A. Garcia, C. Alcaraz, and J. Lopez, "MAS para la convergencia de opiniones y deteccin de anomalas en sistemas ciberfsicos distribuidos ", *VIII Jornadas Nacionales de Investigacin en Ciberseguridad (JNIC)*, 2023.
NICS Lab. Publications: `https://www.nics.uma.es/publications`

MAS para la convergencia de opiniones y detección de anomalías en sistemas ciberfísicos distribuidos

Alberto García, Cristina Alcaraz y Javier Lopez

Departmento de Lenguajes y Ciencias de la Computación

Universidad de Málaga, Spain

{albertogr, alcaraz, javierlopez}@uma.es

Resumen

Los grandes avances que estamos viviendo actualmente en las Tecnologías de la Información están revolucionando todas las industrias, haciendo que gran parte de los procesos operacionales de estas sufran cambios radicales. Esto propicia la aparición de los llamados sistemas ciberfísicos o *Cyber-Physical Systems*, CPS). Sin embargo, esta nueva realidad también hace a estos sistemas se vuelvan más complejos y vulnerables a los ciberataques tradicionalmente asociados a los sistemas informáticos. Es por ello, que en este paper se propone el uso de sistemas multi-agentes para la monitorización, gestión y trazabilidad de amenazas que puedan surgir en los CPS. Esto puede facilitar las labores de los administradores de sistemas y los analistas de seguridad encargados de proteger y garantizar el correcto funcionamiento de estas infraestructuras tan críticas para la sociedad actual. Asimismo, se analizan las ventajas e inconvenientes que plantean este tipo de arquitecturas y su aplicabilidad en distintos escenarios y contextos como las plantas de manufacturación o las redes de estaciones de carga para vehículos eléctricos.

Keywords: Detección avanzada, Sistemas multi-agente, machine-learning, dinámicas de opiniones, sistemas ciberfísicos.

Tipo de contribución: *Investigación original*

1. Introducción

Cada vez más estamos testimoniando como las nuevas corrientes tecnológicas de información (p. ej. IoT (del inglés, *Internet of Things*), la computación en nube o en borde, sistemas de simulación, etc.) tienden a converger en escenarios equipados principalmente con tecnologías operativas de naturaleza crítica, como pueden ser controladores, sensores y actuadores. El resultado es entonces un sistema complejo, compuesto de elementos tanto lógicos como físicos, conectados a través una red de comunicación para permitir confluir datos entre espacios (p. ej. mediciones e instrucciones de comando y control) y operar de manera más inteligente y distribuida con el mundo físico, e

independientemente de la dimensión del sistema y su alcance. Sin embargo, también estamos testimoniando como las nuevas corrientes tecnológicas conllevan a nuevos riesgos de seguridad. Los sistemas ciberfísicos (CPS, del inglés *Cyber-Physical Systems*) tienden a ser cada vez más susceptibles a múltiples tipos de ataques [1]. Recientemente, la Agencia de la Unión Europea para la Ciberseguridad (ENISA, del inglés *European Union Agency for Cybersecurity*) publicó las amenazas más influyentes y peligrosas hasta la fecha [2], como es el caso del malware, la denegación de servicio, y las amenazas provenientes de Internet y la cadena de suministro (hardware, software y open-source). A esto hay que sumar también los casos de ataques a entornos críticos reales, como son el SolarWinds o el Kaseya - ambos reportados en [3] -, que ya han dejado entrever la necesidad urgente de incorporar nuevos mecanismos de protección basados en tecnologías avanzadas que sirvan de soporte a los mecanismos de seguridad existentes.

Por tecnologías avanzadas destacamos aquellas relacionadas con el procesamiento automático de datos, como son la Inteligencia Artificial (IA) y el aprendizaje automático (ML, del inglés *Machine-Learning*), pero también el despliegue de agentes software capaces de retroalimentar todo este procesamiento de datos y dar contexto a lo que está pasando. Dicho de otra manera, la sensorización del entorno y sus recursos mediante el despliegue de sondas específicas pueden ayudar a tener una mejor visión y entendimiento de lo que está ocurriendo y en todo momento, apoyando con esto la "consciencia situacional". Es más, Mica R. Endsley, madre del concepto, ya indicaba que la consciencia situacional debe estar fundamentada en mecanismos que den soporte a la percepción, a la comprensión de la situación, y a la proyección de los estados en un tiempo próximo, especialmente cuando dichos entornos son de naturaleza dinámica [4]. Por esta razón, este artículo aborda las tres fases de la consciencia situacional, donde la sensorización es llevada a cabo por agentes software. Cada agente es capaz de medir un conjunto de variables de un recurso y estimar anomalías mediante el uso de modelos de ML (a ser posible no supervisados [5] para garantizar autonomía y rapidez en la toma de decisión) y, de manera colaborativa, consensuar junto con sus agentes vecinos el estado de salud del sistema para un momento particular en el tiempo. Es aquí donde aparece el concepto de **opinión** entendido como un valor que aporta información sobre el estado contextual, como, por ejemplo, la tasa porcentual de anomalías detectadas, o simplemente un valor que representa un determinado nivel de anomalía. La correlación de todas estas opiniones puede dar lugar a una opinión más global que muestre y monitorice estados contextuales que van más allá de mirar en un sólo nodo o en una única área concreta del sistema.

El resultado es, por tanto, un sistema de multi-agente (MAS, *Multi-Agent System*), cuyas funciones son la de percibir el contexto, derivar anomalías y compartir opiniones para tener una interpretación y proyección más precisa y fiel de la situación. Hasta la fecha, se han propuesto varios trabajos relacionados. Mirando en la literatura, observamos que efectivamente el uso de MAS en CPS complejos está cada vez más extendido, y pueden encontrarse aplicaciones en sistemas de transporte [6], en redes de tipo SG [7, 8, 9] y en el campo de la sanidad [10]. Todos ellos tienen la particularidad de ser infraestructuras críticas que presentan un elevado número de componentes, subsistemas e interacciones, lo que incrementa la superficie de ataque y los hace especialmente vulnerables a ciberataques, accidentes causales y otros eventos inesperados [11]. Sin

embargo, ninguno de los trabajos recopilados se centra en la importancia que tiene crear opiniones completas que ayuden a conducir correlaciones más significativas y válidas: "*cuanto más información se tenga de un contexto, más significativo será su valor*". Por ejemplo, el trabajo [12] muestra la viabilidad del MAS simulado y los modelos de ML para proteger sistemas críticos frente a fallos inesperados, pero sin abordar anomalías causadas por ciber-ataques. Este último aspecto, sí que lo considera el trabajo [8], pero centrando el estudio en ataques a nivel de comunicaciones (denegación de servicio, reenvío, modificaciones en los datos, e inyección de datos falsos) sin mirar otros tipos de amenazas más específicos al rendimiento físico de los nodos operativos. Igualmente, el trabajo [9] define un MAS que hace uso de un modelo de clasificación multi-clase supervisado para detectar anomalías en un CPS de redes eléctricas inteligentes (SG, del inglés *Smart Grid*), pero sin explorar la relevancia que tiene la compartición de opiniones y su correlación, y el rol que tiene los modelos de ML no supervisados para contextos complejos y distribuidos como son los CPS de SG. Por tanto, las **principales contribuciones** de este trabajo son: (1) apoyar y potenciar la consciencia situacional mediante el uso combinado de tecnologías avanzadas como son el MAS, los modelos de ML (especialmente los no supervisados) y los modelos de convergencia de opiniones para garantizar la correlación de estados; (2) extraer posibles indicadores de salud que permitan ofrecer una opinión más fidedigna a la situación y específica para sistemas ciberfísicos; (3) identificar posibles modelos de convergencia de opiniones que puedan ayudar a correlacionar estados contextuales y contribuir a una opinión única; y, por último, (4) analizar posibles aplicaciones y limitaciones que pueden conllevar el despliegue de MAS en sistemas ciberfísicos distribuidos en entornos críticos.

El artículo se organiza de la siguiente forma. En la sección 2 se introduce el concepto de MAS para la gestión y monitorización de sistemas ciberfísicos. En la sección 3 se propone un MAS y se profundiza en su funcionamiento interno, con énfasis en la detección de anomalías y en la síntesis de opiniones. En la sección 4, se exponen algunos escenarios en los que estos sistemas de agentes son aplicables, mientras que en la sección 5 se detallan las limitaciones y problemas que pueden derivar del uso de MAS, siempre desde el punto de vista de la seguridad del CPS. Se concluye en la sección 6 con un listado de las implicaciones que acarrea la integración de MAS en sistemas con componentes ciberfísicos.

2. Despliegue de un MAS para generar opiniones

Para la gestión y monitorización de sistema ciberfísicos complejos en términos de alcance y dimensión, se propone en este trabajo un sistema MAS distribuido donde los diversos agentes son capaces de instalarse y configurarse a lo largo del entero sistema. De esta forma se consigue una mejor trazabilidad de los eventos de seguridad de interés para el estudio y una visión más completa de todos los recursos que componen el CPS, incluso estando éstos en distintos emplazamientos. Para ello, se debe implementar mecanismos de monitorización que reporten periódicamente las métricas de rendimiento, los eventos de seguridad y las alarmas generadas a un sistema central desde el que se lleve el registro y el control de todos los datos recolectados. Como ya se ha mencionado en la introducción, los MAS se caracterizan por estar compuestos de

entidades software autónomas denominadas agentes, que poseen un objetivo individual y que cooperan entre ellos con tal de alcanzar un objetivo grupal [13]. Los MAS, por su propia naturaleza, son particularmente útiles en tareas de monitorización de redes de dispositivos ciberfísicos distribuidos, en los que tienen lugar eventos complejos que suponen la interacción de varios nodos del sistema [8].

En este sentido, el uso de MAS para la monitorización y gestión íntegra de estados de CPS proporciona, frente a las técnicas de monitorización tradicionales, un mayor grado de flexibilidad y adaptabilidad para la detección (temprana) de anomalías y la respuesta a éstas, al favorecer la generación de fuentes valiosas de información y útiles para estimar o detectar estados anómalos. En relación con esto, los MAS también mejoran y facilitan la toma de decisiones por parte de los administradores del CPS, ya que hacen uso de datos e información obtenida de multitud de fuentes que se toman de forma local al recurso observado, para razonar una mejor estrategia de defensa y respuesta temprana a incidentes. A su vez, esta forma de gestionar situaciones también puede ayudar a actualizar las políticas de seguridad y los procedimientos de actuación, si mecanismos adicionales al MAS se aplican para valorar el grado de precisión y acierto de la detección y la respuesta, como puede ser el F-Measure/F-Score, lo que implica también computar de manera automática y permanente la tasa real de falsos positivos y negativos. Luego, está claro que el uso de agentes distribuidos y coordinados, afecta positivamente a la resiliencia y la seguridad del CPS, a la adaptabilidad del sistema tras una amenaza potencial y la gobernancia de la organización, consecuentemente, afectando a la protección de los servicios esenciales y su disponibilidad. Diríamos entonces que todo va en cadena y cualquier efecto (positivo o negativo) tiene un impacto en la/s infraestructura/s crítica/s.

Técnicamente, cada agente dentro de un MAS debe/puede tener un objetivo individual diferente. En los sistemas de monitorización distribuidos para CPS, podemos diferenciar principalmente dos tipos de agentes según la función que desempeñen: los *agentes de monitorización* y el *agente recolector*. Los primeros son las entidades software que se instalan en cada uno de los nodos que se encuentran distribuidos por toda la zona por la que se extiende el CPS. Su cometido es el de tomar datos de diagnóstico y operacionales del dispositivo monitorizado de forma periódica. A partir de estas mediciones, los agentes computan un estado de salud del dispositivo, al que se le da el nombre de *opinión*. Para ello, pueden valerse de distintas técnicas de detección que se expondrán en la sección 3.1. Para obtener un estimación consensuada del estado de salud de un contexto de área, los agentes de monitorización cuentan con un mecanismo de intercambio de opiniones, con el que informan al resto de agentes de monitorización que se encuentren físicamente próximos, o en la misma área dentro del CPS. Esto permite obtener una estimación del estado del sistema más fiable y que, además, puede ser trazado por zonas, de forma que se pueda identificar rápidamente el segmento de la infraestructura afectado por fallos hardware/software, o eventos o incidentes de seguridad.

El segundo tipo de agente presente en nuestro MAS de monitorización es el agente recolector. Su propósito fundamental es el de centralizar todas las opiniones sobre el estado del sistema generadas por los agentes de monitorización desplegados por toda la infraestructura. Una vez en posesión de toda la información, este agente procesa y correlaciona las opiniones locales y las opiniones por área para obtener una opinión

global. Esto propicia una visión global y completa de todo el sistema, desde un único punto y en tiempo real, lo que permite a los administradores y a los analistas de seguridad realizar una toma de decisiones más rápida y precisa, algo que puede suponer un ahorro de tiempo clave en respuesta a incidentes en entornos críticos.

3. Modelo de convergencia de opiniones

En esta sección se propone una implementación de un modelo MAS para la monitorización de CPS donde el flujo de opiniones comienza en cada uno de los agentes de monitorización. Cada uno de éstos, en base a ciertos indicadores obtenidos de forma local, ejecutan una detección de anomalías y generan una opinión local, representando el estado de salud del dispositivo concreto en el que se ejecuta el agente. Esta opinión la concebimos como una variable numérica continua, que fluctúa entre cero (estado anómalo) y uno (estado normal del sistema), es decir: $[0 - 1]$. Sin embargo, esto no se queda aquí, pues los agentes que forman parte de una misma zona del CPS, intercambian sus opiniones como parte de un proceso denominado *Opinion Dynamics* (OD, dinámicos de opinión), que trata de formar una opinión consensuada entre agentes vecinos. Por último, estas opiniones locales, ya influenciadas por el OD, son recolectadas por el agente recolector, que las usa para promediarlas y obtener las opiniones globales.

Este procedimiento, aparentemente simple, requiere del uso de técnicas y mecanismos adicionales para computar el valor de la opinión, lo que le hace ser, posteriormente, un procedimiento complejo y computacionalmente costoso, como veremos en las siguientes secciones.

3.1. Detección de anomalía e indicadores de salud

La detección de anomalías se basa en comparar el comportamiento actual del sistema con el comportamiento que ha sido previamente registrado, de tal manera que cualquier desviación del comportamiento habitual se considerará como un estado anómalo, y, por lo tanto, dará lugar a un descenso en la opinión local (estado de salud del nodo). Existen numerosas técnicas de detección de anomalías que han ido surgiendo con el tiempo y que siguen surgiendo actualmente, aunque las más comunes son las que emplean técnicas estadísticas, algoritmos de Big Data y los avances en IA, en particular, en aprendizaje automático o ML [5, 14, 15].

Los detectores basados en ML se fundamentan en la elaboración de modelos matemáticos que, con el tiempo, van aprendiendo y mejorando su capacidad de detección gracias a la obtención de datos sobre el sistema a monitorizar. Tradicionalmente, dentro del ML se distinguen tres enfoques con respecto al funcionamiento del algoritmo y los datos empleados para su entrenamiento: el aprendizaje supervisado, el aprendizaje no-supervisado y el aprendizaje semi-supervisado. Cuando se trata de detección de anomalías, la mayoría de técnicas se asocian al aprendizaje no-supervisado, pues lo que se intenta es identificar patrones y relaciones entre los datos de entrada para poder agruparlos en categorías (anómalo/no-anómalo). La principal ventaja de este tipo de detecciones es que son capaces de identificar amenazas y eventos que no se habían observado anteriormente en comparación con los sistemas de detección tradicionales

basado en firma, además de favorecer a la autonomía de la propia detección y la toma de decisión. Sin embargo, como contrapartida, los detectores no-supervisados tienen un ratio de falsos positivos sensiblemente superior al no haber una supervisión objetiva del problema.

Los agentes de monitorización computan la opinión local del dispositivo en el que se alojan a partir de una serie de mediciones que realizan, y que sirven como indicadores del estado actual del nodo. Estos indicadores se pueden clasificar en distintos grupos dependiendo de su naturaleza:

- *Indicadores de medición*: se trata de aquellos dinámicos que dependen de la aplicación concreta del CPS. Por ejemplo, si el agente de monitorización se ejecuta sobre un controlador de carga de una batería, el agente deberá ser capaz de detectar desviaciones en las mediciones habituales de tensión, corriente, potencia, etc. Sin embargo, si el agente se aloja en un brazo robótico de una fábrica, deberá hacer lo mismo con valores de velocidad, aceleración y rotación del brazo. En la figura 1 pueden verse algunos ejemplos de dinámicos monitorizados según el escenario y el contexto de aplicación del MAS.

- *Indicadores de nodo*: en esta categoría se encuentran los indicadores de rendimiento y diagnóstico del propio módulo de computación del dispositivo. Es decir, los agentes monitorizan el uso del procesador, el uso de la memoria, el almacenamiento disponible, las lecturas/escrituras en disco, el número de procesos, los accesos al sistema. En esta categoría también se consideran aquellos logs que pudieran encontrarse en un dispositivo que contenga o se produce, además, por alguna herramienta convencional de seguridad como pueden ser firewalls, sistemas de detección de intrusiones (IDS/IPS, del inglés *Intrusion Detection/Prevention Systems*), software antivirus o anti-malware con firmas de ataques potencialmente peligrosos y conocidos, registros de acceso por cuenta de usuario y sistema operativo, etc. En este sentido, es ampliamente aconsejable mantener actualizado los mecanismos detección tradicionales, teniendo presente la conexión con fuentes externas como es MITRE ATT&CK conteniendo tácticas y técnicas de ataque para sistemas de control industrial [16].

- *Indicadores de red*: monitorizar la red se vuelve de vital importancia, ya que estamos hablando de un sistema interconectado en el que las comunicaciones entre nodos juegan un papel fundamental. Es por ello que los agentes deben tomar periódicamente medidas acerca del uso de los interfaces de red del dispositivo, las latencias observadas hacia los nodos vecinos, el ancho de banda disponible de éstos y la pérdida de conexión con otros dispositivos de la red [17].

- *Indicadores de posicionamiento físico y lógico*: en este tipo de indicadores sólo se encuentran presente en los CPS basados en sistemas móviles de los que se habla en la sección 4.2. Estos se caracterizan por sus continuos cambios en su topología física y, por tanto, también pueden cambiar su topología lógica a nivel del red (un dispositivo cambia de subred en un entorno de comunicaciones inalámbricas). En estos contextos, un dispositivo ciberfísico puede estar configurado para desplazarse por ciertos sectores (por ejemplo, un robot dentro de un

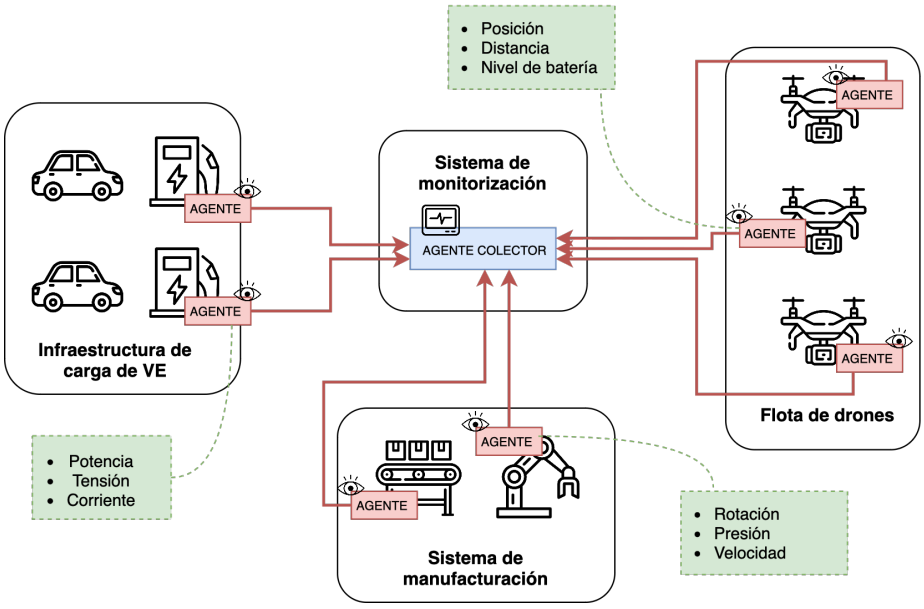

Figura 1: Ejemplos de dinámicos monitorizados por contexto de aplicación

almacén logístico), por lo que cambios en la topología que supongan que dicho dispositivo salga de estos límites, deben ser detectados por los agentes y reportados al sistema central. Además, este tipo de indicadores son útiles para realizar correctamente el intercambio de opiniones entre agentes de una misma área o para la recolección de opiniones por el agente recolector.

Todos estos indicadores se usan como fuentes de datos de entrada al detector de anomalías ML, para obtener de forma periódica las opiniones de un agente. Cuando hay varios agentes por área, es posible identificar el estado de salud de la área por aplicarse alguna técnica específica de consenso, ya sea por votación o por técnicas más específicas al uso y a agentes dinámicos como son OD [18]. En la siguiente sección se expone el funcionamiento de dicha técnica y cómo se usa para trazar amenazas dentro del CPS distribuido.

3.2. Convergencia de opiniones mediante OD

El consenso por OD, en general, corresponde normalmente a un técnica soportada por algún modelo matemático, utilizado para predecir la evolución de opiniones o creencias en un sistema compuesto por individuos que interactúan entre sí [17]. Este modelo estudia el proceso de la formación de opinión en una sociedad, mediante la fusión de las opiniones individuales de cada agente. Dicho proceso de fusión puede realizarse de múltiples formas, en función del contexto y la situación que se esté modelando, entre las que destacan [19]:

- *El modelo DeGroot*: es considerado el modelo clásico, en el que se realiza una media ponderada de las opiniones de los agentes. El peso otorgado a cada una de las opiniones es invariable en el tiempo.
- *El modelo de confianza limitada*: aquí el peso que se otorga a cada opinión a la hora de hacer la media puede cambiar en función del tiempo o del propio valor de opinión. En este caso, la opinión de un agente solo será influenciada por aquellos agentes, cuya opinión difiera no más de un cierto umbral de confianza.
- *El modelo del votante*: aquí las opiniones son binarias (o cero o uno), y cada agente adopta la misma opinión que la de un agente seleccionado aleatoriamente de entre sus vecinos.

Considerando la naturaleza dinámica del MAS, la integración de uno de estos modelos requerirá que cada uno de estos individuos (mencionados anteriormente) sea un agente de monitorización que interactuará con los agentes próximos a él, intercambiando sus opiniones con ellos. Las opiniones de los agentes vecinos las usará para promediarlas con la suya propia, de forma que solo las opiniones parecidas se usarán para el cómputo del promedio (modelo de confianza limitada). La idea es que tras varias iteraciones del algoritmo, los agentes de una misma área lleguen a un consenso y sus opiniones converjan hacia un mismo valor, y lo que correspondería con: la opinión de área o grupal. La técnica de OD permite obtener una imagen de todo el sistema por zonas, y detectar qué zonas están siendo afectadas por un evento significativo o un amenaza, de forma que se pueda trazar y seguir la evolución de ésta y tomar acciones más acertadas con el objetivo de mitigarla [20]. Por tanto, y por simplificar, existen tres tipos de opiniones: local (al nodo), grupal (por área) y global (de todo el sistema).

El despliegue tecnológico del MAS no necesariamente debe seguir las mismas pautas de diseño, es decir, todo centralizado en un servidor. El cómputo de la OD se puede: (i) descentralizar en otros tipos de infraestructuras específicas, como la computación borde (en el Edge); (ii) aprovechar los recursos computaciones de la computación en nube (en el Cloud), con el fin de computar largos volúmenes de datos; o (iii) distribuir la computación en todo el sistema mediante la configuración de sistemas híbridos [21]. Es por ello que la construcción de un MAS y su núcleo de computación puede ser muy selectiva, y esto, puede, incluso, aliviar la carga de trabajo de los nodos de la red que, en ocasiones, pueden presentar recursos muy limitados por tratarse de dispositivos embebidos o de propósito específico. Además, estas configuraciones son preferibles en situaciones en las que los dispositivos ejecutan procesos críticos y el cálculo de la OD puede afectar a su desempeño.

En la figura 2, se muestra un diagrama de flujo con la metodología seguida por los agentes del MAS para la formación de las distintas opiniones (local, grupal y global) que reflejan el estado del CPS.

4. Caso de usos y beneficios en CPS

Para mostrar la utilidad de los modelos de convergencia de opiniones y la integración de los MAS para su generación, esta sección analiza varios casos de uso y su

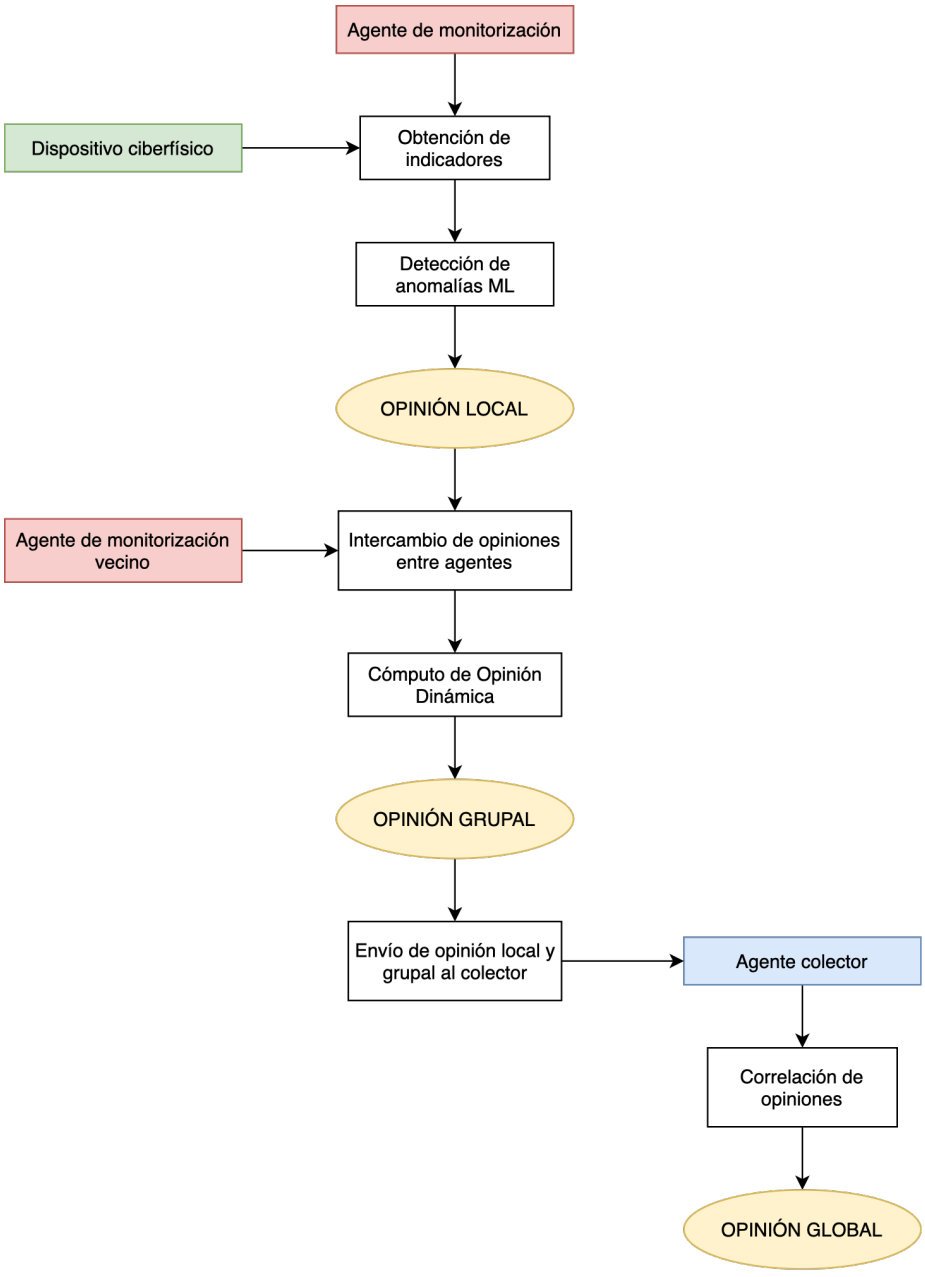

Figura 2: Diagrama de flujo de la metodología del MAS

aplicación, como pueden ser la monitorización constante de sistemas ciberíficos estáticos o móviles desplegados en entornos industriales, la trazabilidad y el seguimiento en vivo de ataques en Centros de Operaciones de Seguridad (SOC, del inglés *Security Operations Center*), y la concienciación y validación de sistemas mediante la integración de MAS en gemelos digitales (DT, del inglés *Digital Twins*). Para los análisis se asume la presencia de entidades maliciosas (*insiders* u *outsiders*) capaces de no solo acceder al escenario y su sistemas integrados (ej. sistemas ciberfísicos o el MAS) sino también interactuar con estos para perturbar su comportamiento, su seguridad y estado natural.

4.1. Monitorización para sistemas ciberfísicos estáticos

Muchos de los sistemas ciberfisicos tradicionales (p. ej. sistemas SCADA (Supervisory Control And Data Acquisition) con sus controladores, sensores y actuadores) se despliegan en infraestructuras críticas consideradas esenciales para la sociedad y su economía, como pueden ser, entre muchas, las infraestructuras de carga de VE (Vehículos Eléctricos) y sus microgrids, así como los sistemas de manufacturación junto con su cadena de suministro. En ambos contextos, observamos que el rol del MAS puede ser fundamental para la monitorización en tiempo real de los principales recursos operativos, cuyas despliegues se realizan normalmente de una manera distribuida y esparcida en varios puntos de localización del sistema. Es por ello, que la propia automatización del MAS es lo que añade sencillez a las tareas de diagnóstico, al suprimir todas aquellas acciones de supervisión manual y local al recurso observado.

En el caso específico de las redes de cargadores de VE, éstas se componen de multitud de estaciones de carga que actúan como los componentes ciberfísicos que aportan la energía eléctrica al VE, y que son gobernados por un sistema de gestión central a través del protocolo OCPP (Open Charge Point Protocol). La integración de agentes de monitorización en cada una de las estaciones permite al MAS, conocer el consumo eléctrico observado por cada una de ellas (potencia, intensidad de corriente, etc.), además de visualizar en tiempo real el estado de los conectores de la estación y las sesiones de carga que realizan los usuarios. Este hecho, posibilita la detección de abusos por parte de usuarios que acaparan los servicios de carga, dejando al resto de usuarios sin energía. También, ayuda a verificar que se cobra correctamente a los usuarios en función de su consumo, evitando, así, el fraude energético. Por su parte, el agente recolector se aloja en el sistema de gestión central de OCPP desde el que tiene visión de todos los agentes de monitorización de la infraestructura. En este tipo de escenarios se vuelven de vital importancia los indicadores de anomalías de red, pues pueden detectar si alguna estación de carga deja de estar conectada. Esta situación es buscada por los atacantes, que aíslan una estación para que pase a un modo de operación *offline* desde el que autoriza todas las sesiones de carga de los usuarios por motivos de continuidad de negocio [22, 23, 24].

También el MAS puede integrarse en escenarios específicos de manufacturación para estimar desviaciones cometidas por sus sistemas ciberfísicos distribuidos a lo largo de una planta o entre plantas. En este sentido, las funciones de los agentes serían equivalentes a las ya descrita en el ejemplo anterior, donde el objetivo es medir el rendimiento real de los elementos operativos, pero con la salvedad de que para ciertas

opiniones, sus valores van a depender de las características del escenario y del contexto de aplicación. En este caso, los indicadores de medición y sus correspondientes dinámicos pueden variar de acuerdo a la aplicación del sistema de manufacturación, estado del producto producido o transportado (ej. nivel de temperatura, presión, etc.), pero también de aquellos indicadores específicos de los nodos observados, como puede ser la velocidad de la cinta transportadora, el nivel de calibración y rotación del brazo robot, etc., y de los indicadores de red en el que los protocolos de comunicación varían de acuerdo a la aplicación, como, por ejemplo, TCPROS (TCP based Robot Operating System protocol), ModbusTCP y OCP-UA, entre otros. Por tanto, los beneficios serían equivalentes a los ya mencionados arriba y no solo en términos de malfunciones y diagnostico, también de abusos deliberados (ej. acceso continuado a recursos específicos), de accesos ilícitos a dominios operativos y críticos, o la manipulación ilegítima de procesos, estados, funciones o evidencias, entre otras muchas detecciones.

4.2. Monitorización para sistemas móviles autónomos

En la subsección anterior se ha abordado el rol de los MAS para el diagnóstico en tiempo real de elementos ciberfísicos desplegados de manera estática en determinados verticales. Ahora extendemos las discusiones para considerar la naturaleza móvil y autónoma de algunos de estos elementos ciberíficos, como pueden ser los vehículos autónomos, drones o robots móviles. En entornos industriales y operativos, estos dispositivos están siendo cada vez más demandados por permitir llevar a cabo las tareas más repetitivas e inasumible por el ser humano (ej. transporte de carga pesada de manera automática, acceso a pozos, áreas de alta radiactividad, etc.), cumpliendo con esto, uno de los principales objetivos de la Industria 5.0: "*crear centralidad en el ser humano, incluyendo su seguridad física*". Esto también significa que cualquier error imprevisto o un ciber-ataque puede suponer una desviación en el comportamiento natural de dichos elementos, con impacto probablemente en el ser humano y su bienestar.

Es por ello que el MAS para el diagnóstico continuado es clave en estos tipos de sistemas, pero también el nivel de movilidad de los elementos observados, ya que obliga al MAS a no solo calcular los nuevos dinámicos del contexto, p. ej. la velocidad del dispositivo, sino también estimar la nueva opinión de acuerdo al posicionamiento físico y lógico de dicho dispositivo. Como es obvio, en estos tipos de entornos, las diversas casuísticas existen (p. ej. clusterización de dispositivos, aislamientos inesperados, o la inhabilitación para acceder temporalmente al medio), afectando con ello, y principalmente, al intercambio de opiniones con los vecinos más próximos para calcular el estado contextual, local al nodo, y con el sistema central para la correlación de éstas. Es más, la naturaleza dinámica del escenario puede también conllevar al despliegue de servicios complejos en el MAS, como pueden ser la gestión confiable de vecinos (nuevos y antiguos) o el descubrimiento continuado de servicios y nodos, resultando en nuevos retos de investigación para poder balancear rendimiento computacional con respecto a seguridad (tanto física como lógica).

4.3. Trazabilidad y seguimiento de ataques potenciales

En todos los escenarios anteriores y casos de uso, cada agente de monitorización se integraría como parte de las tecnologías operativas a analizar para que colaborativamente puedan computar estados de salud, ya sea a nivel local, grupal y global. Esta computación puede ayudar al sistema de monitorización: a derivar malfunciones por área o áreas, e incluso, explotaciones coordinadas (ej. botnets distribuidos), y a trazar el avance de una amenaza en tiempo real, ya sea por subestación o entre subestaciones. Ambas características pueden ser, además, útiles para los SOC, al favorecer no solo la "consciencia situacional" de los expertos sino también la detección avanzada de ataques potenciales de tipo Advanced Persistent Threat (APT) por añadir valor e información adicional en los procesos de análisis.

Movimientos laterales y ataques sigilosos (ej. canales en cubierto) podrían derivarse tras analizar el estado del sistema en su conjunto, que iría más allá de analizar el cumplimiento de patrones o reglas preestablecidas o el análisis de registros típicos de seguridad como son los logs. Es más, el simple hecho de ofrecer opiniones computadas en base a un conjunto importante de indicadores (a nivel de medición, de nodo, de red, de posicionamiento y de seguridad) y compartidas por varios agentes vecinos, es lo que favorece a su vez la identificación local y rápida de un problema y el seguimiento de la amenaza, favoreciendo la trazabiliad en vivo de ataques.

4.4. Conscienciación y validación mediante DT

Técnicamente, un gemelo digital consiste en la representación virtual de un objeto, proceso o sistema físico, capaz de caracterizar, mediante modelos digitales, conceptuales o matemáticos, las funciones primarias, estados y comportamientos que toma su correspondiente homólogo físico. La retroalimentación de ambos espacios, el digital y el físico, debe ser constante para garantizar, por un lado, la sincronización y fidelidad de las representaciones, y, por otro lado, la interacción con su correspondiente homólogo físico en caso de necesidad. Para tener este nivel de autonomía, el DT debe estar dotado de modelos cognitivos que favorezcan la toma de decisión, no solo para la validación de sistemas y el mantenimiento continuado y predictivo, sino también para dar garantías de ciber-defensa, en donde es posible no solo detectar y trazar acciones maliciosas sino también anular eventos anómalos y su efecto. Si, adicionalmente, estos modelos digitales y cognitivos integran información adicional relacionada con las opiniones gestionadas por el MAS en el espacio físico y lógico, es posible también intensificar el nivel de concienciación y conocer de primera mano el estado "real" del sistema. El DT puede en este caso verificar y contrastar en vivo las opiniones generadas por el MAS "virtual" (desplegado en el espacio digital) con respecto a las opiniones del MAS "físico" desplegado en el mundo real.

Por otro lado, el DT puede servir como una herramienta de validación del propio sistema de diagnóstico. El MAS virtual debe corresponder a una copia exacta de los agentes desplegados en el espacio físico, cuya misión es computar los mismos estados contextuales, ya sea a nivel de nodo como de red. Si los estados del objeto físico y digital son estables y están dentro de un umbral de normalidad, sus valores deben coincidir. En caso contrario, el DT puede estimar: (1) una anomalía en el propio proceso

Tabla 1: Resumen sobre los beneficios y limitaciones de los MAS aplicados a CPS

Beneficios para el CPS	Limitaciones para el CPS
Monitorización constante de CPS complejos y dimensionados	Impacto en rendimiento de dispositivos limitados en recursos
Mantenimiento predictivo	Confidencialidad y privacidad de los datos procesados por los agentes
Actualización de políticas de seguridad y procedimiento de actuación	Aislamientos de agentes en la infraestructura
Intensificación de la consciencia situacional	Punto de fallo en el colector
Mejor gobernanza, y control de abusos de recursos y accesos	Problemas de confianza y credibilidad entre los agentes
Trazabilidad y seguimiento de ataques en vivo	Ataques de AML a los detectores de anomalías
Seguridad física de operarios humanos	

de diagnóstico y su algoritmo que puede conllevar, sino se detecta bien, a tomas de decisiones inadecuadas en el espacio físico, o (2) una anomalía por intrusión (que se modifiquen de manera ilícita opiniones de agentes o se generen opiniones fraudulentas durante la convergencia). Por tanto, la replicación de sistemas multiagentes (en el DT) puede ser clave para verificar que el propio modelo de convergencia de opiniones se adecúa correctamente a las condiciones del escenario y no existe desviaciones que puedan corromper la seguridad y el mantenimiento de los sistemas ciberfísicos.

5. Limitaciones técnicas en CPS

Aunque la integración de MAS con sistemas que involucran nodos ciberfísicos puede automatizar los procesos de monitorización y favorecer a la detección temprana de amenazas, también existen una serie de complicaciones o contraindicaciones que dificultan su adopción, y que aún necesitan de una especial atención por la comunidad científica. A continuación, se detallan un conjunto de factores que hace complicado el proceso de integración del MAS en sistemas altamente críticos y limitados en cuanto recursos computacionales:

- *Rendimiento*: tal vez el problema más obvio al que se enfrenta los MAS es a la sobrecarga de trabajo que pueden ocasionar en dispositivos con poca capacidad o pocos recursos. Esto es debido a la necesidad de ejecutar un proceso más para el agente de monitorización en dispositivos empotrados, nodos ciberfísicos (controladores, sensores, actuadores) o nodos de IIoT (Industriales IoT), que suelen estar considerablemente limitados en términos de memoria y procesamiento [5], pues están pensados para dedicarse a una tarea específica.

- *Confidencialidad y privacidad*: dentro del contexto de la seguridad de estos sistemas de monitorización, debemos poner especial énfasis en asegurar que todas los datos usados por los agentes están protegidos frente a escuchas de terceras entidades que puedan o no tener objetivos malintencionados. Los datos que el MAS gestiona son especialmente sensibles, pues dan información precisa sobre el estado del CPS en todo momento, algo que podría ser usado en contra de la organización y provocar graves daños en el sistema. También, la privacidad es un asunto a cuidar en estos entornos, desde que largos volúmenes de datos se

producen, y aunque estén técnicamente protegidos frente a lecturas ilegítimas, si el atacante aplica técnicas para corromper la privacidad, puede derivar vulnerabilidades especificas por área o por nodo. Es por ello, que hay que investigar en técnicas que favorezcan el procesamiento de datos sensibles de una forma segura y privada, y que no suponga una violación a la reputación e integridad de la organización.

- *Aislamiento*: por la propia naturaleza distribuida de los MAS, puede darse el caso en que un agente quede aislado del resto, por pérdidas de conexión provocadas o casuales (ej. por la naturaleza móvil del CPS), lo que convierte a ese nodo en un potencial objetivo, especialmente vulnerable a ataques. El MAS debe contemplar estas situaciones, y proveer de mecanismos de contingencia que minimicen el impacto y que limiten las opciones de los atacantes.
- *Punto de fallo en el colector*: otro aspecto a tener en cuenta es que la propia arquitectura jerárquica que presenta el MAS de monitorización, convierte al agente colector en un elemento angular para todo el sistema, pues es donde se termina concentrando toda la información generada. Es por ello, que este agente debe tener disponibilidad muy elevada y alojarse en un entorno con importantes medidas de seguridad. También se deberían considerar estrategias de redundancia y copias de *backup* de los datos.
- *Confianza*: todo modelo de amenaza que analice el uso de MAS para tareas de monitorización también debe considerar a los propios agentes como vector de ataque para ocasionar falta de disponibilidad, violación a la integridad y engaños en los sistemas CPS. Como se explica en [25], los agentes de monitorización no tienen una visión completa de toda la red CPS y se ejecutan en un entorno no confiable, es por ello que deben ser capaces de medir la credibilidad del resto de agentes con los que interaccionan y gestionar la confianza depositada en ellos a partir de estas interacciones. Es por ello que el algoritmo de consenso de OD propuesto es el de confianza limitada, pues permite a un agente modificar la influencia de sus vecinos en base a las opiniones que reportan.
- *Precisión y fidelidad*: el uso de algoritmos de ML también puede tener sus efectos adversos, sobre todo, cuando actores maliciosos se aprovechan de esto para lanzar ataques de AML (*Adversarial Machine Learning*), en los que tratan de modificar el proceso de aprendizaje de estos sistemas [26]. Como resultado, los detectores ML no funcionan conforme a lo esperado, provocando retrasos a la hora de detectar amenazas, aumentando el número de falsos positivos, o simplemente favoreciendo la evasión de las técnicas de los atacantes.

Todas estas limitaciones de los MAS de monitorización se condensan en la tabla 1, así como los beneficios que aporta su adopción en entornos CPS.

6. Conclusiones

Como puede observarse el uso de tecnologías basada en agentes software tiene gran aplicabilidad en este campo, además de una trayectoria y un desarrollo futuro pro-

metedor. Sin embargo, su uso no está exento de consideraciones e implicaciones de seguridad que estos sistemas traen consigo, por lo que siempre es recomendable estudiar su efecto en el contexto de aplicación concreto de forma previa a su despliegue. Igualmente, también animamos a la comunidad científica a poner su enfoque en los problemas mencionados, y mejorar muchos de los puntos analizados y discutidos en este trabajo. A su vez, creemos que la estandarización del enfoque MAS para la monitorización de sistemas ciberfísicos es clave también para encontrar un claro equilibrio entre funcionalidad, interoperabilidad y seguridad.

En un futuro próximo pretendemos extender nuestro estudio para llevarlo a un entorno más práctico, en este caso, al laboratorio Urban Lab II [27], desplegado en los dominios de la Universidad de Málaga, para la gestión eficiente, sostenible y segura de la energía en infraestructuras de carga de vehículos eléctricos, así como en los desarrollos llevados a cabo dentro del proyecto SecTwin 5.0 [28], en el que se pretende desplegar los agentes dentro de un gemelo digital y dentro de su propia plataforma de interconexión.

Agradecimientos

Trabajo ha sido parcialmente financiado por el proyecto "Smart and Secure EV Urban Lab II", perteneciente al II Plan Propio Smart Campus de la Universidad de Málaga; por el proyecto SecTwin 5.0 (TED2021-129830B-I00) financiado por el Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación (10.13039/501100011033), and European Union "NextGenerationEU'/Plan de Recuperación, Transformación y Resiliencia; y por el proyecto BIGPrivDATA (UMA20-FEDERJA-082), destacando la colaboración económica del Fondo Europeo de Desarrollo Regional (FEDER) y de la Consejería de Economía, Conocimiento, Empresas y Universidad de la Junta de Andalucía.

Referencias

[1] I. Stellios, P. Kotzanikolaou, M. Psarakis, C. Alcaraz, and J. Lopez, "Survey of iot-enabled cyberattacks: Assessing attack paths to critical infrastructures and services," *IEEE Communications Surveys and Tutorials*, vol. 20, pp. 3453–3495, 07/2018 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8410404

[2] L. Ifigeneia, T. Eleni, S. N. Rossen, C. Cosmin, A. Malatras, and T. Marianthi, "ENISA threat landscape 2022," ENISA, ISBN: 978-92-9204-588-3, https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022, accedido abril 2023, 2022.

[3] L. Ifigeneia, T. Marianthi, T. Eleni, A. Malatras, G. Sebastian, and V. Veronica, "ENISA threat landscape for supply chain attacks," ENISA, ISBN: 978-92-9204-509-8, https://www.enisa.europa.eu/publications/threat-landscape-for-supply-chain-attacks, accedido abril 2023, 2021.

[4] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human factors*, vol. 37, no. 1, pp. 32–64, 1995.

[5] C. Alcaraz, L. Cazorla, and G. Fernandez, "Context-awareness using anomaly-based detectors for smart grid domains," in *9th International Conference on Risks and Security of Internet and Systems*, vol. 8924, Springer International Publishing. Trento: Springer International Publishing, 04/2015 2015, pp. 17–34. [Online]. Available: http://link.springer.com/chapter/10.1007\%2F978-3-319-17127-2_2$\#$

[6] A. L. Bazzan and F. Klügl, *Multi-agent systems for traffic and transportation engineering*. IGI Global, 2009.

[7] A. S. Nair, T. Hossen, M. Campion, D. F. Selvaraj, N. Goveas, N. Kaabouch, and P. Ranganathan, "Multi-agent systems for resource allocation and scheduling in a smart grid," *Technology and Economics of Smart Grids and Sustainable Energy*, vol. 3, pp. 1–15, 2018.

[8] M. S. Rahman, M. A. Mahmud, A. M. T. Oo, and H. R. Pota, "Multi-agent approach for enhancing security of protection schemes in cyber-physical energy systems," *IEEE transactions on industrial informatics*, vol. 13, no. 2, pp. 436–447, 2016.

[9] P. Wang and M. Govindarasu, "Multi-agent based attack-resilient system integrity protection for smart grid," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3447–3456, 2020.

[10] E. Shakshuki and M. Reid, "Multi-agent system applications in healthcare: current technology and future roadmap," *Procedia Computer Science*, vol. 52, pp. 252–261, 2015.

[11] ENISA, "Good practices for security of internet of things in the context of smart manufacturing," ISBN: 978-92-9204-261-5, https://www.enisa.europa.eu/publications/good-practices-for-security-of-iot, accedido abril 2023, 2018.

[12] R. Alexander and T. Kelly, "Supporting systems of systems hazard analysis using multi-agent simulation," *Safety Science*, vol. 51, no. 1, pp. 302–318, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925753512002032

[13] M. Wooldridge, N. R. Jennings, and D. Kinny, "The gaia methodology for agent-oriented analysis and design," *Autonomous Agents and multi-agent systems*, vol. 3, pp. 285–312, 2000.

[14] T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *IEEE communications surveys & tutorials*, vol. 10, no. 4, pp. 56–76, 2008.

[15] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications surveys & tutorials*, vol. 18, no. 2, pp. 1153–1176, 2015.

[16] MITRE, "ICS tactics," https://attack.mitre.org/tactics/ics/, accedido abril 2023, 2015-2022.

[17] J. E. Rubio, M. Manulis, C. Alcaraz, and J. Lopez, "Enhancing security and dependability of industrial networks with opinion dynamics," in *Computer Security–ESORICS 2019: 24th European Symposium on Research in Computer Security, Luxembourg, September 23–27, 2019, Proceedings, Part II 24*. Springer, 2019, pp. 263–280.

[18] J. E. Rubio, R. Roman, C. Alcaraz, and Y. Zhang, "Tracking advanced persistent threats in critical infrastructures through opinion dynamics," in *European Symposium on Research in Computer Security (ESORICS 2018)*, vol. 11098, Springer. Barcelona, Spain: Springer, 08/2018 2018, pp. 555–574. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-99073-6_27

[19] Y. Dong, M. Zhan, G. Kou, Z. Ding, and H. Liang, "A survey on the fusion process in opinion dynamics," *Information Fusion*, vol. 43, pp. 57–65, 2018.

[20] J. E. Rubio, R. Roman, C. Alcaraz, and Y. Zhang, "Tracking apts in industrial ecosystems: A proof of concept," *Journal of Computer Security*, vol. 27, pp. 521–546, 09/2019 2019.

[21] C. Alcaraz, "Cloud-assisted dynamic resilience for cyber-physical control systems," *IEEE Wireless Communications*, vol. 25, no. 1, pp. 76–82, 2018.

[22] C. Alcaraz, A. Garcia, and J. Lopez, "Implicaciones de seguridad en mas desplegados en infraestructuras de carga basadas en ocpp," *VII Jornadas Nacionales de Investigación en Ciberseguridad*, 2022.

[23] C. Alcaraz, J. Lopez, and S. Wolthunsen, "Ocpp protocol: Security threats and challenges," *IEEE Transactions on Smart Grid*, vol. 8, pp. 2452 – 2459, 02/2017 2017.

[24] J. E. Rubio, C. Alcaraz, and J. Lopez, "Addressing security in ocpp: Protection against man-in-the-middle attacks," in *2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, 2018, pp. 1–5.

[25] M. S. Hegde and S. Singh, "Safe buzz: adaptive security for multi-agents through situational awareness," *International Journal of Autonomous and Adaptive Communications Systems*, vol. 10, no. 2, pp. 192–212, 2017.

[26] E. Anthi, L. Williams, M. Rhode, P. Burnap, and A. Wedgbury, "Adversarial attacks on machine learning cybersecurity defences in industrial control systems," *Journal of Information Security and Applications*, vol. 58, p. 102717, 2021.

[27] Cristina Alcaraz and Alicia Triviño, “Smart and Secure EV Urban II - Plan Propio de Smart-Campus,” https://eventos.uma.es/63025/detail/smart-and-secure-ev-urban-ii-plan-propio-de-smart-campus.html?private=1da34ddc2ebc057ac5b3, accedido abril 2023, 2021-2023.

[28] Cristina Alcaraz and Javier Lopez, “Cybersecurity Platform based on Digital Twins for Industry 5.0,” https://www.nics.uma.es/projects/sectwin-50, accedido abril 2023, 2022-2024.