspectos de las BCIs, obviando otras preocupaciones dentro de estos dispositivos. Basándose en estas limitaciones, la literatura reciente ha propuesto una nueva familia de ciberataques llamada ciberataques neuronales, después de que varios estudios hayan documentado riesgos de ciberseguridad previamente no considerados [2], [3].

En concreto, la literatura ha propuesto una topología de ocho ciberataques, inspirados en ataques conocidos en el área de la seguridad en comunicaciones, capaces de alterar la actividad neuronal espontánea con diferentes comportamientos maliciosos [4]. En una primera instancia, dichos ciberataques neuronales fueron simulados sobre una topología neuronal no realista obtenida de entrenar una red neuronal convolucional simulando la corteza visual de un ratón. Este enfoque se planteó debido a una falta de topologías neuronales realistas en el momento de realizar la investigación. Además, se plantearon métricas capaces de cuantificar el impacto causado por dichos ataques en relación con la actividad neuronal. Sin embargo, la literatura reciente ha propuesto la aplicación de ciberataques neuronales sobre una topología neuronal realista simplificada de 450 neuronas [5]. El modelo usado consiste en la reconstrucción de la corteza visual primaria del ratón, concretamente, de la capa cuatro. El estudio cuantifica el impacto que podrían tener estos ciberataques sobre la visión. Además, los autores compararon sus resultados con los de la literatura, destacando, principalmente, las diferencias entre modelos neuronales, el número de neuronas afectadas y el

impacto generado por los ciberataques. Aunque la investigación introduce un paso más hacia un escenario realista en comparación con la literatura, existe todavía una falta de estudios sobre el impacto de estas amenazas en el mundo real que podrían afectar a las funciones neuronales, como la visión. Además, el estado del arte destaca la necesidad de utilizar topologías grandes y complejas que consideren una relación con su medio externo, en su caso, estímulos visuales. El presente estudio trata de superar dichas limitaciones y ofrecer nuevas orientaciones sobre los impactos que conllevan los ciberataques neuronales. En base a ello, este trabajo presenta las siguientes contribuciones principales:

- Implementación de los ciberataques Neuronal Flooding (FLO) y Neuronal Jamming (JAM), previamente definidos en la literatura, en una topología neuronal consistente en la reconstrucción de la corteza visual primaria (V1) del ratón de aproximadamente 230.000 neuronas distribuidas en las seis capas de V1. Por el contrario, la literatura existente empleaba una taxonomía simplificada de 450 neuronas distribuidas en la capa cuatro de V1. Además, en esta investigación se simula a nivel neuronal que el ratón ve un flash de luz a nivel de estímulo visual para estudiar el impacto que tienen los ataques de forma realista. Sin embargo, para la red de 450 neuronas no se utilizaron estímulos realistas a nivel de simulación.
- Análisis de los resultados en base a diferentes experimentos. Se hace uso de dos métricas, el número spikes y la dispersión temporal. La primera indica el número de activaciones (potenciales de acción, o spikes) por neurona, mientras que la segunda muestra el porcentaje de instantes de tiempo en los que hubo spikes. Usando FLO, se consigue un aumento de spikes de entre 5.000 y 53.000 sobre el comportamiento espontáneo, según el número de neuronas e instantes de tiempo atacados. Por su parte, JAM disminuye entre diez spikes, como máximo y 2864 spikes como mínimo, dependiendo de las neuronas afectadas y el intervalo bajo ataque. Aunque ambos tienen un impacto considerable, el más efectivo es JAM respecto al número de spikes. Por otro lado, el tiempo de propagación del ciberataque varía según el instante de ataque, ya que las neuronas afectadas tratan de volver a su comportamiento espontáneo tras el ataque.
- Estudio comparativo entre los resultados de esta investigación y los de la literatura. Este trabajo presenta un análisis de tres instantes de tiempo diferentes, mientras que la literatura solo se centraba en un instante específico. Por lo tanto, el efecto de propagación de los ciberataques varía en ambos estudios. En la literatura, FLO tiene una duración de 500 ms y JAM de 600 ms, mientras que en este estudio, la duración del primer ciberataque varía entre 200 y 300 ms, y en el segundo se obtuvo una duración de 800 ms o 400 ms, dependiendo del momento del ataque. Por otro lado, en la literatura se observa que a mayor número de neuronas afectadas, mayor número de spikes. Sin embargo, este trabajo ha demostrado que estos ciberataques pueden generar un gran impacto al atacar un menor número de neuronas, evidenciando así el impacto que los ciberataques neuronales pueden causar en la actividad neuronal espontánea.

Este artículo está organizado de la siguiente forma. La Sección II documenta brevemente el estado del arte sobre la ciberseguridad de las BCIs y la literatura sobre ciberataques en topologías neuronales realistas. A continuación, la Sección III detalla minuciosamente la taxonomía del escenario utilizado y sus componentes fundamentales en la actual investigación. Posteriormente, la Sección IV muestra la configuración de los ciberataques implementados con su explicación correspondiente y, seguidamente, los resultados obtenidos de cada experimento, analizados en base a las métricas definidas. En la Sección V se comparan los resultados de ambos ciberataques y, después, con los presentados en la literatura. Finalmente, la Sección VI presenta las conclusiones y trabajo futuro.

II. ESTADO DEL ARTE

La mayoría de estudios que abordan la ciberseguridad de las BCIs se han centrado en los problemas que afectan a la integridad, la confidencialidad, la disponibilidad y la seguridad física de los usuarios. Comenzando por la integridad, Li et al. [6], detectaron que es posible recopilar señales cerebrales para, posteriormente, suplantarlas por otras. Así mismo, Martínez Beltrán et al. [7] demostraron que al introducir ruido a las ondas cerebrales se puede confundir a los clasificadores usados para identificar aspectos relevantes de las señales cerebrales. Respecto a la confidencialidad, Martinovic et al. [8], aprovechando las respuestas cerebrales de los usuarios tras visualizar estímulos visuales maliciosos, obtuvieron información sensible como tarjetas de débito, contraseñas, zona de residencia y creencia religiosa. Frank et al. [9] presentaron estímulos visuales subliminales incluidos en un vídeo, adquiriendo información de los usuarios, y vulnerando su privacidad. Por su parte, Takabi et al. [10] identificaron las vulnerabilidades más comunes en las aplicaciones para las BCIs, permitiendo acceder a los datos cerebrales sin ningún tipo de restricción. A su vez, Ienca et al. [11] y Li et al. [6] identificaron la posibilidad de vulnerar la disponibilidad del servicio, al alterar la adquisición de datos mediante el uso de diferentes vectores de ataque. En relación con la seguridad física y, concretamente, los dispositivos médicos implantables, Pycroft et al. [12] destacaron el posible daño en el tejido cerebral o efecto rebote al sobreestimar el cerebro, alterando el tratamiento del usuario. Por otro lado, Marin et al. [13] remarcaron que la manipulación de dichos dispositivos podría afectar al habla o al movimiento causando daño cerebral, impactando incluso a la propia vida. En la Tabla I se muestran los trabajos mencionados, destacando la dimensión del impacto que pueden generar (integridad, confidencialidad, disponibilidad o seguridad física) y una breve descripción de las amenazas. A pesar de la variedad de trabajos centrados en la ciberseguridad en BCIs, la literatura es escasa y solo cubre algunos aspectos de las BCIs. Por ello, López Bernal et al. [2], [3] realizaron un análisis exhaustivo de la literatura existente para identificar los problemas de seguridad, las limitaciones y retos futuros de los sistemas de neuroestimulación de nueva generación, siendo Neuralink [1] un ejemplo representativo. Asimismo, la falta de trabajos relacionados con las vulnerabilidades de estos sistemas ofreció una oportunidad a los autores para desarrollar una nueva familia de amenazas denominada ciberataques neuronales. Estos ataques son capaces de sobreestimar o inhibir de

Tabla I: Comparativa de artículos centrados en la ciberseguridad de las BCIs, categorizadas por dimensión del impacto causado.

Autores	Impacto	Amenaza
Martinovic et al. [8]	Confidencialidad	Estímulos visuales maliciosos
Li et al. [6]	Disponibilidad Integridad	Alterar adquisición de datos Suplantar señales cerebrales
Ienca et al. [11]	Disponibilidad	Alterar adquisición de datos
Pycroft et al. [12]	Seguridad física	Daño en el tejido cerebral
Takabi et al. [10]	Confidencialidad	Acceder a los datos cerebrales
Frank et al. [9]	Confidencialidad	Estímulos visuales maliciosos
Marin et al. [13]	Seguridad física	Daño en el tejido cerebral
Martínez Beltrán et al. [7]	Integridad	Introducir ruido en ondas cerebrales

forma individual aquellas neuronas accesibles por las BCIs. Al principio se diseñaron e implementaron dos ciberataques, Neuronal Flooding (FLO) y Neuronal Scanning (SCA) [3], capaces de sobreestimar un conjunto de neuronas en un instante concreto, o bien estimular secuencialmente a lo largo de un periodo de tiempo, respectivamente. Posteriormente, presentaron Neuronal Jamming (JAM) [14] como un nuevo ciberataque neuronal capaz de impedir la actividad neuronal durante un intervalo de tiempo. Por último, los mismos autores definieron una taxonomía de ocho ciberataques neuronales [4], incluidos los mencionados anteriormente, con diferentes comportamientos. Todos ellos se validaron sobre una topología artificial simplificada de la corteza visual del ratón. Debido a una falta de modelos neuronales realistas en el momento de realizar la investigación, los autores optaron por entrenar una red neuronal convolucional (CNN, del inglés, Convolutional Neural Network), cuyas conexiones y pesos se tradujeron posteriormente a parámetros de una red neuronal biológica simulada. La decisión de usar de base una CNN vino motivada por las similitudes a nivel de estructura y función que tienen las redes mencionadas en relación a la corteza visual biológica. Además, estos trabajos solo consideraron una población neuronal de tipo excitadora, utilizando una simplificación del proceso visual como entrada de la simulación neuronal.

Sin embargo, han surgido en los últimos años trabajos en la literatura centrados en reconstruir de forma realista topologías neuronales concretas del cerebro. Debido a ello, el estudio de López Madejska et al. [5] implementó ciberataques neuronales sobre una reconstrucción de la corteza visual primaria (V1) del ratón, concretamente de la capa cuatro, publicada por Arkhipov et al. [15]. La red está compuesta por aproximadamente 45.000 neuronas con dos niveles diferentes de detalle en cuanto a la simulación de las neuronas, utilizando como simulador NEURON [16] o NEST [17] en función de la resolución usada. Por otro lado, los modelos fueron creados y gestionados por la herramienta Brain Modeling Toolkit (BMTK), desarrollada por *Allen Institute* [18]. Para simplificar el problema de realizar los experimentos, los autores del estudio escogieron una topología pequeña de 450 neuronas proporcionada por los creadores del modelo completo. Además, utilizaron el simulador NEST [17] y valores estáticos facilitados por BMTK [18] como estímulo visual externo. Los resultados obtenidos tras implementar los ataques FLO y JAM aportaron nuevos enfoques en el campo de la

ciberseguridad de las BCIs y podrían servir para mejorar en el futuro los tratamientos de enfermedades neurodegenerativas, concretamente en el campo de la visión.

III. DISEÑO GENERAL DE LA SOLUCIÓN PROPUESTA

Esta sección presenta la topología usada en este trabajo, destacando sus componentes fundamentales y los pasos seguidos para obtener y analizar la actividad neuronal resultante después de los ciberataques.

La topología neuronal corresponde con la documentada por Billeh et al. [19], quienes ofrecen una reconstrucción completa de un microcircuito de la corteza visual primaria del ratón, compuesta por unas 230.000 neuronas con diferentes comportamientos, tanto excitadores como inhibidores. Al igual que en el estudio de Arkhipov et al. [15] mencionado en la Sección II, ofrecen dos niveles de granularidad. La primera alternativa ofrece un enfoque detallado a nivel biofisiológico, mientras que la segunda plantea una simplificación del comportamiento de las neuronas siguiendo un enfoque point-neuron en base a neuronas Gated Leaky Integrate and Fire (GLIF) [20]. Ambos modelos tienen que ser creados y generados con la herramienta BMTK [18]. Respecto a su simulación, se ejecutan en NEURON [16] y en NEST [17], respectivamente. También, este modelo simula el núcleo geniculado lateral (LGN, del inglés Lateral Geniculate Nucleus), componente cerebral encargado de procesar el estímulo externo utilizado y remitirlo a V1. En este estudio se escogió el modelo simplificado simulado en NEST [17] puesto que requiere menos recursos computacionales y ofrece una aproximación suficientemente buena. La duración de la simulación está fijada en tres segundos (3000 ms) con una resolución de 0.25 ms, siendo suficiente para evaluar el impacto y la propagación de los ciberataques. Por otro lado, esta topología utiliza estímulos visuales realistas, contando con imágenes, películas, efecto flash, efecto de un disco acercándose y, por último, drifting gratings, que corresponden a efectos visuales en forma de rejilla disponible en ocho direcciones y cinco frecuencias temporales.

Además, esta simulación utiliza otra entrada llamada *background* que representa la información recibida por todas las neuronas simuladas de regiones del cerebro externas a la simulada. Esta información sigue una distribución de *Poisson* a 1kHz. Los autores proporcionan diez ensayos por cada estímulo visual mencionado en el anterior párrafo, y otros 100 ensayos para el *background*, ya que cada ensayo del estímulo está asociada a una prueba del *background*. Estos ensayos tienen como objetivo ofrecer una variabilidad suficiente de ejecuciones de la red neuronal. Para este estudio se utilizó solo un estímulo visual de tipo flash, debido a que ofrece muchas posibilidades de experimentación y explica el comportamiento de las neuronas ante estímulos simples, concretamente, estímulos visuales blancos, negros y grises. El efecto del estímulo total dura 2500 ms, empezando por una pantalla gris durante 500 ms, seguido por 250 ms de estímulo visual blanco (ON-flash). Tras ello se vuelve a mostrar la pantalla gris por 1000 ms, continuando con 250 ms de pantalla negra (OFF-flash). Finalmente, se presenta una pantalla gris de 500 ms. A continuación se muestran los ciberataques implementados definidos en la literatura [4]:

- El ciberataque **FLO**: Consiste en la sobreestimulación

de un conjunto de neuronas en un determinado instante de tiempo.

- El ciberataque **JAM**: Impide la actividad neuronal de un número de neuronas objetivo durante una ventana temporal.

La Figura 1 muestra una representación visual de la topología usada, destacando las poblaciones neuronales involucradas, la capa de la corteza V1 a la que pertenecen y el comportamiento de las neuronas (excitadoras o inhibitorias). Por otro lado, esta figura especifica la implementación de los ciberataques neuronales y cómo estos afectan a la actividad de la corteza visual primaria. Cabe destacar que la simulación de V1 ofrece como resultados los spikes generados, guardados en dos formatos diferentes, CSV y SONATA, siendo este último un formato creado por los propios autores. Para poder visualizar toda la actividad neuronal resultante, la herramienta BMTK [18] ofrece crear una gráfica de tipo “raster plot”, donde cada spike es representado por un punto y está ordenado por capas de V1 y población neuronal. En esta representación, el eje X indica el tiempo de la simulación en milisegundos y el eje Y muestra los identificadores (Id) de las neuronas.

IV. CIBERATAQUES NEURONALES SOBRE LA TOPOLOGÍA NEURONAL ESTUDIADA

Esta sección define la implementación de los ciberataques mencionados en la Sección III y, a su vez, la explicación de los parámetros utilizados. Posteriormente, se analiza el impacto causado sobre los comportamientos espontáneos basados en las métricas usadas en este estudio. Es relevante destacar que los experimentos realizados han sido validados por expertos en anatomía cerebral de la Universidad de Murcia, destacando la viabilidad de dichas amenazas en entornos realistas.

IV-A. Configuración de los ciberataques

Para analizar la efectividad de los ataques en la topología usada, se ha configurado los ciberataques con parámetros preestablecidos, siguiendo la definición de la literatura [4]. En ambos se estudia el número de neuronas atacadas de forma aleatoria. Debido a la complejidad del propio cerebro, se seleccionaron la mitad (115.462) y un cuarto (57.731) del total de las neuronas de la topología, causado por la imposibilidad de atacar a todas ellas desde el punto de vista de un atacante. Para estudiar la variabilidad de la red neuronal tras realizar uno de los ataques, se repite cada experimento diez veces con el objetivo de examinar si entre ejecuciones existe una diferencia notable. A su vez, se estudió el comportamiento espontáneo de un solo ensayo del estímulo flash (trial número 9) con las posibles diez pruebas del *background* (trials desde 90 hasta 99) para comprobar si existe una diferencia apreciable entre ellos o si, por el contrario, son tan similares que sería posible seleccionar cualquiera. Como se aprecia en la Figura 2, la mediana de la distribución del número de spikes varía entre las pruebas, pero todas ellas oscilan dentro de un rango específico (10.000-20.000 spikes). En base a ello, se ha escogido el último ensayo, dado que presenta un mayor espectro de spikes y corresponde al ensayo usado por los autores de la topología para ejemplificar el comportamiento del efecto flash seleccionado. En cuanto a los instantes de ataque, es necesario hacer una distinción entre los ciberataques neuronales. Empezando por FLO, se realiza

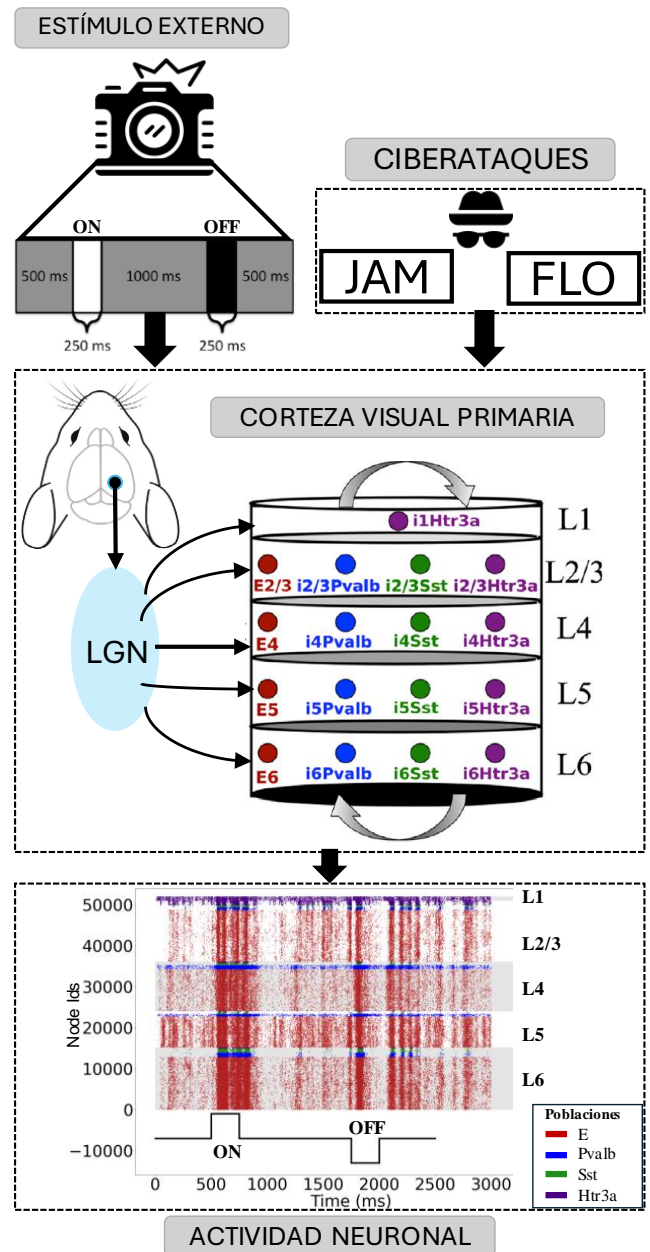


Figura 1: Representación gráfica de la topología utilizada, del estímulo visual externo, de la actividad neuronal obtenida, y de los ciberataques neuronales implementados.

durante tres instantes diferentes, en la pantalla blanca, negra y gris. El primero corresponde al ON-flash que empieza en el instante 500 ms y termina en el 750 ms, siendo el instante de ataque el 625 ms. Por su parte, la pantalla negra (OFF-flash) comprende desde el instante 1750 ms hasta los 2000 ms, por lo que el momento de ataque se ajustó al 1875 ms. En caso de la pantalla gris, el estímulo externo alberga tres, pero se seleccionó el que se encuentra entre el ON/OFF-flash debido a que tiene una mayor duración y, en consecuencia, permite evaluar mejor la propagación del ciberataque. En base a ello, se fijó el instante de ataque para este estímulo en 1000 ms. Por otro lado, los ataques de JAM están comprendidos en una

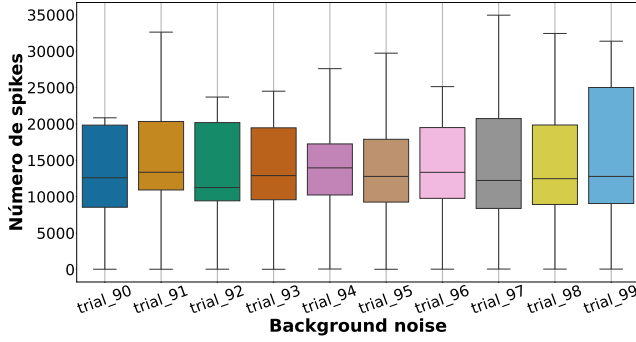


Figura 2: Distribución del número de spikes por cada prueba del *background* ofrecida por los autores de la topología neuronal.

determinada ventana temporal. Por ello, al igual que en el FLO, se escogieron los mismos momentos, con la diferencia de que el ataque se realiza durante un intervalo de 250 ms. En este caso, los intervalos bajo ataque son durante la ejecución del ON-flash (500-750 ms), del OFF-flash (1750-2000 ms) y dentro de la pantalla gris intermedia (1000-1250 ms). Por último, los ataques FLO realizados consisten en establecer el voltaje de las neuronas afectadas al voltaje umbral (V_{th}) en el que la neurona produce un spike. En el caso de JAM, se establece el voltaje en el valor mínimo posible (V_{reset}) para evitar la actividad de las neuronas objetivo durante la ventana de acción del ataque. Estos últimos parámetros son escogidos tras los experimentos realizados por la literatura [3] [5], que indican que los valores de voltaje más cercanos al umbral en caso de FLO y más próximos al voltaje de mínimo en JAM, son los idóneas para causa un efecto mayor. La Tabla II muestra, de forma resumida, los parámetros escogidos en esta investigación por cada ciberataque utilizado.

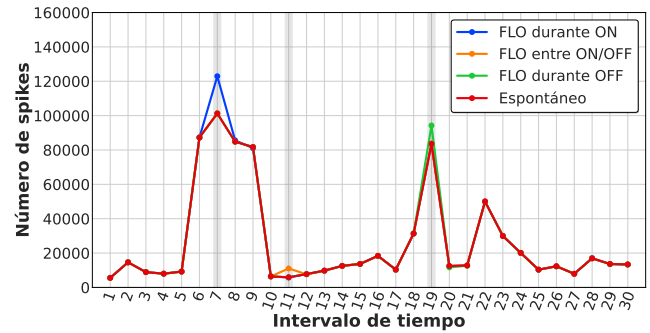
Tabla II: Configuraciones implementadas por cada ciberataque neuronal.

Ataque	Estímulo visual	Background	Instantes bajo ataque	Número de ejecuciones	Número de neuronas atacadas	Voltaje de ataque
FLO	Flash Prueba 9	Prueba 99	625 ms 1000 ms 1875 ms	10	57.731 115.462	V_{th}
JAM	Flash Prueba 9	Prueba 99	500-750 ms 1000-1250 ms 1750-2000 ms	10	57.731 115.462	V_{reset}

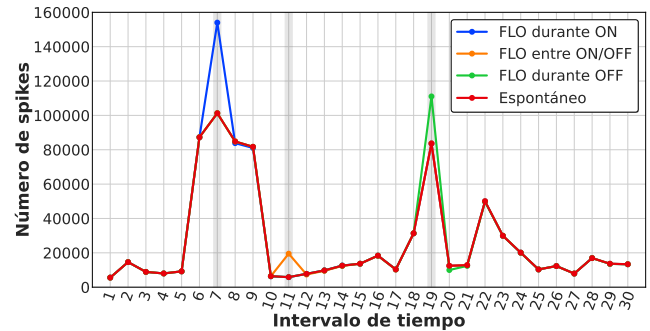
IV-B. Análisis de resultados

Para verificar el impacto de estos ciberataques, en este trabajo se utilizan dos métricas, el número de spikes y la dispersión temporal, ambas definidas en la literatura. La primera consiste en cuantificar el número de activaciones (potenciales de acción) de las neuronas, mientras que la segunda estudia el porcentaje de instantes de tiempo en los que hubo spikes. La Figura 3 muestra el comportamiento de FLO sobre los dos tamaños de neuronas atacadas en los instantes 625 ms, 1000 ms y 1875 ms. La gráfica representa la evolución del ataque ejecutado en el momento donde aparece la pantalla blanca, la pantalla gris intermedia y la pantalla negra, distinguidas por una línea azul, verde y amarilla, respectivamente. Además, se comparan con el comportamiento espontáneo (línea naranja).

Por otra parte, se ha simplificado el eje X a 30 intervalos de 100 ms cada uno para mejorar la visualización de los resultados. Observando dichos resultados, en todos los instantes de ataque el número de spikes aumenta proporcionalmente con respecto al espontáneo. Pese a que, a simple vista, el ataque tiene mayor impacto en el ON-flash, en cada instante las neuronas se comportan de formas diferentes según el estímulo visual presentado, lo que puede causar que reaccionen en mayor o menor medida ante el ataque. Esta situación es común en ambos números de neuronas atacadas. Sin embargo, atacar a la mitad de las neuronas afecta a una cantidad mayor de spikes. Tras el ciberataque, las neuronas tardan en volver a su comportamiento normal alrededor de dos o tres intervalos, unos 200-300 ms. La recuperación del estado espontáneo ocurre en todos los experimentos realizados, dado que la topología está compuesta por neuronas individuales realistas, que tienden a sincronizarse de nuevo con su estado normal en base al estímulo visual mostrado.



(a) Número total de spikes al atacar al cuarto de las neuronas de forma aleatoria en tres instantes diferentes.

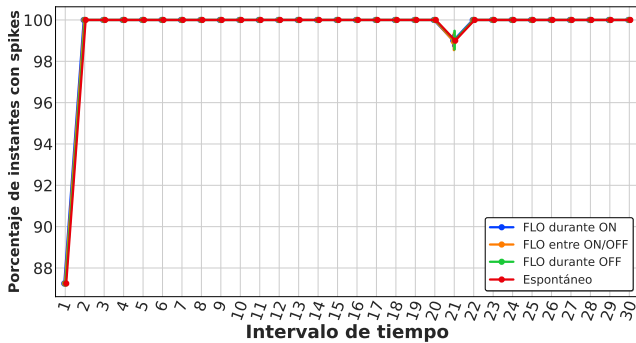


(b) Número total de spikes al atacar a la mitad de las neuronas de forma aleatoria en tres instantes diferentes.

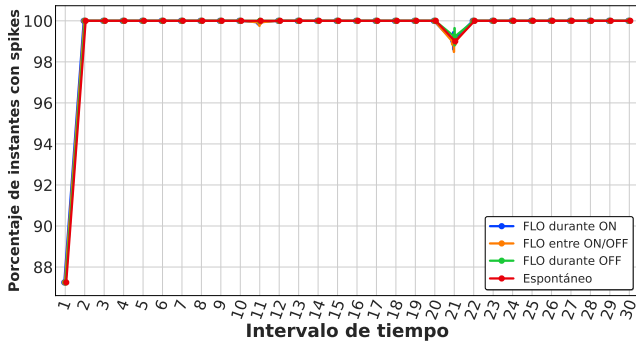
Figura 3: Impacto del ciberataque FLO en base a la métrica del número de spikes, en diferentes instantes de tiempo, y diferentes número de neuronas afectadas, ejecutado diez veces de forma aleatoria por experimento.

En cuanto a la Figura 4, prácticamente en cada intervalo de tiempo, todas las neuronas hacen spikes, a excepción de los intervalos 1 y 21. Además, hay una ligera variabilidad en el intervalo 11 solamente cuando se ejecuta el ataque sobre la mitad del total de las neuronas. Como en la mayoría de los instantes el estado espontáneo presenta un 100% de instantes con spikes, y FLO trata de sobreestimar las neuronas atacadas, se mantienen valores similares al caso espontáneo. La única variabilidad notoria es en el intervalo

21 en que, dependiendo del instante bajo ataque y el número de neuronas atacadas, el porcentaje cambia ligeramente. Estos resultados demuestran que FLO, en un entorno realista, es capaz de aumentar la dispersión temporal, aunque con un impacto limitado.



(a) Número total de spikes al atacar al cuarto de las neuronas de forma aleatoria en tres instantes diferentes.

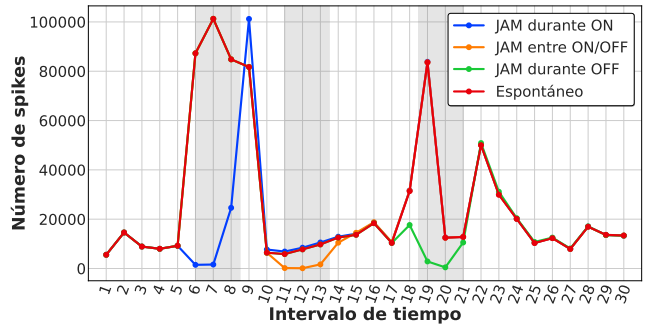


(b) Número total de spikes al atacar a la mitad de las neuronas de forma aleatoria en tres instantes diferentes.

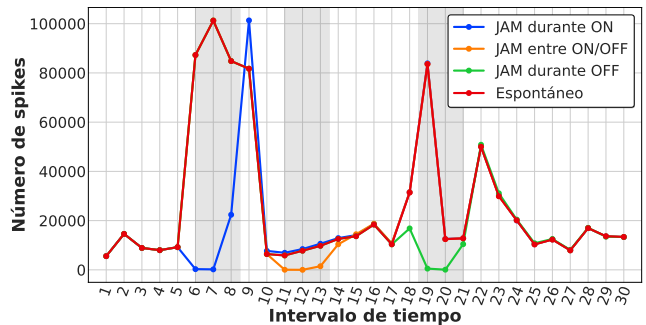
Figura 4: Impacto del ciberataque FLO en base a la métrica de dispersión temporal, en diferentes instantes de tiempo, y diferentes número de neuronas afectadas, ejecutado diez veces de forma aleatoria por experimento.

Por otro lado, el impacto generado por el ciberataque JAM está reflejado en la Figura 5. De igual manera que en el anterior ciberataque, se estudia en base a la cantidad de neuronas atacadas y los instantes bajo ataque. JAM inhibe la actividad neuronal durante 250 ms en los intervalos 500-750 ms (ON-flash), 1000-1250 ms (estímulo gris) y 1750-2000 ms (OFF-flash). Estos intervalos están representados en la gráfica con un sombreado gris para visualizar la duración del ciberataque. En esta figura se puede observar que JAM reduce considerablemente el número de spikes durante el tiempo de actividad del ataque. Una vez que ha finalizado, todas las neuronas inhibidas generan un pico elevado de spikes en el intervalo nueve, debido a que son influidas por el estímulo visual blanco. Posteriormente, se sincronizan con su comportamiento espontáneo, llegando a tardar aproximadamente ocho intervalos en estabilizarse (800 ms). Por contra, si se ataca durante el estímulo gris, tarda en recuperarse unos cuatro intervalos (400 ms), al igual que en el intervalo OFF-flash, donde las neuronas se sincronizan tras cuatro intervalos. Esto se debe a que las neuronas reaccionan según el estímulo visual presentado y además, tras el ciberataque, intentan

sincronizarse con el comportamiento espontáneo hasta poder alcanzarlo. Respecto a la dispersión temporal de JAM, está



(a) Número total de spikes al atacar al cuarto de las neuronas de forma aleatoria en tres instantes diferentes.



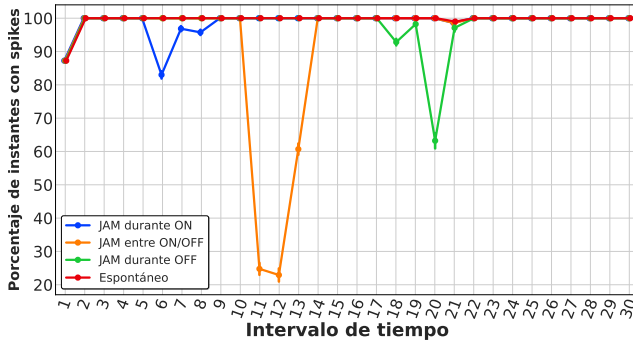
(b) Número total de spikes al atacar a la mitad de las neuronas de forma aleatoria en tres instantes diferentes.

Figura 5: Impacto del ciberataque JAM en base a la métrica del número de spikes, en diferentes instantes de tiempo, y diferentes número de neuronas afectadas, ejecutado diez veces de forma aleatoria por experimento.

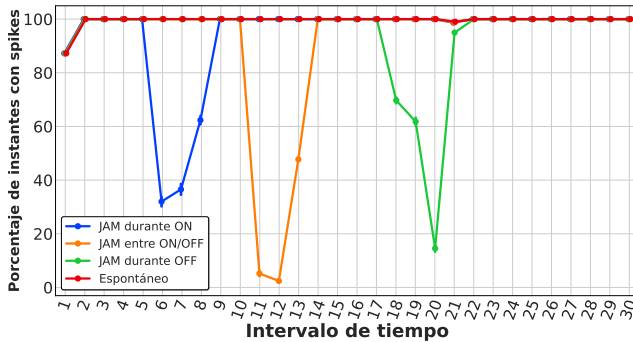
representada en la Figura 6. En el primer intervalo, según el número de neuronas atacadas, los límites de instantes con spikes son el 83 % y el 32 %. Seguidamente, en el intervalo correspondiente a la pantalla gris, el mínimo para el primer conjunto de neuronas (cuarto del total) es de 22.9 %, mientras que en el segundo conjunto (mitad del total) es de 2.4 %. Por su parte, en el intervalo OFF, atacar a la cuarta parte de las neuronas provoca que la simulación llegue a mínimos del 63.2 % de spikes, aproximadamente. En cambio, si se ejecuta sobre la mitad de las neuronas, el mínimo está sobre el 14.5 %.

V. DISCUSIÓN

En esta sección se comparan y discuten los resultados mostrados en la Sección IV, además de contrastar dichos resultados con los existentes en la literatura. Respecto a la experimentación presentada en este trabajo, se ha considerado el número de spikes y la dispersión temporal de spikes como métricas para el estudio del impacto de los ciberataques neuronales. En FLO, el número de spikes en el intervalo ON-flash incrementa alrededor de 53.000 y 22.000 spikes respecto al comportamiento espontáneo según el número de neuronas atacadas. Por el contrario, durante la pantalla gris el aumento ha sido de 14.000 y 5.000 spikes, mientras que en el OFF-flash ha sido de 27.000 y 11.000



(a) Número total de spikes al atacar al cuarto de las neuronas de forma aleatoria en tres instantes diferentes.



(b) Número total de spikes al atacar a la mitad de las neuronas de forma aleatoria en tres instantes diferentes.

Figura 6: Impacto del ciberataque JAM en base a la métrica de dispersión temporal, en diferentes instantes de tiempo, y diferentes número de neuronas afectadas, ejecutado diez veces de forma aleatoria por experimento.

spikes respecto al espontáneo. En relación con la dispersión temporal, se pone de manifiesto que en la mayoría de instantes de tiempo, la mayoría de neuronas hacen spikes, excepto en el intervalo 21, donde existe una ligera variabilidad.

Por su parte, en JAM, cuando se inhibe la mitad del total de las neuronas durante el ON-flash, hay un máximo de 288 y 193 spikes en los intervalos seis y siete, respectivamente. Sin embargo, si se ataca la cuarta parte, se obtienen 1485 y 1613 spikes. El intervalo entre ON/OFF-flash bajo ataque, en los intervalos 11 y 12, llega a diez y 1458 spikes respectivamente, al aplicar el ciberataque a la mitad de las neuronas. En cambio, afectando a la cuarta parte de las neuronas en los mismos intervalos se consiguen 176 y 106 spikes. En cuanto al OFF-flash, en el primer conjunto de neuronas objetivo, los intervalos 19 y 20 tienen 503 y 67 spikes respectivamente, mientras que en el segundo conjunto son 2864 y 462 spikes. Respecto a la dispersión temporal, cuanto mayor sea el número de neuronas atacadas, menor será el porcentaje de instantes con spikes. Por otro lado, dicho porcentaje está también influido por el estímulo visual presentado a las neuronas, por lo que, si en el caso espontáneo reaccionan pocas neuronas ante este, el ciberataque causará que el porcentaje disminuya considerablemente. Por contraste, si un gran número de neuronas responden a un determinado estímulo, como en el ON-flash, el porcentaje es mayor. Los resultados indican que ambos ciberataques son efectivos

en función del comportamiento que se ejecuta, es decir, estimular o inhibir a las neuronas objetivo. Sin embargo, se ha demostrado que JAM es más efectivo que FLO respecto a la variabilidad de spikes por cantidad de neuronas atacadas. En relación al número de neuronas afectadas se ha observado que los ataques han tenido un gran impacto tanto si se ataca a la mitad de todas las neuronas o la cuarta parte. Por lo tanto, el atacante podría ejecutar un ciberataque sobre un pequeño número de neuronas y conseguir un impacto similar. Por su parte, es importante analizar el impacto de los ciberataques en diferentes instantes de tiempo, sobre todo en estímulos visuales realistas. En este caso, se ha observado que las neuronas en su comportamiento espontáneo reaccionan de diferente manera si se les transmite un estímulo visual blanco, negro, o gris. Por ello, los ciberataques se ejecutan en tres instantes distintos, correspondientes a los estímulos visuales mencionados, por lo que, no es equivalente un ataque en el ON-flash al de OFF-flash, ni tampoco al de la pantalla gris. Además, tras el ciberataque, las neuronas tienden a estabilizarse hacia su comportamiento espontáneo, lo que puede resultar en un aumento de spikes en algunos instantes con el fin de volver al espontáneo.

Una vez realizado el análisis de los resultados es necesario realizar una comparación con los que se han obtenido en la literatura. En la topología de 450 neuronas presentada por López Madejska et al. [5], se realizaron dos experimentos para FLO en el que estudiaron el impacto al utilizar diferentes voltajes y números de neuronas. Respecto al voltaje, se concluía que cuanto mayor sea el voltaje, mayor efectividad tendrá el ataque. Por ello, en esta investigación se han usado los parámetros más efectivos documentados en la publicación mencionada y en el estudio de López Bernal et al. [2]. En cuanto al número de spikes se llegan a las mismas conclusiones, donde a mayor número de neuronas afectadas, mayor número de spikes. Sin embargo, en esta investigación se ha observado que no se necesita una gran cantidad de neuronas atacadas para causar un considerable impacto sobre el comportamiento neuronal. En el caso del JAM sucede lo mismo que en el anterior ciberataque, en referencia al número de spikes, donde es incluso más evidente que no se precisa atacar muchas neuronas para tener un gran impacto. En relación con el tiempo de propagación según el ciberataque, en la literatura, FLO dura alrededor de 500 ms, en cambio en este estudio son entre 200 y 300 ms. Por su parte, JAM tiene un tiempo de propagación de 600 ms en la literatura, mientras que en este trabajo se obtuvo una duración de 800 ms o 400 ms, según el instante de ataque.

Este trabajo introduce un nivel más de realismo al analizar el impacto de los ciberataques neuronales de forma cuantitativa. A pesar de ello, se necesita más estudios centrados en implementar nuevos ciberataques que afecten otros componentes de la red neuronal, por ejemplo, las sinapsis entre neuronas. Por otro lado, para analizar la efectividad de los ciberataques, se requiere investigar los diferentes instantes de tiempo bajo ataque ante distintos estímulos visuales externos (película, drifting grating y efecto de un disco acercándose) además del que se utiliza en este trabajo (efecto flash). De igual forma, en este estudio se realiza el ataque sin hacer distinción entre poblaciones neuronales, capas de V1 y según si las neuronas

actúan como excitadoras o inhibitoras, desconociendo su efectividad conforme a esta diferenciación.

VI. CONCLUSIONES

Este estudio analiza el impacto de los dos ciberataques neuronales, JAM y FLO, implementados en una topología neuronal compleja de 230.000 neuronas, siendo una reconstrucción del VI de un ratón compuesto por seis capas, con diferentes poblaciones neuronales y distintos comportamientos. Además, se utilizó un estímulo visual realista, consistente en un efecto de flash de luz. Para poder estudiar los impactos, se configuraron los ciberataques considerando el número de neuronas objetivo, el número de repeticiones por cada experimento, el voltaje utilizado y los instantes de ataque.

Después de ejecutar los experimentos, tanto FLO como JAM han generado un gran impacto respecto al número de spikes, siendo el más efectivo según las métricas utilizadas JAM. De igual forma, el tiempo de propagación cambia según el instante de ataque, siendo JAM el ciberataque con mayor duración. Por otro lado, se ha observado que no es necesaria una cantidad considerable de neuronas atacadas para causar un impacto sustancial sobre el comportamiento neuronal. Además, los resultados muestran que las neuronas están influidas por el estímulo visual, incluso después de los ciberataques, por lo que se sincronizarán conforme al estímulo presentado tras un tiempo después del ataque. En el caso de FLO, tardan entre 200 y 300 ms, mientras que en el JAM unos 800 ms o 400 ms según el instante atacado.

Como trabajo futuro, se identificarán nuevos ciberataques que se centren en otros componentes del modelo, por ejemplo, las sinapsis entre neuronas. Por otro lado, con el objetivo de analizar la efectividad de los ciberataques, se estudiarán los instantes críticos donde los ciberataques tienen mayor impacto, así como utilizar otros estímulos visuales externos. Además, se plantearán otras configuraciones de ataques, por ejemplo, ataques consecutivos de FLO o cambiar la duración de la ventana temporal de JAM. De igual manera, se buscará determinar si el ataque es efectivo distinguiendo entre los comportamientos y poblaciones neuronales, así como entre las capas. Por otro lado, se realizará un estudio centrado en aplicar determinados voltajes igual de eficientes que los estudiados en la literatura hacia poblaciones neuronales concretas. Finalmente, para elaborar conclusiones respecto al impacto que podrían generar estos ciberataques en el mundo real, es necesario seguir colaborando con especialistas en neurociencia para recrear situaciones concretas como ceguera temporal o generar estímulos donde no los había realmente.

FUNDING

This work has been partially supported by (a) the strategic project CDL-TALENTUM from the Spanish National Institute of Cybersecurity (INCIBE) and by the Recovery, Transformation and Resilience Plan, Next Generation EU, (b) the Swiss Federal Office for Defense Procurement (armasuisse) with the CyberTracer (CYD-C-2020003) project, and (c) the University of Zürich UZH.

REFERENCIAS

[1] E. Musk, "An integrated brain-machine interface platform with thousands of channels," *J Med Internet Res*, vol. 21, no. 10, p. e16194, Oct 2019.

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

[3] S. López Bernal, A. Huertas Celdrán, G. Martínez Pérez, M. T. Barros, and S. Balasubramaniam, "Security in brain-computer interfaces: State-of-the-art, opportunities, and future challenges," *ACM Computing Surveys*, vol. 54, no. 1, Jan. 2021.

[4] S. López Bernal, A. Huertas Celdrán, and G. Martínez Pérez, "Eight reasons to prioritize brain-computer interface cybersecurity," *Communications of the ACM*, vol. 66, no. 4, p. 68–78, mar 2023.

[5] V. M. López Madejska, S. López Bernal, G. Martínez Pérez, and A. Huertas Celdrán, "Impact of neural cyberattacks on a realistic neuronal topology from the primary visual cortex of mice," *Wireless Networks*, vol. 30, no. 1, Jan. 2024.

[6] 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, pp. 663–666.

[7] E. T. Martínez Beltrán, M. Quiles Pérez, S. López Bernal, A. Huertas Celdrán, and G. Martínez Pérez, "Noise-based cyberattacks generating fake p300 waves in brain-computer interfaces," *Cluster Computing*, vol. 25, no. 1, pp. 33–48, Feb 2022.

[8] 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, pp. 143–158.

[9] 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, pp. 133–136.

[10] H. Takabi, A. Bhalotiya, and M. Alohaly, "Brain computer interface (BCI) applications: Privacy threats and countermeasures," in *IEEE 2nd International Conference on Collaboration and Internet Computing*. Pittsburgh, PA, USA: IEEE, Nov 2016, pp. 102–111.

[11] 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.

[12] L. Pycroft, S. G. Bocard, S. L. Owen, J. F. Stein, J. J. Fitzgerald, A. L. Green, and T. Z. Aziz, "Brainjacking: Implant Security Issues in Invasive Neuromodulation," *World Neurosurgery*, vol. 92, pp. 454–462, Aug 2016.

[13] E. Marin, D. Singelé, B. Yang, V. Volski, G. A. E. Vandenbosch, B. Nuttin, and B. Preneel, "Securing wireless neurostimulators," in *Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy*, ser. CODASPY '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 287–298.

[14] S. López Bernal, A. Huertas Celdrán, and G. Martínez Pérez, "Neuronal jamming cyberattack over invasive bcis affecting the resolution of tasks requiring visual capabilities," *Computers & Security*, vol. 112, p. 102534, 2022.

[15] A. Arkhipov, N. W. Gouwens, Y. N. Billeh, S. Gratiy, R. Iyer, Z. Wei, Z. Xu, R. Abbasi-Asl, J. Berg, M. Buice, N. Cain, N. da Costa, S. de Vries, D. Denman, S. Durand, D. Feng, T. Jarsky, J. Lecoq, B. Lee, L. Li, S. Mihalas, G. K. Ocker, S. R. Olsen, R. C. Reid, G. Soler-Llavina, S. A. Sorensen, Q. Wang, J. Waters, M. Scanziani, and C. Koch, "Visual physiology of the layer 4 cortical circuit in silico," *PLOS Computational Biology*, vol. 14, no. 11, pp. 1–47, 11 2018.

[16] M. L. Hines and N. T. Carnevale, "The neuron simulation environment," *Neural Computation*, vol. 9, no. 6, pp. 1179–1209, 1997.

[17] M.-O. Gewaltig and M. Diesmann, "Nest (neural simulation tool)," *Scholarpedia*, vol. 2, no. 4, p. 1430, 2007.

[18] K. Dai, S. L. Gratiy, Y. N. Billeh, R. Xu, B. Cai, N. Cain, A. E. Rimehaug, A. J. Stasik, G. T. Einevoll, S. Mihalas, C. Koch, and A. Arkhipov, "Brain modeling toolkit: An open source software suite for multiscale modeling of brain circuits," *PLOS Computational Biology*, vol. 16, no. 11, pp. 1–23, 11 2020.

[19] Y. N. Billeh, B. Cai, S. L. Gratiy, K. Dai, R. Iyer, N. W. Gouwens, R. Abbasi-Asl, X. Jia, J. H. Siegle, S. R. Olsen, C. Koch, S. Mihalas, and A. Arkhipov, "Systematic integration of structural and functional data into multi-scale models of mouse primary visual cortex," *Neuron*, vol. 106, pp. 388–403.e18, 5 2020.

[20] C. Teeter, R. Iyer, V. Menon, N. Gouwens, D. Feng, J. Berg, A. Szafer, N. Cain, H. Zeng, M. Hawrylycz, C. Koch, and S. Mihalas, "Generalized leaky integrate-and-fire models classify multiple neuron types," *Nature Communications*, vol. 9, no. 1, p. 709, Feb 2018.