ATAQUES CONTRA SISTEMAS ROBÓTICOS Y AUTÓNOMOS

EVA MARTÍN IBÁÑEZ

DOCTORA POR LA UCM (CIENCIAS DE LA INFORMACIÓN)

RESUMEN

Los sistemas robóticos y autónomos (RAS) son sistemas ciberfísicos, en los que la computación, las comunicaciones y los procesos físicos dependen unos de otros y están estrechamente acoplados. Las amenazas explotan tanto su naturaleza ciber como la física. Los atacantes pretenden aprovechar puntos de entrada desde donde pueden comunicar directamente con sensores y efectores, o desde donde pueden afectar indirectamente su operación manipulando las infraestructuras de control y comunicación. Se puede distinguir entre ataques ciber-físicos (originados en el ciberespacio y con consecuencias negativas en el espacio físico) y ataques físico-ciber (desarrollados en el espacio físico y con perjuicios en el ciberespacio).

Palabras clave: Sistemas Robóticos y Autónomos, RAS, sistemas no tripulados, seguridad, sistemas ciberfísicos, ataques.

ABSTRACT

Robotic and autonomous systems (RAS) are cyberphysical systems, where computation, communications and physical processes are interdependent and tightly coupled. Threats to security exploit both its cyber and physical properties. Attackers try to identify entry points to communicate directly with sensors and actuators, or to affect indirectly their operations manipulating control and communication infrastructures. It is useful to distinguish between cyber-physical attacks (originated in cyberspace with negative consequences to physical space) and physical-cyber attacks (started in physical space with damages in cyberspace).

Keywords: Robotic and Autonomous Systems, RAS, unmanned systems, security, cyberphysical systems, attacks

1. INTRODUCCIÓN

Los sistemas robóticos y autónomos o RAS (*Robotic and Autonomous Systems*) son aquellos que tienen un elemento robótico, un elemento autónomo o, frecuentemente, ambos.

Los RAS están dotados de controladores, sensores y efectores; son sistemas ciberfísicos, donde la computación, las comunicaciones y los procesos físicos están estrechamente acoplados y dependen unos de otros.

Estos sistemas ciberfísicos presentan vulnerabilidades únicas derivadas del estrecho acoplamiento entre sensores, efectores, interfaces humano-máquina y software de procesamiento de información. Esas vulnerabilidades proceden tanto de su naturaleza ciber como de la física.

En cuanto a la estructura de este artículo, en primer lugar, analizamos los RAS como sistemas ciberfísicos. A continuación, distinguimos entre ataques ciber-físicos (originados en el ciberespacio y con efectos negativos en el espacio físico) y ataques físico-ciber (desarrollados en el espacio físico y con consecuencias perjudiciales en el ciberespacio). Después, identificamos los potenciales puntos de entrada en los ataques ciber-físicos. Posteriormente, ofrecemos una taxonomía de los métodos de ataque ciber-físicos y físico-ciber más comunes. Como cierre, están las conclusiones.

2. SISTEMAS ROBÓTICOS Y AUTÓNOMOS COMO SISTEMAS CIBERFÍSICOS

El término Sistemas Robóticos y Autónomos (RAS) subraya los aspectos físicos (robóticos) y los aspectos cognitivos (autónomos). Describe aquellos sistemas que tienen un elemento robótico, un elemento autónomo o, a menudo, ambos. Según avanza la tecnología, cada vez hay más sistemas robóticos con capacidades autónomas (US Army, 2016, p. 22).

Los sistemas robóticos y autónomos, dotados de controladores, sensores y efectores, pueden considerarse sistemas ciberfísicos, donde la computación, las comunicaciones y los procesos físicos están estrechamente acoplados y dependen unos de otros.

Los sistemas ciberfísicos pueden definirse como aquellos que integran computación, comunicaciones, sensores y efectores con sistemas físicos para desempeñar funciones sensibles al tiempo con diversos grados de interacción con el entorno, incluyendo la interacción humana (NIST, 2017a, p. 5). Entre las aplicaciones actuales de los sistemas ciberfísicos se encuentran los coches inteligentes, los edificios inteligentes, los robots, los vehículos no tripulados y los dispositivos médicos (NIST, 2017b, p. 1).

Un sistema ciberfísico integra procesos físicos, computacionales y de comunicación. Una de sus características clave es la integración continua de los recursos de hardware y de software con finalidades computacionales, de comunicación y control, todos ellos diseñados conjuntamente con los componentes físicos (Lun et al., 2016, p. 1).

Para comprender cómo es un sistema ciberfísico conviene conocer someramente los principales elementos involucrados: sensores, efectores, controladores y sistemas embebidos. Están recogidos en la Tabla 1, donde además se indica porqué tienen interés para los ciberatacantes.

	Definición	Interés para los ciberatacantes
Sensores	Dispositivos que transforman los datos del mundo real en una forma eléctrica con el objetivo de medir u observar el entorno físico. La cantidad, propiedad o condición medida es el estímulo, que puede ser acústico, biológico, químico, eléctrico, magnético, mecánico, radiante o térmico. Pueden incluir varios transductores que conviertan una forma de energía en otra hasta que se produzca una señal eléctrica que un sistema de proceso de información pueda interpretar. Un sensor puede ser natural, del propio organismo vivo.	Los sensores son de interés para los ciberatacantes, porque ganando acceso al ordenador que los controla pueden observar un entorno físico remoto.
Efectores	Su labor es iniciar una acción física cuando reciben la instrucción de hacerlo mediante una señal eléctrica. Es cualquier dispositivo capaz de iniciar una acción física en su entorno.	Los efectores son incluso más atractivos que los sensores, porque permiten alterar el entorno físico.
Controladores	Muchos controladores son mecánicos, hidráulicos o neumáticos, pero los electrónicos basados en ordenadores y sistemas embebidos son los relevantes en este caso. En los electrónicos el software está constantemente procesando las mediciones procedentes de los sensores y determina los parámetros de los efectores.	Los controladores crean un enlace directo entre sensores y efectores, que puede ser explotado por un adversario. Un error en el proceso de percepción, ya sea natural o resultado de un ataque intencionado, puede producir una actuación no deseada.
Sistemas embebidos	Desde la perspectiva de la seguridad ciberfísica, son ordenadores camuflados, escondidos dentro de otros dispositivos. Están programados para realizar un conjunto específico de funciones requeridas por el sistema donde están incluidos.	Para los ciberatacantes, los sistemas embebidos son ordenadores que ejecutan alguna forma de software y que a menudo presentan capacidades de comunicación en red.

Tabla 1. Principales dispositivos y sistemas relacionados con los sistemas ciberfísicos. Fuente: Elaboración propia a partir de Loukas , 2015, pp. 4-7.

Sistemas ciberfísicos como los UAV (vehículos aéreos no tripulados), que crean un bucle cerrado de flujo de datos, ilustran cómo los componentes del ciberdominio y los componentes del dominio físico están estrechamente acoplados. El efecto del acoplamiento en este caso tiene dos significados. El primero se refiere a los flujos de datos macroscópicos entre el ciberdominio y el dominio físico. Los flujos de datos se introducen en el ciberdominio mediante los sensores desde el mundo físico, y finalmente son alimentados a los efectores que producen efectos en el mundo físico. El segundo implica influencias mutuas y dependencias entre cada uno de los componentes del dominio ciberfísico a nivel micro (Wang et al., 2018, p. 26).

La seguridad de los sistemas ciberfísicos presenta peculiaridades que la distinguen de otros sistemas de tecnologías de la información convencionales, como los sistemas informáticos empresariales. La Tabla 2 muestra esas características distintivas.

Todas esas peculiaridades deben ser tenidas en cuenta, desde la identificación de activos hasta la detección de amenazas. Eso puede resultar difícil en sistemas extensos donde deben considerarse todos los elementos y sus dependencias, junto con todas las posibles interacciones entre distintas infraestructuras (Cheminod et al., 2013, p. 282).

Los sistemas ciberfísicos combinan componentes físicos y ciber-componentes. Los componentes físicos son sistemas existentes en la naturaleza, como los seres biológicos, o aquellos desarrollados por humanos, como los sistemas de generación de energía. Estos componentes físicos existen, operan e interactúan con su entorno en tiempo continuo u ordinario. Los componentes computacionales son sistemas y entidades encargados de procesar, comunicar y controlar la información vía medios informáticos. Eso incluye algoritmos implementados en el software y en los sistemas digitales, con interfaces a los componentes físicos mediante convertidores de analógico a digital, convertidores de digital a analógico y redes de comunicaciones digitales. Todos esos componentes computacionales operan en tiempo discreto o de modo dirigido por eventos. La complejidad de integración de los sistemas ciberfísicos surge de que los componentes computacionales están distribuidos por el sistema y estrechamente acoplados con los componentes físicos. Como consecuencia, los sistemas ciberfísicos están muy interconectados para combinar dinámicas continuas y discretas (Sanfelice, 2015, pp. 3-4).

El comportamiento de los sistemas ciberfísicos es heterogéneo, en el sentido de que combina variables continuas y variables discretas. El estado de los componentes físicos suele determinarse mediante variables continuas, que cambian según el tiempo físico (ordinario) y toman valores de un conjunto denso. Por su parte, el estado de los ciber-componentes normalmente está definido por variables discretas, que cambian dentro del código, que es ejecutado en eventos de tiempo discreto, y que toman valores de conjuntos discretos. Inevitablemente, esta combinación heterogénea de variables y nociones de tiempo requiere modelos dinámicos que combinen variables continuas y discretas, junto con nociones de tiempo. Los modelos híbridos se adaptan bien a los sistemas ciberfísicos, porque imponen automáticamente la noción de tiempo a los ciber-componentes (Sanfelice, 2015, pp. 5-6).

Requerimientos de tiempo real	Cuando el tiempo de respuesta es crítico, un ancho de banda modesto es aceptable, pero un alto retardo y/o una alta variabilidad de frecuencia (jitter) no lo son. Además es necesaria la respuesta de un humano u otras interacciones de emergencia. Asimismo, realizar acciones asíncronas o esporádicas, como las actualizaciones de antivirus, son muy complicadas e incluso imposibles de completar. Y la adopción de cortafuegos y filtros complejos pueden introducir retrasos inaceptables o impredecibles en las redes de control y proceso.
Las prácticas típicas de seguridad no son tolerables	Las prácticas rutinarias de parchear y actualizar el hardware y el software suelen requerir que al menos parte del sistema esté temporalmente fuera de línea. Los requisitos críticos de disponibilidad propios de los sistemas ciberfísicos son difícilmente compatibles.
Distintas consecuencias de los fallos	En los sistemas informáticos empresariales y de consumo suelen estar limitadas a pérdidas financieras y/o reputacionales, porque solo los datos (la información) suelen necesitar protección. Eso también puede ser importante en los sistemas ciberfísicos, pero, en este caso, los fallos pueden causar daños catastróficos al medioambiente, junto con lesiones a seres humanos y pérdida de vidas.
Recursos limitados	El rendimiento y la energía son críticos en los sistemas industriales de control de automatización, donde muchos dispositivos de campo y de control tienen capacidades de computación reducidas o limitadas por la disponibilidad de alimentación. Esto hace poco factible usar mecanismos sofisticados de cifrado y protocolos de seguridad que consuman muchos recursos de computación.
Significados diferentes de las dimensiones de ciberseguridad (disponibilidad, integridad y confidencialidad)	Los sistemas ciberfísicos están diseñados para satisfacer ciertas metas operativas, por lo que la disponibilidad consiste en mantener esas metas, resistiendo ataques de denegación de servicio (DoS) a la información recogida por los sensores, los comandos de los controladores y las acciones físicas realizadas por los efectores. La integridad en los sistemas ciberfísicos pretende preservar las metas operativas evitando, detectando o sobreviviendo a ataques contra la información transmitida a y desde sensores, controladores y efectores. La confidencialidad en los sistemas ciberfísicos pretende impedir que un adversario pueda inferir el estado del sistema husmeando en los canales de comunicación entre sensores y controladores, y entre controladores y efectores, o por medio de ataques de canal lateral a sensores, controladores y efectores.
Prioridades de seguridad diferentes	En los sistemas ciberfísicos el orden de prioridad es disponibilidad, integridad y confidencialidad. Por el contrario, en los sistemas de TIC convencionales lo primero suele ser la confidencialidad, seguida de la integridad y la disponibilidad.

Tabla 2. Peculiaridades de la seguridad de los sistemas ciberfísicos. Fuente: Elaboración propia a partir de Cheminod et al., 2013, p. 279 y Lun et al., 2016, p. 3.

El bucle cerrado de muchos sistemas ciberfísicos como los vehículos no tripulados contiene múltiples componentes (sensores, comunicación, computación y control) que los hace vulnerables a ataques externos, porque los problemas de seguridad en cualquiera de esos componentes paralizarán todo el sistema. Los ataques se pueden ejecutar tanto desde la capa física como desde la capa ciber. Por ejemplo, en el primer caso, los ataques se pueden desplegar a través de malware, consiguiendo acceder a los elementos de la red de comunicación o falsificando la información de los sensores. En el segundo, los ataques manipulan directamente los elementos físicos del sistema, por ejemplo, para ocultar objetivos o modificar la fuente de energía (Wang et al., 2018, p. 34).

3. ATAQUES CIBER-FÍSICOS Y ATAQUES FÍSICO-CIBER

Para analizar la seguridad de los sistemas robóticos y autónomos (RAS) conviene distinguir entre ataques ciber-físicos y ataques físico-ciber.

3.1. ATAQUES CIBER-FÍSICOS

Un ataque ciber-físico es una brecha de seguridad en el ciberespacio que afecta negativamente al espacio físico. Es una categoría particular de ciberataque que, intencionadamente o no, también perjudica al espacio físico apuntando a la infraestructura computacional y de comunicaciones que permite que las personas y los sistemas monitoricen y controlen los sensores y los efectores.

Un sistema ciberfísico suele ser un sistema de bucle cerrado formado por una red de sensores y efectores, donde los datos recopilados por los sensores son comunicados a los controladores (generalmente sistemas embebidos) que ajustan la operación del sistema a través de los efectores (Loukas, 2015, pp. 8, 12).

3.2. ATAQUES FÍSICO-CIBER

Un ataque físico-ciber es aquel desarrollado en el espacio físico que afecta negativamente en el ciberespacio (Loukas, 2015, p. 222).

Al igual que los ataques ciber-físicos, los ataques físico-ciber explotan las interacciones entre el espacio físico y el ciberespacio. Los sensores, los efectores, los controladores, los ordenadores, los dispositivos de red y otros componentes de las infraestructuras de control o de comunicación son elementos físicos que existen en el espacio físico. Basta con dañarlos físicamente para interrumpir o impedir su operación. También pueden aprovechar las entradas físicas a los sensores, las emanaciones que emiten los equipos y el modo en que se implementan las técnicas criptográficas en ordenadores y otros dispositivos.

4. PUNTOS DE ENTRADA EN LOS ATAQUES CIBER-FÍSICOS

Las amenazas a la seguridad de los sistemas ciberfísicos presentan similitudes en todas las plataformas. En todos los casos -incluyendo los sistemas robóticos y autónomos- el atacante pretende identificar puntos de entrada desde donde es posible comunicar directamente con sensores y efectores, o desde donde puede afectar indirectamente su operación manipulando las infraestructuras de control y comunicaciones. En la mayoría de los casos, el descubrimiento y la explotación de vulnerabilidades relevantes requiere una investigación y una planificación considerables.

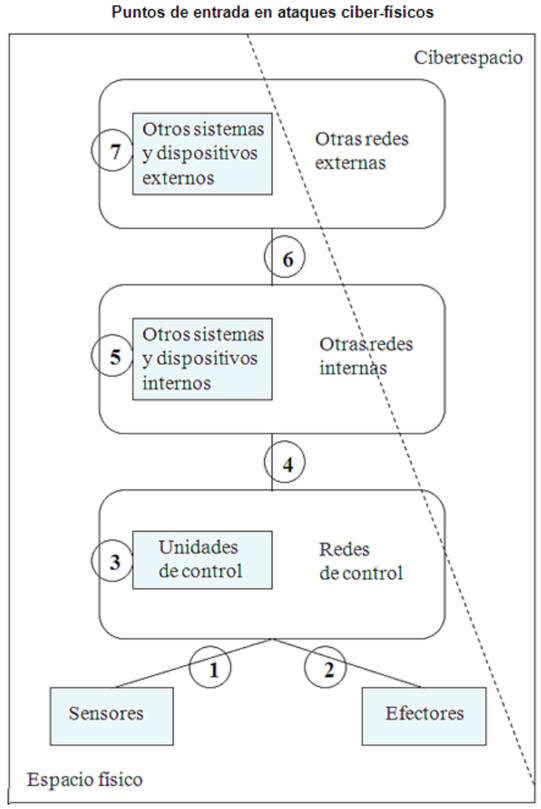

Figura 2. Puntos de entrada en ataques ciber-físicos. Fuente: Elaboración propia.

La Figura 2 muestra los potenciales puntos de entrada en los ataques ciber-físicos desde donde se puede efectuar una intrusión. Hemos elaborado un diagrama genérico que resulta aplicable a cualquier sistema ciberfísico. Eso incluye la gran variedad de sistemas robóticos y autónomos, independientemente del dominio (tierra, mar, aire o espacio) donde operen.

Los posibles puntos de entrada para un ataque ciberfísico, es decir, aquellos con origen en el ciberespacio y con efectos sobre el espacio físico, son los siguientes:

1 Canales de comunicación con los sensores

El atacante aprovecha el canal de comunicación utilizado para recopilar los datos que recogen los sensores y acceder a ellos.

2 Canales de control de los efectores

El atacante explota el canal empleado para enviar instrucciones a los efectores y alterar sus acciones.

3 Redes de control para unidades de control

Las redes de control pueden dar acceso a las unidades de control que además están enlazadas a sensores y efectores.

4 Canales de comunicación entre redes de control y otras redes internas

El sistema ciberfísico incluye otros sistemas y dispositivos aparte de las unidades de control. Un atacante puede explotar el canal de comunicación entre las redes de control y el resto de las redes internas para llegar hasta esos otros equipos o a las unidades de control.

5 Canales de comunicación entre otras redes internas y otros sistemas y dispositivos internos

A través de esos canales, los adversarios pueden acceder a otros sistemas y dispositivos internos, por ejemplo, los de gestión.

6 Canales de comunicación entre redes internas y externas

La red interna del sistema ciberfísico puede estar conecta a Internet y a otras redes externas, como pueden ser las de otros sistemas o las de proveedores o contratistas. Eso amplía la superficie de ataque.

7 Canales de comunicación entre otros sistemas y dispositivos externos y redes externas

Las redes externas pueden constituir una puerta de entrada a otros sistemas y dispositivos externos que a su vez estén conectados con el sistema ciberfísico. Incluyen cualquier red externa fuera del control de los gestores del sistema ciberfísico, por ejemplo, las de los aliados o las de proveedores de comunicaciones y de servicios de almacenamiento, entre otros.

5. MÉTODOS DE ATAQUE CIBER-FÍSICOS Y FÍSICO-CIBER MÁS COMUNES

A continuación, figuran los métodos más comunes usados en los ataques ciber-físicos y en los ataques físico-ciber, que se aplican sobre uno o sobre varios puntos de entrada.

5.1. MÉTODOS DE ATAQUE CIBERFÍSICOS

Como antes se ha comentado, los ataques ciberfísicos se originan en el ciberespacio y causan impacto en el espacio físico.

5.1.1. Agujero gris (grey hole) y agujero negro (black hole)

El adversario altera la disponibilidad de los datos comprometiendo un nodo de red y la transmisión de paquetes. Si deja caer los paquetes selectivamente para evitar ser detectado sería un agujero gris y, si tira todos, sería un agujero negro. Evita o retrasa la actuación del sistema interrumpiendo la comunicación con los efectores o la recopilación de datos desde los sensores (Loukas, 2015, p. 158).

5.1.2. Aislamiento de red

Destinado a sistemas ciberfísicos extensos. Su objetivo es aislar un área geográfica física de una red. Se puede lograr comprometiendo los nodos de la red dentro del área objetivo, retrasando o tirando los paquetes transmitidos dentro y fuera de ella. En la práctica, es un ataque de agujero negro coordinado donde se seleccionan los objetivos según el área geográfica física en la que operan. Representa una brecha a la integridad de la red en los nodos comprometidos y una pérdida de disponibilidad de red. El impacto físico es similar al de un ataque de agujero negro, pero centrado en los sensores y los actuadores de un área geográfica específica (Loukas, 2015, p. 171).

5.1.3. Ataques a criptosistemas

Los sistemas ciberfísicos que emplean técnicas criptográficas pueden estar sujetos a los ataques habituales contra criptosistemas, incluyendo los que utilizan criptografía ligera, apta para dispositivos y sistemas embebidos con recursos limitados. En cuanto a su impacto, pueden producir brechas de confidencialidad, de integridad y de autenticación, aparte de posibilitar la manipulación de los activos que intentan proteger.

Algunos de los ataques más conocidos y que son aplicables incluso aunque se usen primitivas criptográficas bien implementadas son los siguientes (Shostack, 2014, pp. 342-343):

- Texto cifrado conocido. El adversario prueba todas las posibles combinaciones de claves para intentar descifrar el mensaje (fuerza bruta).
- Texto cifrado elegido. El atacante puede insertar un texto cifrado de su elección. Normalmente se consigue menos información que con un texto claro completo, pero puede servir de palanca para un ataque mayor.
- Texto claro elegido. Resulta fácil conseguir texto claro en los protocolos modernos, porque usan programas que carecen de la habilidad de detectar entradas extrañas.
- Adaptativo elegido. Variante de los ataques de texto cifrado y texto plano elegido en la que el adversario inyecta algo, observa la respuesta y luego inyecta algo nuevo.
- Man-in-the-Middle (MitM). El atacante (Mallory) se coloca en la red y altera el tráfico según pasa y evita que las dos partes de la comunicación (Alice y Bob) vean los mensajes reales.
- Retransmisión (relay). Alice podría ser un navegador web y Bob un servidor web; el atacante presenta un servidor web a Alice, que introduce su clave y el atacante lo reenvía a Bob.
- Reenvío. El adversario captura mensajes y los reenvía.
- Reflejo. El atacante vuelve a reproducir contenidos de Alice a la propia Alice. Si el atacante quiere hacerse pasar por Alice, vigila a Bob enviar un mensaje y luego pide a Bob que se autentique con el mismo mensaje. Bob envía al atacante el mensaje cifrado creado para la ocasión, que después usa para realizar envíos simulando que es Alice.

- Bajada de categoría (downgrade). Son ataques contra protocolos usando un MitM. Suceden cuando una versión insegura de un protocolo es actualizada pero el cliente o el servidor no está seguro de qué versión entiende el otro extremo. El atacante se sitúa en el medio, haciéndose pasar por las dos partes y forzándolas a usar una versión menos segura.
- Ataque de cumpleaños. Un atacante genera un conjunto de documentos, buscando dos que den el mismo valor hash[1]. Los hashes, igual que las fechas de cumpleaños, tienen un conjunto fijo de posibles valores. Cuando un atacante tiene dos documentos con el mismo hash, puede sustituir uno por otro. Si los dos dan el mismo hash, difícilmente podrá el software detectar la sustitución.
- Ataque de análisis de tráfico. Supone estudiar una serie de mensajes para extraer patrones sin realmente comprender mensaje. El tamaño del mensaje es importante en el análisis de tráfico.
- Ataque de prolongación de longitud. Por ejemplo, si un atacante conoce el hash de 'foo', le resultará trivial calcular el hash de 'foobar'. Funcionan todavía mejor en aquellos supuestos, como el uso de hashes para autenticarse en sitios web, donde los atacantes pueden añadir parámetros adicionales.
- Ataque de sincronización. Diferencias sutiles del tiempo necesario para realizar operaciones criptográficas pueden revelar información como la longitud de un mensaje, en qué parte del código falla un código o el peso de una clave privada.

5.1.4. Ataques a la cadena de suministro

El atacante lograr acceder a los sistemas informáticos y a las redes de los proveedores, modifica el firmware y preinstala puertas traseras (software que facilita el acceso no autorizado al sistema) y otros tipos de malware dentro los dispositivos antes de que se despachen. Comprende cualquier modificación maliciosa del hardware o del software que suceda antes de la adquisición por un comprador legítimo. El impacto físico son actuaciones no autorizadas, incorrectas, retrasadas o impedidas, o brechas de privacidad física a través de la extracción de los datos de los sensores a personas no autorizadas (Loukas, 2015, p. 176).

5.1.5. Ataques a redes acústicas acuáticas

Las redes bajo el agua se utilizan sobre todo para recoger datos de sensores y para comunicaciones. Son especialmente relevantes en los UMS (sistemas marítimos no tripulados). Esas infraestructuras acústicas acuáticas presentan diferencias con respecto a sus equivalentes terrestres. El rendimiento de las comunicaciones acústicas acuáticas está limitado por las características distintivas del canal de sonido: baja velocidad de propagación, estrecho ancho de banda, atenuación de frecuencias y severa propagación multi-ruta (Dong y Liu, 2010, pp. 1-3).

1 Hash o función resumen es aquella que sirve para esquematizar datos de tamaño arbitrario en datos de tamaño fijo. Su resultado son los valores hash, códigos hash o hash a secas.

Algunos ejemplos de ataques a las redes acústicas acuáticas inalámbricas son los siguientes:

- Interferencias (jamming). Para interferir en el canal físico se ponen portadoras en los nodos de las frecuencias cercanas usadas para comunicar. Estas redes suelen ser vulnerables a interferencias de banda estrecha. La localización de los nodos puede verse afectada por ataques de reenvío cuando el atacante interfiere en la comunicación entre emisor y receptor, y el segundo reenvía el mismo mensaje con información desfasada haciéndose pasar por el emisor.
- Agujero de gusano. Un agujero de gusano es una conexión fuera de banda creada por un adversario entre dos localizaciones físicas en una red con un retardo más bajo y mayor ancho de banda que las conexiones corrientes. Esa conexión se realiza mediante radio rápida (por encima de la superficie del agua) o mediante enlaces cableados fijos. El nodo malicioso transfiere paquetes seleccionados recibidos en un extremo del agujero de gusano al otro extremo usando conexiones fuera de banda y los reinyecta dentro de la red. El efecto es que se establecen falsas relaciones de cercanía, porque dos nodos fuera del rango cada uno del otro pueden creer erróneamente que están cerca por culpa del agujero de gusano. Los efectos son devastadores. Los protocolos de enrrutado eligen las rutas que contienen enlaces al agujero de gusano porque parecen más cortas. Así, el adversario puede monitorizar el tráfico de red y retrasar o dejar caer paquetes enviados a través del agujero de gusano. Los protocolos de localización también pueden verse afectados cuando los nodos maliciosos declaran posiciones erróneas y confunden a otros nodos.
- Ataque de sumidero. El nodo atacante intenta atraer el tráfico de un área concreta hacia sí mismo. Puede lograrlo anunciando que su ruta es de alta calidad. De esa forma el tráfico se redirige erróneamente hacia el nodo o los nodos maliciosos o comprometidos como sumidero.
- Inundación con mensajes HELLO. Un nodo que recibe un paquete HELLO de un nodo malicioso puede interpretar que el adversario es un vecino. Es una presunción falsa si el adversario usa una transmisión de alta potencia.
- Ataque Sybil. Un atacante con múltiples identidades puede pretender que está en varios sitios a la vez. Los protocolos de enrrutado geográfico también se pueden equivocar, porque el atacante con múltiples identidades afirma estar en varios lugares simultáneamente (Domingo, 2011, pp. 23-24).

5.1.6. Ataques a sistemas de aprendizaje automático (machine learning)

Pueden suceder en cualquier sistema con aprendizaje automático, incluyendo los ciberfísicos. El aprendizaje automático se entiende en el sentido de generar comportamientos automáticos a partir de conjuntos de ejemplos. Cuanto mayor sea el nivel de autonomía de un sistema ciberfísico, mayor es el papel del aprendizaje. Si un atacante lograr afectar por distintos medios el aprendizaje automático, podría modificar el comportamiento del sistema ciberfísico, con impactos tanto en el ciberespacio como en el espacio físico.

- Ataques de ejemplos adversos. Los ejemplos adversos son entradas sutilmente modificadas -a menudo indistinguibles para los seres humanos- especialmente creadas para comprometer la integridad de sus salidas. Se usan para manipular el comportamiento del sistema. Este método suele constar de dos fases. Primero se entrena el modelo sustitutivo y luego se elaboran los ejemplos adversos (Papernot et al., 2016, pp. 1, 4).

- Ataques de caja negra[2] usando ejemplos adversos. Se pueden usar contra el deep learning[3] (aprendizaje profundo). El adversario puede llegar a controlar a distancia una DNN (red neuronal) sin acceder al modelo, a sus parámetros, ni al conjunto de datos de entrenamiento. La estrategia de ataque es introducir un modelo sustituto en las parejas de entrada-salida, para después crear ejemplos adversos basados en ese modelo auxiliar. Consigue cegar a distancia los oráculos[4] de aprendizaje automático construidos con deep learning con el resultado de clasificar erróneamente los ejemplos adversos (Papernot et al., 2016, pp. 1, 13). Además, es posible transferir los ejemplos adversos dentro de la misma técnica de aprendizaje automático, y entre técnicas como redes neuronales, regresión logística, máquinas de soporte vectorial (SVM) o árboles de decisión (Papernot, McDaniel y Goodfellow, 2016, pp. 1, 8).

- Ataques de extracción del modelo. Un adversario con acceso a la caja negra, sin conocimiento previo de los parámetros del modelo de aprendizaje automático ni del conjunto de datos de entrenamiento, se dirige a duplicar la funcionalidad del modelo. Son unos ataques simples y eficientes contra modelos populares como los de regresión logística, redes neuronales y árboles de decisión (Tramèr et al., 2016, p. 601).

- Ataques de evasión en aprendizaje automático. El adversario intenta evitar ser detectado manipulando ejemplos o muestras de test maliciosos. Daña la integridad del sistema. El objetivo es manipular una sola muestra (sin pérdidas de generalidad) para que sea clasificada erróneamente. En lugar de ajustar ligeramente el umbral de decisión, se crea un ejemplo que es clasificado erróneamente con una alta confianza. Los conocimientos del adversario sobre el sistema objetivo pueden ser: el conjunto de entrenamiento o parte de él; la representación de características de cada muestra; el tipo de algoritmo de aprendizaje y la forma de su función de decisión; el modelo de clasificador ya entrenado; o los resultados del clasificador (Biggio et al., 2013, pp. 338-340).

- Ataques de envenenamiento en aprendizaje automático. Funcionan inyectando ejemplos en el conjunto de datos de entrenamiento. Es factible llegar a confundir a algunos algoritmos de aprendizaje manipulando únicamente una pequeña parte de los datos de entrenamiento (Biggio, 2016, p. 1). El adversario tiene acceso a los datos de entrenamiento y los contamina para subvertir o controlar

2 Caja negra (en aprendizaje automático): algoritmo con una implementación opaca en el sentido de que se desconoce su funcionamiento interno, y solo se conocen las entradas y las salidas.

3 Deep learning (aprendizaje profundo): técnica de aprendizaje automático que usa DNN (redes neuronales profundas) para registrar una jerarquía de conceptos incrementalmente complejos. Las redes neuronales profundas se utilizan en clasificación y en aprendizaje reforzado, y forman parte de las técnicas de aprendizaje no supervisado (los datos no están etiquetados).

4 Oráculo (en informática) es un mecanismo para determinar si una prueba ha tenido éxito o ha fallado.

la selección de un conjunto reducido de características. La meta es modificar los resultados de la clasificación (Xiao et al., 2015, pp. 2-3).

5.1.7. Ataques de dinámica cero

El adversario puede confundir al sistema de control haciéndole creer que está en un estado diferente sin necesidad de comprometer los sensores. En estos ataques de dinámica cero[5], basta con que comprometa los efectores.

Cuando la señal de una o más actuaciones está comprometida, ya sea porque el propio efector lo está o porque acepta comandos de control de una entidad no confiable, la actuación respecto al fenómeno de interés en el espacio físico será distinta de la deseada por el controlador. Esa falsa actuación a su vez afecta a las variables medidas por el sistema, que a su vez influyen en las mediciones de los sensores realimentadas al controlador (Urbina et al., 2016, pp. 2-3, 10).

5.1.8. Ataques de planificación de paquetes

En los sistemas ciberfísicos en red, múltiples sensores envían información a los controladores a través de un canal de comunicación compartido y los controladores transmiten paquetes de control a los efectores que están conectados al sistema físico. Se planifica la transmisión de los paquetes de datos que deben alcanzar su nodo de destino antes de un plazo límite permitido.

Los ataques de planificación de paquetes se producen en el camino entre el controlador y los sensores. Suelen ser furtivos y fáciles de ejecutar tanto en canales de comunicación cableados como inalámbricos usando distintas técnicas. Los efectos de estos son añadir tiempo para variar el retardo de la red y alterar el orden en el que los paquetes son recibidos por el controlador (Shoukry et al, 2013, p. 2).

5.1.9. Ataques de reenvío (replay)

Apto para cualquier sistema donde se transmitan datos de sensores y de control de forma remota que necesiten estar actualizados. El atacante observa y graba una secuencia de comunicación para volver a reproducirla posteriormente. Su impacto físico primario es una actuación no autorizada si se trata de reenviar un comando y una actuación incorrecta si son datos anticuados del sensor. También puede impedir y retrasar actuaciones (Loukas, 2015, p. 174).

Puede funcionar como lanzadera de malware o de una escalera lógica maliciosa, que coloque al controlador objetivo en un bucle infinito dejándolo inservible. Para efectuar manipulaciones sutiles del sistema, el adversario debe contar con conocimientos específicos sobre las operaciones. Por el contrario, si la meta es sabotearlo, resulta mucho más sencillo interrumpir los procesos. Si solo se manipulan los valores de lectura, el dispositivo transmitirá valores falsos, pero si además se manipulan los valores

5 En la teoría de control, dinámica cero es la dinámica interna cuando la salida y todas sus derivadas son nulas.

de escritura, se puede conseguir que la funcionalidad del protocolo quede inservible para ese dispositivo (Knapp y Langill, 2015, p. 189).

5.1.10. Ataques de retransmisión (relay)

En los sistemas de autenticación basados en tokens, las restricciones de alcance físico del enlace de comunicación se suelen emplear como prueba implícita de la proximidad de un token a su correspondiente lector de token legítimo. Sin embargo, un adversario puede capturar la señal transmitida de uno y retransmitirla al otro a través de otros enlaces de comunicación de mayor alcance que está bajo su control. Así puede engañar al lector de token haciéndole creer que está cerca del token. Se produce una brecha de autenticación que puede conducir a actuaciones no autorizadas, como la apertura de puertas (Loukas, 2015, p. 173).

5.1.11. Ataques Sybil y clonado de nodos

La gestión de identidades en redes es compleja, sobre todo en redes ad hoc inalámbricas, como las desplegadas durante las operaciones de respuesta a emergencias. En un ataque Sybil el adversario roba o fabrica identidades usando pocos recursos computacionales. Los nodos Sybil minan la confianza ligada a una identidad y posibilitan actuaciones no fiables degradando la calidad del sistema. Y pueden llegar a colapsar una red ad hoc inalámbrica (Casey et al., 2016, p. 1).

Las redes inalámbricas de sensores son vulnerables a estos ataques. En los nodos de sensores desplegados en entornos hostiles, el adversario puede inyectar fácilmente datos maliciosos o alterar el contenido de los mensajes legítimos. Puede hacerlo mediante el clonado de nodos o con un ataque Sybil.

El clonado de nodos consiste en capturar unos cuantos nodos, extraer el código y las credenciales secretas y luego usar esos materiales para clonar muchos nodos a partir de hardware de sensores. Esos nodos clonados, que parecen legítimos, son capaces de unirse a la red y causar daños severos.

En el ataque Sybil, los nodos maliciosos enmascaran otros nodos o reclaman una identidad falsa en la red. En el peor escenario, el atacante puede generar un número arbitrario de identidades de nodo adicionales y usar un dispositivo físico para lanzar ataques DoS a los nodos legítimos, reduciendo su cuota de recursos y proporcionando más recursos para realizar otros ataques (Wen, 2013, p. 59).

Ambos ataques además causan un impacto físico de aumento del consumo de energía, que puede agotar las baterías de los dispositivos.

5.1.12. Denegación de servicio (DoS)

Amplio conjunto de técnicas cuyo objetivo es dejar inoperativo un servicio, impidiendo que esté disponible. Puede afectar a cualquier sistema ciberfísico que esté conectado a Internet o que dependa de redes que son utilizadas por usuarios externos. Aumenta la carga de procesado y de red. Retrasa o impide la actuación interrumpiendo la comunicación de instrucciones a los efectores o la recogida de datos desde los

sensores. Los retrasos en la comunicación con los sensores también pueden provocar actuaciones incorrectas indirectamente, al estar basadas en datos antiguos. Otro posible impacto es llevar al sistema de control a un estado no seguro. Igualmente puede aumentar el consumo de energía (Loukas, 2015, pp. 161-163).

- Ataques DoS a controladores. Cuando la denegación de servicio está destinada a sistemas de automatización que monitorizan o controlan procesos puede causar dos problemas en los controladores: pérdida de control (Loss of Control o LoC) y pérdida de control del DCS (Loss of View o LoV).

  La pérdida de control (LoC) suele tener como resultado que el proceso físico sea colocado en un estado 'seguro' (apagado). Eso implica que una simple interrupción en las funciones de control puede traducirse en graves consecuencias físicas.

  La pérdida de control del DCS (sistema de control distribuido) o LoV está relacionada con la interfaz humano-máquina (HMI), que no está directamente conectada al equipamiento mecánico. Sin embargo, si esa interfaz no es capaz de realizar sus funciones, se puede producir una pérdida LoV, que a menudo requiere que todo el proceso se apague si la vista de los datos no se puede restablecer a tiempo. Por ejemplo, en el caso de un sistema de control de ignición de un motor, si el controlador se para también lo hace el motor (Knapp y Langill, 2015, pp. 187-188).

- Ataque DoS a robots teledirigidos. Los robots portátiles para realizar intervenciones quirúrgicas pueden ser vulnerables a ataques DoS. Esos sistemas se pueden desplegar en campos de batalla, en zonas devastadas por desastres y en áreas rurales lejanas. Sin embargo, la naturaleza abierta e incontrolable de esos medios de comunicación entre robots y operadores hace que esos sistemas ciberfísicos sean vulnerables a gran variedad de amenazas de ciberseguridad, que no pueden ser evitadas usando métodos de criptografía tradicional (Bonaci et al., 2015).

- Ataques repentinos (rushing attacks). Son una categoría emergente de DoS que puede suceder en redes ad hoc vehiculares (VANET) sin conductor, con efectos directos y negativos sobre los protocolos de enrrutado. El vehículo fuente inunda la carretera con solicitudes al vehículo destinatario a través de la VANET. El vehículo que está siendo apresurado recibe una solicitud de carretera y mueve el paquete directamente a los vehículos de destino sin retardo. El paquete original es automáticamente descartado por el nodo destinatario como una copia del paquete, porque el nodo ya ha aceptado el paquete procedente del ataque repentino. Tales ataques son particularmente efectivos cuando están cerca del vehículo fuente o del vehículo destinatario (Alheeti, Gruebler, y McDonald-Maier, 2016, p. 3).

5.1.13. Comprometer la interfaz humano-máquina

En ocasiones, la manera más fácil de conseguir el control no autorizado de un sistema es servirse de las capacidades de una consola de interfaz humano-máquina (HMI o Human-Machine Interface). El atacante puede aprovechar vulnerabilidades

conocidas de un dispositivo y explotarlas para instalar un acceso remoto a la consola y comprometer un host. No hacen falta conocimientos específicos sobre las operaciones de control del sistema. Basta con la habilidad de interpretar una interfaz gráfica de usuario, pulsar botones y cambiar valores utilizando una interfaz especialmente diseñada para ser fácil de usar (Knapp y Langill, 2015, p. 189).

5.1.14. Esnifado de paquetes (packet sniffing)

Aplicable a cualquier sistema ciberfísico con redes inalámbricas o con cualquier clase de red potencialmente accesible a usuarios externos. Una vez ganado el acceso a una red, el adversario utiliza software de esnifado de paquetes para husmear en los mensajes transmitidos a través de la red. Se quiebra la confidencialidad de las comunicaciones y la privacidad física (Loukas, 2015, p. 172).

5.1.15. Fuzzing

Tras haber logrado acceder a una red, el atacante la bombardea con mensajes aleatorios y observa cuáles causan efectos físicos. Es una técnica efectiva en sistemas con autenticación débil o sin ella. Puede acabar desactivando motores o bloqueando frenos. También puede formar parte de la fase inicial de reconocimiento (Loukas, 2015, p. 166).

Hay tres variantes de fuzzing: normal, de protocolo y de máquina de estado. En el fuzzing normal, el adversario elabora los datos de entrada cambiando parte de una entrada correcta que ha grabado previamente. En el fuzzing de protocolo la entrada está generada según las especificaciones del protocolo en cuanto a formación de paquetes y dependencias entre campo. En el fuzzing de estado el objetivo no es encontrar errores y vulnerabilidades cambiando el contenido de los paquetes, sino alterar el estado de la máquina del software (Domin, Marin y Symeonidis, 2016, p. 2).

Los ataques de fuzzing se pueden aplicar a protocolos de comunicaciones como MAVLink (Micro Air Vehicle Communication Protocol), un protocolo de comunicación bidireccional entre un UAV y su estación terrestre de control, que va a convertirse en estándar mundial (Domin, Marin y Symeonidis, 2016, p. 2).

5.1.16. Interferencias (jamming) en comunicaciones

Actividades para producir interferencias intencionadas en la recepción de comunicaciones por ondas de radio. Sus formas más simples no son ataques ciberfísicos estrictos, porque se originan en el espacio físico. Su primer impacto en el ciberespacio es la interrupción de las comunicaciones, que puede afectar a procesos físicos, como retrasar o impedir actuaciones. Igualmente puede llegar a posibilitar actuaciones incorrectas o no autorizadas (Loukas, 2015, p. 160). Por ejemplo, interferir en los enlaces de comunicación entre un UAV y la estación terrestre de control puede imposibilitar el cumplimiento de la misión en los sistemas no tripulados teledirigidos.

Los vehículos no tripulados de tamaño más pequeño suelen ser más vulnerables a las interferencias. La mayoría de los UAV portables están alimentados por baterías

y no tiene capacidad de producir su propia energía. Cualquier energía redirigida para su autoprotección electrónica puede afectar drásticamente a su misión. Los UAS tácticos pueden generar la energía necesaria para operar sus sistemas, pero su capacidad para soportar requerimientos energéticos adicionales es cuestionable (Yochim, 2010, p. 76).

5.1.17. Interferencias (jamming) en GPS

El objetivo puede ser cualquier sistema ciberfísico que dependa de la disponibilidad de GPS (Global Positioning System) para posicionamiento, navegación y sincronización de tiempo, por ejemplo, un UAV o un UGV. El atacante transmite señales de alta potencia para impedir la recepción de las señales GPS legítimas. Incluso puede emitir señales GPS especialmente creadas para dañar el receptor GPS de manera que el vehículo no tripulado quede desorientado y no pueda regresar a la base (Loukas, 2015, pp. 82, 167).

Una idea errónea bastante difundida es que los receptores GPS militares son inmunes a las interferencias, incluso aunque la señal esté cifrada. Las señales de los satélites son tan débiles que incluso un inhibidor de frecuencias de 1 a 10 vatios puede negar la cobertura GPS en un área extensa, tanto para señales civiles como militares (Cole, 2016).

En diciembre de 2011, Irán capturaba un UAV de vigilancia de la CIA estadounidense. Un ingeniero iraní afirmaba que Irán había suplantado (spoofing) el GPS del drone con falsas coordenadas, engañándolo para que creyera que estaba cerca de casa y aterrizara en suelo iraní. Esas afirmaciones son cuestionables para muchos especialistas, porque el GPS militar está cifrado y es muy complicado de suplantar. Otros expertos sugieren que pudo ser un ataque combinado de interferencias al GPS militar y una suplantación de las señales del GPS civil (Shepard, Bhatti, Humphreys y Fansler, 2012, p. 3).

5.1.18. Infección de malware

Su objetivo principal son los sistemas ciberfísicos conectados a redes corporativas o a dispositivos convencionales y móviles. Producen una brecha en la integridad del sistema seguida de otras brechas potenciales, según la clase de malware. Puede producir todo tipo de impactos físicos. Por ejemplo, se puede introducir malware en dispositivos móviles que sirvan para interactuar con estaciones terrestres de control de un UAV (Loukas, 2015, pp. 82, 170).

Otro caso es la infección con malware de varios UAV del ejército estadounidense en 2011, cuyo resultado fue la instalación de un capturador de teclado. El motivo más probable es la creación un mapa entre las señales emitidas por las pulsaciones de teclado del piloto y las piezas correspondientes de la aeronave que eran activadas (Vuong, Filippoupolitis, Loukas y Gan, 2014, p. 2).

5.1.19. Inyección de código

Se introducen instrucciones adicionales en el código de un programa informático con intenciones maliciosas. Uno de los más populares es la inyección SQL, que explota los

fallos de diseño en los sitios web con bases de datos para facilitar el acceso completo a la base de datos. Puede utilizarse para ejercer un control no autorizado y transmitir información a un adversario, interrumpiendo la disponibilidad del sistema o causando daños físicos en ese u otro sistema (Loukas, 2015, p. 159).

5.1.20. Inyección de comandos

Puede afectar a cualquier dispositivo que controle un efector directa o indirectamente. El atacante accede a la unidad de control o una red para ejecutar un comando con intenciones maliciosas. El código que se ejecuta está definido en el propio sistema objetivo, por ejemplo, el comando frenar. Facilita actuaciones no autorizadas o incorrectas o también puede impedir ciertas actuaciones. Si el comando es activar un sensor o transmitir datos desde un sensor se produce una brecha de privacidad física (Loukas, 2015, p. 159).

5.1.21. Inyección de datos falsos en bases de datos, en comunicaciones y en sensores

En las bases de datos se explotan vulnerabilidades para borrar registros o añadir registros erróneos. Así se logran actuaciones incorrectas causadas por datos incorrectos.

En el supuesto de las comunicaciones, el objetivo puede ser cualquier sistema ciberfísico que dependa de datos de sensores recibidos a través de un canal de comunicación vulnerable. El atacante compromete el canal, bloquea los datos legítimos y transmite sus propios datos. Si la comunicación se realiza a través de una red, el adversario puede lograrlo secuestrando un nodo intermedio para usarlo como relé de datos hacia el centro de control. El impacto físico son actuaciones incorrectas causadas por datos erróneos.

En los sensores, el atacante compromete el controlador de un sensor para transmitir datos falsos en lugar de las medidas realmente tomadas por el sensor. El resultado son actuaciones incorrectas por estar basadas en datos equivocados (Loukas, 2015, pp. 163-164).

5.1.22. Man-in-the-Middle (MitM)

El adversario se sitúa en el canal de comunicación para intervenir en los mensajes intercambiados; puede tanto husmear activamente como inyectar o manipular mensajes. Una vía de aproximación habitual es primero comprometer un nodo existente o conectar un nodo deshonesto (rogue node) en la red objetivo. Luego explota las debilidades de los mecanismos de comunicación usados. Una de las técnicas más efectivas es el ARP spoofing o suplantar el protocolo de resolución de direcciones, en la que el adversario emite respuestas falsas a las solicitudes de la dirección física de un nodo legítimo, haciendo que otros nodos de la red crean que tiene la dirección del nodo deshonesto. Provoca brechas de confidencialidad, autenticidad e integridad. La brecha de disponibilidad es igualmente posible si el atacante dejar caer los mensajes intercambiados entre los dos ordenadores. Todas las formas de impactos físicos son posibles (Loukas, 2015, p. 170).

Cuando hay cifrado y autenticación, un ataque MitM puede escuchar los intercambios de claves y pasar la clave del adversario en lugar de la clave legítima. En muchos sistemas ciberfísicos la situación se agrava porque los dispositivos se pueden comunicar a través de sesiones que se establecen y permanecen intactas durante largos periodos de tiempo. El principal reto de es insertarse en el flujo de mensajes, que requiere convencer a ambos extremos de la conexión de que es el destinatario legítimo (Knapp y Langill, 2015, p. 187).

5.1.23. Modificación de firmware

Los sistemas embebidos, como muchos controladores, y el hardware de red son los objetivos primarios. Puede considerarse una subcategoría de los ataques de inyección de código. El atacante no necesita modificar el firmware de su objetivo final para alterarlo; basta con que modifique el firmware de otros dispositivos como un router o una impresora que estén en la misma red. Entre los posibles impactos físicos está conseguir el control no autorizado para transmitir información, causar daños físicos al sistema o hacer que el sistema deje de estar disponible (Loukas, 2015, p. 165).

5.1.24. Nodo fraudulento (rogue node)

Para sistemas con redes de controladores o de sensores. El atacante, utilizando equipamiento inalámbrico o cableado compatible, introduce un dispositivo fraudulento en una red para hacerlo pasar por un nodo legítimo. En ciertos casos se puede conseguir conectándolo físicamente a un puerto. El nodo fraudulento puede leer todas las comunicaciones en la red y generar sus propios mensajes, incluyendo comandos para los efectores. Esta brecha a la integridad del sistema, puede ser un paso previo que posibilite otros ataques, como DoS, MitM, inyección de comandos o esnifado de paquetes, entre otros. No existe un impacto físico directo, porque depende del ataque posterior que se lance desde el nodo fraudulento. El nodo fraudulento puede ser un UAV que vuele en las proximidades de un grupo de UAV, y que se conecte a su red para husmear en sus comunicaciones o inyectar información falsa (Loukas, 2015, pp. 82-175).

5.1.25. Privación de sueño

Su objetivo primario son los sistemas y dispositivos alimentados con baterías, como vehículos no tripulados, robots móviles o redes de sensores. La autonomía de los sensores y de los efectores está limitada por la carga de sus baterías. Por eso suelen estar configurados para desarrollar sus tareas con eficiencia energética. Un atacante puede agotar rápidamente sus baterías forzándoles para que nunca se pongan en espera (sleep) y estén constantemente en acción, o recibiendo, procesando o enviado datos. El impacto físico es aumentar el consumo de energía e incluso impedir la actuación de los dispositivos (Loukas, 2015, p. 175).

5.1.26. Rotura de contraseñas

La mayoría de los sistemas ciberfísicos tienen algún elemento protegido por contraseña. El atacante puede obtenerla usando técnicas de ingeniería social,

encontrándola en documentos o realizando un ataque de fuerza bruta valiéndose de diccionarios. Si el mecanismo de autenticación por contraseña emplea hashes, el adversario puede recurrir a las tablas rainbow. El impacto físico depende de la funcionalidad del sistema cuya contraseña se haya roto. Por ejemplo, se puede emplear para sortear el sistema de inmovilización de un vehículo no tripulado (Loukas, 2015, pp. 101-102, 172).

5.1.27. Suplantación de GPS (GPS spoofing)

Afecta a aquellos sistemas que requieren tener GPS para navegar, posicionarse y sincronizar el tiempo, caso de los vehículos no tripulados. El adversario sintetiza y transmite una señal GPS falsa (spoofing) para engañar a un receptor GPS sobre su posición real. En otras ocasiones el atacante captura y una señal GPS legítima y la reemite con un ligero retardo (meaconing); ese retardo afecta a la estimación de distancias respecto al satélite que realiza el receptor y, en consecuencia, a su propia posición. Es una brecha en la integridad de los datos GPS. Suele causar actuaciones incorrectas, interrumpiendo la capacidad del sistema para establecer su posición y navegar de forma autónoma. Puede llegar a facilitar el control no autorizado de objetivos en movimiento si son totalmente autónomos y dependen solamente del GPS (Loukas, 2015, pp. 167-168). El GPS militar está cifrado, por eso es más resiliente a estos ataques de suplantación.

5.2. MÉTODOS DE ATAQUE FÍSICO-CIBER

Los ataques físico-ciber ocurren en el espacio físico y producen consecuencias negativas en el ciberespacio. A continuación se recogen ejemplos de posibles métodos.

5.2.1. Ataques mediante daños físicos directos a equipos e infraestructuras

La disponibilidad de los datos en el ciberespacio depende del bienestar de la infraestructura física, como los discos duros donde están almacenados, los dispositivos de red y los cables a través de los que se transmiten, los procesadores o las fuentes de alimentación, entre otros (Loukas, 2015, p. 222). Los ataques se pueden realizar causando daños físicos directos a los equipos e infraestructuras.

5.2.2. Ataques mediante pulsos electromagnéticos (EMP)

Los ataques físicos directos mediante pulsos electromagnéticos son particularmente efectivos contra los dispositivos electrónicos comerciales que son muy vulnerables a los excesos de corriente o de voltaje. En cambio, los dispositivos de grado militar suelen estar apantallados contra interferencias electromagnéticas para protegerlos contra interferencias intencionadas (jamming) y el husmeo, y esa protección también es bastante eficaz contra los pulsos electromagnéticos. Por otro lado, los cables de fibra óptica son naturalmente resilientes contra EMP. Así, otra técnica de protección es reemplazar los cables de cobre por otros de fibra óptica (Loukas, 2015, pp. 225-226).

5.2.3. Ataques Emsec

Los ordenadores están formados por una multitud de componentes mecánicos y electrónicos. Cuando está operativos, y especialmente cuando procesan información, producen calor, sonidos, pérdidas de señales eléctricas a través de los cables de alimentación y radiaciones electromagnéticas. Esas emanaciones pueden ser valiosas fuentes de información.

Este tipo de ataques se suelen denominar Emsec (seguridad de las emanaciones), TEMPEST o de canal lateral. Las diferencias entre estos tres términos son difusas. Emsec generalmente suele emplearse en la mayoría de los casos. TEMPEST suele usarse sobre todo para las filtraciones de emanaciones electromagnéticas. Los ataques de canal lateral son habituales en el contexto de los análisis criptográficos (criptoanálisis de canal lateral), pero el término asimismo se aplica relacionado con emanaciones ópticas y acústicas o con el husmeo basado en sensores de movimiento (Loukas, 2015, pp. 233-247).

- Ataques de Emsec electromagnético. Se aprovechan de las pérdidas de radiaciones electromagnéticas para conseguir información sensible. Otras veces consisten en realizar mediciones sobre la operación de un dispositivo. Pueden revelar información sobre operaciones criptográficas y sobre otras configuraciones.
- Ataques de Emsec óptico. Tratan la luz como fuente de transmisión y representación de información. A veces basta con la observación directa del sistema, por ejemplo, con cámaras de vídeo o con cámaras térmicas.
- Ataques de Emsec acústico. Explotan las emanaciones de sonido, como el ruido de un teclado o los ruidos que produce el procesador o la placa base de un ordenador.
- Ataques de canal lateral contra sensores de movimiento en dispositivos móviles. Están dirigidos a dispositivos móviles que incorporan giróscopos y acelerómetros. Así se pueden detectar los movimientos del dispositivo. Y las interacciones con las pantallas táctiles también causan movimientos en el dispositivo que pueden servir para inferir qué se ha introducido.
- Ataques Emsec activos o ataques teapot[6]. El adversario estimula artificialmente las emanaciones del sistema para que resulten más fáciles de explotar.

5.2.4. Ataques físicos a sensores

El atacante modifica el entorno donde está el sensor. El método de aproximación depende del tipo de sensor, por ejemplo, espejos o cristales reflectantes para alterar un lidar[7], radiaciones infrarrojas contra sensores térmicos, generadores de ruido contra micrófonos o superficies anecoicas para engañar a sensores de ultrasonidos (Loukas, 2015, pp. 228-230).

6 Teapot es un alias derivado del modismo inglés “Tempest in a teapot”, que significa exagerar un suceso nimio. En círculos militares se usa para referirse a los ataques Emsec activos.

7 Lidar (Light Detection and Ranging o Laser Imaging Detection and Ranging), tecnología basada en haces de pulsos de láser empleada para calcular distancias.

Manipulando físicamente el entorno que monitorizan los sensores es posible introducir mediciones falsas. Los datos incorrectos de los sensores causan actuaciones incorrectas. Por ejemplo, esto es relevante para sistemas no tripulados que usan sensores continuamente para mapear el entorno y evitar obstáculos.

Un ejemplo son los ataques físicos a los sistemas antibloqueo de frenos (ABS) de vehículos a través de los sensores analógicos. El atacante usa las interacciones entre el entorno físico y los componentes de proceso embebidos para alterar el comportamiento de los cibercomponentes. La información recogida del entorno físico influye en las decisiones del controlador. Explota las debilidades de los sensores magnéticos de velocidad de las ruedas para inyectar medidas arbitrarias al ordenador que controla el ABS. Se coloca un fino efector electromagnético cerca de esos sensores que genera campos magnéticos para cancelar la señal medida verdadera e inyectar una señal maliciosa, suplantando la velocidad de las ruedas. El montaje del ataque es no invasivo (Shoukry, Martin, Tabuada y Srivastava, 2015).

5.2.5. Ataques de hardware

Explotan las vulnerabilidades del hardware. Sus objetivos pueden ser muy variados, pero el efecto final es el mismo: el comportamiento final del sistema es diferente del deseado por sus usuarios. Existen varios mecanismos para realizar un ataque de hardware. Algunos ejemplos son:

- Un troyano de hardware, que incluya un bloque hardware malicioso en el sistema para alterar su comportamiento global.
- Ataques de reenvío donde se suplanta la identidad de un bloque autorizado por otro bloque no autorizado en un proceso de comunicación.
- Ataques de inyección de fallos para aumentar su vulnerabilidad.
- La ingeniería inversa para obtener el comportamiento de los valores de señal (Gomez-Bravo et al., 2015, p. 148).

Otro ejemplo serían los ataques al hardware del piloto automático de un UAV (vehículo aéreo no tripulado). Pueden suceder cuando un atacante tiene acceso directo a cualquiera de los componentes del piloto automático y corrompe los datos almacenados en la placa del piloto automático o instala componentes adicionales que pueden corromper el flujo de datos. Estos ataques pueden tener lugar durante la fabricación y la entrega del aparato, o durante el mantenimiento o almacenamiento. El atacante también puede conectarse directamente con el piloto automático para dañarlo o reprogramarlo si tiene los medios, los sustituye o añade componentes que le proporcionen el control sobre el UAV y/o sobre los datos tácticos recopilados. Los ataques de hardware pueden afectar a la supervivencia del UAV, comprometer el control del UAV y comprometer datos tácticos recogidos por el UAV (Kim, Wampler, Goppert y Hwang, 2012, p. 6).

5.2.6. Ataques físicos a criptosistemas

Engloban aquellos ataques originados en el espacio físico que buscan romper sistemas criptográficos en el ciberespacio.

- Ataques de canal lateral[8] a criptosistemas

Estos ataques se centran en extraer información del estado físico de un sistema. En varios sistemas de criptografía cuántica se emplean espejos, y las posiciones y los movimientos de los espejos pueden revelar información. Aparte es posible realizar criptoanálisis a partir de pistas acústicas (Shostack, 2014, p. 343).

En el modelo de seguridad tradicional, los adversarios explotan las especificaciones matemáticas del protocolo. En cambio, los ataques de canal lateral aprovechan propiedades específicas de la implementación y del entorno operativo. Por eso también son conocidos como ataques de implementación. En la práctica, los algoritmos criptográficos siempre están implementados en hardware o software en dispositivos físicos que interactúan con el entorno. Esas interacciones físicas pueden ser instigadas y monitorizadas por adversarios, que pueden extraer información útil para el criptoanálisis.

En el extremo de Alice (remitente del mensaje cifrado), Eve (atacante pasivo) puede sacar información del sonido, de las radiaciones electromagnéticas, de la luz visible y del calor que emiten los dispositivos. En el extremo de Bob (receptor que descifra el mensaje), el adversario (Eve) puede explotar información del consumo de energía, del tiempo de ejecución, de las salidas fallidas, de los mensajes de error y de la frecuencia (Zhou y Feng, 2005, pp. 2, 6).

- Criptoanálisis de “manguera de goma”

Dañar físicamente a las personas que conocen la clave constituiría otro posible ataque. Una manera de romper un criptosistema sería golpear con una manguera de goma a la persona que sabe la clave secreta hasta que la revele (Shostack, 2014, p. 343).

5.2.7. Ataques físicos a sistemas de aprendizaje automático (machine learning)

Los robots perciben el mundo a través de cámaras y otros sensores, sistemas de video vigilancia y aplicaciones móviles para clasificar imágenes o sonidos. En ese escenario es posible crear ejemplos adversos que sirvan para realizar ataques adversos en los sistemas de aprendizaje automático que operan en el mundo físico y que perciben los datos a través de sensores, en lugar de a través de una representación digital.

Un caso serían los ataques con entradas de audio que los dispositivos móviles pueden reconocer, pero que resultan ininteligibles para los seres humanos. Otro sería usar órdenes de voz grabadas que simularan una canción, pero que contuvieran comandos que el algoritmo de aprendizaje automático puede reconocer. Un tercer caso sería un ejemplo adverso relacionado con el reconocimiento facial, que consistiría en aplicar marcas de maquillaje muy sutiles al rostro de una persona de manera que pasaran desapercibidas para un observador humano, pero que el sistema de aprendizaje automático le reconocería como una persona distinta. Generar ejemplos adversos de imágenes es otro procedimiento sencillo, que puede hacerse tomando fotos con la cámara de un teléfono móvil. Después esas imágenes eran alimentadas a un sistema de clasificación basado en una red neural, con el resultado de que una parte significativa era incorrectamente clasificada (Kurakin, Goodfellow y Bengio, 2016, pp. 2, 14).

8 En criptografía, canales laterales son aquellos canales con salidas involuntarias de un sistema.

6. CONCLUSIONES

La seguridad de los sistemas robóticos y autónomos (RAS) se puede analizar viéndolos como sistemas ciberfísicos. Los RAS son sistemas ciberfísicos, donde la computación, las comunicaciones y los procesos físicos están estrechamente acoplados y dependen unos de otros. Una característica esencial es la integración continua de los recursos de hardware y software con finalidades computacionales, de comunicación y de control, todos ellos diseñados conjuntamente con los componentes físicos.

La seguridad de los sistemas ciberfísicos presenta varias peculiaridades que la distinguen de la de otros sistemas de tecnologías de la información convencionales. En primer lugar, existen requerimientos de tiempo real, con tiempos de respuestas críticos. Además, ciertas prácticas típicas de seguridad, como el parcheo, no son tolerables, porque son incompatibles con la disponibilidad. En tercer lugar, las consecuencias de los fallos en los sistemas ciberfísicos pueden ser catastróficas. En cuarto, muchos dispositivos de estos sistemas tienen recursos limitados de computación y energía. En quinto, los elementos esenciales de la ciberseguridad (disponibilidad, integridad y confidencialidad) tienen significados ligeramente diferentes. Por último, el orden de prioridades de esos elementos es distinto; lo primero es la disponibilidad, seguida de la integridad y la confidencialidad.

Para analizar la seguridad de esos sistemas distinguimos entre ataques ciber-físicos y ataques físico-ciber.

Un ataque ciber-físico es una brecha de seguridad en el ciberespacio que afecta negativamente al espacio físico. Es una categoría particular de ciberataque que, intencionadamente o no, también perjudica al espacio físico apuntando a la infraestructura computacional y de comunicaciones que permite que las personas y los sistemas monitoricen y controlen los sensores y los efectores.

Un ataque físico-ciber es aquel desarrollado en el espacio físico que afecta negativamente en el ciberespacio. Explota las interacciones entre el espacio físico y el ciberespacio. Los sensores, los efectores, los controladores, los dispositivos de red y otros componentes de las infraestructuras de control o de comunicación son elementos físicos que existen en el espacio físico. Basta con dañarlos físicamente para interrumpir o impedir su operación. También se pueden aprovechar las entradas físicas a los sensores, las emanaciones que emiten los ordenadores y otros dispositivos, y el modo en que se implementan las técnicas criptográficas.

Las amenazas a la seguridad de los sistemas ciberfísicos, incluyendo los RAS, presentan similitudes: el atacante pretende identificar puntos de entrada desde donde es posible comunicar directamente con sensores y efectores, o desde donde puede afectar indirectamente su operación manipulando las infraestructuras de control y comunicaciones.

Hemos identificado los potenciales puntos de entrada en los ataques ciber-físicos desde donde se puede efectuar una intrusión. Hemos elaborado un diagrama genérico que resulta aplicable a cualquier sistema ciberfísico, incluyendo la gran variedad de sistemas robóticos y autónomos, independientemente del dominio donde operen.

Otra de nuestras aportaciones es una taxonomía de los métodos de ataque más comunes tanto ciber-físicos como físico-ciber, que es aplicable tanto a sistemas robóticos y autónomos como a otros sistemas ciberfísicos.

Finalmente, los sistemas robóticos y autónomos (RAS) ofrecen posibilidades fascinantes, pero también desafíos formidables. Ahora es el momento oportuno de analizar cuidadosamente los riesgos que presentan, antes de su despliegue masivo. Cuando se introducen los requerimientos de seguridad a posteriori en vez de al principio, la tecnología subyacente debe ser costosamente modificada para solucionar los problemas. Experiencias pasadas muestran que la seguridad funciona mejor como objetivo de diseño prefijado que como ocurrencia tardía.

REFERENCIAS BIBLIOGRÁFICAS

Alheeti, K. M. A., Gruebler, A., y McDonald-Maier, K. (2016). Intelligent Intrusion Detection of Grey Hole and Rushing Attacks in Self-Driving Vehicular Networks. Computers, 5(16).

Biggio, B. (2016). Machine Learning under Attack: Vulnerability Exploitation and Security Measures. IH&MMSec 2016, Vigo, España.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Nedim, S., Laskov, P., Giacinto, G., y Roli, F. (2013). Evasion Attacks Against Machine Learning at Test Time. En Blockeel, H., Kersting, K., Nijssen, S. y Železný, F. (Eds.), Machine Learning and Knowledge Discovery in Databases, Volume 8190 of the series Lecture Notes in Computer Science pp 387-402. Berlín: Springer.

Bonaci, T., Yan, J., Herron, J., Kohno, T., y Chizeck, H. J. (2015). Experimental Analysis of Denial-of-Service Attacks on Teleoperated Robotic Systems. ICCPS '15, Seattle, WA.

Casey, W., Memarmoshrefi, P., Kellner, A., Morales, J. A., y Bud Mishra, B. (2016). Identity Deception and Game Deterrence via Signaling Games. Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS), Nueva York.

Cheminod, M., Durante, L., y Valenzano, A. (2013). Review of Security Issues in Industrial Networks, IEEE Transactions On Industrial Informatics, 9(1), 277-293.

Cole, S. (2016, septiembre 19). Securing military GPS from spoofing and jamming vulnerabilities. Recuperado de http://mil-embedded.com/articles/securing-military-gps-spoofing-jamming-vulnerabilities/

Domin, K., Marin, E. y Symeonidis, I. (2016). Security Analysis of the Drone Communication Protocol: Fuzzing the MAVLink protocol. Recuperado de https://securewww.esat.kuleuven.be/cosic/publications/article-2667.pdf

Domingo, M. C. (2011). Securing Underwater Wireless Communication Networks. IEEE Wireless Communications, febrero, 22-28.

Dong, Y., y Pingxiang Liu, P. (2010). Security Considerations of Underwater Acoustic Networks. Proceedings of 20th International Congress on Acoustics, ICA 2010, Sydney, Australia.

Gomez-Bravo, F., Jiménez Naharro, R. Medina García, J., Gómez Galán, J., y Raya, M.S. (2015). Hardware Attacks on Mobile Robots: I2C Clock Attacking. En Reis, L. P., Moreira, A. P., Lima, P. U., Montano, L., y Muñoz-Martinez, V. (Eds.) Robot 2015: Second Iberian Robotics Conference - Advances in Robotics, Volume 2, Cham: Springer, pp. 147-159.

Kim, A., Wampler, B., Goppert, J., y Hwang, I. (2012). Cyber Attack Vulnerabilities Analysis for Unmanned Aerial Vehicles. Infotech@Aerospace 2012, Garden Grove, California, Estados Unidos.

Knapp, E. D. y Langill, J. T. (2015). Industrial Network Security: Securing Critical Infrastructure Networks for Smart Grid, SCADA, and Other Industrial Control Systems. Amsterdam: Elsevier.

Loukas, G. (2015). *Cyber-Physical Attacks: A Growing Invisible Threat*. Amsterdam: Elsevier.

Lun, Y. Z., D'Innocenzo, A., Malavolta, I., Di Benedetto, M. D., "Cyber-Physical Systems Security: a Systematic Mapping Study", 2016. Recuperado de https://arxiv.org/pdf/1605.09641v1.pdf

National Institute of Standards and Technology (NIST). (2017a, junio). NIST Special Publication 1500-201. Framework for Cyber-Physical Systems: Volume 1, Overview, Version 1.0. Recuperado de https://doi.org/10.6028/NIST.SP.1500-201

National Institute of Standards and Technology (NIST). (2017b, junio). NIST Special Publication 1500-201. Framework for Cyber-Physical Systems: Volume 2, Working Group Reports, Version 1.0. Recuperado de https://doi.org/10.6028/NIST.SP.1500-202

Papernot, N., McDaniel, P. y Goodfellow, I. (2016, mayo). Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. Recuperado de https://arxiv.org/pdf/1605.07277v1.pdf

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., y Swami, A. (2016, febrero). Practical Black-Box Attacks against Deep Learning Systems using Adversarial Examples. Recuperado de https://arxiv.org/pdf/1602.02697v3.pdf

Sanfelice, R. G., Analysis and Design of Cyber-Physical Systems: A Hybrid Control Systems Approach. En Rawat, D. B., Rodrigues, J. y Stojmenovic, I. (Eds.), *Cyber Physical Systems: From Theory to Practice*. Boca Raton: CRC Press, pp. 3-31.

Shepard, D. P., Bhatti, J. A., Humphreys, T. E., y Fansler, A. A. (2012). Evaluation of Smart Grid and Civilian UAV Vulnerability to GPS Spoofing Attacks. 2012 ION GNSS Conference, Nashville, TN.

Shostack, A. (2014). Threat Modeling: Designing for Security. Indianapolis: Wiley.

Shoukry, Y., Araujo, J., Tabuada, P., Srivastava, M., y Johansson, K. H. (2013). Minimax Control For Cyber-Physical Systems under Network Packet Scheduling Attacks. HiCoNS'13, Philadelphia, Estados Unidos.

Shoukry, Y., Martin, P., Tabuada, P., y Srivastava, M. (2015). Non-invasive Spoofing Attacks for Anti-lock Braking Systems. IACR 2015. Recuperado de https://eprint.iacr.org/2015/419.pdf

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., y Ristenpart, T. (2016). Stealing Machine Learning Models via Prediction APIs. Proceedings of the 25th USENIX Security Symposium, Austin, TX. Recuperado de https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/tramer

Urbina, D. I., Giraldo. J., Cardenas, A. A., Valente, J. y Faisal, M. (2016, noviembre).

Survey and New Directions for Physics-Based Attack Detection in Control Systems. NIST (National Institute of Standards and Technology), NIST GCR 16-010.

US Army. (2016, 30 septiembre). *The US Army Robotic and Autonomous Systems Strategy*.

Vuong, T., Filippoupolitis, A, Loukas, G. y Gan, D. (2014). Physical Indicators of Cyber Attacks Against a Rescue Robot. 2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS), 338-343.

Wang, H., Zhao, H., Zhang, J., Ma, D., Li, J. y Wei, J. (2018). Survey on Unmanned Aerial Vehicle Networks: A Cyber Physical System Perspective [Preimpresión].

Wen, H. (2013). Physical Layer Approaches for Securing Wireless Communication Systems. Nueva York: Springer.

Xiao, H., Biggio, B., Brown, G., Fumera, G., Eckert, C., y Roli, F. (2015). Is Feature Selection Secure against Training Data Poisoning? Proceedings of the 32nd International Conference on Machine Learning, Lille, Francia. Recuperado de http://www.jmlr.org/proceedings/papers/v37/xiao15.pdf

Yochim, J. A. (2010). The Vulnerabilities of Unmanned Aircraft System Common Data Links to Electronic Attack. Master's thesis. Recuperado de https://fas.org/irp/program/collect/uas-vuln.pdf

Zhou, Y. y Feng, D. (2005). Side-Channel Attacks: Ten Years After Its Publication and the Impacts on Cryptographic Module Security Testing. Recuperado de http://csrc.nist.gov/groups/STM/cmvp/documents/fips140-3/physec/papers/physecpaper19.pdf

Fecha de recepción: 16/01/2019. Fecha de aceptación: 20/06/2019